Outcome Effects, Moral Luck and the Hindsight Bias

Markus Kneer (markus.kneer@uzh.ch)
Izabela Skoczen (izaskoczen@gmail.com)
Draft. Please cite the final version.

Abstract

In a series of ten preregistered experiments (N=2043), we investigate the effect of outcome valence on judgments of probability, negligence, and culpability – a phenomenon sometimes labelled moral (and legal) luck. We found that harmful outcomes, when contrasted with neutral outcomes, lead to increased perceived probability of harm *ex post*, and consequently to increased attribution of negligence and culpability. Rather than simply postulating a hindsight bias (as is common), we employ a variety of empirical means to demonstrate that the outcome-driven asymmetry across perceived probabilities constitutes a systematic cognitive distortion. We then explore three distinct strategies to alleviate the hindsight bias and its downstream effects on mens rea and culpability ascriptions. Not all are successful, but at least some prove promising. They should, we argue, be taken into consideration in criminal jurisprudence, where distortions due to the hindsight bias are likely considerable and deeply disconcerting.

1. Introduction

1.1 Outcome Effects on Culpability

Frank and Su drive to work. They are well-rested, alert and stick to the speed limit. A child jumps in front of Frank's car and dies, whereas Su arrives at work without incident. Who is more to blame? In between-subjects designs, a pronounced outcome effect tends to arise: Frank is judged morally and legally more culpable than Su (henceforth the *Outcome Effect*). This might strike us as unjust, if we hold, with Kant (1978), that agents are morally responsible only for features of their actions over which they have control (the *Control Principle*).

Philosophers assume that a difference in moral judgment arises even within-subjects, i.e. when people directly compare Frank's and Su's case (the *Difference Intuition*). This would give rise to the *Problem of Resultant Moral Luck* (cf. Williams, 1981, Nagel, 1979, Nelkin, 2004, 2019, 2021, Hartman, 2017, Kamtekar & Nichols, 2019; for empirical work on moral luck, see e.g. Spranca et al. 1998, Cushman, 2008, Young et al. 2010, Nichols et al. 2014, Kneer & Machery, 2019): We must square the consequentialist Difference Intuition with the Kantian Control Principle, but the two are fundamentally inconsistent. Importantly, however, Folk Morality disagrees: When presented with Frank and Su's cases side by side, the vast majority of participants evaluate the two agents identically. Western Criminal Law, with its deep distaste for strict liability, sides with the Folk in this regard. So there might not be a complex *philosophical* problem (the within-subjects Difference Intuition assumed by philosophers seems to be empirically mistaken).

The *practical* problem, however, must be taken seriously: In everyday life, we are not confronted with two neat cases side-by-side. Usually, we assess situations where a concrete harm has occurred and here outcome is likely to have a distorting effect on our judgment, violating the Control Principle to which both the Law and the Folk are committed.

1.2 The Mechanics of the Outcome Effect

How can we alleviate the outcome effect? This depends, in parts, on its more intricate mechanics. There is some evidence in favour of a *probabilistic* account of moral luck-type phenomena (Kamin & Rachlinski, 1995; Kneer & Machery, 2019). On this account, the post-hoc probability of harming a child is perceived higher for Frank than for Su. It thus seems more appropriate to judge that Frank incurred a substantial risk than that Su did, which, in turn would mean he was more *reckless* or *negligent* than Su.[1] If this account is on the right track, then a perceived difference in probability and risk drives an asymmetry of risk-related inculpating mental states and hence moral (and legal) evaluation (see Figure 1). The whole series of inferences from descriptive features to normative evaluation is innocuous, except for the first step, which is affected by the hindsight bias: in Frank's case, people tend to exaggerate the degree to which a harmful outcome could, or should, have been anticipated (Fischhoff, 1975; 1980). To address the distorting effect of outcome on culpability judgments, this suggests, we must find ways to alleviate the hindsight bias – which is the topic of the paper.

The paper proceeds as follows: We first explore whether the probabilistic account of the effect of outcome on culpability replicates (section 2). Our experiments are the first to control explicitly for the distinction between objective probability (probability from the perspective of the universe) and subjective probability (as perceived from the agent's context). Having replicated the outcome effect on probability, mens rea and moral judgment, we *show* – rather than just *assume*, as is standardly the case – that it must be considered a bias. The effect of outcome is much more pronounced in between-subjects designs than in within-subjects designs, in which participants have the possibility to reflect on whether outcome *should* make a difference to their assessment of probability, mens rea and guilt (section 3). Once the process is clearly laid bare (see Figure 7), we turn to the core objective of the paper: *Debiasing strategies*. The first such strategy investigated is *probability anchoring* (section 4), in which we test whether giving participants the possibility to evaluate the likelihood of a harmful outcome before the consequences are revealed has an impact on their probability assessments *ex post*. The next strategy is *counterfactual priming* (section 5), where we investigate whether entertaining alternative outcomes reduces the outcome effect on probability, mens rea and moral judgments. Finally, we turn to *probability stabilizing* (section 6), in which an expert provides the actual ex ante probability of a harmful outcome from the point of view of a scientifically informed perspective. Figure 1 visualizes the different debiasing strategies.

[1] For the effect of outcome on possibly inculpating mental states more generally, see the literature on the Knobe effect (Knobe, 2003a, 2003b, 2010; for reviews, see Feltz, 2007, Cova, 2016) and the epistemic side-effect effect (Beebe & Buckwalter, 2010; Beebe & Jensen, 2012, Alfano et al. 2012, Kneer 2018). For empirical studies regarding the Knobe Effect and the Severity Bias conducted with legal experts (judges, lawyers and law students), see Kneer & Bourgeois-Gironde (2017), Bourgeois-Gironde & Kneer (2018), Prochownik et al. (2020), Tobia (ms).

Probability anchoring and counterfactual priming attempt to prevent inappropriate inferences from outcome information to probability *ex post* in indirect fashion. By contrast, explicit probability stabilizing, for instance by invoking an expert, makes short shrift of the problem by directly stipulating the probability ex post so as to prevent inadequate downstream consequences on mens rea and culpability assessment.

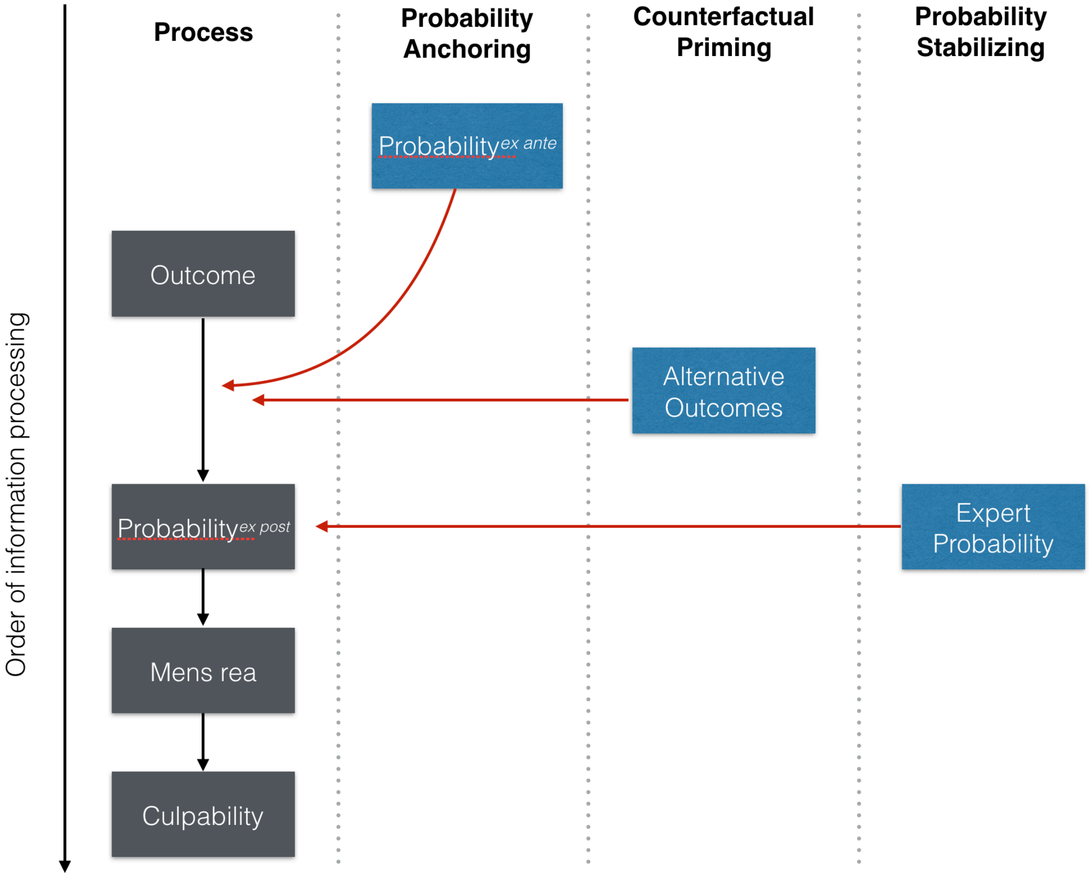

Figure 1: The order of information processing in negligence cases and possible debiasing strategies.

To anticipate the findings, which we consider of considerable importance both for moral psychology and criminal jurisprudence: Consistent with previous research, the effects of outcome on probability post hoc and downstream variables such as mens rea and culpability are persistent and robust across experiments with different scenarios. These effects, we demonstrate, are the results of a cognitive bias (though for judgments concerning deserved punishment, they are not – a fact on which we will elaborate at length). Probability anchoring and counterfactual priming succeed in mitigating the outcome bias somewhat. However, neither strategy fully eradicates inappropriate inferences from outcome to probability and distorted downstream effects on mens rea and culpability judgments thus remain. What works best is probability stabilizing, which is indeed a means courts sometimes resort to (though all too frequently they do not cf. for example Lee, 1988).

2. Experiment 1: Outcome Effects

Whereas there is no lack of literature concerning the hindsight bias (for a review, see e.g. Roese & Vohs, 2012), few studies explore the downstream effects on moral and legal culpability and the mechanism by way of which probability affects the latter. Exceptions are Kamin & Rachlinsky (1995), who show that perceived probability post hoc has an effect on perceived culpability. Kneer & Machery (2019) go one step further in demonstrating that the relation between outcome-driven perceived probability and culpability is itself mediated by perceived negligence of the agent.

Our first experiment attempts to replicate these findings with a new scenario. It also introduces a methodological novelty: Rather than asking for the probability or likelihood of a harmful outcome simpliciter, we disambiguate the notion of probability into two kinds: Objective probability, i.e. the actual likelihood of an accident independent of potential epistemic distortions, and subjective probability, i.e. the probability of a bad outcome as perceived from the agent's particular epistemic situation. The first question employed the locution "how likely was it from an objective point of view that [X would occur]". To minimize confusion between types of probability among the participants, the subjective probability question was phrased in terms of the agent's "having good reasons to believe that [X] would occur". Evidently, subjective probability and reasons for belief are not perfectly coextensive. However, given the features that were salient in our scenarios, the two DVs are similar enough: A high subjective probability of a flood corresponds to good reasons for believing there will be a flood and vice versa.

> **Objective probability:** On a scale from 0 (completely unlikely) to 100 (certain) how likely was it from an objective point of view that there would be a flood this year?
> **Reasons (subjective probability):** To what extent do you agree with the following statement: Ms. Russel had good reasons to believe that there would be no flood this year. (0 = completely disagree; 100 = completely agree)

2.1 Participants

We recruited 195 participants online via Amazon Mechanical Turk. The IP address location was restricted to the USA. In line with the preregistered criteria,[2] participants who failed an attention check, were not native speakers of the English language, took less than two minutes to complete the entire survey or failed the comprehension question were excluded, leaving a sample of 169 participants (female: 47 %; mean age: 43 years, SD = 12 years, range: 19–82 years).[3]

[2] https://aspredicted.org/blind.php?x=3kq5bn. Preregistrations, stimuli and data for this and all further experiments can be found on the project's OSF site under the following link: https://osf.io/e2u8q/.

[3] Experiments 1 and 3 are very similar, differing only with respect to a minor design choice (ex ante assessment of probability). Since we had planned from the outset to compare the results across designs, we preregistered and ran the

2.2 Methods and Materials

Participants were shown a vignette (see Appendix section 1.1 for detail) in which a strawberry farmer, Ms. Russel, hosts workers on her farm during harvest time. The lodgings, which are on Ms. Russel's grounds, are close to a river, which flooded two years ago. Though Ms. Russel took precautions the previous years against potential flooding (none occurred), this year she believes there will be no flood and uses the budget to refurbish the kitchens of the workers' houses instead. The vignette came with one of two endings (labels in bold omitted):

> **Neutral:** As during the previous years, the river's water supply is low all season and it never overflows. The fruit pickers are glad that the money has been invested into the refurbishment of the kitchens.
>
> **Bad:** It just so happens that there is a torrential downpour one night that nobody saw coming. The lodgings are flooded within hours. Several fruit pickers are severely injured and one worker and his two children die a slow and painful death as they get trapped in a flooded house.

Thereafter, participants were asked to answer two questions concerning objective and subjective probability (on a scale from 0 to 100), as formulated above.

