Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Outcome effects, moral luck and the hindsight bias

Markus Kneer [a,*], Izabela Skoczeń [a,b]

[a] Department of Philosophy, University of Zurich, 8008 Zurich, Switzerland
[b] Faculty of Law and Administration, Jagiellonian University, Kraków, Poland

ARTICLE INFO

Keywords:
Moral luck
Moral judgment
Outcome effect
Hindsight bias
Probability judgment
Negligence

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 an increased perceived probability of harm *ex post*, and consequently, to a greater attribution of negligence and culpability. Rather than simply postulating 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 strategies are successful, but some prove very promising. They should, we argue, be considered 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 separately. 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 the office 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 assessment might strike us as unjust, if we hold, with Kant (1787), that agents are morally responsible only for 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' and Su's cases (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, Minsk, & Baron, 1991, Cushman, 2008, Young, Nichols, & Saxe, 2010, Nichols, Timmons, & Lopez, 2014, Kneer & Machery, 2019, Frisch, Kneer, Krueger, & Ullrich, 2022, for a recent overview see Malle, 2021): 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 (Kneer & Machery, 2019, see also Nichols, 2009, Schwitzgebel & Cushman, 2012, Lench, Domsky, Smallman, & Darbor, 2015). Western criminal law, with its deep distaste for strict liability, sides with the Folk in this regard. Thus, there might not be a complex *philosophical* problem (the within-subjects Difference Intuition assumed by philosophers seems to be empirically incorrect). 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 distort our judgment, violating the Control Principle to which both the law and the folk are committed.

* Corresponding author
E-mail address: markus.kneer@uzh.ch (M. Kneer).

https://doi.org/10.1016/j.cognition.2022.105258
Received 15 March 2021; Received in revised form 3 August 2022; Accepted 16 August 2022

1.2. The mechanics of the outcome effect

How can we alleviate the outcome effect? This depends, in part, 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 as higher for Frank than for Su. It thus seems more appropriate to judge that Frank incurred a substantial risk as opposed to Su, which, 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. 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, for meta-analyses, see Christensen-Szalanski & Willham, 1991 and Guilbault, Bryant, Brockway, & Posavac, 2004). To address the distorting effect of outcome on culpability judgments, we must find ways to alleviate the hindsight bias. This is the central topic of the paper.

This paper, which aspires to make a contribution both in moral psychology and experimental jurisprudence,[2] 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 a universal perspective) 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 of judgment has been clearly uncovered, we turn to the core objective of the paper: *Debiasing strategies*, which are summarized in Fig. 1. 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 a scientifically-informed perspective. Fig. 1 visualizes the different debiasing strategies. 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.

We consider the findings 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; Arkes & Schipani, 1994; Jurs, 2013).

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, for reviews of the hindsight bias in the context of the law, see Harley, 2007 and Giroux, Coburn, Harley, Connolly, & Bernstein, 2016, for discussion in the context of the law see Rachlinski, 1998, Teichman, 2014, Wittlin, 2016), few studies explore the downstream effects on moral and legal culpability and the mechanism by way of which probability affects the latter. An exception is Kamin and Rachlinski (1995), who show that perceived probability *post hoc* has an effect on perceived culpability. Kneer and Machery (2019) go one step further in demonstrating that the relation between outcome-driven perceived probability and culpability is itself mediated by the perceived negligence of the agent.

Our first experiment attempts to replicate these findings with a new scenario. It also introduces a novel methodological approach. 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,[3] 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

[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, Lantian, & Boudesseul, 2016), the epistemic side-effect effect (Alfano, Beebe, & Robinson, 2012; Beebe & Buckwalter, 2010; Beebe & Jensen, 2012; Kneer, 2018). For empirical studies regarding *mens rea* attribution conducted with legal experts (judges, lawyers or law students), see Kneer and Bourgeois-Gironde (2017), Bourgeois-Gironde and Kneer (2018), Prochownik et al. (2020), Tobia (2020a), and Kneer et al. (2022). For mock juror studies on *mens rea* attribution, see inter alia Shen et al. (2011), Ginther et al. (2014), Mott and Heiphetz (2022).

[2] Experimental jurisprudence is a nascent discipline which employs empirical strategies to explore topics in (or related to) philosophy of law. The topics are quite diverse. Beyond theory of mind and *mens rea* attribution, which is the subject of this paper, they include inter alia consent (Sommers, 2019; Sommers & Bohns, 2018), purpose (Almeida, Knobe, Struchiner, & Hannikainen, 2021), perjury (Skoczeń, 2021, 2022; Skoczeń & Smywiński-Pohl, 2022), legal causation (Macleod, 2019, Knobe & Shapiro, 2021, Güver & Kneer, 2022, Prochownik, 2022), legal explanation (Liefgreen & Lagnado, 2021), law and morality (Donelson & Hannikainen, 2020, Flanagan & Hannikainen, 2022, Hannikainen et al., 2021, Kirfel and Hannikainen, 2022, Macleod, 2015), the reasonable person standard (Jaeger, 2020; Kneer, 2022; Tobia, 2018), product liability (Gill & Keil, 2022), determinants of judicial decision making (Engel & Rahal, 2020; Liu, 2018; Spamann & Klöhn, 2016), bias in legal decision making (Engel & Glöckner, 2013, Lidén, Gräns, & Juslin, 2019, and Strohmaier, Pluut, van den Boos, Adriaanse, & Vriesendorp, 2021), legal interpretation (Bystranowski, Janik, Próchnicki, Hannikainen, & Struchiner, 2021, Hannikainen et al., 2022, Macleod, 2021, Pirker & Skoczeń, 2022, Tobia, 2020b). Several of these topics have been explored across different jurisdictions and cultures (see the OSF project of Hannikainen, Kneer, et al., 2021). For reviews and discussion of experimental jurisprudence, see Prochownik (2021), Tobia (2022) and Jiménez (2022).

[3] We thank an anonymous reviewer for drawing our attention to skeptical arguments concerning the existence of objective probability (de Finetti, 1974; de Finetti, 1992). Distinguishing between subjective and objective probability, however, is the orthodoxy in the literature (for a review, see Hájek, 2019). What is more, we doubt that the folk assumes a deterministic world view, which is presupposed by accounts like de Finetti's.

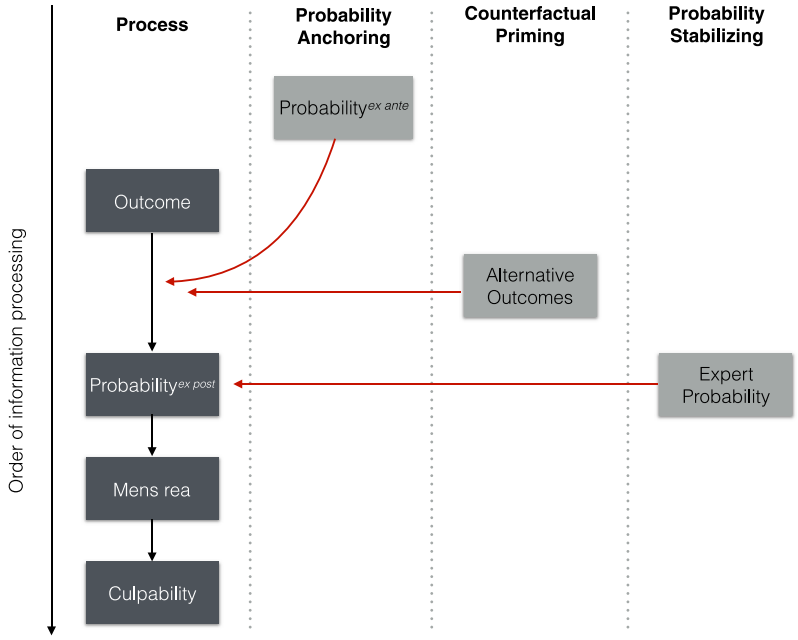

Fig. 1. The order of information processing in negligence cases and possible debiasing strategies.

agent's "having good reasons to believe that [X] would occur". Evidently, subjective probability and having good 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,[4] 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).[5]

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 incurs an unjustifiable and substantial risk while being aware of such a risk. An agent acts negligently, if she *should have* been aware of a substantial risk.[6] 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 risks as substantial. Nonetheless,

[4] https://aspredicted.org/wu2ki.pdf. 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/.

[5] 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 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.

[6] We work with the standard formulation of negligence in US law familiar from the Model Penal Code. Alternative possible formulations might invoke the *reasonable person standard* (see US Model Penal Code as well as e.g. Gardner, 2001, 2015 and Zipursky, 2014, for empirical work see Tobia, 2018 and Kneer, 2022) or the notion of *due care* (see e.g. Margoni & Surian, 2021).

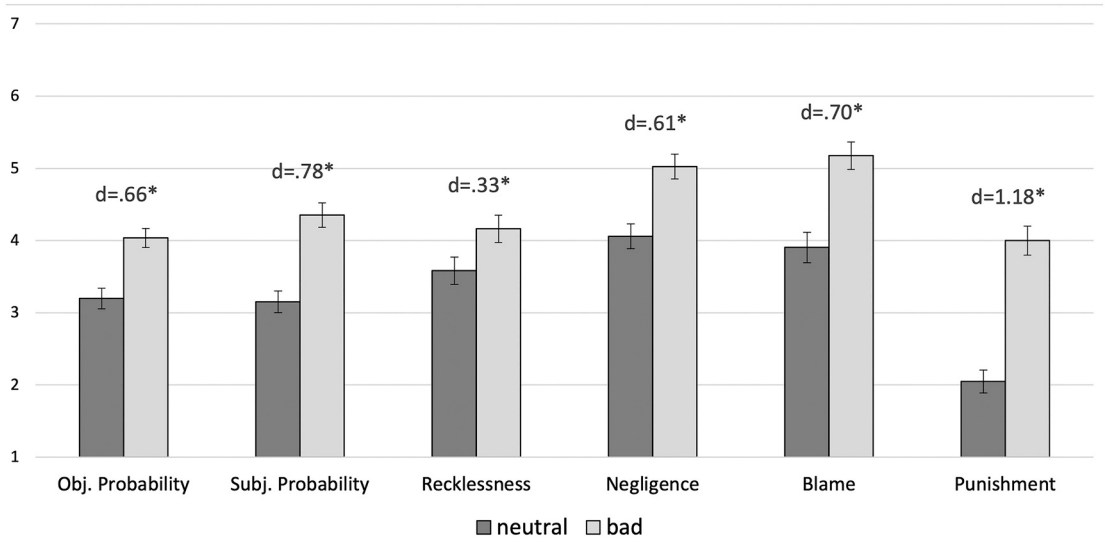

Fig. 2. Mean ratings for probability, *mens rea* and moral judgment for the between-subjects design (outcome: neutral v. bad). Effect sizes are given in terms of Cohen's *d*, significance is reported at the $p < .05$ threshold (for details, see Appendix section 1.2.1). Error bars denote the standard error of the mean.

since there is evidence that laypeople frequently have difficulties differentiating between these two types of *mens rea* (Shen, Hoffman, Jones & Greene, 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):

> **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, deserved 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).

