Moral Luck: Mechanisms, Robustness, and Prevalence[*]

Armin Falk Sven Heuser David Huffman

Abstract

In many situations people can take moral or immoral actions that influence the probability that someone is harmed, but chance still plays a role in determining what happens. It has been shown that the punishment actors receive in such cases tends to be influenced by the (randomly determined) occurrence of harm, a tendency known as "moral luck." This paper investigates a long-standing open question, of whether moral luck reflects a considered moral judgement, i.e., a preference, or whether it is a bias. Understanding the mechanism is important, because moral luck is potentially very prevalent, arising in legal, employment, and political spheres, and because it violates a property of optimal incentives. We use a controlled experiment that can cleanly identify moral luck, but which also involves a consequential moral outcome: Life or death (for a mouse). We show that moral luck in punishment is at least partially a bias: Random outcomes influence judgements and incentivized beliefs about the nature of the actor, despite containing zero information; a treatment that shuts down this channel, by providing information about the actor, reduces moral luck in punishment, establishing a causal link from biased beliefs to punishment. A treatment that encourages deliberation also reduces moral luck, consistent with a bias mechanism. Additional results suggest that the bias is at least partly due to emotion, and suggest that actors also internalize moral luck in self-judgements. Novel implications include moral luck being stronger for actors who are less well known, and for outcomes that are more upsetting.

[*] This project has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7-2007-2013) (Grant agreement No. 340950). Funding was also provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2126/1-390838866. The main experiment was pre-registered at the AEA RCT Registry (AEARCTR-0008260). Falk: Institute for Behavior and Inequality; Armin.Falk@briq-institute.org. Heuser: University of Bonn; sven.heuser@uni-bonn.de. Huffman: University of Pittsburgh; huffmand@pitt.edu.

1 Introduction

In many situations, individuals can can choose between moral or immoral actions that affect the probability of harm to others, but which leave a role for chance in determining what ultimately happens. For example, driving under the influence of alcohol may increase the probability of subsequently hitting and killing a pedestrian if a pedestrian crosses the street, but the presence of a pedestrian depends on chance. Likewise, an employee can take self-interested actions that expose the employer to increased risk of a loss, but chance will ultimately determine whether the loss occurs. More generally, in meritocratic societies, individuals can have a strong work ethic and exert high effort, but due to bad luck still end up being unsuccessful. There is evidence that in such cases, the punishments and rewards that actors receive are sensitive to whether harm occurs, despite this being randomly determined, a tendency sometimes called "moral luck."[1]

What is less understood, however, is the mechanism underlying moral luck; specifically, whether it reflects a considered moral judgement, i.e., a preference, or is instead a mistake or bias, and if the latter, what is the source of the bias. Understanding the mechanism is important because moral luck is potentially very prevalent, arising in legal, employment, and political spheres, and because it violates a property of optimal incentives, sometimes called the "informativeness" principle (Holmström, 1979; Bolton and Dewatripont, 2004). Of course, in many real-world applications, actions of agents may be only imperfectly observed, so that outcomes are a signal of hidden action, in which case optimal incentives do involve conditioning on outcomes; this does not mean moral luck is irrelevant, however, as it can cause punishments to be *more sensitive* to outcomes than can be justified by their information content, thereby leading to suboptimal incentives. If moral luck is a preference, then there is a potential tradeoff, between inefficiency, and satisfying some form of moral preference that calls for greater punishment when harm is greater. If moral luck is a bias, however, this makes a clearer case for developing interventions and institutions to mitigate the tendency.