In the experiment, we tested for two types of mens rea: recklessness and negligence (see MPC 2.02 (c) and (d), for the US, see e.g. Fletcher, 2000; for the UK, see e.g. Herring, 2012). An agent acts recklessly, if she knowingly incurs a substantial risk. An agent acts negligently, if she *should have* been aware of a substantial risk. The scenario for our first experiments concerns the failure to install a protection against river flooding. Further down, we report experiments with a second scenario that focuses on speeding at an intersection (see section 7, and Appendix sections 6-10). In principle, both scenarios could be treated either as recklessness or negligence cases. However, given the details of the situations described, they are best interpreted as negligence cases, because in both scenarios the agents evaluated the risk at hand as unsubstantial, whereas a reasonable person would have considered the said risks as substantial. Nonetheless, since there is evidence that laypeople frequently have difficulties differentiating between these two types of mens rea (Shen et al. 2011), we ran questions focusing on both negligence and recklessness. On a 7-point Likert scale, participants had to report their agreement and disagreement with the following claims (labels in bold omitted):

two experiments together (from a single Qualtrics survey). Given that prevention of ballot-box stuffing was turned on in Qualtrics, no participant could participate both in Experiment 1 and Experiment 3.

Recklessness: Ms. Russel was aware of a substantial risk of a flood occurring this year. (1= completely disagree; 7= completely agree)
Negligence: Ms. Russel should have been aware of a substantial risk of a flood occurring this year. (1= completely disagree; 7= completely agree)

Finally, we tested two types of moral judgment, blame and deserved punishment (cf. Cushman, 2008; Kneer & Machery, 2019). The reason for this was twofold: First, punishment is known to be considerably more sensitive to outcomes than blame and, second, it is a variable which is directly relevant for legal contexts. The questions read (labels in bold omitted):

Blame: To what extent is Ms. Russel blameworthy for not installing the flood protection this year? (1 = not at all blameworthy; 7 = extremely blameworthy)
Punishment: How much punishment does Ms. Russel deserve for not installing the flood protection this year? (1 = no punishment at all; 7 = very severe punishment)

2.3 Results

2.3.1 Main Results

Probabilities are most naturally reported in percentages (following our ordinary practices), rather than 7-point Likert scales. To improve ease of presentation, we rescaled all probabilities to fit the 7-point Likert scales which we employ for the measurement of *mens rea* and the moral variables. The mean ratings for all dependent variables are presented in Figure 2.

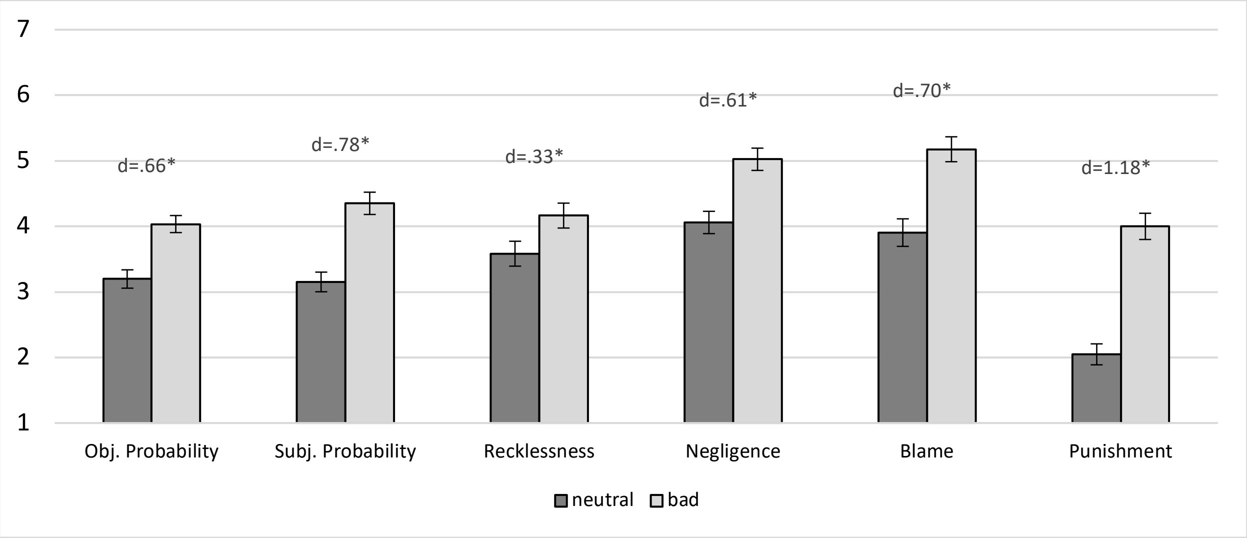

Figure 2: Mean ratings for probability, mens rea and moral judgment for the between-subjects design (outcome: neutral v. bad). Error bars denote standard error of the mean.

Consistent with previous research (Cushman 2008; Cushman et al, 2009; Gino et al, 2009; Gino et al, 2010; Young et al, 2010; Schwitzgebel & Cushman, 2012; Lench et al, 2015), there is a significant main effect of outcome (all *ps*<.035) on the moral variables (blame, punishment, see

Appendix section 1.2), mens rea (see Kneer & Bourgeois-Gironde, 2017, Kneer & Machery, 2019) and perceived probability (Arkes et al, 1981, Dawson et al, 1988; Christensen-Szalanski & Willham, 1991; Kamin & Rachlinski, 1995; Kneer, 2021) both when assessed in objective and subjective terms (see Appendix, B.1.1).

2.3.2 Mediation Analyses for Blame

In order to explore whether the pattern proposed by Kneer & Machery (2019) replicates (see Figure 2 above), we conducted a series of mediation analyses. A key novelty of our experiment is that we differentiate between subjective and objective probability, and that this might help to reveal the precise mechanics of the hindsight bias in more detail. We first conducted a multiple mediation analysis to explore which of the potential factors mediates the relation between outcome and blame. As shown by Figure 3, recklessness and objective probability proved nonsignificant. However, both subjective probability and negligence were significant mediators, and taking them into account rendered the impact of outcome on blame nonsignificant (*p*=.16).

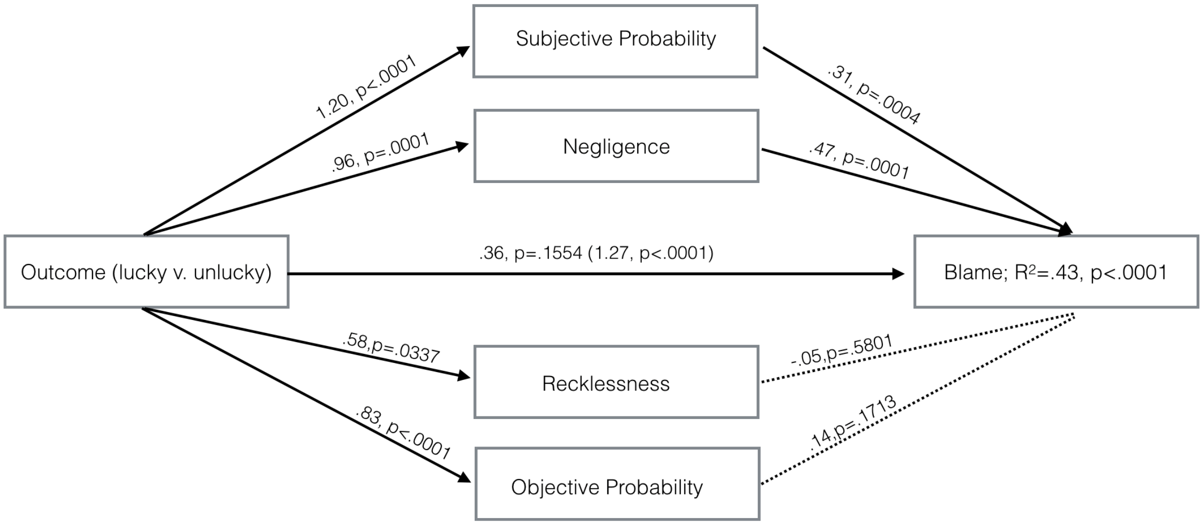

Figure 3: Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and blame judgments by probability (objective and subjective), negligence and recklessness.

A serial mediation analysis with subjective probability and negligence provides more clarity: the relation between outcome and blame is not mediated by negligence *per se* (the a^2 path is nonsignificant). Instead, mediation through negligence travels through subjective probability (the a^1db^2 path is significant) and some of the mediation occurs via subjective probability independently of negligence (the a^1b^1 path is significant).

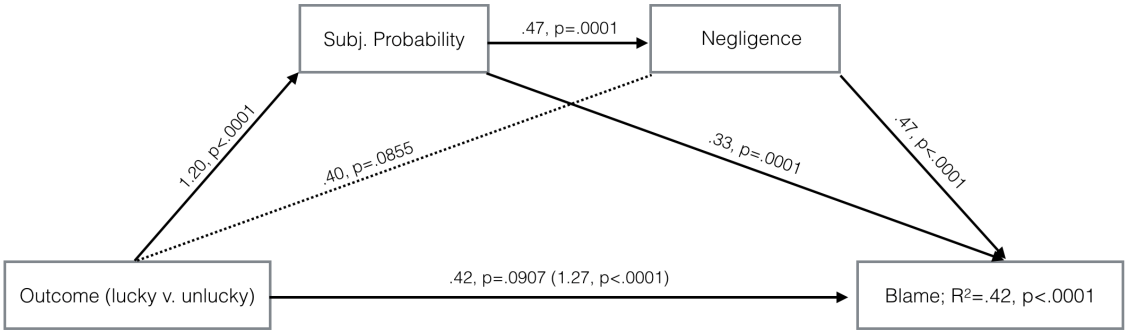

Figure 4: Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and blame judgments by subjective probability and negligence.

2.3.3 Mediation Analysis for Punishment

Things are different concerning the moral DV of punishment. *First*, whereas accounting for the mediators in the blame analysis renders the *c-path* nonsignificant, suggesting near-complete mediation, mediation accounts only for about a third of the total effect of outcome on punishment, which remains significant (p<.001), see Figure 5. In contrast to blame, this suggests, punishment is strongly sensitive to outcome itself. *Second*, whereas in the blame analysis subjective probability played a key role besides negligence, it proves nonsignificant for punishment. Here, however, objective probability is a significant mediator besides negligence. A serial mediation model shows that all three mediation paths are significant, confirming that mediation accounts for about a third of the total effect of outcome on punishment. Objective probability by itself (the a^1b^1 path) accounts for more than half of the mediation (54%) cf. Figure 6.

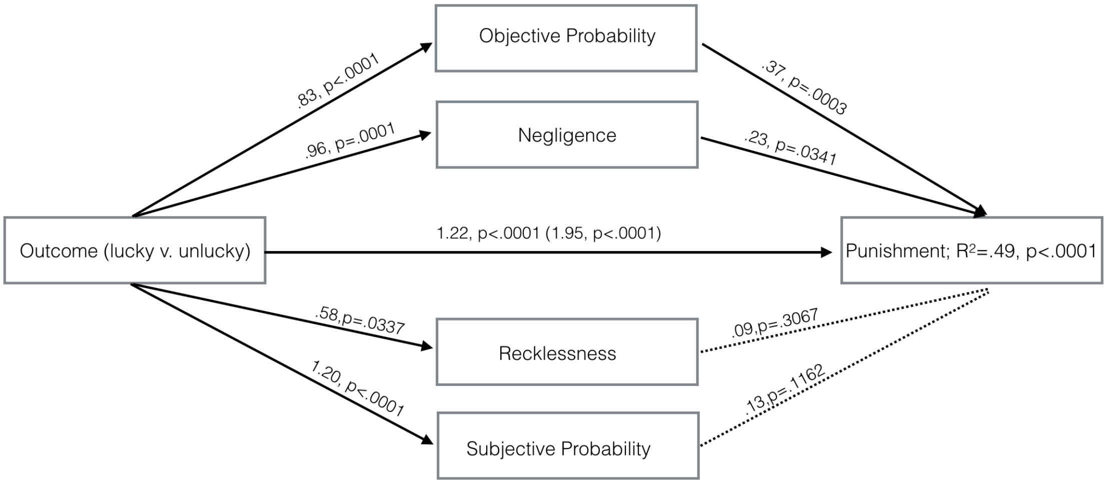

Figure 5: Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and punishment judgments by probability (objective and subjective), negligence and recklessness.

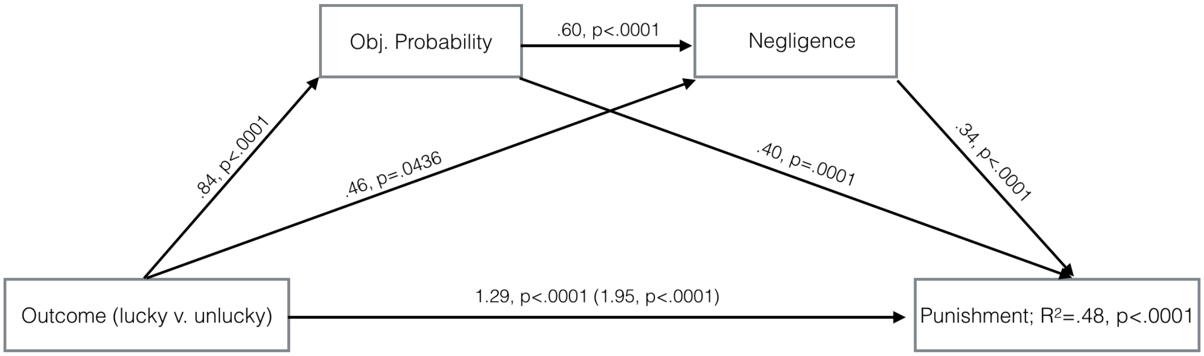

Figure 6: Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and punishment judgments by objective probability and negligence.