We kept the order of questions fixed in all experiments here reported. This approach has the advantage of roughly tracking the order of the legal adjudication of culpability, where *actus reus* and *mens rea* are determined before liability and damages are decided.

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 the 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 (for this and the following experiments, the nonconverted mean probabilities are provided in the Appendix section 1.2.2). Fig. 2 graphically represents the mean ratings for all dependent variables.

Consistent with previous research (Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009; Gino, Moore, & Bazerman, 2009; Gino, Shu, & Bazerman, 2010; Lench et al., 2015; Schwitzgebel & Cushman, 2012; Young et al., 2010), there is a significant main effect of outcome (all *ps* $< .035$) on the moral variables (i.e. blame, punishment, see Appendix section 1.2), *mens rea* (see Kneer & Bourgeois-Gironde, 2017; Kneer & Machery, 2019) and perceived probability (Arkes, Wortmann, Saville, & Harkness, 1981, Dawson et al., 1988; Christensen-Szalanski & Willham, 1991; Kamin & Rachlinski, 1995; Kneer, 2022) both when assessed in objective and subjective terms (see Appendix, 1.2.1).

2.3.2. Mediation analyses for blame

In order to explore whether the mediation results reported by Kneer and Machery (2019) replicate, 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 in Fig. 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 ($p = .160$).

A serial mediation analysis with subjective probability and negligence provides more clarity (Fig. 4): The relation between outcome and blame is not mediated by negligence *per se* (the a^2 path between outcome and negligence is nonsignificant). Instead, mediation through negligence travels via 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).

2.3.3. Mediation analyses for punishment

The results are different concerning the moral DV of deserved 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 one-third of the total effect of outcome on punishment, which remains significant ($p < .001$), see Fig. 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

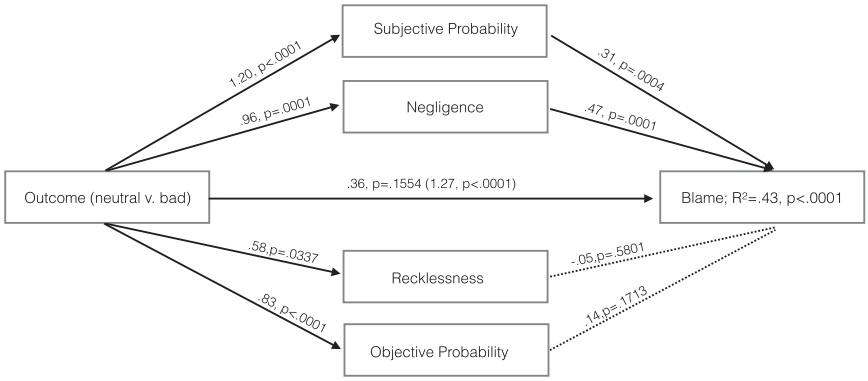

Fig. 3. Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and blame judgments by probability (objective and subjective), negligence and recklessness.

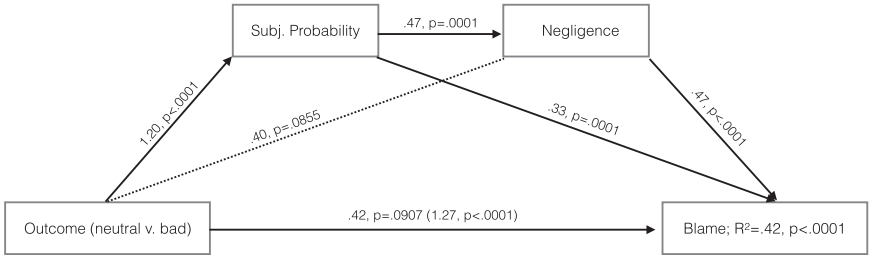

Fig. 4. Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and blame judgments by subjective probability and negligence.

accounts for about one-third of the total effect of outcome on punishment. Objective probability (the a^1b^1 path) accounts for more than half of the mediation (54%) cf. Fig. 6.

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 deserved punishment, by contrast, is only partially mediated by objective probability and negligence. Importantly, about two-thirds of the effect of outcome on deserved 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.[7] Deserved 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.[8]

A bias is a disposition to systematically diverge from a particular standard of judgment or behavior. As a relational concept, a bias is thus always relative to a certain standard of adequate judgment or behavior. Differently put, a certain disposition might constitute a bias with respect to some standard, rule or norm S_1 yet not with respect to another standard, rule or norm S_2 (cf. Hahn & Harris, 2014; Kahneman, 2000). People's propensity to judge an outcome that has occurred more likely *ex post* than *ex ante*, or "creeping determinism" as Fischhoff (1982) calls it, is near-universally considered a *bias* (Walster, 1967; Fischhoff, 1975, 1980; Hoch & Loewenstein, 1989; Agans & Shaffer, 1994; Hertwig, Gigerenzer, & Hoffrage, 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 et al., 1981; Casper, Benedict, & Perry, 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; Frisch et al., 2022; Kneer & Machery, 2019) strongly

[7] 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.

[8] Cushman argues that judgments concerning permissibility and wrongness depend primarily on mental states, whereas blame and punishment depend strongly on non-mental features. Like Young et al. (2010), Kneer and Machery (2019), and Frisch et al. (2022), we find blame to be more in the former category, i.e. mainly sensitive to mental states. Two things bear mentioning however: First, the exact status of blame is still contentious (see e.g. Prochownik and Cushman, 2018; Frisch et al., 2022), in particular since the exact formulation seems to matter. Cushman (2008) uses "blame", the diverging studies use the expression "blameworthy". 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.

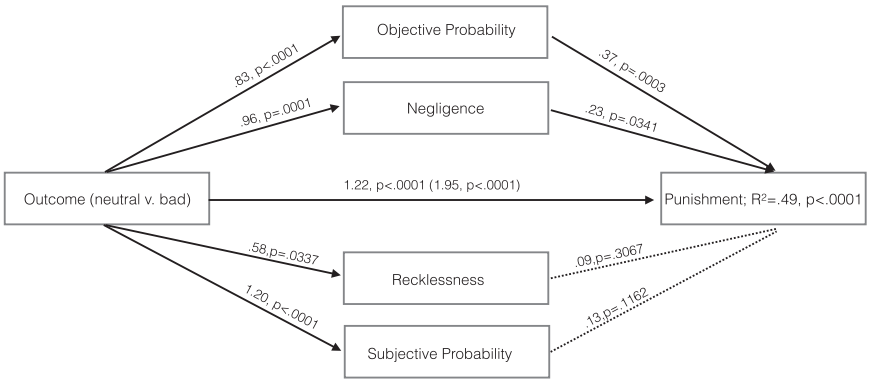

Fig. 5. Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and punishment judgments by probability (objective and subjective), negligence and recklessness.

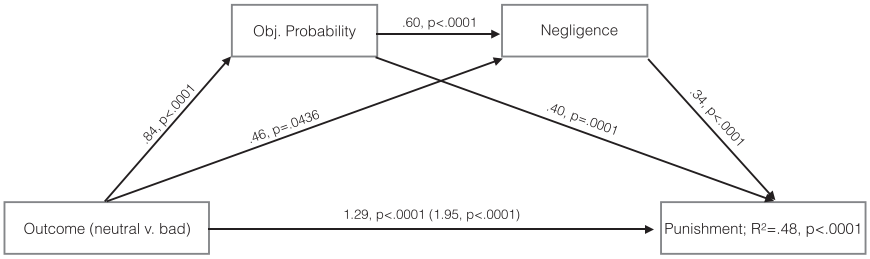

Fig. 6. Mediation analysis with 5000 bootstrap samples of the relationship between outcome (neutral v. bad) and punishment judgments by objective probability and negligence.

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. But given that the notion of a *risk* is defined as a function of an event's probability and its severity of harm, where severity is held fixed, an increase in probability means an increase in risk. Thus, if the perceived probability of Frank's action were justly judged higher *post hoc* than that of Su's action, the risk incurred by Frank would be higher than the risk incurred by Su. Consequently, it can appear that Frank should have been aware of a substantial risk (i.e., he was negligent), though Su need not have been, for the simple reason that only the higher risk incurred by Frank can be judged *substantial* in the first place. But once Frank is judged more negligent, it makes sense to consider him more deserving of blame and punishment.

How to adjudicate between the two views? A bias – a systematic distortion of judgment – arises when the *application* or use of a concept is frequently at odds with certain features of the *concept as such* (thus violating rules of its correct application). In order to postulate a bias, one must know the *correct* application of the concept, and hence the nature of the concept itself. Consequently, if we suspect that a certain concept's application (probability judgments, say) is distorted by a certain feature *F* (outcome information), we must first establish whether *F* is indeed not part of the concept. One good way of doing so, we take it, consists in presenting people with scenarios in which only the crucial feature *F* is manipulated side-by-side (i.e. with a within-subjects experiment where *F* is the sole factor). In a nutshell, then, if the application of a certain concept is systematically sensitive to a feature *F* in between-subjects experiments to which it is *not* sensitive in within-subjects experiments, there are at least preliminary grounds for a bias. That said, it is of course at least possible that certain particularly sticky biases arise even in within-subjects designs.

Let's make these somewhat abstract considerations a little more concrete. 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 (no bias). 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 finding 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 into 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 (Baron, 2000; Baron & Ritov, 2004; Hsee, 1996; Hsee & Zhang, 2004; Rachlinski, 1998, 2000). 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,[9]

[9] Link to preregistration: https://aspredicted.org/hr3bb.pdf.

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 who finished the entire survey (including the 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 outcomes, as well as effect sizes and results of paired-samples *t*-tests for all six DVs are presented in Fig. 7 (for *t*-test details, see Appendix 2.2.1).[10] Except for objective probability, we found a significant difference for all dependent variables (all *ps* < .049), though subjective probability just barely met 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 $d = .63$.