This paper provides evidence that moral luck in punishment is at least partly caused by a bias. Our findings indicate that random outcomes influence spectator judgements, and incentivized beliefs, about the nature and preferences of the actor. Because the random outcomes contain zero information in the controlled setting of our study, this is a clear bias in percpetions. To test for a causal impact on punishment behavior, we conduct a treatment that provides additional information about the preferences of the actor, thereby partially shutting down the potential for outcomes to influence beliefs about the actor. This causes punishment to vary less with random outcomes, showing that moral luck in punishment is at least partly caused by biased perceptions. An-

[1] See, e.g., Gurdal et al., 2013; Brownback and Kuhn, 2019; Kneer and Skoczeń (2023).

other treatment, which encourages deliberation, also reduces moral luck, consistent with a bias mechanism, although moral luck is not eliminated, suggesting that the bias is strong. Exploring underlying mechanisms, we find that a key determinant of moral luck is how consequential the spectator views the outcome to be. Those who care about the outcome have a stronger reported emotional reaction, and exhibit stronger moral luck, consistent with an emotional component to the bias. We find less evidence for a cognitive component, in that the bias is orthogonal to measures of cognitive ability or education, and factors such as hindsight bias are ruled out by the design of our study.

Our paper uses an approach that helps address some methodological challenges to studying moral luck. Using observational data, it is difficult to cleanly identify moral luck, because the role of action versus chance, as perceived by the punishers, is hard to establish. An approach often taken in psychology, of using vignettes that describe an actor taking an action, and vary the severity of harm that occurs, also face a similar challenge.[2] A solution is to use controlled laboratory experiments, in which the role of action versus chance is made explicit. Specifically, an actor makes a choice associated with different, explicitly given probabilities of harm (see Gino et al., 2008; Gurdal et al., 2013; Brownback and Kuhn, 2019). A challenge with experiments, however, is having consequential moral outcomes. We adopt the paradigm developed in Falk and Szech (2013), in which choices in the experiment determine whether a third party (a mouse) lives or dies. The crucial feature of the paradigm is that it uses a population of "surplus mice" owned by research laboratories, who will die by default; research money from our study is used to buy and rescue mice, and therefore our study improves the welfare of this population of mice. Having consequential outcomes is desirable in order to recruit potential mechanisms, e.g., strong emotion, in a realistic way, and thus have a way to gauge of how strong the bias will be in real world applications, and how resistant to interventions, e.g., ones encouraging deliberation. The value of the life of a mouse is also quit heterogeneous across individuals, in contrast to money or other rewards, allowing an exploration of how the strength of moral luck varies with the consequentiality of the outcome.

The findings in our paper are relevant for several literatures. They contribute to the large theoretical literature on optimal incentives, by showing how principals may deviate from providing optimal incentives, and identifying novel factors that can matter for the extent of the deviation. Our findings also complement a previous literature in psychology and behavioral economics on moral luck and the related idea of "outcome bias" (e.g., Robbenholt, 2020; Martin and Cushman, 2016; Gurdal et al., 2013; Brownback and Kuhn, 2019; Kneer and Skoczeń (2023).). While previous studies have shown that

[2] Gino et al., 2008 discuss the "insider information" problem of such vignette approaches, in which subjects may infer from the outcome something about the efforts or precautions of the actor.

punishment varies with random outcomes, and that judgements and beliefs are also influenced by such outcomes, our paper contributes evidence on the causal link between these two, helping to establish that moral luck in punishment is at least partly a bias. Our findings also have novel implications, e.g., showing that providing information about an actor can mute moral luck, and that moral luck is stronger for more upsetting or consequential outcomes. The findings of our paper are also relevant for theories of social preferences. Models of intentions-based reciprocity predict that individuals will engage in costly punishment of actions that reveal bad or selfish intentions (e.g., Rabin, 1993; Levine, 1998; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). Our evidence shows how random outcomes can influence punishments through a type of biased reciprocity. Such models could be augmented to allow for perceptions of intentions being skewed by outcomes, which in turn causes punishment to vary.