2.4 Discussion

There is a pronounced outcome effect on both types of probability, mens rea, and the moral variables. The outcome effect on blame is mediated completely by subjective probability and negligence: People consider the harmful outcome more likely if it occurs and thus the unlucky agent more negligent and blameworthy than the lucky one. The effect of outcome on punishment, by contrast, is only partially mediated by objective probability and negligence. Importantly, about two thirds of the effect of outcome on punishment is direct (at least given the mediators we tested), and its impact remains significant even when taking the mediators into account.

The findings not only reveal the mechanics of the outcome effect on two different measures of culpability, but they also shed light on Cushman's (2008) influential *Dual Process Model of Moral Judgment.* According to this model, one process of moral judgment is strongly sensitive to mental states, whereas the other is predominantly sensitive to non-mental features of the action sequence. This is precisely what we find: For blame, what matters is the agent's subjective situation, and its attribution is entirely mediated by the inculpating mental state of negligence.[4] Punishment, by contrast, is strongly sensitive to outcome. What is more, in so far as probability matters, it is not the likelihood of a harmful outcome as envisioned by the agent (a mental representation), but the *objective* probability that drives punishment judgments.[5]

[4] In the legal literature, negligence – i.e. that the agent should have been aware of a substantial risk – is frequently considered an "objective state" and distinguished from the "subjective states" of intention, knowledge and recklessness. Whereas there is, of course, an important difference between holding someone culpable for the *presence* of inappropriate mental states v. the *absence* of appropriate ones, what matters is still their *mental state*, the attribution of which, furthermore, is contingent on their particular epistemic context.

[5] Cushman argues that judgments concerning permissibility and wrongness depend primarily on mental states, whereas blame and punishment depend strongly on non-mental features. Like Kneer & Machery, 2019 we find blame to be more in the former category. Two things bear mentioning however: First, the exact status of blame is still contentious (see e.g. Prochownik & Cushman, ms; Frisch et al., ms). Second, and more importantly, no matter on which side of the fence blame falls, our findings show that the central thrust of Cushman's Dual Process Theory of Moral Judgment is correct – there are two very distinct processes of moral judgment.

People's propensity to judge an outcome that has occurred more likely ex post than ex ante, or "creeping determinism" as Fischoff (1982) calls it, is near-universally considered a *bias* (Walster, 1967; Fischhoff, 1975 and 1980; Hoch & Loewenstein, 1989; Agans & Shaffer, 1994; Hertwig et al, 1997; in the legal literature: Arkes & Schipani, 1994; Lowe & Reckers, 1994; Buchman, 2002). A similar assessment regards its downstream effects on inculpating mental states (Kneer & Machery, 2019) and judgments of culpability (Arkes, 1981; Casper et al., 1989; Wexler & Schopp, 1989; Bodenhausen, 1990; Kamin & Rachlinski, 1995, Rachlinski, 1998; more generally, see also Alicke, 2000). However, one should tread carefully here: Perhaps the folk *concept* of probability simply is, like the concept of punishment (Cushman, 2008; Kneer & Machery, 2019) strongly outcome-sensitive – without this constituting a bias. Differently put, perhaps the folk think that it is *appropriate* to take outcome into account when assessing probability, and hence doing so does not constitute a performance error. Returning to our opening example, one way to construe such a *rationality view of outcome effects* would be this: the likelihood of a fatal accident in Frank's situation is reasonably judged higher *post hoc* than its probability in Su's situation since Frank just *did have* an accident, whereas Su did not. Given that the notion of a *risk* is defined as the product of an event's probability and the degree of harm occasioned, it thus follows that the risk incurred by Frank was higher than the risk incurred by Su. But if this is so, then it seems more warranted to contend that Frank should have been aware of a substantial risk than Su, *inter alia* because the risk incurred by Su might not have been substantial in the first place. So Frank is deemed more negligent, and consequently judged as more deserving of blame and punishment.

How to adjudicate between the two views? In principle, if the folk concept of probability were outcome-dependent, then assessing two in all respects identical scenarios that differ only in terms of outcome side by side should lead to an asymmetry in perceived probability. For punishment, for instance, this is exactly what we tend to find. In within-subjects designs, in which the situational and mental features of two agents are held fixed and in which only outcomes differ, a robust outcome effect on punishment can be found. The folk concept of punishment, this implies, simply is outcome-dependent or consequentialist. For wrongness of an action or deserved blame, by contrast, a robust between-subjects difference across outcomes tends to vanish in within-subjects experiments (Kneer & Machery, 2019). This suggests that the folk concepts of wrongness or blame are *not* consequentialist, though when evaluating just a single case, we tend to draw strong, most likely inappropriate, inferences from outcome to wrongness or blame. In the following experiment, we will put all three types of variables so far explored to the test in a within-subjects design to gain some insight as to the bias question.

3. Experiment 2 Within-subjects design

The goal of experiment 2 was to explore whether the effect of outcome on probability constitutes a bias, as is near-universally assumed, or whether the folk concept of probability might be sensitive to the (occurrence and nonoccurrence) of outcomes (Hsee, 1996, Rachlinski, 1998, 2000, Baron

& Ritov, 2004; Hsee & Zhan, 2004, Baron, 2010). To do this, we ran the *Flood Scenario* in a within-subjects design.

3.1 Participants

96 participants were recruited online via Amazon Mechanical Turk. The IP address location was restricted to the USA. As preregistered,[6] participants who failed the attention check or the comprehension question were excluded, as well as those who were not native speakers of the English language or finished the entire survey (including demographic questionnaire) in under two minutes. A sample of 84 participants remained (female: 51%; age M=46 years, SD = 14 years, range: 23-74 years).

3.2 Methods and Materials

Participants were presented with both outcome conditions of the *Flood* vignette side-by-side. To facilitate presentation, one farm owner was called Ms. Russel, the other Ms. Miller. Having read both vignettes, participants had to rate probabilities (subjective and objective), mens rea (negligence and recklessness) and moral judgment (blame and punishment) for both agents. To encourage a comparative assessment, the questions always mentioned both agents. The blame questions, for instance, read "To what extent are Ms. Russel and Ms. Miller blameworthy for their actions, if at all?". Participants had to rate Ms. Russel's action, and thereafter Ms. Miller's action, on separate Likert scales ranging from 1 ("not at all blameworthy") to 7 ("extremely blameworthy").

3.3 Results

Mean ratings across outcome, effect sizes and results of paired-samples t-tests for all six DVs are presented in Figure 7 (for t-test details, see Appendix 2.2). Except for objective probability, we found a significant difference for all dependent variables (all *ps*<.049), though subjective probability just barely makes the threshold. Importantly, however, the effects sizes for all variables are much lower than a between-subjects design, and small for all variables except blame and punishment. For blame, the effect size decreased from *d*=.70 (between-subjects) to *d*=.49 (within-subjects), for punishment it decreased from *d*=1.18 to *d*=.63.

[6] Link to preregistration: https://aspredicted.org/blind.php?x=42tv63

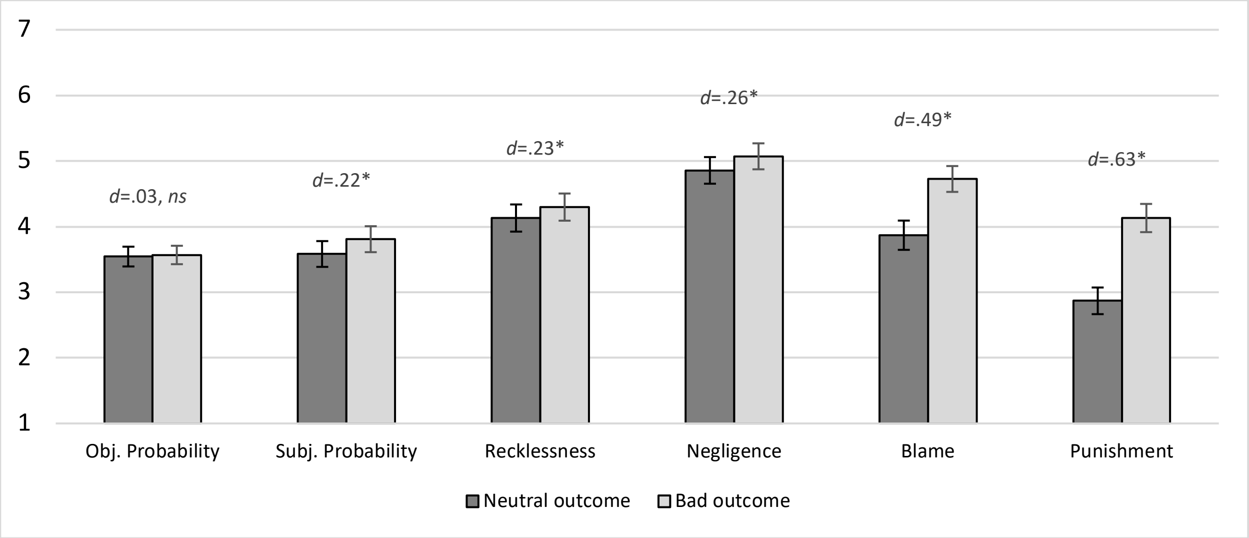

Figure 7: Mean probability, mens rea and moral responsibility judgments for the within-subjects design in the two conditions (neutral vs. bad outcome). Effect sizes are given in terms of Cohen's *d*, significance is reported at the *p*<.05 threshold (for details, see Appendix). Error bars denote standard error of the mean.

Despite the considerable decrease in outcome effect size, one might be astonished by the fact that the effect of outcome on mens rea and moral judgment is *still* significant in the within-subjects design. As argued by Kneer & Machery (2019), however, it might be instructive to look at the proportions of participants who manifest a *Difference Intuition* across the two situations (neutral v. bad) in the within-subjects design, i.e. who judge the two situations and agents differently with respect to probability, mens rea and morality. As Figure 8 illustrates, a substantial majority (more than 60%) judges the two situations/agents identically across all variables except punishment (49%); all being significantly above chance (binomial tests, all *ps*<.022, two-tailed) except punishment and objective probability (both *ps*>.062). As concerns punishment, this is no surprise. It is well established (Cushman, 2008, Kneer & Machery, 2019, see also mediation analyses above) that there is a strong, direct effect of outcome on punishment.

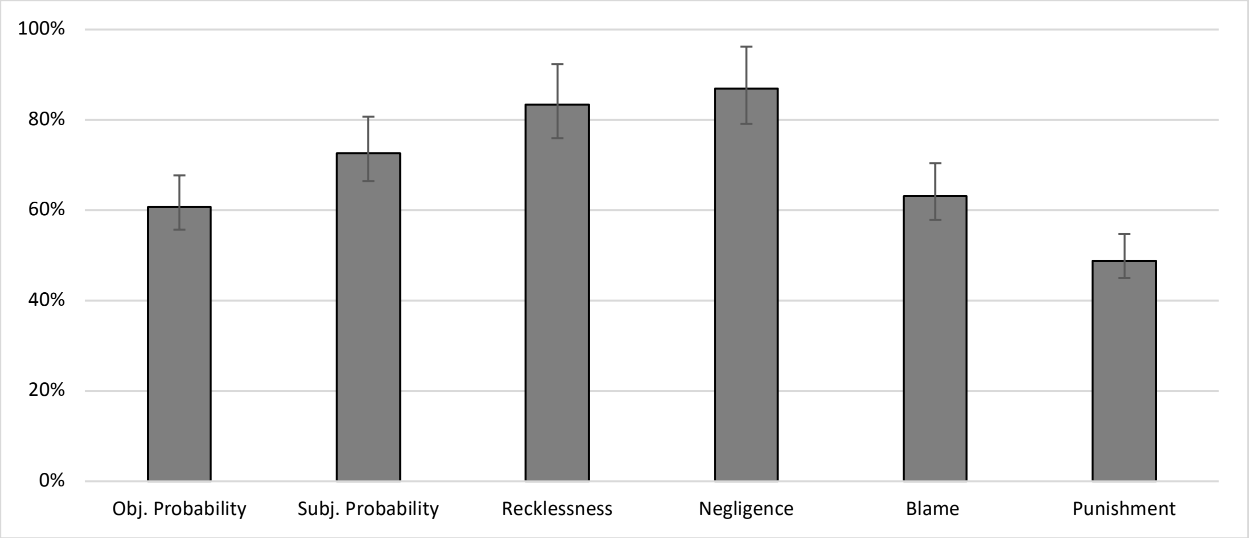

Figure 8: Proportions of participants who judged probabilities, mens rea, blame and punishment identically (no Difference Intuition) across scenarios. Error bars denote 95% confidence intervals, Wilson method, see Brown, Cai, & DasGupta, 2001.