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 and 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 Fig. 8 illustrates, a substantial majority (>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, Frisch et al., 2022, see also the mediation analyses above) that there is a strong, direct effect of outcome on punishment.[11]

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, 2.2.3) 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, *ps* < .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 current scenario.[12]

3.4. Discussion

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; Frisch et al., 2022; Kneer & Machery, 2019; Martin & Cushman, 2015, 2016; Nobes & Martin, 2022), this suggests that the folk concept of deserved punishment is outcome-sensitive. Note that, according to 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 design, amounting to a very large effect – might indeed be taken to be, at least in part, 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 both types of *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 inculpating mental states 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, with a medium-sized effect ($d = .49$). However, here, too, a significant majority holds that the two agents deserve the same amount of blame. This is consistent with Kneer & Machery's (2019, Experiment 2) within-subjects results: Here too, there remains a significant effect of outcome on blame (though also smaller than in the between-subjects design) although the vast majority of participants judged the lucky and unlucky agent identically with regards to blameworthiness. 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. Gilbert, Tenney, Holland, & Spellman, 2014; Schauer & Spellman, 2020; Spellman & Kincannon, 2001), and tentative evidence for the outcome-insensitivity of the folk-concept of blame. Consequently, we suggest that the substantial outcome effects on all dependent variables – except punishment – in the between-subjects design are 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

[10] According to a *post hoc* power analysis with an error probability of α = .05 achieved power to detect a midsized effect ($d = .50$) was very high (.99) for the *t*-test. The analysis was conducted with G*Power 3.1 (see Faul, Erdfelder, Buchner, & Lang, 2009).

[11] As one of the reviewers helpfully pointed out, Margoni, Geipel, Hadjichristidis, and Surian (2018, 2019) report interesting effects, according to which the influence of outcome on moral judgment is more pronounced among older than younger adults. In our sample, we did not find a significant correlation between age and the difference in blame or punishment across conditions (all *ps* > .095, see Appendix, Section 2.2.4). However, following Margoni et al. (2018) in contrasting a younger subsample (21–39 years) and an older one (63–90 years), we found a more pronounced difference in blame for the older subsample ($d = .63$, $p = .023$) than for the younger one ($d = .37$, $p = .066$), where the effect did not reach significance. The effect of outcome on mean punishment was also somewhat more pronounced for the older sample ($d = .80$, $p = .006$) than for the younger sample ($d = .55$, $p = .003$).

[12] 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.

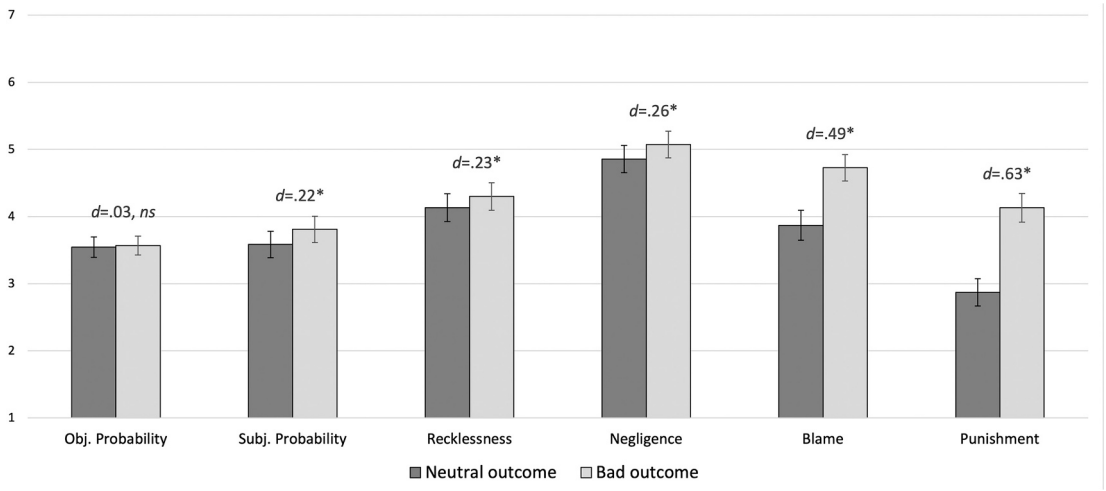

Fig. 7. Mean probability, *mens rea* and moral responsibility judgments for the within-subjects design in the two conditions (neutral vs. bad outcome). Effect sizes are given in terms of Cohen's *d*, significance is reported at the $p < .05$ threshold (for details, see Appendix section 2.2.1). Error bars denote standard error of the mean.

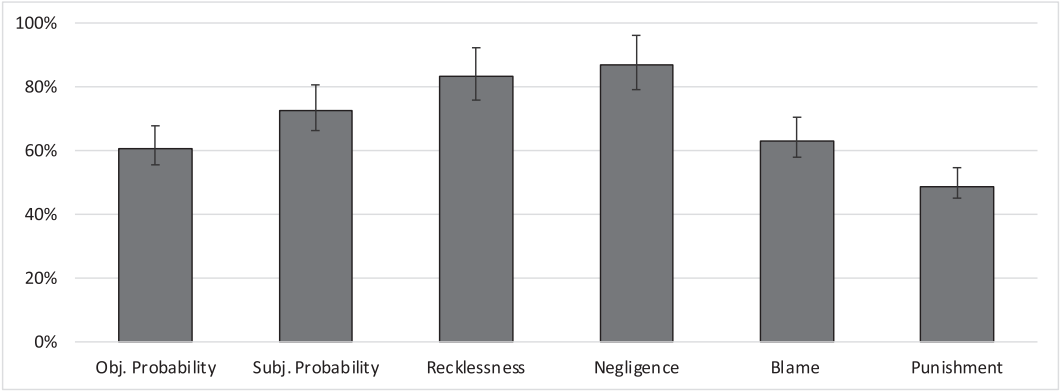

Fig. 8. Proportions of participants who judged probabilities, *mens rea*, blame and punishment identically (no Difference Intuition) across scenarios. Error bars denote 95% confidence intervals, Wilson method, see Brown, Cai, & DasGupta, 2001.

this approach 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. A recent meta-analysis by Bystranowski et al. (2021) suggests that anchoring (broadly conceived) can affect legal decision making, though the authors point out that their findings might be subject to publication bias, possibly exaggerating the effect.

4.1. Participants

We recruited 199 participants online via Amazon Mechanical Turk. The IP address location was restricted to the USA. As preregistered,[13] 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, age 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, the outcomes were revealed (between-subjects), 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 Fig. 9.

4.3. Results

Expectedly, perceived objective and subjective probabilities *ex ante* (i.e. before the outcome was revealed) across conditions did not differ significantly (independent samples *t*-test $ps > .235$, see Appendix, 3.2.1). 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. Fig. 10 and Table 1 (anchoring – contrasted with the results of Experiment 1). Only for recklessness did we not find a significant effect ($p = .435$).

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 approach 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, find a small reduction in effect size in

[13] https://aspredicted.org/wu2ki.pdf.

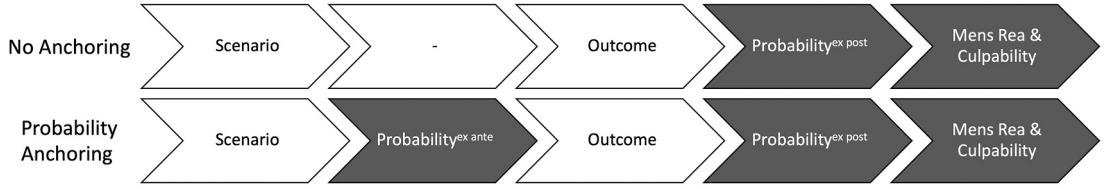

Fig. 9. Experimental design for Experiment 1 (no anchoring) and for Experiment 3 (probability anchoring). The question regarding the probabilities of the bad outcome's possible occurrence were asked before the outcome was revealed.

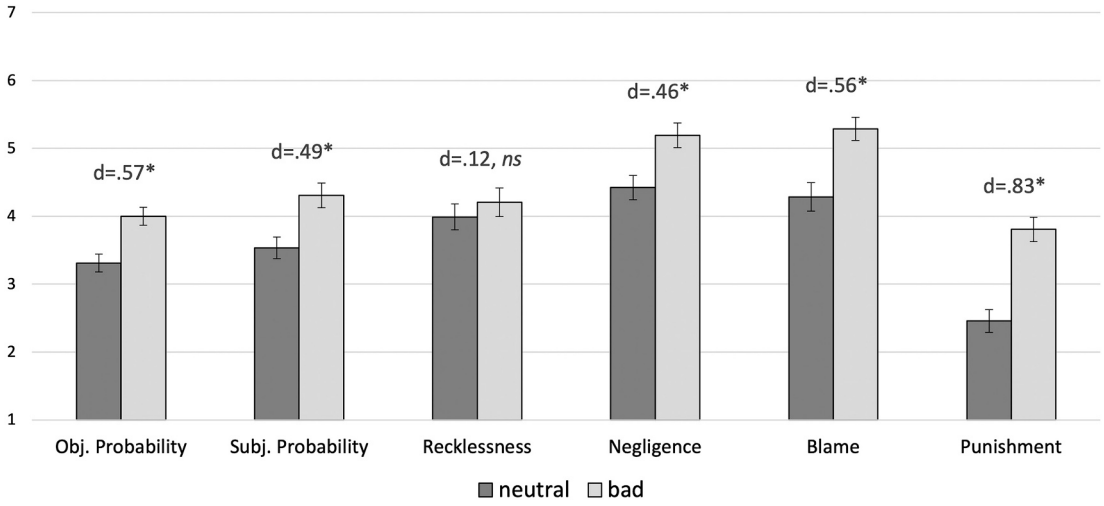

Fig. 10. Mean ratings of probabilities *ex post*, *mens rea* and moral judgments across outcomes (neutral v. bad). Effect sizes are given in terms of Cohen's *d*, significance is reported at the $p < .05$ threshold (for details, see Appendix section 2.2.1). Error bars denote the standard error of the mean.

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).

	No anchoring				Anchoring			
	t(167)	*p*	95% CI	*Cohen's d*	*t(173)*	*p*	95% CI	*Cohen's d*
Obj. Probability	4.31	<.001	[.45;1.22]	.66	3.72	<.001	[.33; 1.06]	.57
Subj. Probability	5.07	<.001	[.73;1.67]	.78	3.23	.001	[.30;1.24]	.49
Recklessness	2.14	.034	[.45;1.12]	.33	.78	.435	[−.33;.76]	.12
Negligence	3.96	<.001	[.48;1.45]	.61	3.02	.003	[.27;1.27]	.46
Blame	4.53	<.001	[.72;1.83]	.70	3.71	<.001	[.47;1.55]	.56
Punishment	3.82	<.001	[1.45;2.45]	1.18	5.47	<.001	[.86;1.84]	.83

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 & Rachlinski, 1995). Even with anchoring, outcome still has medium-sized effect on both kinds of probability, 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 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). For related interesting work exploring counterfactual thinking and blame attributions, see Murray, Krasich, Irving, Nadelhoffer, and De Brigard (2022).

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).[14, 15]

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 told participants which outcome actually occurs. The experiment 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). Note that beyond the priming conditions, we thus effectively gathered data for the ordinary between-subjects conditions of Experiment 1 one more time (the results are largely the same).

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 Fig. 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 (all $ps < .034$, see Table 2). There was no significant main effect of priming on any dependent variables (all $ps > .058$) 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 ($p = .003$) see Table 2.