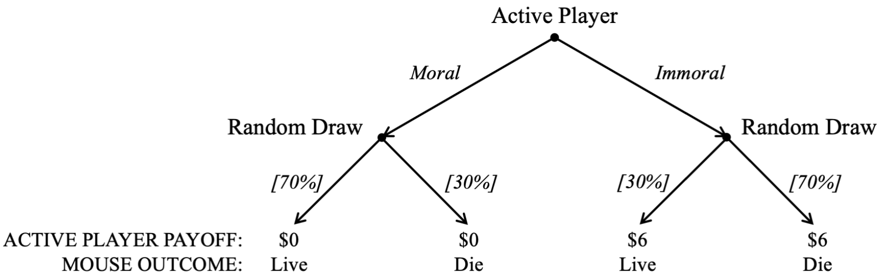
2 Design of the experiments

In a first stage of our experiments, shown in Figure 1, subjects in the role of active players make a choice between two lotteries, denoted the moral lottery and the immoral lottery, where outcomes are consequential in that they involve life or death for a third party. Specifically, the immoral lottery involves a 70% chance that a mouse dies, and a 30% chance that a mouse is instead rescued from death. The immoral lottery gives the active player $6 for themselves, regardless of the outcome for the mouse. The moral lottery, by contrast, involves only a 30% chance of death for the mouse, and a 70% chance that it is rescued, but gives the active player no money. An active player who chooses the immoral lottery thus indicates a willingness to increase the risk of death for the mouse, in order to achieve personal financial gain, whereas choice of the moral lottery reflects a willingness to sacrifice personal gain, in order to reduce likelihood of death for the mouse.

Our study uses the mouse paradigm developed in Falk and Szech (2013), where a key feature is that the population of mice used will be killed by default, in the absence of intervention through the study, and thus the scientific study can only improve welfare for the mice. The mice in question are ordinary laboratory mice, bred by a company for, e.g., medical research, but slated to be euthanized by the company to do lack of demand. If it is determined in our study that a mouse should be rescued, our research money is used to purchase one of these "surplus mice" from the company, and allow the mouse to live out the rest of its natural life in a hygienic environment with other mice.

The first stage of our study also elicits traits and judgements of the active players.

Figure 1: Stage 1 of experiment: Active player choices and outcomes

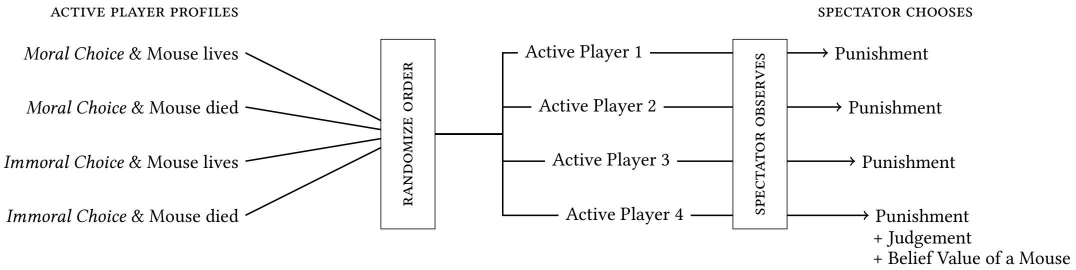


Notes: The active player first chooses one of two options, shown in the figure as *moral* or *immoral*, although more neutral, factual labels "option likely live" and "option likely die" were used with subjects. The moral choice leads to a subsequent random draw with a low probability of death for the mouse, 30%, and gives the active player no money regardless of what happens to the mouse. The immoral choice leads to a random draw with a high probability of death for the mouse, 70%, and gives the active player $6 regardless of what happens to the mouse. Note that the default for such surplus laboratory mice is to be killed, so the study is rescuing mice.

Specifically, we measure an active player's "value of the life of a mouse" using a question asking how much they would need to be paid, in order to allow a mouse to die for sure. In addition, the study measures active players' judgements about, e.g., the morality of their own choice, and whether they see themselves as a good person, after learning what happens to their mouse. The active players also have an additional, "pending payment" of $12; how much of this they receive depends on the choices of spectators in stage 2 of our experiment. We use university students as active players (N=562).