We would have expected the proportions of identical ratings of subjective probability, and particularly objective probability, to be higher. The reason, we'd like to suggest, is simply the response mechanism. We used the Qualtrics slider-scale (pictured in the Appendix, 2.2.2) and it is quite hard to indicate a *precise* probability, in particular on a mobile device. Once the criterion for identical probabilities is relaxed to include probabilities with a maximum 5-point (out of 100) difference – which would be nonsignificant – the proportion of identical assessments for objective probability is 79% and for subjective probability it is 80%, both significantly above chance (binomial tests, *p*s<.001). These figures – roughly four of five participants – squares with the proportions of identical assessment of mens rea, which are the same. Note that if perceived probabilities were indeed quite different, it would make little sense to rate mens rea identically in the scenario at hand.[7]

3.4 Discussion

The results of the experiment are relatively clear. Even in a design where people see both scenarios side-by-side, and should thus become aware that the only difference consists in outcome, punishment ratings across cases differ significantly and manifest a medium-sized effect (*d*=.63). In line with previous findings (Cushman, 2008; Martin & Cushman, 2015, 2016, Kneer & Machery, 2019), this suggests that the folk concept of punishment is outcome-sensitive. Note that, on this view, the outcome effect (neutral v. bad) *per se* in the between-subjects design should not be regarded as a bias. However, its *size* – it's nearly twice as pronounced in a between-subjects

[7] Naturally, this would not hold for a case where probabilities are extremely low or high, yet different, since then we would have a clear-cut case of mens rea or clear-cut case of absence thereof for both conditions.

design, amounting to a very large effect – might indeed be taken to be, at least in parts, the consequence of a performance error.

Things are different as regards objective probability (not significant) and subjective probability (barely significant). For negligence and recklessness, we find a significant effect, though for both probabilities and mens rea, the effect sizes are very small. What is more, once we have corrected for the technical problem of the slider scale, about 80% of participants rate the probabilities and mens reas identically across cases, which suggests that the folk concepts of objective and subjective probability, as well as those of recklessness and negligence are outcome-independent. For blame, the findings are not quite as clear. The difference is significant, and there's a medium-sized effect (*d*=.49). However, here, too, a significant majority holds that the two agents deserve the same amount of blame. Taken together, we consider the results of Experiment 2 to constitute evidence in favour of the view that the folk concepts of probability (both objective and subjective) and mens rea are outcome-independent (cf. Spellman & Kincannon, 2001; Gilbert et al, 2015; Schauer & Spellman, 2020), and tentative evidence for the outcome-insensitivity of the folk-concept of blame. Consequently, we suggest to regard the substantial outcome effects on all dependent variables – except punishment – in between-subjects design as performance errors.

4. Study 3: Anchoring Probability

In the next study, we explore whether the hindsight bias and its downstream effects can be mitigated. One way to do this consists in having participants assess the probability of a potentially harmful consequence *ex ante*, that is, before the outcome is revealed. The point of this is to anchor people's probability perception to a level unbiased by outcome, and to explore whether a priming strategy of this sort reduces the hindsight bias in *ex post* assessments of probability, mens rea and culpability.

4.1 Participants

We recruited 199 participants online via Amazon Mechanical Turk. The IP address location was restricted to the USA. As preregistered,[8] participants who were not native speakers of the English language, took less than two minutes to complete the entire survey, failed the attention check or the comprehension question were excluded, leaving a sample of 175 participants (female: 62%; mean age: 42 years, SD = 12 years, range: 19–82 years).

4.2 Methods and Materials

The experimental design was identical to the one familiar from Experiment 1, except for one small change: Participants first read the general scenario and had to rate the objective and subjective probability of a flood that year. Subsequently, outcomes were revealed, and participants had to rate the objective and subjective probabilities again, as well as mens rea (recklessness and negligence) and culpability (blame and punishment), see Figure 9.

[8] https://aspredicted.org/blind.php?x=3kq5bn

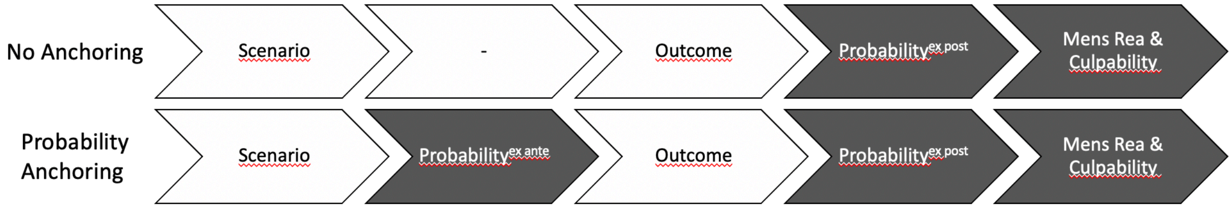

Figure 9: Experimental design for Experiment 1 (no anchoring) and for Experiment 3 (probability anchoring). The question regarding the probabilities of the bad outcome's possible occurrence were asked before the outcome was revealed.

4.3 Results

Expectedly, perceived objective and subjective probabilities *ex ante* (i.e. before the outcome was revealed) across conditions do not differ significantly (independent samples t-test *ps*>.640). We found a significant main effect of outcome (all *ps*<.004) on the moral variables (blame, punishment), mens rea (negligence) and perceived probability, both when assessed in objective and subjective terms, cf. Figure 10 and Table 1 (anchoring). Only for recklessness was the main effect nonsignificant (*p*=. 435).

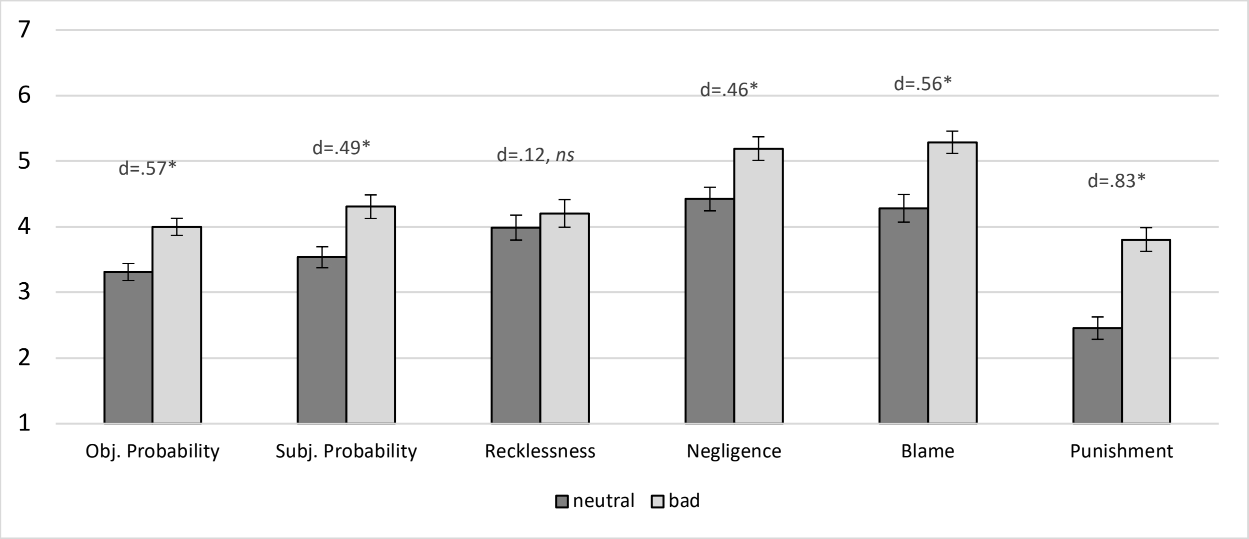

Figure 10: Mean ratings of probabilities, mens rea and moral judgments across outcomes (neutral v. bad). Error bars denote the standard error of the mean.

	No anchoring				Anchoring			
	t(167)	*p*	*Cohen's d*	*95% CI*	*t(173)*	*p*	*Cohen's d*	*95% CI*
Obj. Probability	4.31	<.001	.66	[.45;1.22]	- 3.72	<.001	.57	[-1.06;-.33]
Subj. Probability	5.07	<.001	.78	[.73;1.67]	- 3.23	.001	.49	[-1.24;-.30]
Recklessness	2.14	.034	.33	[.45;1.12]	- .78	.435	.12	[-.76;.33]
Negligence	3.96	<.001	.61	[.48;1.45]	- 3.02	.003	.46	[-1.27;-.27]
Blame	4.53	<.001	.70	[.72;1.83]	- 3.71	<.001	.56	[-1.54;-.47]
Punishment	3.82	<.001	1.18	[1.45;2.45]	- 5.47	<.001	.83	[-1.84;-.86]

Table 1: Effect of outcome on probabilities, mens rea and moral judgment in the no anchoring design (Experiment 1) and with probability anchoring (Experiment 3).

The data for Experiment 3 was purposefully gathered jointly with the data for Experiment 1. Given that people were randomly assigned to one of the four conditions of the two experiments (ballot-box stuffing was prevented), nobody had seen any other condition before. This allowed us to explore whether anchoring reduced the outcome effect on probability judgments in contrast to the results where participants did not have to rate subjective and objective probability ex ante. For none of the six DVs could we find a significant main effect of anchoring (all *ps*>.130), or a significant anchoring*outcome interaction (all *ps*>.089), see Appendix 3.2.2. We do, however, see a small reduction in effect size in the anchoring design for probability ratings in contrast to an anchoring-free design (Experiment 1). Consistent with our previous findings according to which outcome effects on mens rea and moral culpability are mediated by probability, the effect sizes for negligence decrease and turn nonsignificant for recklessness; they also decrease for blame and punishment, see Table 1.

4.4 Discussion

Anchoring, we have shown, is not a quick fix to the distorting effects of outcome on perceived probability, and the latter's downstream effects on mens rea and moral judgment (contrary to the findings of Karlovac & Darley, 1988, and in line with the results of Kamin & Rachlinsky, 1995). Even with anchoring, outcome still has medium-sized effects on both kinds of probability, as well as negligence and blame. Expectedly, the effect size of outcome on punishment remains large even with anchoring, since outcome has a strong direct effect on punishment (see section 4.3).

5. Experiment 4 Counterfactual priming

So far, a number of things have been established: *First*, in between-subjects experiments, outcome has a significant effect on perceived probability, mens rea and moral judgment. *Second*, mediation analyses suggest that the difference in perceived subjective probability drives the asymmetry in the downstream assessment of negligence and blame. *Third*, the hindsight effect must be considered a bias: In within-subjects designs, a significant majority of participants does not draw an inference from outcome to its likelihood, and the differences in mens rea and blame are strongly reduced. *Fourth*, probability anchoring is only moderately successful in mitigating the hindsight bias and its downstream effects.

With this knowledge at hand, we will turn to another potential debiasing strategy. Developing on Experiment 2, we will take a cue from the within-subjects design: Although it basically presents two distinct *actual* outcomes that have come to pass side-by-side, perhaps a similar result can be found when people are simply encouraged to consider the relevant *counterfactuals*. Moral luck experiments by Lench et al. (2015), for instance, suggest, that this strategy can have an effect on moral judgment (probability and mens rea were not tested).

5.1 Participants

396 participants were recruited online via Amazon Mechanical Turk. The IP address location was restricted to the USA. Participants who were not native speakers of the English language, failed the attention check, the comprehension question or took less than two minutes to complete the whole survey (including demographics) were excluded. The remaining sample comprised of 321 participants (female: 47%; mean age: 43 years, SD =12 years, range: 22-88 years).[9]

5.2 Methods and Materials

Lench and colleagues asked participants to imagine *some* alternative outcome. However, content *type* – and in particular the *severity* of the counterfactual outcome – are best controlled tightly. Using the *Flood Scenario*, we thus imitated a design by Spranca et al. (1991), who give two different possible endings to a story (one neutral, one bad), and tell people which outcome actually comes to pass. The experiment thus took a 2 (outcome: neutral v. bad) x 2 (counterfactual priming: no v. yes) design. Participants saw one out of 4 conditions: A story with two endings, one being specified as the actual one (neutral v. bad); or else a story with just a single ending (neutral v. bad).

5.3 Results

Contrasting the results of the neutral v. bad outcome in the plain conditions (i.e. no counterfactual priming) replicates the findings from Experiment 1: We find a significant, and pronounced outcome effect on all DVs (all *ps*<.005, all *ds*>.45), see Appendix 4.3.1 and Figure 11. A series of 2 outcome (neutral v. bad) x 2 priming (yes v. no) ANOVAs revealed a significant main effect of outcome on all the dependent variables (Table 2),[10], a nonsignificant main effect of priming (all *ps*>.169) except for recklessness (*p*=.022). The outcome*priming interactions were significant for subjective probability (*p*=.006), the two types of mens rea (*ps*<.040), and punishment (*p*=.003).

[9] Though we originally planned to report each outcome as a separate experiment, for ease of exposition we'll report them together. The preregistration links are https://aspredicted.org/blind.php?x=e2hq4x and https://aspredicted.org/blind.php?x=n8iq3t. Participants who took both surveys were excluded. For detailed documentation, see Appendix, Section 4.

[10] For details, see Appendix, Section 4.3.1.

The effect sizes for subjective probability and punishment were substantial (η_p^2>.22). Figure 11 graphically represents the findings.

	outcome				*priming*				*outcome*priming*			
	df	F	*p*	η_p^2	*df*	F	*p*	η_p^2	*df*	F	*p*	η_p^2
Obj. probability	1	13.88	<.001	.042	1	3.28	.170	.006	1	1.27	.261	.004
Subj. probability	1	22.14	<.001	.065	1	1.55	.433	.002	1	7.61	.006	.023
Recklessness	1	4.59	.033	.014	1	17.63	.022	.016	1	4.31	.039	.013
Negligence	1	10.14	.002	.031	1	3.15	.322	.003	1	4.68	.031	.015
Blame	1	39.86	<.001	.112	1	4.35	.251	.004	1	3.40	.066	.011
Punishment	1	52.68	<.001	.143	1	10.45	.059	.011	1	9.21	.003	.028

Table 2: Results of the 2 *outcome* (neutral v. bad) x 2 *priming* (yes v. no) ANOVAs for probabilities, mens rea and moral judgment.