Fig. 11 graphically represents the findings (see also Appendix 4.3.1). Counterfactual priming decreases the difference across outcomes, 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 on 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 Fig. 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. Fletcher, 2000; Herring, 2012). 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,[16] 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 169 participants (female: 47%; mean age: 42 years, SD =11 years, range: 22–74 years).

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. By contrast, 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.

[14] 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/ei5bd.pdf and https://aspredicted.org/qm8pb.pdf. Participants who took both surveys were excluded. For detailed documentation, see Appendix, Section 4.2.

[15] A note on achieved power: Experiments 3–5 explore strategies to decrease the impact of outcome on probability and its downstream effects in contrast to the levels reported for Experiment 1 (objective probability $d = .66$, subjective probability $d = .78$). According to *post hoc* analyses with an error probability of $\alpha = .05$ and a midsize effect ($d = .50$), achieved power was high (>.88) for the *t*-tests of all three experiments. Running the analyses with the actual effect sizes measured for probability in Experiment 1 (objective probability $d = .66$, subjective probability $d = .78$) shows that achieved power was very high (>.98 with $d = .66$) for the t-tests of all three experiments. The analyses were conducted with G*Power 3.1, see Faul et al. (2009).

[16] https://aspredicted.org/f2ne9.pdf.

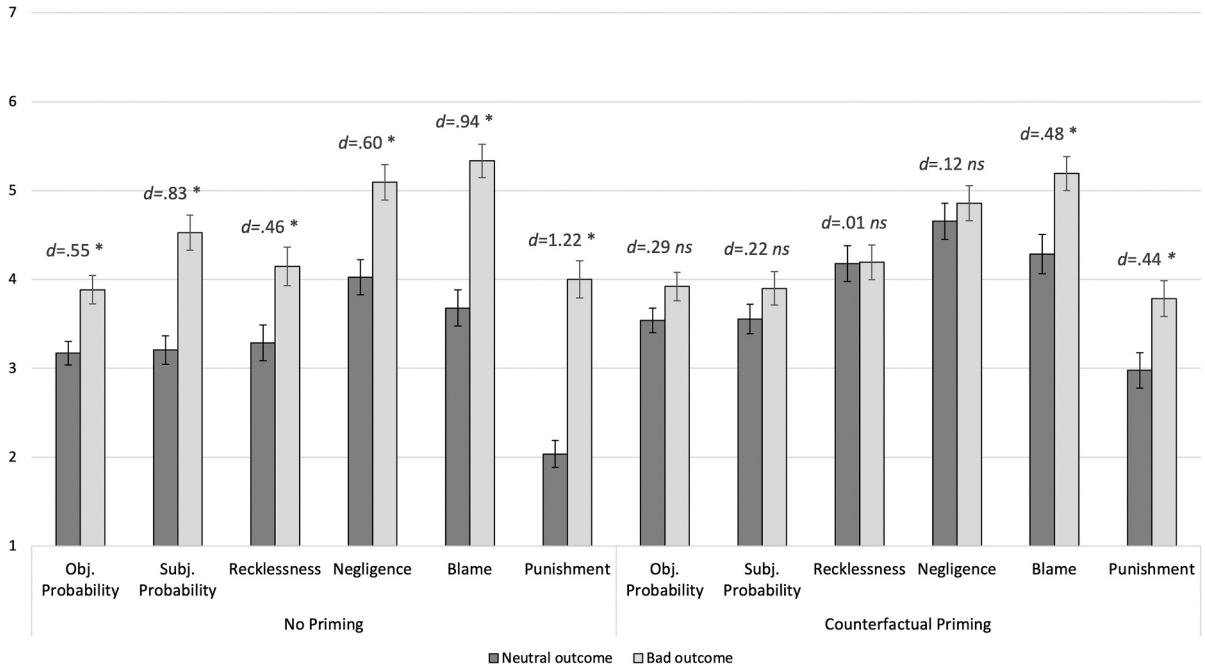

Fig. 11. Mean ratings for probabilities, *mens rea* and moral judgment across outcomes for the priming and no priming conditions. Effect sizes are given in terms of Cohen's *d*, significance is reported at the $p < .05$ threshold (for details, see Appendix section 4.3.1). Error bars denote standard error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

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

	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

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, section 5.1).

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 Fig. 12 (detailed test results in Appendix, section 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. Furthermore, there was a substantial proportion of participants (over half) whose *ex post* objective probability ratings exceeded the specified 5% (see Appendix, section 5.2.3). This suggests that participants have difficulties correctly assessing probability *post hoc*, even when explicit information is provided. It also suggests that the debiasing effect of probability stabilizing could be even more pronounced, if the information were rendered more salient such that all participants take it clearly into account.

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 the inculpating mental states, 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

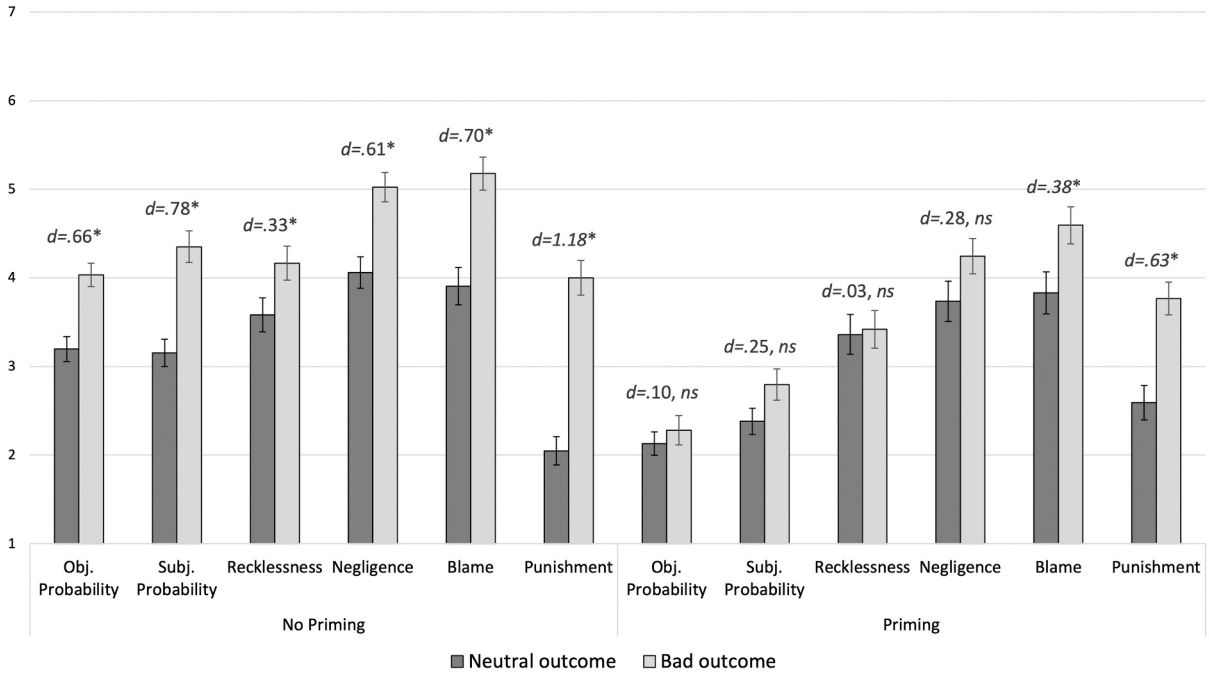

Fig. 12. Mean ratings for probabilities, *mens rea* and moral judgment across outcomes for the priming and no priming conditions. Effect sizes are given in terms of Cohen's *d*, significance is reported at the $p < .05$ threshold (for details, see Appendix 5.2.1). Error bars denote standard error of the mean.

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. Fig. 13 reports outcome effects on perceived probabilities and *mens rea*, Fig. 14 reports outcome effects on perceived probabilities, blame and punishment.

Replicating Experiments 1 and 4 (between-subjects data, see Appendix 9.3.2), 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 6.3.3–6.3.4): 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).[17] 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, see Appendix, section 7.3).

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), see Appendix, section 8.3.2.

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

[17] 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.

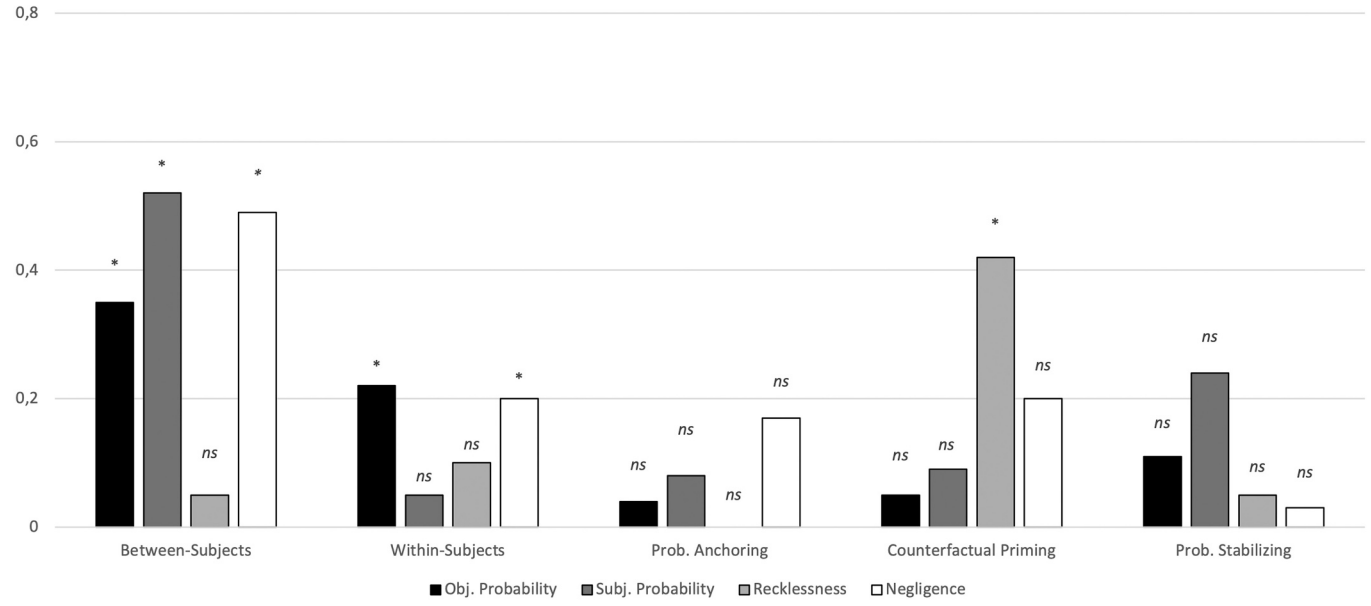

Fig. 13. Effect sizes and significance of the difference between the assessment of perceived probabilities and *mens rea* across outcomes (neutral v. bad) for all five Intersection experiments. Effect sizes are given in terms of Cohen's *ds*, significance is reported at the $p < .05$ threshold.