In the second stage of our experiment, which was pre-registered, we recruit a large sample of US adults to participate in online experiments in the role of spectators; our main treatment, Treatment Main, has N=2,200. We explain the concept of surplus mice to spectators, and elicit their (hypothetical) value of a life of a mouse. As was pre-registered, our analysis focuses only on spectators who have more than a minimal value for mice, to eliminate those who might dislike mice and thus not view active players as facing a moral dilemma. Spectators are given an endowment of $6, and can choose how much of this to spend, in order to reduce the pending payment of an active player. As shown in Figure 2, our design matches a given spectator with a sequence of four active players, so that they see each possible combination of choice and outcome for the mouse. The order of seeing the different active players with different possible choices and outcomes is randomized across spectators, to address any possible order effects. Spectators make a choice of how much money to deduct from each of the four active players, knowing that only one of the four choices will be randomly selected to potentially be implemented. In this sense, our design is an example of the "strategy method," where subjects make choices without knowing for sure which case will be realized. Spectators knew that multiple spectators might be matched to a given active player, in which case it would be randomly determined which spectator's choice was used to determine the active player's payoff. This design allows a within-subject analysis. It can thus can speak to individual heterogeneity in a tendency to condition punishments on random outcomes, as well as the robustness of such a tendency to making the different possible choices and outcomes of actors salient to the spectator.

Figure 2: Stage 2 of experiment: Spectator punishment choices, judgements, and beliefs



Notes: The spectators see a sequence of four different active players, with each possible combination of choice and outcome. The order is randomized across spectators. For each active player, the spectator has $6 to spend on punishment, with each dollar spent deducting two dollars from a pending $12 payoff of the active player. Spectators are asked for judgements and beliefs about the fourth active player that they see. Spectators know that one of the four active players will be randomly selected, and their punishment choice in that case will affect their payoff and potentially the payoff of the active player.

The study also elicited judgements of the spectators about the fourth active player they saw, e.g., in terms of morality of the choice, and whether the active player was a good person. The elicitation asks only about the final active player a spectator saw, to reduce complexity of asking about all previous active players, and to focus on the one that was discussed most recently. We can compare across spectators how choices and outcomes affect judgements and emotions, because order is randomized. We also elicit incentivized beliefs of the spectator, about how the active player answered the question about value of the life of a mouse, paying the spectator for correctly guessing the money range indicated by the active player.

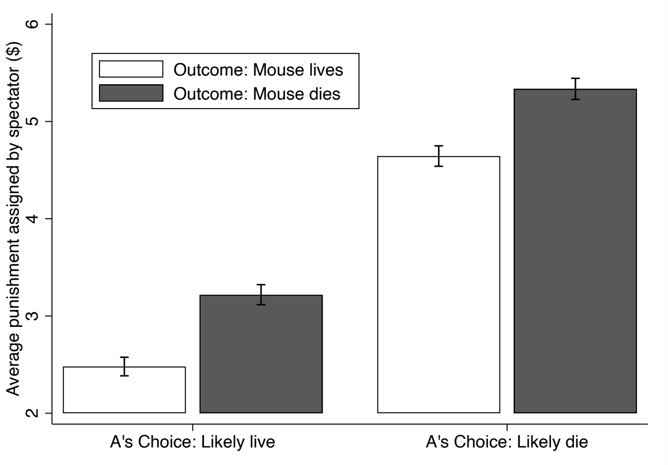
The rest of the study measures additional traits of the active player, and also assesses whether spectators exhibit moral luck in their judgements of hypothetical scenarios that span a range of contexts from crime, to politics, to economic interactions. Key traits that are measured include cognitive ability, captured by the cognitive reflection test (CRT) and a subset of Raven's progressive matrixes. We also ask about educational attainment. The questionnaire elicits agreement with the control principle, beliefs about the role of chance in determining outcomes like poverty in the US, and political affiliation and self-reported conservatism. Additional demographics include traits such as gender, age, and religion.

3 Results

3.1 Moral luck in punishment

Figure 3 shows our first set of results from Treatment Main, on whether there is moral luck in punishment choices. The figure shows average punishment levels by choice of the active player and outcome for the mouse, using all choices of spectators for a within-subject analysis. We see that punishments are on average significantly higher for active players who choose the immoral lottery, compared to those who choose the moral lottery, consistent with spectators sanctioning an immoral choice (OLS; s.e. clustering on spectator; $p < 0.001$). Punishments also vary significantly, however, with the outcome for the mouse, conditional on the active player's choice. For both the moral choice and the immoral choice, active players are punished significantly less if the mouse lives than if the mouse dies (OLS; s.e. clustering on spectator; $p < 0.001$, $p < 0.001$). Punishment choices thus violate the informativeness principle, in that active players are not being punished solely based on factors under their control. Results are similar and also statistically significant in a between-subject comparison, using only first choices of spectators (see Figure A1). This shows that the result is robust in the sense that it is not confined to within-subject contrasts. These findings raise the question whether moral luck in punishment reflects some alternative moral principle, or whether instead it is a mistake or bias.