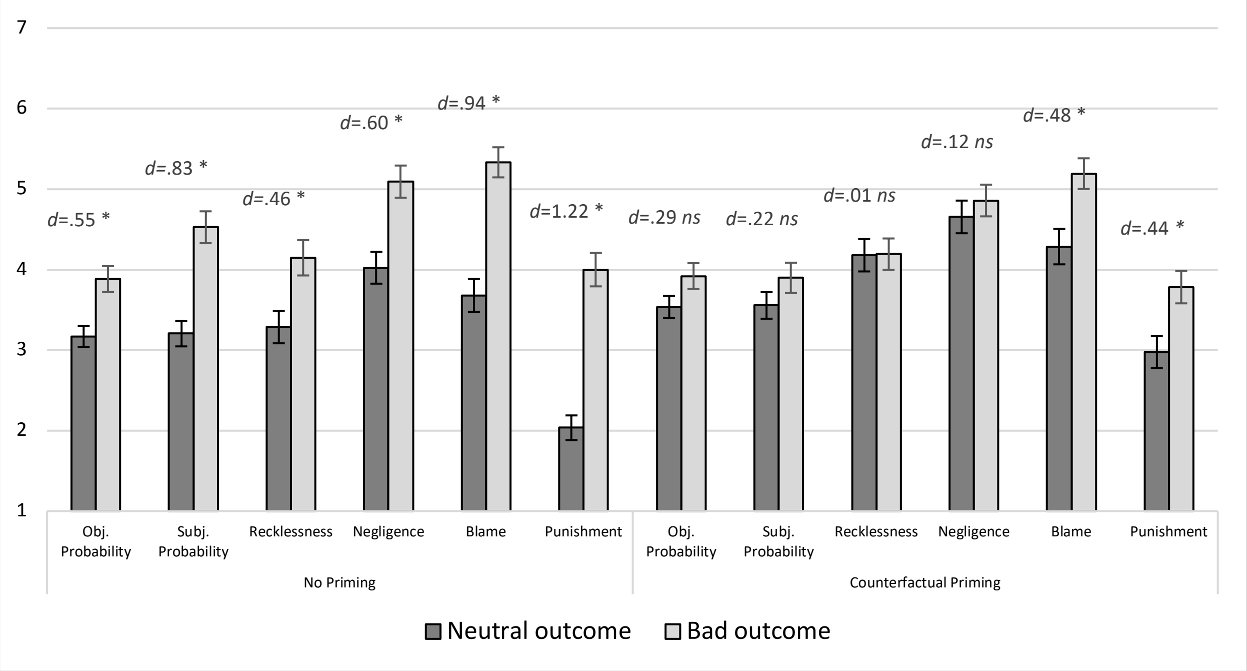

Figure 11: Mean ratings for probabilities, mens rea and moral judgment across outcomes for the priming and no priming conditions. Effect sizes are given in terms of Cohen's *d*, significance is reported at the *p*<.05 threshold (for details, see Appendix section 4.3.1). Error bars denote standard error of the mean.

Counterfactual priming decreases the difference across outcomes (Figure 11, for details see Appendix 4.3.1), rendering the outcome effect nonsignificant for all variables, except for blame and punishment (both *ps*=.002). Importantly, however, counterfactual priming also decreases the effect size of outcome on moral judgment dramatically in comparison to the between-subjects design (for punishment from *d*=1.22 to *d*=.44, for blame from *d*=.94 to *d*=.48).

5.4 Discussion

Asking people to imagine a counterfactual outcome strongly reduces the outcome effect on blame and punishment, and renders it nonsignificant for objective and subjective probability, recklessness and negligence. Counterfactual priming does not completely eradicate the outcome effect of either moral variable tested, though the effect is much smaller for both blame and punishment. What is notable again is that, despite the fact that the folk concept of punishment does seem outcome-dependent (see Experiment 2), there might yet be a bias in the *extent* to which outcome drives punishment judgments when unchecked. As in the within-subjects design, entertaining counterfactual consequences reduces the size of the outcome effect on punishment by more than half (from *d*=1.22 in a between-subjects design to *d*=.44).

As discussed (and graphically represented in Figure 1), our first two debiasing experiments attempted to reduce the impact of outcome on perceived probability *ex post*. Differently put, we explored *indirect* mechanisms to reduce the hindsight bias and its downstream effects on mens rea attribution and moral judgment. In certain contexts, however, one could attempt to *directly* influence perceived probability *ex post*. The evident way to do this is by consulting an expert. The question then arises whether probability *stabilizing* of this sort does indeed mitigate the outcome effect on mens rea and moral judgment, as our mediation results (as well as those reported by Kneer & Machery, 2019) would suggest. To this final debiasing strategy we now turn.

6. Experiment 5: Stabilizing Probability

Many of our decisions are characterized by *uncertainty* – that is, undertaken in circumstances where it is impossible to quantify the probabilities of an event (be it *ex ante* or *ex post*). In *contexts of risk*, by contrast, the relevant probabilities of an event's coming to pass can, at least in principle, and roughly, be specified ex ante. In situations where serious harm has occurred – i.e. cases that tend to end up in court – one might thus want to consult an expert to determine the probability of harm engendered by the agent's actions. Although not a standard procedure in risk-related cases in court, the law *sometimes* resorts to experts, e.g. for assessing risk of harm in road traffic offenses or recidivism of those with mental health disorders (cf. Herring 2012; Fletcher 2000). Given the robustness of the hindsight bias, and the fact that it is quite resistant against certain debiasing strategies such as the above-reported probability *anchoring* (Experiment 3), our final experiment explores whether *probability stabilizing by expert testimony* is indeed a promising strategy to keep creeping determinism and its downstream consequences at bay.

6.1 Participants

238 participants were recruited online via Amazon Mechanical Turk. The IP address location was restricted to the USA. As preregistered,[11] participants who were not native English speakers, failed the attention check, the comprehension question, or took less than two minutes to complete the whole survey (including demographics), were excluded. The remaining sample comprised of 169 participants (female: 47%; mean age: 42 years, SD =11 years, range: 22-74 years).

[11] https://aspredicted.org/blind.php?x=ve5ei4

6.2 Methods and Materials

We used the same scenario as in Experiment 1: Ms. Russel did not install temporary flood barriers to protect her workers' homes so as to refurbish their kitchens instead. In the original version, participants were presented with either the neutral or the severe outcome, and then asked to rate probability, mens rea and culpability. But in this version, both groups were presented with the following additional information before responding to the questions:

> The case of Ms. Russel not installing the temporary flood barriers is brought to court. An expert witness states that there was a 5% chance that there would be a flood this year.

The questions, focusing on objective and subjective probability, mens rea and moral judgment, were the same as in the previous experiments (full details in the Appendix 5.2.1).

6.3 Results

Probability stabilizing via expert testimony works: When there is an explicit specification of the flood's likelihood at the context of action, people view objective probability identically across outcomes (*p*=.487, *d*=.10), and the same holds for subjective probability (*p*=.074, *d*=.25), see Figure 12 (detailed test results in Appendix 5.2.1). Consistent with the mediation analyses from Experiment 1, ensuring that perceived probability is fixed across conditions cancels out the outcome effect on the two types of mens rea (recklessness: *p*=.853, *d*=.03; negligence: *p*=.094, *d*=.28). Expectedly, the ratings for punishment, a DV which is strongly and directly sensitive to outcome, remained significant across conditions (*p*<.001, *d*=.63). Less expectedly, the outcome effect on blame *also* remained significant (*p*=.017, *d*=.38), however its effect size was small. Notably, the effect size for both moral DVs was cut *in half* by probability stabilizing.

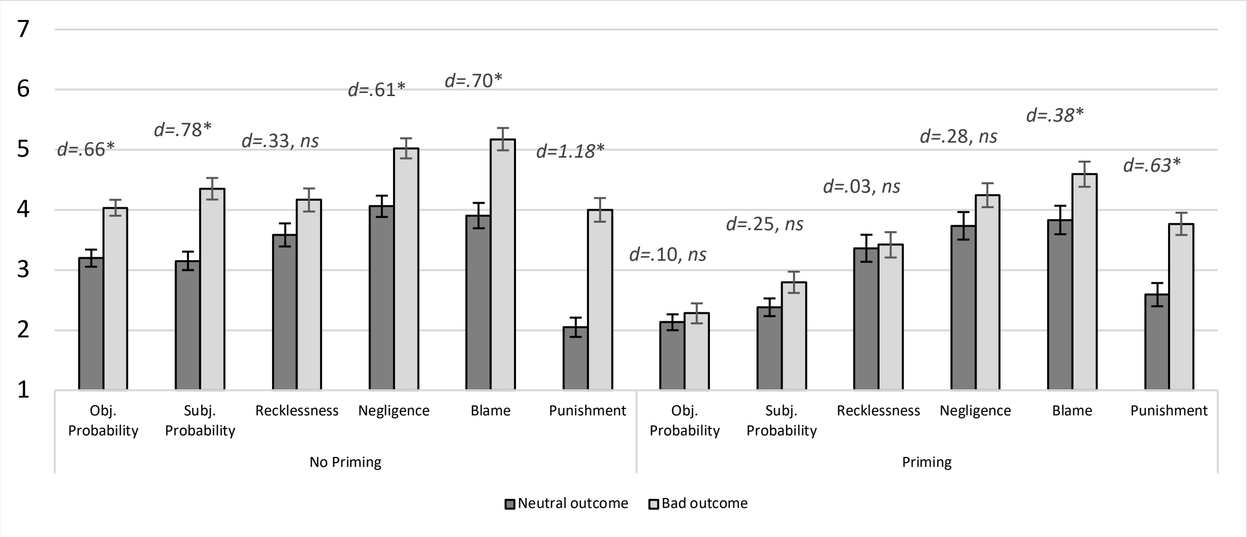

Figure 12: Mean ratings for probabilities, mens rea and moral judgment across outcomes for the priming and no priming conditions. Effect sizes are given in terms of Cohen's *d*, significance is

reported at the *p*<.05 threshold (for details, see Appendix 5.2.1). Error bars denote standard error of the mean.

6.4 Discussion

An expert assessment of the actual *ex ante* probability of a harmful outcome cancels out the hindsight bias. Since (at least in the experiments at hand) it is distorted post-hoc probability that mediates the outcome effect on mens rea, we would expect that these, too, are now assessed identically across conditions. And indeed they are – we found no significant difference across negligence or recklessness ascriptions. As predicted, judgments of deserved punishment differed significantly across outcomes even after probability stabilizing. Somewhat astonishingly, blame was *also* significant across outcomes (neutral v. bad), though this effect cannot be due to diverging assessments of probability or mens rea. Blame, this suggests (and the mediation analysis from Experiment 1 does, too), is to some, relatively small, extent also directly sensitive to outcome (at least in a between-subjects experiment of this sort). Note, however, that for both punishment and blame probability stabilizing reduced the outcome effect by about 50%, and the remaining effect of outcome on blame was small (*d*=.38). As in the previous experiments, the story regarding punishment replicates: The folk concept of punishment, we said, is outcome-sensitive (Experiments 1 and 2). However, it is likely that folk judgments of punishment can easily fall prey to a bias when it comes to the *extent* to which outcome information is taken into consideration. Probability stabilizing offsets much of that bias, and thus reduces the impact of outcome on punishment significantly in contrast to a standard between-subjects design.

7. Replications

For ease of exposition, we have worked with a single root scenario throughout this paper. However, we have run the entire suite of experiments with a different scenario, keeping all the parameters (question phrasing, design, exclusion criteria etc.) the same. In the vignette, adapted from Spranca et al. (1991, Experiment 1, p. 83), John tests a recently fixed car on a standardly deserted highway, speeding through an intersection. In the neutral outcome condition, no other cars are in sight. In the bad outcome condition, John hits another car and injures the driver. The questions once again focused on subjective and objective probability, negligence, recklessness, blame and punishment. Full details of the scenario, questions and results are provided in the Appendix (sections 6-10). Since pretty much everything replicated perfectly, we'll here limit ourselves to a short overview of the findings. To facilitate a quick grasp of the results for the reader, we have produced two figures that graphically represent the effect sizes in terms of Cohen's *d* and state significance across conditions for all five experiments. Figure 13 reports outcome effects on perceived probabilities and mens rea, Figure 14 reports outcome effects on perceived probabilities, blame and punishment.

Replicating Experiments 1 and 4 (between-subjects data, see Appendix 7.3.1), we found a significant impact of outcome for all DVs except recklessness (*p*=.769), and similar effect sizes. The mediation analyses also replicated well (see Appendix, section 6.3.2-6.3.3): A serial mediation model suggests that the effect of outcome on blame travels entirely via subjective probability first and negligence thereafter (the *individual* mediating paths of subjective probability and negligence proved nonsignificant).[12] As in Experiment 1, once mediation is taken into account, the effect of outcome on blame turns nonsignificant. Also replicating the findings from Experiment 1, the mediation analyses for punishment differed from blame in two regards: First, it was *objective* probability which played a role (directly and indirectly via negligence) whereas *subjective* probability did not. Second, most of the effect – about three quarters – of outcome on punishment is direct. Once again, these findings confirm Cushman's proposal that there are two different processes of moral judgment, one more dependent on mental factors (of which subjective probability is a part), and one more dependent on causal factors (objective probability and outcome *per se*).