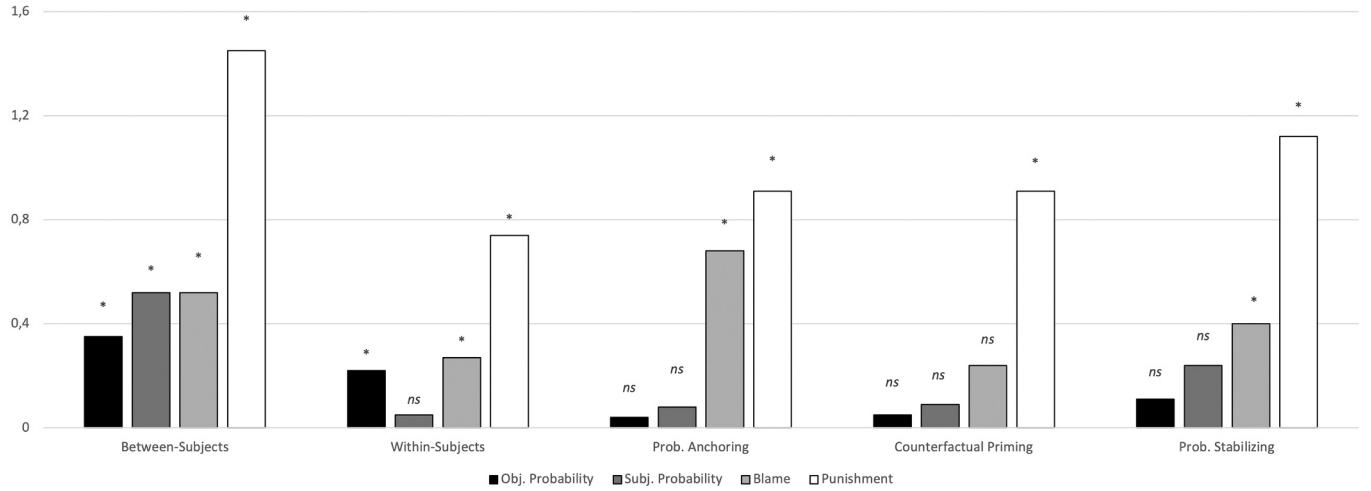

Fig. 14. Effect sizes and significance of the difference between the assessment of probabilities, blame and punishment across outcomes (neutral v. bad) for all five Intersection experiments. Effect sizes are given in terms of Cohen's *ds*, significance is reported at the $p < .05$ threshold.

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, see Appendix, section 9.3.2.

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 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 $d = 1.12$), see Appendix, section 10.3.2.

8. Meta-analyses

We have argued that (a) the considerable difference in effect size across designs (between-subjects v. within-subjects) suggests that the impact of outcome on probability, *mens rea* and moral judgment constitutes a bias, and (b) that some, though not all, of the alleviation strategies do reduce the bias quite effectively.

In order to provide more statistical support for these claims, we combined the data across scenarios for each design (or alleviation strategy) and ran meta-analyses for all DVs.[18] Beyond the between-subjects baseline (data from Experiments 1, 4 and 6), there were four treatment groups: (i) the within-subjects design (Experiments 2 and 7), (ii) probability anchoring (Experiments 3 and 8), (iii) counterfactual priming (Experiments 4 and 9) and (iv) probability stabilizing

[18] We would like to thank an anonymous reviewer for their helpful suggestion to provide further support for our claims.

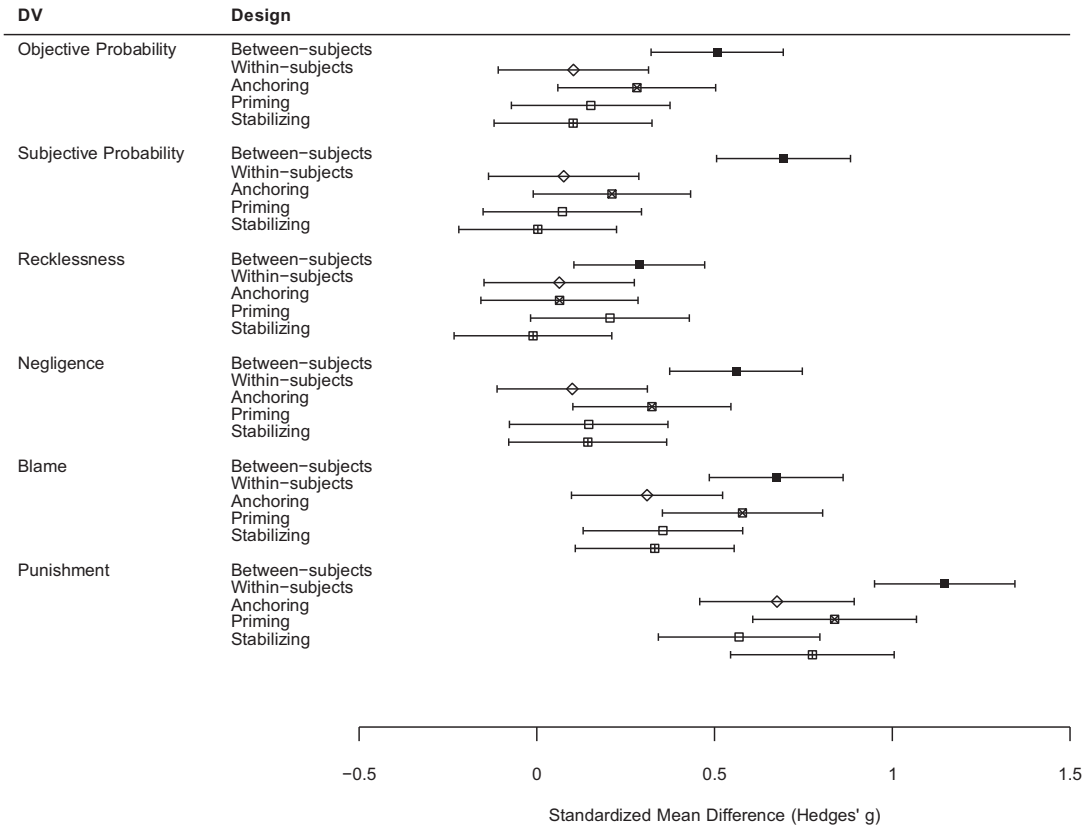

Fig. 15. Effects of outcome on Dependent Variables (DVs) across designs in a random effects model in terms of Hedges' *g* (Hedges, 1981, Hedges & Olkin, 1985). Error bars denote 95% confidence intervals.

(Experiments 5 and 10). Fig. 15 presents the mean effects of outcome on all DVs, estimated with the restricted maximum-likelihood method based on a random effects model (see Viechtbauer, 2010). On the basis of the 95% CIs, one can thus grasp at a glance that, say, the impact of outcome on objective probability judgments was considerably and significantly smaller in the *probability stabilizing* treatment than in the between-subjects baseline design, whereas this was not the case for *anchoring*.

For each of the four contrasts comparing baseline (between-subjects data) and treatment (i-iv), we also tested the interaction of condition and design. To do so, we ran a linear mixed-effects model with the "lmerTest" package (Kuznetsova, Brockhoff, & Christensen, 2017), regressing the dependent variables on condition and design while controlling for the random effect of the scenario. The interaction was then explored by means of an ANOVA (type 2 and F-test). Table 3 summarizes the results. From the *p*-values one can infer whether the effect size in a certain bias alleviation design was significantly lower than in the baseline design (between-subjects). To return to our example from the previous paragraph: As regards objective probability, we find for instance that *anchoring* did not significantly decrease the effect of outcome ($p = .140$), whereas *probability stabilizing* ($p = .011$) did.

Taking stock: In contrast to the between-subjects design, the impact of outcome was significantly smaller for probability, negligence and moral judgment in the within-subjects design (all $ps < .008$). This, as we have argued at length, suggests that the impact of outcome on the tested DVs constitutes a bias. Furthermore, a significant reduction in outcome effect can be found for counterfactual priming with $ps < .025$ for all DVs, and for probability stabilizing with $ps < .028$ for all DVs. Probability anchoring, by contrast, did not significantly decrease the effect size for outcome on objective probability, negligence or blame, though it did for subjective probability and deserved punishment ($ps < .016$). Overall then, counterfactual priming and probability stabilizing are relatively effective methods to reduce the hindsight bias and its downstream effects on *mens rea* and moral judgment, whereas anchoring holds less promise.

9. General discussion

9.1. Outcome effects on punishment and blame

Across all ten experiments, punishment proved strongly sensitive to outcome. This is consistent with previous findings (Cushman, 2008; Frisch et al., 2022; Kneer & Machery, 2019; Martin & Cushman, 2015, 2016) and suggests that the folk *concept* of punishment is outcome-dependent: Even in our within-subjects designs, severity of outcome generates at least a medium-large effect (Experiment 2, $d = .63$, Experiment 7, $d = .74$), and in the meta-analysis, we found at least an effect of a Hedges' $g > .56$ for each of the five designs. Furthermore, in both within-subjects experiments, the majority of participants judges the lucky and unlucky agent distinctly in terms of deserved punishment. The findings for blame differ considerably from the results for punishment. The impact of outcome on blame is significantly and strongly reduced in within-subjects designs as compared to between-subjects designs (Table 3). The remaining effects in the within-subjects designs are driven by a minority of participants, since the vast majority judges the blameworthiness of both agents identically. 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 process which is more sensitive to causal factors such as outcome (judging deserved punishment), and another which is less sensitive to them (blame) yet

Table 3
Contrasts of design/alleviation strategy with between-subjects results across Dependent Variables (DVs).