Figure 3: Punishment levels by active player choice and outcome in Treatment Main



Notes: Average punishment levels for each of the four cases. "A's choice" refers to Active Player's choice. Each spectator chooses punishment for all four cases so there are four observations per spectator (within-subject comparison). Figure shows standard error bars clustering on spectator.

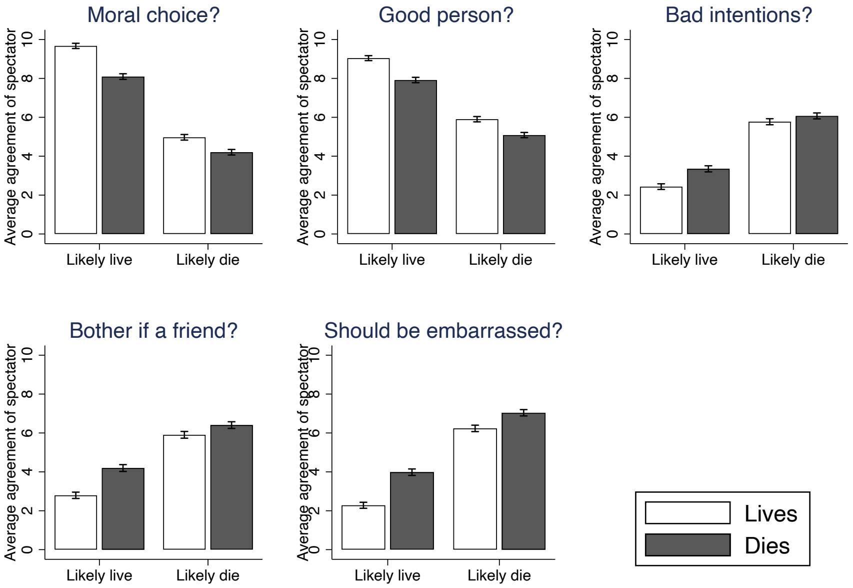
3.2 Biased judgements and beliefs about the active player

Figures 4 and 5 explore one possible explanation (pre-registered), which is that punishments might vary with random outcomes because these influence judgements and beliefs about the nature of the active player, despite the fact that the outcome conveys zero information above and beyond the observed choice. Figure 4 shows that for both the moral and immoral choice, the mouse dying causes spectators to judge the active player's choice as relatively more immoral, and to agree less that the active player is a good person. It also causes spectators to agree more strongly that the active player should be embarrassed, and that the active player had bad intentions. The impact of the mouse dying on each of these judgements is statistically significant, for both the moral and immoral choice, although interestingly, the effect of outcomes on most judgements is significantly stronger for the moral choice (see Table A1). Results are robust to controlling for spectator characteristics (see Table A2). We also elicited spectator beliefs about the active player's value of a mouse, because this might be a belief that is biased by whether the mouse dies, and potentially relevant for spectator punishment decisions. The belief measure also has the advantage that it can be incentivized for accuracy. Panel (a) of Figure 5 shows that the mouse dying significantly influences spectator beliefs about the active player's value of the life of a mouse (t-test; $p < 0.001$), particularly for the moral choice, where the effect is very large and individually significant (t-test; $p < 0.001$). The effect for the immoral choice is also positive, but smaller and not statistically significant individually (t-test; $p < 0.13$), and the effect is also significantly smaller than for the moral choice (OLS; $p < 0.001$). Finding a smaller effect for the immoral choice is consistent with the results on judgements. One explanation for these findings could be that spectators find the immoral choice as relatively more informative about the nature of the active player, and also recognize that it is consistent with only a relatively narrow range of values for life of a mouse, and this acts as a constraint on how much outcomes bias judgments and beliefs.