Experiment 7 successfully replicated Experiment 2, in which we explored whether between-subjects outcome effects on probability, mens rea and blame (though not punishment) are best understood as a bias. When people see both outcomes side-by-side, the rationale was, they are aware that they differ only in terms of outcome, and will only assign different probabilities, mens rea or blame if they think the latter *should* be sensitive to outcome. Experiment 7 confirmed that, by and large, they do *not* think the respective DVs should be sensitive to outcome. Though some variables just made the significance threshold, all effect sizes were very small (all *ds*<.28). Importantly (and as in Kneer & Machery, 2019), the effects were driven by a small minority of participants, since the vast majority judged the two situations (neutral v. bad) identically with respect to objective probability (72%), subjective probability (78%), recklessness (93%), negligence (90%) and blame (86%). Only punishment proved – expectedly – outcome-sensitive properly conceived. There was a significant effect of outcome (*p*<.001, *d*=.74), and only 48% of the participants judged the two agents as deserving the same punishment.

Replicating Experiment 3, having people reflect on the probabilities before outcomes were revealed decreased the outcome effect. In fact, anchoring worked a little better than in Experiment 3, since the effect of outcome turned nonsignificant for all but the moral variables. Anchoring reduced the outcome effect on punishment (from *d*=1.45 to *d*=.91), whereas the effect remained roughly the same for blame (*d*=.52 v. *d*=.68 with anchoring).

Following Lench et al. (2015), Experiment 9 explored whether asking people to entertain an alternative outcome is a helpful strategy to mitigate the hindsight bias. Once again, we found that it is: The outcome effect on both types of probability, negligence and blame disappears entirely (all *ps*>.155). As expected, punishment remained significant (*p*<.001, *d*=.91). Curiously, recklessness was also significant (*p*=.013, *d*=.42). Given that recklessness was not significant in any of the other *Intersection* experiments, including the between-subjects design (*p*=.769, *d*=.05), we think this might perhaps just be an oddity in the data.

Finally, we reran the probability stabilizing strategy explored in Experiment 5 with the *Intersection* scenario (Experiment 10). Replicating the exact same pattern which we found in

[12] In Experiment 1, subjective probability picked up a bit of the indirect effect by itself, but the bulk of the mediation occurred via probability-and-negligence (i.e. the a^1db^2 path), as in Experiment 6.

Experiment 8, probabilities and negligence turned nonsignificant (all *ps*>.149). Once again, blame remained significant (*p*=.018) and – expectedly – so did punishment (*p*<.001). Importantly, though, here too, the effect sizes decreased in comparison to the between-subjects (priming-free) experiment for blame (no priming: *d*=.52, probability stabilizing. *d*=.40) and punishment (no priming *d*=1.45, probability stabilizing *d*=1.12).

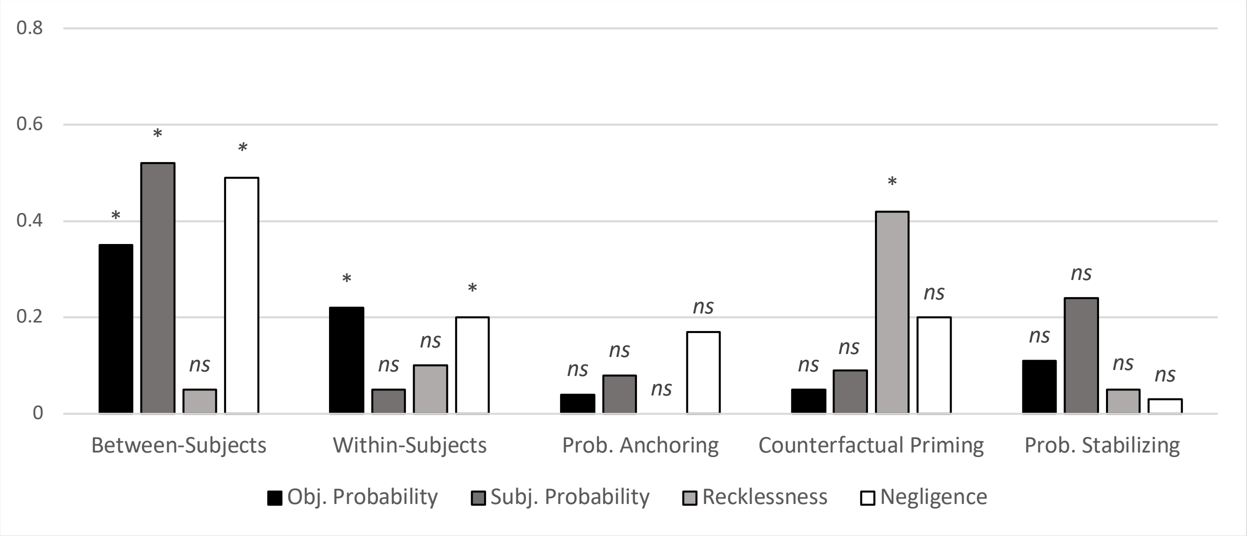

Figure 13: Effect sizes and significance of the difference between the assessment of perceived probabilities and mens rea across outcomes (neutral v. bad) for all five Intersection experiments. Effect sizes are given in terms of Cohen's ds, significance is reported at the *p*<.05 threshold.

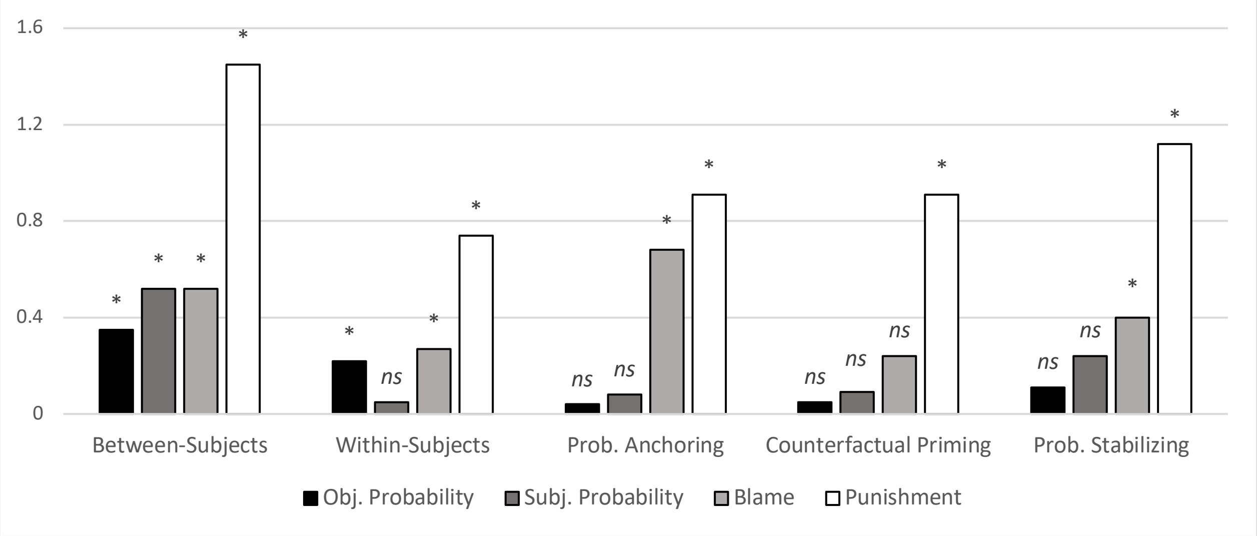

Figure 14: Effect sizes and significance of the difference between the assessment of probabilities, blame and punishment across outcomes (neutral v. bad) for all five Intersection experiments. Effect sizes are given in terms of Cohen's ds, significance is reported at the *p*<.05 threshold.

8. General Discussion

8.1 Outcome Effects on Punishment and Blame

Across all ten experiments, punishment proved strongly sensitive to outcome. This is consistent with previous findings (Cushman, 2008; Martin & Cushman, 2015, 2016, Kneer & Machery, 2019) and suggests that the folk *concept* of punishment is outcome-dependent: Even in our within-subjects designs, the severity of the consequence is a decisive factor in adjudicating deserved punishment. The effect of outcome on blame, by contrast is nonsignificant (Experiment 2) or marginal (Experiment 6, *d*=.27) in within-subjects designs. From this we can draw several conclusions.

First, the findings confirm Cushman's *Dual Process Model of Moral Judgment*. There are two distinct moral processes, one which is fundamentally more sensitive to causal factors such as outcome (judging deserved punishment), and another which is fundamentally less sensitive to them (blame) yet more sensitive to mental factors (see mediation analyses in Experiments 1 and 6).[13]

Second, the fact that the pronounced between-subjects outcome effect on blame disappears by and large in the within-subjects design suggests it's a bias. While this is commonly claimed,[14] few authors back this claim up convincingly. Our within-subjects data demonstrates that the folk *concept* of probability is *not* outcome-dependent, and that the effect of outcome in between-subjects data constitutes a distortion even from the folk perspective (not just from a perspective of rational choice theory or some such).

Third, and in line with previous arguments (Kneer & Machery, 2019; Frisch et al. ms), there is no philosophical puzzle of moral luck. What puzzles philosophers is that, on the one hand, we do *not* want to hold people morally responsible for consequences beyond their control. On the other, however, we allegedly blame unlucky agents more than lucky ones when directly contrasting the two cases (Nagel, 1979, Williams, 1981; Hartman,2017, for a review see Nelkin, 2006). But – as the within-subjects design data shows – most of us simply do not blame the two agents differently (for similar results see e.g. Nichols, 2009; Schwitzgebel & Cushman, 2012; Lench et al. 2015, Kneer & Machery, 2019). So there's no puzzle. Or is there? Perhaps on the basis that, even in within-subjects designs, we find a strong outcome effect on *punishment* (Kumar, 2019, for instance always speaks of "blame *and punishment*" in the same breath)? We doubt it. As argued by Enoch & Marmor (2007), not just any vaguely "blame-related" moral variable is suited to get a substantial philosophical puzzle of moral luck off the ground. Punishment has a host of pragmatic functions (e.g. the deterrence of potential offenders, as well as the incapacitation and/or

[13] Cushman (2008) tested four types of moral judgment: Wrongness and permissibility of an action, as well as the blame and punishment the agent deserves. In his experiments, wrongness and permissibility are predominantly sensitive to mental states, whereas blame and punishment are also strongly influenced by causal factors, notably outcome. Here, as in Kneer & Machery (2019), blame seems to fall on the mental, rather than the causal, side of the fence. The difference could be due to the formulations of the blame question ("is blameworthy" v. "deserves blame"), or its focus (it can focus on agent, action or consequence), a topic which merits further investigation (see Prochownik & Cushman, ms; Björnson & Kneer, ms).

[14] For work on the hindsight bias, see e.g. Fischoff, 1977; Alicke, 1989, 1994, 2008; Arkes et al 1981; Arkes et al 1988; Bukszar & Conolly, 1988; Spranca et al.1989; Hawkins & Hastie, 1990; LaBine & LaBine, 1996; Bradfield & Wells, 2005, Wood, 1978; Rachlinski, 1998, 2000; Baron, 2006, Blank et al. 2008, Ackerman et al. 2020.

rehabilitation of previous offenders, see e.g. Duff, 2001) that go beyond moral assessment, and it is quite likely those factors that make the concept of punishment sensitive to outcome.[15]

8.2 Alleviating the outcome bias

We have argued that the folk-concept of punishment is outcome-sensitive, whereas the concept of blame is not. In ordinary life situations and in court, outcome information might thus distort ascriptions of moral or legal culpability. What, exactly, is it that drives the outcome effect? Consistent with previous findings, the mediation analysis suggests that the outcome effect on culpability is in large parts a consequence of the hindsight bias (see Kamin & Rachlinski, 1995, Rachlinski, 1998, 2000; Kneer & Machery, 2019): Participants view the subjective and objective likelihood of possible events that actually come to pass as higher ex post than those that do not come to pass. In virtue of the higher perceived risk in the unlucky cases (where the harm does occur), people judge the agents as more negligent and (sometimes) more reckless. Consequently (and reasonably), they deem the agent who is viewed as acting more negligently as more blameworthy. Once subjective probability and negligence are accounted for as mediators, however, no direct effects of outcome on blame remain.

We have explored three distinct ways to alleviate the hindsight bias and its downstream effects on mens rea ascription and blame. What worked best was *probability stabilizing*. Lawsuits where the focus lies on the question whether the agent should have avoided a *substantial* risk, for instance cases of medical malpractice,[16] sometimes employ experts to establish whether there was a substantial risk in the first place (and how pronounced it was). We tested explicit probability stabilizing and found that it blocks the asymmetric assessments of the perceived likelihood of a harmful outcome across cases. Once the hindsight bias is thus stopped in its tracks, the outcome effect on mens rea disappears entirely, and only small direct effects of outcome on blame remain (*Flood* d=.38, Intersection d=.40). Interestingly, the effect of outcome on punishment is also reduced for both scenarios (in *Flood* by about 50%, though less in *Intersection*).