DV	Contrast	*F*	Df	Df.res	*p*
Objective Probability	Within-subjects - Between-subjects	7.91	1	798.00	.005
	Anchoring - Between-subjects	2.18	1	770.00	.140
	Priming - Between-subjects	5.10	1	764.00	.024
	Stabilizing - Between-subjects	6.45	1	766.02	.011
Subjective Probability	Within-subjects - Between-subjects	18.76	1	798.00	<.001
	Anchoring - Between-subjects	10.38	1	770.02	.001
	Priming - Between-subjects	17.63	1	764.01	<.001
	Stabilizing - Between-subjects	18.29	1	766.00	<.001
Recklessness	Within-subjects - Between-subjects	2.42	1	798.00	.120
	Anchoring - Between-subjects	2.37	1	770.02	.124
	Priming - Between-subjects	.38	1	764.01	.536
	Stabilizing - Between-subjects	3.93	1	766.01	.048
Negligence	Within-subjects - Between-subjects	9.78	1	798.00	.002
	Anchoring - Between-subjects	2.70	1	770.00	.101
	Priming - Between-subjects	8.35	1	764.00	.004
	Stabilizing - Between-subjects	6.88	1	766.00	.009
Blame	Within-subjects - Between-subjects	7.42	1	798.00	.007
	Anchoring - Between-subjects	.92	1	770.00	.337
	Priming - Between-subjects	6.78	1	764.00	.009
	Stabilizing - Between-subjects	4.90	1	766.00	.027
Punishment	Within-subjects - Between-subjects	9.76	1	798.00	.002
	Anchoring - Between-subjects	5.96	1	770.00	.015
	Priming - Between-subjects	18.03	1	764.00	<.001
	Stabilizing - Between-subjects	7.84	1	766.00	.005

more sensitive to mental factors (see mediation analyses in Experiments 1 and 6).[19]

Second, the fact that the pronounced between-subjects outcome effect on blame is significantly reduced in the within-subjects design and driven by a minority suggests it is a bias. While this is commonly claimed, few authors back this claim up convincingly. Our within-subjects data demonstrates that the folk *concept* of probability is largely outcome-insensitive, 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 (Frisch et al., 2022; Kneer & Machery, 2019), 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; see also Hartman, 2017, for a review see Nelkin, 2004). 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 not much of a puzzle here (though there is a puzzled minority of about 25% across Experiments 2 and 7). Or is there? Perhaps on the basis that, even in within-subjects designs, we find a strong outcome effect on deserved *punishment* (Kumar, 2019, for instance always speaks of "blame *and punishment*" in the same breath)? We doubt it. As argued by Enoch and 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 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.[20]

9.2. *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; Kneer & Machery, 2019; Rachlinski, 1998, 2000): 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 significant direct effects of outcome on blame* remain.[21]

[19] 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 and 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, 2018; Björnsson and Kneer, 2022).

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

[21] This point can also be connected to Monroe and Malle's (2017) influential model of blame. It states that negligence (unintentionality) and blame ascriptions are closely connected to considerations regarding 'whether the agent could have prevented the norm violating event (capacity to prevent) and should have prevented it (obligation to prevent),' see also Malle, Guglielmo, & Monroe, 2014). Since in the unlucky case, participants perceive the risk as higher due to the hindsight bias, they also perceive the agent as more obligated to foresee and prevent the risk, which in turn might drive higher ascriptions of negligence and blame (for further empirical work supporting the model cf. Margoni & Surian, 2021).

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,[22] 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 (and significantly reduced) direct effects of outcome on blame remain (*Flood* $d = .38$, *Intersection* $d = .40$). Interestingly, the effect of outcome on punishment is also significantly reduced (see Table 3), and substantially so (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. The impact of *probability anchoring,* as we called it, was significant for some DVs (Table 3, Fig. 15 – meta-analysis), though not all, and its effect was not particularly pronounced.

Taking inspiration from Lench et al. (2015), we explored whether entertaining *counterfactuals* (i.e. alternative outcomes) 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 reviews, see Rachlinski, 1998 and Wittlin, 2016, for an unsuccessful attempt to reduce the hindsight bias through explaining it to participants see Pohl & Hell, 1996), 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) and its impact limited, 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 susceptible to bias. In the within-subjects designs, the effect of outcome on punishment is only about half that of the between-subjects designs, and all three alleviation strategies significantly reduce the effect (all $ps < .016$, Table 3).

9.3. 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, King, 2009, Raz, 2010, Husak, 2011, Yaffe, 2012, Amaya, 2022, for interesting recent empirical work see inter alia Murray et al., 2022, Nobes & Martin, 2022, Frisch et al., 2022). 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, Teichman, 2014, Wittlin, 2016 and Kneer, 2022). 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 (for reviews of the hindsight bias in the law, see Giroux et al., 2016 and Wittlin, 2016). Since in many countries, such as inter alia the US and the UK, the *mens rea* question is decided by lay jurors, 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 pronounced influence an expert assessment of *ex ante* likelihood exerts on *mens rea* and culpability judgments, it must be used with care and the procedural conventions for choosing such experts might require more attention.[23] Most of case law in which expert testimony is decisive pertains to medical malpractice. 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).[24] These two cases lead to the formulation of general standards of acceptability of expert witness evidence and influenced thousands of later cases. The *Frye* standard claims that expert witness evidence is admissible only if based on generally accepted views of the scientific community. By contrast, the *Daubert* standard, which replaced the *Frye* standard, states that it is the judge who decides which evidence shall prevail. Given the powerful impact of probability stabilizing via expert testimony, it might stand to reason that evidence of this sort should be consistent with the scientific consensus, and not the opinions of individual judges.

9.4. 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 explore and contrast several strategies to

[22] 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 & Schwartz, 2005; Schofferman, 2007.

[23] Importantly, there are procedural differences across legal systems in who can *choose* and *present* an expert witness in court. In adversarial systems where the judge has a limited procedural role (e.g. the US and the UK Chartered Institute of Arbitrators, 2020), 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 decide that expert testimony is helpful, and determine whom to consult.

[24] As regards the UK, notorious cases where expert testimony was controversial include e.g. '*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).

alleviate the systematic performance error. Further research should examine whether the results replicate with different scenarios, methods,[25] 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, some evidence of a cross-cultural effect of outcome severity on ascriptions of intention and knowledge (Kneer et al., 2022 report findings from 12 countries). Similar effects can be found for legal experts (for France, see e.g. Kneer & Bourgeois-Gironde, 2017, Bourgeois-Gironde & Kneer, 2018, for Germany, see Prochownik, Krebs, Wiegmann, & Horvath, 2020, though see Tobia, 2020a). 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 rea* ascription of *any* kind. The possible threat to just legal ruling – not limited to countries with lay juror systems – should motivate a serious exploration of debiasing strategies along the lines here proposed and beyond.

10. 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 deserved 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 harm 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 where *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 equal.

Credit author statement

MK roles: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Visualisation, Writing - original draft and revision. IS roles: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Visualisation, Writing - review & editing.

Data availability

Data, stimuli and preregistrations are available on the project's OSF site: https://osf.io/e2u8q/.

[25] A successful debiasing method for the hindsight bias, explored by Arkes, Faust, Guilmette, and Hart (1988) in the context of neuropsychological diagnosis, consists in prompting participants to provide reasons for their judgments. Future work should attempt to replicate this strategy for mental state ascriptions in legal contexts.

Acknowledgements

We would like to thank Fiery Cushman, Shaun Nichols, Ivar Hannikainen and the reviewers for generous and very helpful feedback. We are also grateful for comments from the audiences at the IVR World Congress 2019 in Lucerne, Tomas Żuradzki's ERC seminar at the University of Krakow, the Dubrownik Law, Language and Philosophy Summer School organized by the Jagiellonian University, the European Society for Philosophy and Psychology 2021 conference, the Edinburgh Legal Theory Seminar Series and the Agency and Intentions workshop at Harvey Mudd College. Special thanks to the Guilty Minds Lab members Marc-André Zehnder, Levin Güver and Jan Garcia Olier for feedback and support. This project was supported by a Swiss National Science Foundation Grant for the project *Reading Guilty Minds* (PZ00P1_179912) and a grant by the Polish National Science Centre (2018/30/M/HS5/00254). Thanks are also due to the Polish National Agency for Academic Exchange (PPI/APM/2018/1/00022), which funded the research stay of I. Skoczeń at the Guilty Minds Lab (University of Zurich).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105258.

References

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. https://doi.org/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. https://doi.org/10.5840/monist201295215

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

Almeida, G., Knobe, J., Struchiner, N., & Hannikainen, I. (2021). *Purposes in law and in life: An experimental investigation of purpose attribution*. Available at SSRN 3929735 http://dx.doi.org/10.2139/ssrn.3929735.

Amaya, S. (2022). Negligence: Its Moral Significance. In Manuel Vargas, & John M. Doris (Eds.), *The Oxford Handbook of Moral Psychology*. Oxford Handbooks (2022; online edn, Oxford Academic, 20 Apr. 2022) https://doi.org/10.1093/oxfordhb/9780198871712.013.33 accessed 4 Oct. 2022, pages 661-683.

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. https://heinonline.org/HOL/LandingPage?handle=hein.journals/orglr73&div=28&id=&page=.

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

Baron, J. (2000). *Thinking and deciding*. Cambridge University Press. https://www.cambridge.org/pl/academic/subjects/psychology/cognition/thinking-and-deciding-4th-edition?format=HB&isbn=9780521862073.

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

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. https://doi.org/10.1073/pnas.1522070113

Beebe, J., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language, 25*(4), 474–498. https://doi.org/10.1111/j.1468-0017.2010.01398.x

Beebe, J. R., & Jensen, M. (2012). Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. *Philosophical Psychology, 25*(5), 689–715. https://doi.org/10.1080/09515089.2011.622439

Björnsson, G., & Kneer, M. (2022). *The folk concept of blame reexamined*. In preparation.

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. https://doi.org/10.1111/j.1559-1816.1990.tb00394.x

Bourgeois-Gironde, S., & Kneer, M. (2018). Intention, cause et responsabilité: *Mens rea* et effet Knobe. In S. Ferey, & F. G'Sell (Eds.), *Causalité, responsabilité et contribution à la dette* (pp. 117–144). Editions Brylant. https://www.larcier.com/fr/causalite-responsabilite-et-contribution-a-la-dette-2018-9782802752844.html.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*(2), 101–117. https://doi.org/10.1214/ss/1009213286

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

Bystranowski, P., Janik, B., Próchnicki, M., Hannikainen, I. R., & Struchiner, N. (2021). Do formalist judges abide by their abstract principles? A two-country study in adjudication. *International Journal for the Semiotics of Law, 35*(1), 1–33. https://doi.org/10.1007/s11196-021-09846-6

Bystranowski, P., Janik, B., Próchnicki, M., & Skórska, P. (2021). Anchoring effect in legal decision-making: A meta-analysis. *Law and Human Behavior, 45*(1), 1–23. https://doi.org/10.1037/lhb0000438

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

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.

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes, 48*(1), 147–168. https://doi.org/10.1016/0749-5978(91)90010-Q

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

Cova, F., Lantian, A., & Boudesseul, J. (2016). Can the Knobe effect be explained away? Methodological controversies in the study of the relationship between intentionality and morality. *Personality and Social Psychology Bulletin, 42*(10), 1295–1308. https://doi.org/10.1177/0146167216656356

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. https://doi.org/10.1371/journal.pone.0006699

Dawson, N. V., Arkes, H. R., 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), 259–264. https://doi.org/10.1177/0272989X8800800406

Donelson, R., & Hannikainen, I. R. (2020). The inner morality of law revisited, *3. Oxford studies in experimental philosophy* (pp. 6–28). https://doi.org/10.1093/oso/9780198852407.003.0002

Duff, A. (2001). *Punishment, communication, and community*. Oxford University Press. https://global.oup.com/academic/product/punishment-communication-and-community-9780195166668?cc=pl&lang=en&.