3.3 Biased judgements and beliefs causing moral luck in punishment

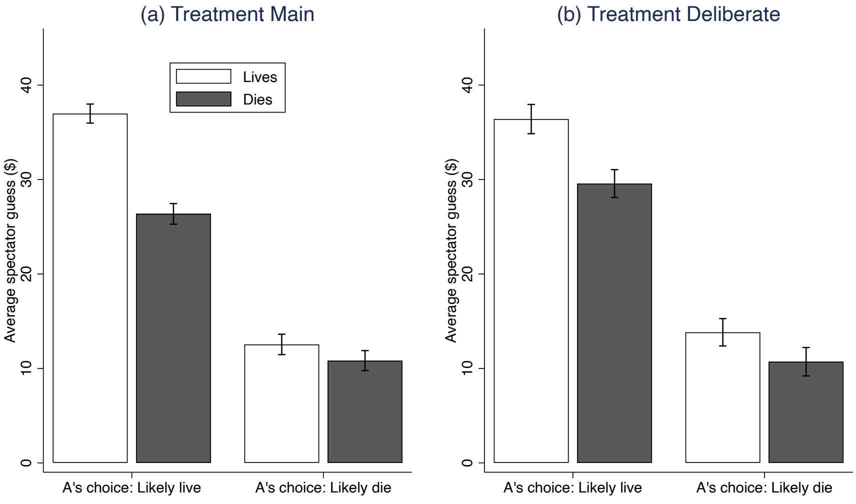
Despite conveying no information, the random outcomes influence judgements and even incentivized beliefs about the active player. If this is a mechanism underlying moral luck in punishment, one would expect punishment choices to be explained by variation in the judgements and beliefs. Table 1 presents evidence that this is the case. Columns (1) and (5) show that spectators who have a less favorable judgements of the active player's choice or nature punish significantly more. Column (6) shows that

Figure 4: Judgements by choice and outcome in Treatment Main



Notes: Average agreement levels for each of the four cases. Each spectator judges one case so there is one observation per spectator (between-subject comparison). Figure shows standard error bars.

Figure 5: Spectator beliefs about the active player's value of the life of a mouse



Notes: Average incentivized guess about the active player's value of the life of a mouse. Each spectator makes a guess for one case so there is one observation per spectator (between-subject comparison). Panel (a) shows results from Treatment Main, Panel (b) from Treatment Deliberation. Figure shows standard error bars.

incentivized beliefs about the active player's value of the life of a mouse are also significantly related to strength of punishment, with punishment decreasing in beliefs about how much the active player values a mouse. The results are robust to adding controls for other factors that matter for punishment choices, including choice and outcome of the active player, and spectator characteristics including own value of a mouse, income, education, gender, and educational attainment (see Table A3). These findings are consistent with the bias in perceptions of the active player, caused by the (uninformative) random outcomes, being a mechanism behind why punishment varies with random outcomes, but the evidence is correlational.

Table 1: Punishment choices in Treatment Main as a function of judgements and beliefs

	Punishment ($)					
	(1)	(2)	(3)	(4)	(5)	(6)
Moral choice	-1.39***					
	(0.10)					
Good person		-1.26***				
		(0.11)				
Bad intentions			1.34***			
			(0.10)			
Bother if a friend				1.27***		
				(0.11)		
Embarassing					1.30***	
					(0.10)	
Belief about active player						-0.77***
						(0.10)
Constant	3.80***	3.84***	3.75***	3.77***	3.75***	4.02***
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)
Observations	1441	1446	1443	1446	1446	1446
Adjusted R^2	0.127	0.100	0.112	0.096	0.105	0.037

Notes: OLS regressions. The dependent variable is punishment in dollars of the fourth active player seen by the spectator. Independent variables include self-reported judgements about the active player given information about the fourth active player's choice and outcome for the mouse: Morality of active player's choice; active player is a good person; it would bother the spectator if active player was a friend; active player had bad intentions. Another independent variable is the spectator's incentivized guess about the active player's value of the life of a mouse, in dollars. All independent variables are standardized, so coefficients give the impact of a one standard deviation increase in the independent variable. Each spectator makes judgements and reports beliefs for one case so there is one observation per spectator (between-subject comparison). Robust standard errors in parentheses.