An alternative strategy we tested was consulting people on probability ex ante (i.e. before the outcome was revealed), so as to *anchor* their perceived probabilities by estimates not yet distorted by outcome. This reduces the effect size of outcome both on subjective probability and

[15] Naturally, judgments of deserved punishment have *some* moral component. But note that the effect sizes across lucky v. unlucky conditions in the within-subjects design, as well as some of the debiasing strategies are only about half as pronounced as in the between-subjects design.

[16] See e.g. Arkes et al., 1981; Cohen 2004; Johnston 2013. Some medical malpractice legal cases of interest are: *Johns Hopkins v. Genda*, 1969 (the court stated that without expert evidence the defendant cannot be convicted and proclaimed him innocent); *Claar v. Burlington*, 1994 (discussing which expert testimony is admissible); *Ambrosini v. Labaraque*, 1996 (the court stated that expert evidence was defective and assessed standards of such evidence); *Navarro v. Austin* 2006 (expert witness testimony helped plaintiff receive one of the largest compensations in history); *Griffen v. Univ. of Pittsburgh Med. Ctr.-Braddock Hospital*, 2008 (discussing what is the probability threshold established by an expert witness which evidence must "reach" in a medical malpractice case, arguing that 51% is not enough); *Day v. Bryant*, 2010 (arguing that it is not enough that an expert establishes that harm is 'more likely than not'). For an assessment of the practice of expert witnesses in medical malpractice cases and a clarification of the guidelines and responsibilities of expert witnesses as well as independent medical evaluators cf. Masella & Meister, 2001; Friston, 2005; Hammond & Schwarz, 2005; Schofferman, 2007.

objective probability (in *Intersection* they turn nonsignificant, in *Flood* they remain significant), and consequently on mens rea (only negligence in *Flood* remains significant, but the effect is small, *d*=.26). In the *Flood* scenario, these decreases in effect size on probability and mens rea trickle down to blame, though not in *Intersection.*

The impact of probability anchoring was significant, though limited. Taking inspiration from Lench et al. (2015), we explored whether entertaining counterfactuals reduces the hindsight bias and its downstream effects on mens rea and blame. The results suggest it does: After counterfactual priming, we can no longer detect a significant difference in objective and subjective probability across cases, or in negligence ascriptions with either scenario. For the *Intersection* scenario, blame, too, turns nonsignificant; for *Flood*, the effect remains significant, though decreases in size *vis-à-vis* the between-subjects experiment.

Taking stock: Although the hindsight bias is robust, pervasive and its consequences can be daunting (for a review, see Rachlinski, 1998), there are measures that can be taken. Whereas the practical import of probability anchoring is small (since it is hard to effect, e.g. in a court case), both probability stabilizing and prompting people to entertain alternative outcomes hold a lot of promise. They block the hindsight bias, the distorted ascription of risk-related types of mens rea (negligence and recklessness) and decrease the outcome bias on blame substantially or cancel it out. Interestingly, in all three bias alleviation experiments, the impact of outcome on punishment is *also* reduced, though a pronounced, and significant effect remains. What this suggests is the following: Although the folk concept *is* sensitive to outcome, and although the asymmetric punishment attributions across cases do thus not constitute a bias *per se*, its *size* might be the consequence of a partial outcome bias. In the within-subjects designs, the effect of outcome on punishment is only about half that of the between-subjects designs, and the alleviation strategies manage to correct the effect downwards to a substantial degree.

8.3 Implications for the Law

The scenarios here tested are negligence cases, and the data suggests that the folk understands them as such (cf. two interesting studies on folk understanding of mens rea terms by Shen et al. 2011 and Ginther et al. 2014). In negligence cases, the question is whether the agent *should have been* aware of a substantial risk of harm or not (see e.g. Model Penal Code, 2.02. (d), and for discussion of negligence more generally, Hall 1963, Hart, 1968, Fletcher, 1971, Simons, 1994, Hurd & Moore, 2002, 2011, King, 2009, Raz, 2010, and Husak, 2011). The law is explicit about the fact that what matters for the assessment of negligence is the risk as assessed from the point of view of a reasonable person at the *context of action*, not the risk as it appears post-hoc, once it is clear what turn the events have taken (for discussion, see Rachlinski, 1998, and Kneer, forthcoming). The hindsight bias, and the here demonstrated pronounced distortive effects on perceived negligence and culpability are thus a serious problem from the legal point of view. Since in many countries, such as *inter alia* the US and the UK, the mens rea question is decided by lay jurors, serious precautions should be taken to minimize the hindsight bias. Our examination of different debiasing strategies constitutes a first step in the quest for offsetting the systematic performance error afflicting probability judgments post hoc, and the unjust rulings they are likely to engender.

One note of caution is, however, in order. Given the powerful influence an expert assessment of ex ante likelihood exerts on mens rea and culpability judgments, it must be used with extreme care and the procedural conventions for choosing such experts might require more attention.[17]

Most of case law in which expert witness is decisive pertains to medical malpractice.[18]For US case law, the two landmark cases where expert witness evidence was decisive are Frye v. US (1923) and Daubert v. Merell Dow Pharmaceuticals (1993). These two cases lead to the formulation of general standards of acceptability of expert witness evidence and influenced thousands of later cases, yet perhaps the standards need to be carefully reconsidered in light of our results. The Frye standard claims that expert witness evidence is admissible only if based on generally accepted theories in the scientific community. By contrast, the Daubert standard, which replaced the Frye standard, states that it is the judge who decides on which evidence shall prevail. If our results are to be taken seriously, then the Frye standard appears to be more reliable, since expert witness evidence can be extremely suggestive, and the judgment of the entire scientific community could be taken as more reliable than the judgement of a single person.

8.4 Future Research

Whereas the hindsight bias is well established, this paper is among the first (i) to examine its downstream effects and their inherent "mechanics" in detail, and (ii) to propose strategies to alleviate the systematic performance error. Further research should examine whether the results replicate with different scenarios, methods, alternative formulations of what we termed "subjective" and "objective" probability, and across different populations – in particular non-WEIRD populations (see e.g. Barrett et al. 2016). There is, for instance, considerable evidence of a strong, cross-cultural effect of outcome on ascriptions of intention and knowledge (Kneer et al. ms). Moreover, legal experts from France (Kneer & Bourgeois-Gironde 2017, Bourgeois-Gironde & Kneer, 2018), Germany (Prochownik et al. 2020), the Netherlands, Brazil, and the UK seem similarly affected as the folk (Kneer et al. ms; though cf. Tobia, ms. for diverging findings for the US). It thus stands to reason to explore whether the effects of outcome on the lower echelons of inculpating mental states – negligence and recklessness – are similarly robust across cultures and expertise. If so, this would suggest a systematic distortive effect of outcome information on mens

[17] Importantly, there are procedural differences across legal systems in who can choose and present an expert witness testimony in court. In adversarial systems where the judge has a limited procedural role (e.g. the US and the UK), expert witnesses are presented exclusively by the parties. By contrast, in inquisitorial systems (mainly continental Europe), the role of the judge in a trial is more active. Here, it is the judge who can choose and ask an expert witness to present an expertise.

[18] Notorious cases where expert testimony was controversial include in the UK '*John Radford* (formerly known as John Worboys) *versus The Parole Board of England and Wales*' (2018); '*Regina versus Georgina Sarah Anne Louise Challen*' (2019); '*Regina versus Sally Clark*' (2003); '*Guinness Plc versus Ernest Saunders Plc*' (1990). In these cases, expert witness testimony pertained to assessing either medical evidence concerning the victims or assessing the mental health of the perpetrator.

rea ascription of *any* kind. The possible threat to just legal ruling – not limited to countries with lay juror systems – would motivate serious exploration of debiasing strategies along the lines here proposed and beyond.

Conclusion

In a series of experiments with 2043 participants, we explored the effect of outcome on judgments of subjective and objective probability, mens rea and culpability. For mens rea and blame attributions (though not for punishment), the outcome effect constitutes a bias. The distorted assessment of mens rea and blame, we showed, is ultimately rooted in the hindsight bias: People tend to assess a potential, harmful outcome as more likely when it does come to pass than when it does not; they therefore ascribe more negligence to the agent, and consequently consider him more culpable.

Echoing the literature from behavioral economics and legal psychology, we argued that the downstream effects of the hindsight bias constitute a serious threat to the just adjudication of legal trials, in particular in countries in which mens rea is determined by lay juries (such as the US and the UK). And although it is well established that the hindsight bias is pervasive and difficult to overcome, we have shown that there *are* measures to reduce its impact. Among a series of different debiasing strategies we have put to the test, we showed that expert probability stabilizing (which, on occasion, is already in use in courts) and entertaining counterfactual outcomes hold considerable promise. We would strongly urge further research conducted jointly with legal practitioners that explores the most suitable ways of introducing (or further implementing) these techniques in the courtroom, so as to make the law more just and fair for all.

Bibliography

Ackerman, R., Bernstein, D.M. & Kumar, R. (2020) Metacognitive hindsight bias. *Memory and Cognition* 48, 731–744. https://doi.org/10.3758/s13421-020-01012-w

Agans R. P. & Shaffer L. S. (1994) The Hindsight Bias: The Role of the Availability Heuristic and Perceived Risk. *Basic and Applied Social Psychology*, 15:4, 439-449. DOI: 10.1207/s15324834basp1504_3

Alfano, M., Beebe, J. R., & Robinson, B. (2012). The centrality of belief and reflection in Knobe-effect cases: A unified account of the data. *The Monist*, *95*(2), 264-289.

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology*, 66(2), 252–254. https://doi.org/10.1037/0021-9010.66.2.252

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, 73(2), 305–307. https://doi.org/10.1037/0021-9010.73.2.305

Arkes, H. R., & Schipani, C. A. (1994). Medical malpractice v. the business judgement rule: Differences in hindsight bias. *Oregon Law Review*, 73(3), 587-638.

Alicke, M. D.& Davis, T. L., (1989). The role of a posteriori victim information in judgments of blame and sanction. *Journal of Experimental Social Psychology,* 25(4), 362-377.

Alicke, M., Davis, T. L., & Pezzo, M. V. (1994). A posteriori adjustment of a priori decision criteria. Social Cognition, 12(4), 281–308.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126(4), 556–574. https://doi.org/10.1037/0033-2909.126.4.556

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

American Law Institute. (1985). Model penal code: official draft and explanatory notes: complete text of the model penal code as adopted at the 1962 annual meeting of the American Law Institute at Washington, D.C., May 24, 1962. Philadelphia.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., & Scelza, B. A. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences, 113*(17), 4688–4693.

Baron, J. (2000). *Thinking and deciding.* Cambridge University Press.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, 94(2), 74–85. https://doi.org/10.1016/j.obhdp.2004.03.003

Baron, J. (2006). Descriptive theory of probability judgment. In *Thinking and Deciding* (pp. 137-160). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511840265.009

Björnsson, G. & Kneer, M. (ms) "The folk concept of blame reexamined".

Blank H., Nestler S., von Collani G., Fischer V. (2007). How many hindsight biases are there?. *Cognition*, 106(3), 1408-1440. 10.1016/j.cognition.2007.07.007

Bodenhausen, G. V. (1990). Second-guessing the jury: Stereotyping and hindsight biases in perceptions of court cases. *Journal of Applied Social Psychology*, 20, 1112-1121.

Bourgeois-Gironde, S. & M. Kneer (2018). Intention, cause et responsabilité: *Mens rea* et effet Knobe. In Ferey, S. & F. G'Sell (eds.), *Causalité, responsabilité et contribution à la dette*. Editions Brylant.

Bradfield, A. & Wells, G. (2005). Not the same old hindsight bias: Outcome information distorts a broad range of retrospective judgments. *Memory & Cognition*, 33(1), 120-130. https://doi.org/10.3758/BF03195302

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 101–117.

Buchman, T. A., (2002). An effect of hindsight on predicting bankruptcy with accounting information. *Accounting, organisations and society,* 10(3), 267-285. https://doi.org/10.1016/0361-3682(85)90020-0

Bukszar, E., & Connolly, T. (1988). Hindsight Bias and Strategic Choice: Some Problems in Learning from Experience. *The Academy of Management Journal*, 31(3), 628-641. Retrieved November 5, 2020, from http://www.jstor.org/stable/256462

Casper, J. D., Benedict ,K., & Perry, J. L. (1989).Juror decision making, attitudes, and the hindsight bias. *Law and Human Behavior*, 13, 291-310.

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 48(1), 147–168.

Cohen, F. (2004). The expert medical witness in legal perspective. *Journal of Legal Medicine*, 25(2), 185-209. 10.1080/01947640490457479

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a “trembling hand” game. *PloS one*, 4(8), e6699.

Chartered Institute of Arbitrators, (2020). Party Appointed and Tribunal Appointed Experts. Available at https://www.ciarb.org/media/4200/guideline-7-party-appointed-and-tribunal-appointed-expert-witnesses-in-international-arbitration-2015.pdf

Dawson NV, Arkes HR, Siciliano C, Blinkhorn R, Lakshmanan M, Petrelli M. (1988). Hindsight bias: an impediment to accurate probability estimation in clinicopathologic conferences. *Medical Decision Making*. 8(4), 259-64. doi: 10.1177/0272989X8800800406. PMID: 3185178.

Duff, A. (2001). *Punishment, communication, and community*. Oxford University Press.

Enoch, D., & Marmor, A. (2007). The case against moral luck. *Law and Philosophy*, 26(4), 405–436.