Engel, C., & Glöckner, A. (2013). Role-induced bias in court: An experimental analysis. *Journal of Behavioral Decision Making, 26*(3), 272–284. https://doi.org/10.1002/bdm.1761

Engel, C., & Rahal, R. M. (2020). What the judge argues is not what the judge thinks-eye tracking evidence about the normative weight of conflicting concerns in a torts case. In *, 3. MPI collective goods discussion paper*. https://ideas.repec.org/p/mpg/wpaper/2020_03.html.

Enoch, D., & Marmor, A. (2007). The case against moral luck. *Law and Philosophy, 26*(4), 405–436. https://doi.org/10.1007/s10982-006-9001-3

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feltz, A. (2007). The Knobe effect: A brief overview. *The Journal of Mind and Behavior, 28*(3), 265–277. https://www.jstor.org/stable/pdf/43854197.pdf.

de Finetti, B. (1974). *A theory of probability*. John Wiley and sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119286387.

de Finetti, B. (1992). Foresight: Its logical Laws, its subjective sources. In S. Kotz, & N. L. Johnson (Eds.), *Springer series in statistics (perspectives in statistics)Breakthroughs in statistics* (pp. 134–175). Springer. https://doi.org/10.1007/978-1-4612-0919-5_10.

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. (1980). *For those condemned to study the past: Reflections on historical judgment*. Decision Research. https://doi.org/10.1017/CBO9780511809477.024

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.032.

Flanagan, B., & Hannikainen, I. R. (2022). The folk concept of law: Law is intrinsically moral. *Australasian Journal of Philosophy, 100*(1), 165–179. https://doi.org/10.1080/00048402.2020.1833953

Fletcher, G. (1971). The theory of criminal negligence: A comparative analysis. *University of Pennsylvania Law Review, 119*(3), 401–438. https://doi.org/10.2307/3311308

Fletcher, G. P. (2000). *Rethinking criminal law (reprint)*. Oxford University Press. https://global.oup.com/academic/product/rethinking-criminal-law-9780195136951?cc=pl&lang=en&.

Frisch, L., Kneer, M., Krueger, J., & Ullrich, J. (2022). Do you feel the same? The effect of outcome severity on moral judgment and interpersonal goals of perpetrators, victims, and bystanders. *European Journal of Social Psychology, 51*(7), 1158–1171. https://doi.org/10.1002/ejsp.2805

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

Gardner, J. (2001). The mysterious case of the reasonable person. *The University of Toronto Law Journal, 51*(3), 273–308. https://doi.org/10.2307/825941

Gardner, J. (2015). The many faces of the reasonable person. *Law Quarterly Review, 131*(1), 563–584. https://doi.org/10.1093/oso/9780198852940.003.0009

Gilbert, E., Tenney, E. R., Holland, C., & Spellman, B. A. (2014). Counterfactuals, control, and causation: Why knowledgeable people get blamed more. *Personality and Social Psychology Bulletin, 41*(5), 643–658. https://doi.org/10.2139/ssrn.2463520

Gill, M., & Keil, F. (2022). What is a consumer product for? How teleology guides judgments of product liability. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the annual meeting of the cognitive science society* (pp. 1019–1024). https://escholarship.org/uc/item/37820879.

Gino, F., Moore, D. A., & Bazerman, M. H. (2009). No harm, no foul: The outcome bias in ethical judgments. In *Harvard business school NOM working paper* (pp. 8–80). https://ideas.repec.org/p/hbs/wpaper/08-080.html.

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes, 111*(2), 93–101. https://doi.org/10.1016/j.obhdp.2009.11.001

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. Vanerbilt. *Law Review, 67*, 1327–1372. https://scholarship.law.vanderbilt.edu/vlr/vol67/iss5/2/.

Giroux, M., Coburn, P., Harley, E., Connolly, D., & Bernstein, D. (2016). Hindsight bias and law. *Zeitschrift für Psychologie, 224*, 190–203. https://doi.org/10.1027/2151-2604/a000253

Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology, 26*(2–3), 103–117. https://doi.org/10.1080/01973533.2004.9646399

Güver, L., & Kneer, M. (2022). Causation and the silly norm effect. In S. Magen, & K. Prochownik (Eds.), *Advances in experimental philosophy of law*. https://ssrn.com/abstract=4047203 (forthcoming).

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of Learning and Motivation, 61*, 41–102. https://www.ucl.ac.uk/lagnado-lab/publications/harris/Hahn_Harris_L&M2014.pdf.

Hájek, A. (2019). Interpretations of probability. In *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/.

Hall, J. (1963). Negligent behavior should be excluded from penal liability. *Columbia Law Review, 63*(4), 632–644. https://doi.org/10.2307/1120580

Hammond, C., & Schwartz, P. (2005). Ethical issues related to medical expert testimony. *Obstetrics and Gynecology, 106*, 1055–1058. https://pubmed.ncbi.nlm.nih.gov/16260525/.

Hannikainen, I. R., Tobia, K. P., de Almeida, G.da F. C. F., Struchiner, N., Kneer, M., Bystranowski, P., Dranseika, V., Strohmaier, N., Bensinger, S., Dolinina, K., Janik, B., Lauraitytė, E., Laakasuo, M., Liefgreen, A., Neiders, I., Próchnicki, M., Rosas, A., Sundvall, J., & Żuradzki, T. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences, 119*(44), Article e2206531119. https://doi.org/10.1073/pnas.2206531119.

Hannikainen, I. R., Kneer, M., Tobia, K., Dranseika, V., Almeida, G. D. F. C. F., Poama, A., … Laakasuo, M. (2021, August 24). *Experimental jurisprudence cross-cultural study swap*. https://doi.org/10.17605/OSF.IO/SK7R3

Hannikainen, I. R., Tobia, K., Almeida, G., Donelson, R., Dranseika, V., Kneer, M., … Struchiner, N. (2021). Are there cross-cultural legal principles? Modal reasoning uncovers procedural constraints on law. *Cognitive Science, 45*(8), Article e13024. https://doi.org/10.1111/cogs.13024

Harley, E. (2007). Hindsight bias in legal decision making. *Social Cognition, 25*, 48–63. https://doi.org/10.1521/soco.2007.25.1.48

Hart, H. L. A. (1968). Negligence, Mens Rea and criminal responsibility. In *Hart, punishment and responsibility: Essays in the philosophy of law* (pp. 136–157). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199534777.003.0006.

Hartman, R. J. (2017). *In defense of moral luck: Why luck often affects praiseworthiness and blameworthiness*. Series number 38. Taylor & Francis https://www.routledge.com/In-Defense-of-Moral-Luck-Why-Luck-Often-Affects-Praiseworthiness-and-Blameworthiness/Hartman/p/book/9780367372415.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/1076998600600210

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press. https://idostatistics.com/hedges-olkin-1985-statistical-methods-for-meta-analysis/.

Herring, J. (2012). *Criminal law: The basics* (1st ed.). Routledge https://www.routledge.com/Criminal-Law-The-Basics/Herring/p/book/9780367626969.

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. https://doi.org/10.1006/obhd.1996.0077

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology, 86*(5), 680–695. https://doi.org/10.1037/0022-3514.86.5.680

Hurd, H. M., & Moore, M. S. (2002). Negligence in the air. *Theoretical Inquiries in Law, 3*(2), 1–80. 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

Jaeger, C. B. (2020). The empirical reasonable person. *Alabama Law Review, 72*, 887–957. https://ssrn.com/abstract=3686146.

Jiménez, F. (2022). The limits of experimental jurisprudence. In *Cambridge handbook of experimental jurisprudence*. Forthcoming http://dx.doi.org/10.2139/ssrn.4148963.

Johnston, J. C. (2013). The expert witness in medical malpractice litigation: Through the looking glass. *Journal of Child Neurology, 28*(4), 484–501. https://doi.org/10.1177/0883073813479669

Jurs, A. W. (2013). Utilization of rules 614 and 706 in fact-finding: A recent study of midwest judges. *Drake University Law School Research Paper, 132*(49), 132–138. Available at SSRN: https://ssrn.com/abstract=2303931.

Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences, 23*(5), 681–683. https://doi.org/10.1017/S0140525X00403432

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

Kamtekar, R., & Nichols, S. (2019). Agent-regret and accidental agency. *Midwest Studies In Philosophy, 43*, 181–202. https://doi.org/10.1111/misp.12112

Kant, I. (1787). *Critique of pure reason (15th reprint, 2009)*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804649

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

Kirfel, L., & Hannikainen, I. R. (2022). *Why blame the ostrich? Understanding culpability for willful ignorance*. in preparation https://psyarxiv.com/kswtu/.

Kneer, M. (2018). Perspective and epistemic state ascriptions. *Review of Philosophy and Psychology, 9*(2), 313–341. https://doi.org/10.1007/s13164-017-0361-4

Kneer, M. (2022). Reasonableness on the Clapham omnibus. In P. Bystranowski, B. Janik, & M. Prochnicki (Eds.), *Judicial decision-making: Integrating empirical and theoretical perspectives*. Springer Publishing. forthcoming https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3800110.

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition, 169*, 139–146. https://doi.org/10.1016/j.cognition.2017.08.008

Kneer, M., Hannikainen, I. R., Zehnder, M.-A., Almeida, G., Aguiar, F., Bystranowski, P., Dranseika, V., Janik, B. M., Garcia Olier, J., Güver, L., Lam, J., Liefgreen, A., Próchnicki, M., Rosas, A., Skoczeń, I., Strohmaier, N., Struchiner, N., & Tobia, K. (2022). The severity effect on intention and knowledge. A cross-cultural study with laypeople and legal experts. in preparation https://bit.ly/3aJBtO9.

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. https://doi.org/10.1111/1467-8284.00419

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*(2), 309–324. https://doi.org/10.1080/09515080307771

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33*(4), 315–329. https://doi.org/10.1017/S0140525X10000907

Knobe, J., & Shapiro, S. (2021). Proximate Cause Explained: An Essay in Experimental Jurisprudence. *The University of Chicago Law Review, 88*(1), 165–236. https://www.jstor.org/stable/26966493.

Kumar, V. (2019). Empirical vindication of moral luck. *Nous, 53*(4), 987–1007. https://doi.org/10.1111/nous.12250

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, T. (1988). Court-appointed experts and judicial reluctance: A proposal to amend rule 706 of the federal rules of evidence. *Yale Law and Policy Review, 480*(6), 480–503. https://www.jstor.org/stable/40239296.