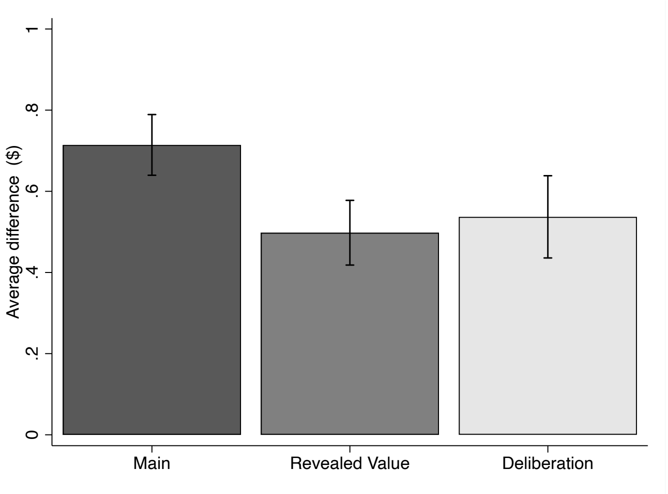
To provide evidence on whether the impact of random outcomes on perceptions of the active player is a causal mechanism explaining moral luck in punishment, we conducted a second treatment, Treatment Revealed Value (N=1,000). In this treatment, spectators learned the active player's value of the life of a mouse, along with the choice and the outcome for the mouse. Each spectator was matched with four active players, two who chose the moral lottery and two who chose the immoral lottery. The active players who chose the moral lottery had different outcomes for the mouse, but had the same (high) value of a mouse, while the active players choosing the immoral lottery had different outcomes but the same (low) value of a mouse.[3] The key feature of the design is that value of a mouse was known to the spectator, and constant across active players

[3] The information conveyed about value of a mouse was calibrated to be line with priors conditional on choices. We used the modal guesses of spectators in Treatment Main, about values of active players choosing the moral or immoral lotteries, respectively, and selected active players with these values to use for the matching.

with different outcomes, so there was no scope for random outcomes to influence beliefs. If part of the reason why punishment varies with outcomes in Treatment Main is the bias in beliefs about the active player value of a mouse, we would expect moral luck to be weaker in Treatment Revealed Value.

As shown in Figure 6, we find that punishment does, indeed, vary substantially less with the random outcome in Treatment Revealed Value compared to Treatment Main, a reduction of about 30 percent. This treatment difference is also statistically significant ($p < 0.05$; see Table A4). Moral luck is still, however, present and significant in Treatment Revealed Value, with punishment significantly stronger for cases when the mouse dies (t-test; $p < 0.001$). This is not unexpected, given that Treatment Revealed Value only eliminated the bias in a single aspect of how the active player is perceived, among a bundle of different types of judgments that are influenced by random outcomes and all seem to matter (correlationally) for punishment. Indeed, we find that judgements about the active player are still significantly skewed by random outcomes in Treatment Revealed Value, and these effects are not significantly different from in Treatment Main (see Table A5).

Figure 6: Punishment of die minus punishment of live: Average difference by treatment



Notes: Average difference in punishment of die minus punishment of live. Each spectator chooses for all four cases so there are four observations per spectator (within-subject comparison). Figure shows standard error bars clustering on spectator.

In another treatment, Treatment Deliberation, we investigate whether moral luck is robust to encouraging deliberative rather than intuitive thinking. We prime individuals to deliberate, through an essay asking about times when deliberation lead to good decisions, and intuition to bad, and also require a minimum time of 30 seconds to assign punishments, make judgements, and form beliefs. This treatment is based on previous approaches to encourage deliberative rather than intuitive decision making

(Rand et al., 2012; Gino et al., 2008). If moral luck is weakened in this condition, it would suggest that violations are at least partly due to a mechanism of intuitive judgements, that are swayed by salient random outcomes when decisions are made quickly and spontaneously.