Feltz, A. (2007). The Knobe effect: A brief overview. *The Journal of Mind and Behavior*, 265-277.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, 3(2), 349–358. https://doi.org/10.1037/0096-1523.3.2.349

Fischhoff, B. (1980). For those condemned to study the past: Reflections on historical judgment. *Decision Research*.

Fletcher, G. (1971). The Theory of Criminal Negligence: A Comparative Analysis. *University of Pennsylvania Law Review*, 119(3), 401-438. doi:10.2307/3311308

Fletcher, G. P. (2000). *Rethinking criminal law* (reprint). Oxford University Press.

Frisch, L. K., Kneer, M., Krueger, J. I., & Ullrich, J. (ms). Do You Feel the Same? The Effect of Outcome Severity on Moral Judgment and Interpersonal Goals of Perpetrators, Victims, and Bystanders.

Friston M. (2005). Roles and responsibilities of medical expert witnesses. *BMJ* (Clinical research ed.), 331(7512), 305–306. https://doi.org/10.1136/bmj.331.7512.305

Gilbert, E., Tenney, E. R., Holland, C. and Spellman, B. A. (2014). Counterfactuals, Control, and Causation: Why Knowledgeable People Get Blamed More. *Personality and Social Psychology Bulletin*, vol 41(5), 643-658. http://dx.doi.org/10.2139/ssrn.2463520

Gino, F., Moore, D. A., & Bazerman, M. H. (2009). No harm, no foul: The outcome bias in ethical judgments. *Harvard Business School NOM Working Paper*, 08–080.

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior.

Ginther, M. R., Shen, F. X., Bonnie, R. J., Hoffman, M. B., Jones, O. D., Marois, R., & Simons, K. W. (2014). The language of mens rea. *Vand. L. Rev.*, *67*, 1327.

Hall, J. (1963). Negligent Behavior Should Be Excluded from Penal Liability. *Columbia Law Review*, 63(4), 632-644. doi:10.2307/1120580

Hammond C, Schwartz P. (2005) Ethical issues related to medical expert testimony. *Obstetrics and Gynecology*, 106, 1055–8.

Hart, H. L. A. (1968). Negligence, mens rea and criminal responsibility. In *Hart, punishment and responsibility: Essays in the philosophy of law*. Oxford: Oxford University Press.

Hartman, R. J. (2017). *In defense of moral luck: Why luck often affects praiseworthiness and blameworthiness* (Vol. 38). Taylor & Francis.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, 107(3), 311–327. https://doi.org/10.1037/0033-2909.107.3.311

Herring, J. (2010). *Criminal law: the basics* (first edition). Routledge.

Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review*, 104(1), 194–202. https://doi.org/10.1037/0033-295X.104.1.194

Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4), 605–619. https://doi.org/10.1037/0278-7393.15.4.605

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.

Hsee, C. K., Zhang J., Distinction Bias: Misprediction and Mischoice due to Joint Evaluation. *Journal of Personality and Social Psychology*,86(5), 680-695. 10.1037/0022-3514.86.5.680

Hurd, H. M., & Moore, M. S. (2002). Negligence in the Air. *Theoretical Inquiries in Law*, 3(2). doi: https://doi.org/10.2202/1565-3404.1054

Husak, D. (2011) Negligence, Belief, Blame and Criminal Liability: The Special Case of Forgetting. *Criminal Law and Philosophy*, 5, 199–218. https://doi.org/10.1007/s11572-011-9115-z

Johnston J. C. (2013). The Expert Witness in Medical Malpractice Litigation: Through the Looking Glass. *Journal of Child Neurology*, 28(4), 484-501.

Kamin, K. A., & Rachlinski, J. J. (1995). Ex post ≠ ex ante: Determining liability in hindsight. *Law and Human Behavior*, 19(1), 89–104. https://doi.org/10.1007/BF01499075

Kant, I. (2009). *Critique of pure reason* (15th reprint). Cambridge University Press.

Kamtekar, R., & Nichols, S. (2019). Agent-Regret and Accidental Agency. *Midwest Studies In Philosophy*, *43*, 181-202.

Karlovac, M., & Darley, J. M. (1988). Attribution of Responsibility for Accidents: A Negligence Law Analogy. *Social Cognition*, 6(4), 287–318. https://doi.org/10.1521/soco.1988.6.4.287

King M. (2009). The Problem with Negligence. *Social Theory and Practice,* 35(4), 577-595. https://doi.org/10.5840/soctheorpract200935433

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, *169*, 139-146.

Kneer, M., Hannikainen, I.R., Almeida, G., Aguiar, F., Bystranowski, P., Dranseika, V., Janik, B. M., Garcia Olier, J., Güver, L., Liefgreen, A., Tobia, K., Próchnicki, M., Rosas, A., Skoczeń, I., Strohmaier, N. & Struchiner, N. (in prep.) Outcome effects on mental state ascriptions across cultures

Kneer, M. (2018). Perspective and epistemic state ascriptions. *Review of Philosophy and Psychology*, *9*(2), 313-341.

Kneer, M. (forthcoming). Reasonableness on the Clapham Omnibus. In Bystranowski, P., Janik, B. and Prochnicki, M. (eds), *Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives*, Springer Publishing.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, 182, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190-194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical psychology*, *16*(2), 309-324.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33(4), 315–329.

Kumar, V. (2019). Empirical vindication of moral luck. *Nous*, 53(4), 987-1007.

LaBine, S. J., & LaBine, G. (1996). Determinations of negligence and the hindsight bias. *Law and Human Behavior*, 20(5), 501–516. https://doi.org/10.1007/BF01499038

Lee T. (1988). Court-Appointed Experts and Judicial Reluctance: A Proposal to Amend Rule 706 of the Federal Rules of Evidence. *Yale Law and Poliscy Review*, 480(6), 480-503.

Lench, H. C., Domsky, D., Smallman, R., & Darbor, K. E. (2015). Beliefs in moral luck: When and why blame hinges on luck. *British Journal of Psychology*, 106(2), 272–287. https://doi.org/10.1111/bjop.12072

Lowe, D.J. and Reckers, P.M. (1994), The Effects of Hindsight Bias on Jurors' Evaluations of Auditor Decisions. *Decision Sciences*, 25, 401-426. https://doi.org/10.1111/j.1540-5915.1994.tb00811.x

Martin, J. W., & Cushman, F. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors. *PloS ONE*, *10*(4).

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition, 147*, 133-143.

Masella R, Meister M. (2001). The ethics of health care professionals' opinions for hire. *Journal of the American Dental Association*, 132, 361–7.

Moore, M.S., Hurd, H.M. (2011) Punishing the Awkward, the Stupid, the Weak, and the Selfish: The Culpability of Negligence. *Criminal Law and Philosophy,* 5, 147–198. https://doi.org/10.1007/s11572-011-9114-0

Nagel, T. (1979). *Mortal Questions* (vol. 89, vol. 3). Cambridge University Press.

Nichols, S., Timmons, M., & Lopez, T. (2014). Using experiments in ethics–ethical conservatism and the psychology of moral luck. In *Empirically informed ethics: Morality between facts and norms* (pp. 159-176). Springer, Cham.

Nelkin, D. K. (2004). *Moral luck*. Stanford Encyclopedia of Philosophy.

Nelkin, D. K. (2019). Thinking Outside the (Traditional) Boxes of Moral Luck. *Midwest Studies In Philosophy*, *43*, 7-23.

Nelkin, D. K. (2021). Liability, culpability, and luck. *Philosophical Studies*, 1-19.

Nichols, S. (2009). *Ethics and the psychology of moral luck*. Presented at the Pacific American Psychological Association, Vancouver, BC.

Prochownik & Cushman (ms). Replication of Kneer and Machery (2018). Exploring the Sensitivity of Judgments of Blame vs. Blameworthiness to Outcomes in Moral Luck Scenarios.

Prochownik, K., Krebs, M., Wiegmann, A., & Horvath, J. (2020). Not as Bad as Painted? Legal Expertise, Intentionality Ascription, and Outcome Effects Revisited. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 1930–1936.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on psychological science*, *7*(5), 411-426.

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

Rachlinski, J. J. (2000). Heuristics and biases in the courts: ignorance or adaptation. *Or. L. Rev.*, *79*, 61.

Raz J., (2010) Responsibility and the Negligence Standard, *Oxford Journal of Legal Studies*, 30(1), 1–18. https://doi.org/10.1093/ojls/gqq002

Schauer, F., Spellman, B. A., (2020). Probabilistic Causation in the Law. *Journal of Institutional and Theoretical Economics*, 176, 4-17.

Schofferman, J., (2007). Opinions and Testimony of Expert Witnesses and Independent Medical Evaluators. *Pain Medicine*, 8(4), 376-382.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language, 27*(2), 135–153.

Shen, F. X., Hoffman, M. B., Jones, O. D., Greene, J. D., & Marois, R. (2011). Sorting Guilty Minds. *New York University Law Review*, 86, 1306-1355.

Simons, K. W. (1994). Culpability and Retributive Theory: The Problem of Criminal Negligence. *Contemporary Legal Issues,* 365.

Spellman, B.A., Kincannon, A. (2001). The Relation Between Counterfactual ("But For") and Causal Reasoning: Experimental Findings and Implications for Jurors' Decisions, *Law and Contemporary Problems* 241-264.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27(1), 76–105. https://doi.org/10.1016/0022-1031(91)90011-T

Tobia, Kevin (ms). Legal Concepts and Legal Expertise (February 10, 2020). Available at SSRN: https://ssrn.com/abstract=3536564 or http://dx.doi.org/10.2139/ssrn.3536564

Turner, B. (2009). Expert Opinion in Court: A Comparison of Approaches. In A. Jamieson & A. Moenssens (Eds.), *Wiley Encyclopedia of Forensic Science*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470061589.fsa001

Wallace, R. J. (2017). *The view from here: On affirmation, attachment, and the limits of regret*. Oxford University Press.

Walster, E. (1967). Second guessing important events. *Human Relations*, 20(3), 239–249. https://doi.org/10.1177/001872676702000302

Wexler, D. B., & Schopp, R. F (1989). How and when to correct for juror hindsight bias in mental health malpractice litigation: Some preliminary observations. *Behavioral Sciences and the Law*, 7, 485-504.

Williams, B. (1981). *Moral luck: Philosophical papers 1973–1980*. Cambridge University Press.

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2), 345–353. https://doi.org/10.1037/0096-1523.4.2.345

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, 1(3), 333–349.

Zimmerman, M. (1986). Negligence and Moral Responsibility. *Noûs*, 20(2), 199-218. doi:10.2307/2215391

United Kingdom Case Law

'John Radford (formerly known as John Worboys) versus The Parole Board of England and Wales' (2018) High Court of Justice, Queen's Bench Division, CO/368/2018, CO/370/2018 and CO/554/2018. Judiciary.uk [Online]. Available at: https://www.judiciary.uk/wp-content/uploads/2018/03/dsd-nbv-v-parole-board-and-ors.pdf (Accessed: 23.10.2020).

'Regina versus Georgina Sarah Anne Louise Challen' (2019) the Court of Appeal criminal division, 201605604 B2. Judiciary.uk [Online]. Available at: https://www.judiciary.uk/wp-content/uploads/2019/06/challen-approved.pdf (Accessed: 23.10.2020).

'Regina versus Sally Clark' (2003) the Court of Appeal criminal division, EWCA Crim 1020. Netk.bet.au [Online]. Available at: http://netk.net.au/UK/SallyClark1.asp (Accessed: 23.10.2020).

'Guinness Plc versus Ernest Saunders Plc' (1990) House of Lords, 2 AC 663. Casemine.com [Online]. Available at: https://www.casemine.com/judgement/uk/5a8ff8c960d03e7f57ecd701 (Accessed: 23.10.2020).

United States Case Law

Ambrosini v. Labarraque, 101 F.3d 129, D.C. Cir. (1996). Available at: https://casetext.com/case/ambrosini-v-labarraque-2 (Accessed: 04.11.2020).

Claar v. Burlington N. R.R., 29 F.3d 499, 9th Cir. (1994). Available at: https://casetext.com/case/claar-v-burlington-northern-r-co (Accessed: 04.11.2020).

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). Available at: https://www.law.cornell.edu/supct/html/92-102.ZS.html (Accessed: 23.10.2020).

Day v. Bryant, 697 S.E.2d 345; (2010). Available at: https://zaytounlaw.com/wp-content/uploads/2014/03/Recent-Decisions-Med-Mal-2010-2011-Final.pdf (Accessed 05.11.2020).

Frye v. United States, 293 F. 1013 D.C. Cir. (1923). Available at: https://en.wikisource.org/wiki/Frye_v._United_States (Accessed: 23.10.2020).

Griffen v. Univ. of Pittsburgh Med. Ctr.-Braddock Hospital, 950 A.2d 996 Superior Ct. Pa. (2008). Available at: https://www.courtlistener.com/opinion/2335632/griffin-v-university-of-pittsburgh-medical/ (Accessed: 04.11.2020).

Johns Hopkins Hospital v. Genda, 258 A.2d 595 Md. (1969). Available at: https://www.courtlistener.com/opinion/1970301/johns-hopkins-hospital-v-genda/ (Accessed: 05.11.2020).

Navarro v. Austin, 928 So.2d 348, (2006). Available at: https://www.morelaw.com/verdicts/case.asp?n=Unknown10/4/2006&s=FL&d=32025 (Accessed: 05.11.2020).