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

Liefgreen, A., & Lagnado, D. (2021). The role of causal models in evaluating simple and complex legal explanations. In , *43. Proceedings of the annual meeting of the cognitive science society* (pp. 2316–2322).

Lidén, M., Gräns, M., & Juslin, P. (2019). Guilty, no doubt': detention provoking confirmation bias in judges' guilt assessments and debiasing techniques. *Psychology, Crime & Law, 25*(3), 219–247. https://doi.org/10.1080/1068316X.2018.1511790

Liu, Z. (2018). Does reason writing reduce decision bias? Experimental evidence from judges in China. *The Journal of Legal Studies, 47*(1), 83–118. https://doi.org/10.1086/696879

Lowe, D. J., & 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

Macleod, J. A. (2015). Belief states in criminal law. *Oklahoma Law Review, 68*, 497–554. https://digitalcommons.law.ou.edu/olr/vol68/iss3/2/.

Macleod, J. A. (2019). Ordinary causation: A study in experimental statutory interpretation. *Indiana Law Journal, 94*, 957–1030. https://www.repository.law.indiana.edu/ilj/vol94/iss3/4.

Macleod, J. A. (2021). Finding original public meaning. *Georgia Law Review, 56*, 1–80. https://brooklynworks.brooklaw.edu/faculty/1311/.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology, 72*(1), 293–318.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Margoni, F., Geipel, J., Hadjichristidis, C., & Surian, L. (2018). Moral judgment in old age: Evidence for an intent-to-outcome shift. *Experimental Psychology, 65*(2), 105–114. https://doi.org/10.1027/1618-3169/a000395

Margoni, F., Geipel, J., Hadjichristidis, C., & Surian, L. (2019). The influence of agents' negligence in shaping younger and older adults' moral judgment. *Cognitive Development, 49*, 116–126. https://doi.org/10.1016/j.cogdev.2018.12.002

Margoni, F., & Surian, L. (2021). Judging accidental harm: Due care and foreseeability of side effects. *Current Psychology*. https://doi.org/10.1007/s12144-020-01334-7

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), e0125193. https://doi.org/10.1371/journal.pone.0125193

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition, 147*, 133–143. https://doi.org/10.1016/j.cognition.2015.11.008

Masella, R., & Meister, M. (2001). The ethics of health care professionals' opinions for hire. *Journal of the American Dental Association, 132*, 361–367. https://doi.org/10.14219/jada.archive.2001.0179

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General, 146*(1), 123–133. https://doi.org/10.1037/xge0000234

Mott, C., & Heiphetz, L. (2022). *Mens rea in criminal cases: How contrast affects attribution of culpable mental states*. In preparation.

Murray, S., Krasich, K., Irving, Z. C., Nadelhoffer, T., & De Brigard, F. (2022). Mental control and attributions of blame for negligent wrongdoing. *Journal of Experimental Psychology: General*. PhilArchive copy https://philarchive.org/archive/MURMCA-5v1.

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

Nelkin, D. K. (2004). Moral luck. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/moral-luck/.

Nelkin, D. K. (2019). Thinking outside the (traditional) boxes of moral luck. *Midwest Studies In Philosophy, 43*, 7–23. https://doi.org/10.1111/misp.12101

Nelkin, D. K. (2021). Liability, culpability, and luck. *Philosophical Studies*, 1–19. https://doi.org/10.1007/s11098-021-01612-5

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

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. https://rdcu.be/cXC7A.

Nobes, G., & Martin, J. W. (2022). They should have known better: The roles of negligence and outcome in moral judgements of accidental actions. *British Journal of Psychology, 113*, 370–395. https://doi.org/10.1111/bjop.12536

Pirker, B., & Skoczeń, I. (2022). Pragmatic inferences and moral factors in treaty interpretation—Applying experimental linguistics to international law. *German Law Journal, 23*(3), 314–332. https://doi.org/10.1017/glj.2022.22

Pohl, R. F., & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes, 67*(1), 49–58. https://doi.org/10.1006/obhd.1996.0064

Prochownik, K. (2022). Causation in the law and experimental philosophy. In P. Willemsen, & A. Wiegmann (Eds.), *Advances in experimental philosophy of causation* (pp. 165–188). Bloomsbury Academic. https://doi.org/10.5040/9781350235830.ch-008.

Prochownik, K., & Cushman, F. (2018). Replication of Kneer and Machery. In *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* (pp. 1930–1936). https://cognitivesciencesociety.org/cogsci20/papers/0437/index.html.

Prochownik, K. M. (2021). The experimental philosophy of law: New ways, old questions, and how not to get lost. *Philosophy Compass, 16*(12), e12791. https://doi.org/10.1111/phc3.12791

Rachlinski, J. J. (1998). A positive psychological theory of judging in hindsight. *The University of Chicago Law Review, 65*(2), 571–625. http://scholarship.law.cornell.edu/facpub/801.

Rachlinski, J. J. (2000). Heuristics and biases in the courts: Ignorance or adaptation. *Oregon Law Review, 79*, 61–102. https://scholarship.law.cornell.edu/facpub/810.

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

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science, 7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Schauer, F., & Spellman, B. A. (2020). Probabilistic causation in the law. *Journal of Institutional and Theoretical Economics, 176*, 4–17. https://EconPapers.repec.org/RePEc:mhr:jinste:urn:doi:10.1628/jite-2020-0003.

Schofferman, J. (2007). Opinions and testimony of expert witnesses and independent medical evaluators. *Pain Medicine, 8*(4), 376–382. https://doi.org/10.1111/j.1526-4637.2007.00318.x

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. https://doi.org/10.1111/j.1468-0017.2012.01438.x

Shen, F. X., Hoffman, M. B., Jones, O. D., & Greene, J. D. (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*, 365–398. https://heinonline.org/HOL/LandingPage?handle=hein.journals/contli5&div=16&id=&page=.

Skoczeń, I. (2021). Modelling perjury: Between trust and blame. *International Journal for the Semiotics of Law, 35*, 771–805. https://doi.org/10.1007/s11196-021-09818-w

Skoczeń, I. (2022). From lying to blaming and perjury: deceptive implicatures in the courtroom and the materiality requirement. (In preparation).
Skoczeń, I., & Smywiński-Pohl, A. (2022). The context of mistrust: Perjury ascriptions in the courtroom. In L. R. Horn (Ed.), *From lying to perjury linguistic and legal perspectives on lies and other falsehoods* (pp. 309–353). Mouton de Gruyter. https://doi.org/10.1515/9783110733730-013.
Sommers, R. (2019). Commonsense consent. *The Yale Law Journal, 129*, 2232–2322. https://www.yalelawjournal.org/pdf/SommersArticle_ho84grf3.pdf.
Sommers, R., & Bohns, V. K. (2018). The voluntariness of voluntary consent: Consent searches and the psychology of compliance. *The Yale Law Journal, 128*, 1966–2020. https://www.yalelawjournal.org/pdf/SommersBohns_w4cmjkwe.pdf.
Spamann, H., & Klöhn, L. (2016). Justice is less blind, and less legalistic, than we thought: Evidence from an experiment with real judges. *The Journal of Legal Studies, 45*(2), 255–280. https://doi.org/10.1086/688861
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. https://scholarship.law.duke.edu/lcp/vol64/iss4/10/.
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
Strohmaier, N., Pluut, H., van den Boos, K., Adriaanse, J., & Vriesendorp, R. (2021). Hindsight bias and outcome bias in judging directors' liability and the role of free will beliefs. *Journal of Applied Social Psychology, 51*, 141–158. In press https://doi.org/10.1111/jasp.12722.
Teichman, D. (2014). The hindsight bias and the law in hindsight. In E. Zamir, & D. Teichman (Eds.), *The Oxford handbook of behavioral economics and the law* (pp. 354–373). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199945474.013.0014.
Tobia, K. (2018). How people judge what is reasonable. *Alabama Law Review, 70*, 293–359. https://doi.org/10.1177/17456916221096110
Tobia, K. (2020a). *Legal concepts and legal expertise*. Available at SSRN: https://ssrn.com/abstract=3536564 or http://dx.doi.org/10.2139/ssrn.3536564.
Tobia, K. (2022). Experimental jurisprudence. *The University of Chicago Law Review, 89*(3), 735–802.
Tobia, K. P. (2020b). Testing ordinary meaning. *Harvard Law Review, 134*, 726–805. https://harvardlawreview.org/wp-content/uploads/2020/11/134-Harv.-L.-Rev.-726.pdf.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03
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 & the Law, 7*, 485–504. https://doi.org/10.1002/bsl.2370070406
Williams, B. (1981). *Moral luck: Philosophical papers 1973–1980*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165860
Wittlin, M. (2016). Hindsight evidence. *Columbia Law Review, 116*, 1323–1394. https://columbialawreview.org/content/hindsight-evidence/.
Yaffe, G. (2012). Intoxication, recklessness and negligence. *Ohio State Journal of Criminal Law, 9*, 545–583. http://hdl.handle.net/20.500.13051/3141.
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. https://doi.org/10.1007/s13164-010-0027-y
Zipursky, B. C. (2014). Reasonableness in and out of negligence law. *University of Pennsylvania Law Review, 163*, 2132–2168. https://ir.lawnet.fordham.edu/faculty_scholarship/625.

United Kingdom case law

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).
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).

United States case law

Ambrosini v. Labarraque. (1996). *101 F.3d 129, D.C. Cir*. Available at: https://casetext.com/case/ambrosini-v-labarraque-2 (Accessed: 04.11.2020).
Claar v. Burlington N. R.R. (1994). *29 F.3d 499, 9th Cir*. Available at: https://casetext.com/case/claar-v-burlington-northern-r-co (Accessed: 04.11.2020).
Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993). *509 U.S. 579*. Available at: https://www.law.cornell.edu/supct/html/92-102.ZS.html (Accessed: 23.10.2020).
Day v. Bryant. (2010). *697 S.E.2d 345*. 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. (1923). *293 F. 1013 D.C. Cir*. Available at: https://en.wikisource.org/wiki/Frye_v._United_States (Accessed: 23.10.2020).
Griffen v. Univ. of Pittsburgh Med. Ctr.-Braddock Hospital. (2008). *950 A.2d 996 Superior Ct. Pa*. Available at: https://www.courtlistener.com/opinion/2335632/griffin-v-university-of-pittsburgh-medical/ (Accessed: 04.11.2020).
Johns Hopkins Hospital v. Genda. (1969). *258 A.2d 595 Md*. Available at: https://www.courtlistener.com/opinion/1970301/johns-hopkins-hospital-v-genda/ (Accessed: 05.11.2020).
Navarro v. Austin. (2006). *928 So.2d 348*. Available at: https://www.morelaw.com/verdicts/case.asp?n=Unknown10/4/2006&s=FL&d=32025 (Accessed: 05.11.2020).