Figure 6 shows that encouraging deliberation does have a directional effect of reducing moral luck, leading to less variation in punishment with random outcomes, but the difference relative to Treatment Main is not statistically significant ($p < 0.16$; Table A4). Indeed, moral luck in punishment is still significant within Treatment Deliberation ($p < 0.001$; Table A4), and we also see signs of the bias mechanisms identified in Treatment Main, albeit somewhat weaker. Panel (b) of Figure 5 shows that beliefs about the active player's value of a mouse are significantly influenced by the mouse dying in Treatment Deliberation, overall (t-test; $p < 0.02$) and for both the moral and immoral choices individually (t-tests; $p < 0.01$, $p < 0.04$), and the effect is weaker for the immoral choice like in Treatment Main, although this difference is not statistically significant in Treatment Deliberation (OLS; $p < 0.22$). Overall, the impact of the mouse dying on beliefs is directionally weaker in Treatment Deliberation compared to Treatment Main, but the difference is not statistically significant (OLS; robust s.e. clustering on spectator; $p < 0.12$). We also see that outcomes significantly influence judgements in Treatment Deliberation. As for beliefs, the effects tend to be weaker than in Treatment Main, although the difference is significantly in the case of some judgements (see Table A5). Thus, it appears that deliberation may be directionally reducing moral luck by reducing, but not eliminating, the impact of outcomes on various judgements and beliefs. The fact that moral luck in punishment and the bias mechanisms are still present and significant in Treatment Deliberation suggest that the bias is relatively deeply-rooted and not easily corrected.

Taken together, our results are consistent with moral luck in punishment being at least partly a bias, with random outcomes biasing judgements and beliefs about the active player, which in turn cause punishments vary with outcomes. These findings raise questions about what might be the source of the bias.

3.4 Investigation of deeper mechanisms

The bias we identify raises questions about what might be the deeper mechanisms that drive the bias; in exploratory analysis, we investigate three possible mechanisms – belief in a just world; hindsight bias or limited salience of counterfactuals; emotional impact of outcomes – and find some support for the final mechanism.

The first two mechanisms would involve spectators viewing the bad outcome as more likely, if it occurs, and thereby potentially viewing the actor's choice as more immoral in that case. Belief in a just world is a type of motivated bias, such that people want to believe that bad things happen to bad people (Rubin and Peplau, 1975). Hindsight bias is a tendency for ex-post beliefs about the likelihood of an outcome to be greater than ex-ante, and has been hypothesized to reflect the fact that outcomes that occur are more salient than counterfactual outcomes (Roese and Vohn, 2012). A factor that works against these biases in our design, however, is the use of explicit probabilities. We also elicited spectator beliefs about the role of chance versus effort in determining inequality and poverty in the United States, as a proxy for belief in a just world, but find that the impact of the mouse dying on punishment is not significantly different depending on the extent of between belief in a just world (see Table A6;$p < 0.89$). The fact that we find strong moral luck in a within-subject design, where spectators make choices for all possible choices and outcomes, and counterfactuals are therefore salient, provides another indication that hindsight bias is not likely to be a key driver of the results.

The third mechanism would involve moral luck being stronger for individuals who have stronger emotional reactions. We first check whether outcomes affected emotions using a survey measure of self-reported emotions "about the fourth active player," on a scale from strongly negative to strongly positive, and find that the mouse dying significantly decreases positive emotions in the case of the moral choice (t-test; $p < 0.001$), and exacerbates negative emotions in the case of the immoral choice (t-test; $p < 0.001$). We hypothesize that spectators who experience stronger emotions in response to outcomes are those who care more about the outcome. As a candidate proxy for a trait of caring more about the outcomes, we take the spectator's own value of a mouse, and find that the impact of the mouse dying on negative emotions about the active player is, indeed, significantly stronger for spectators with a higher value of a mouse (OLS; $p < 0.002$). Turning to mechanisms for moral luck, there is a significantly stronger bias in beliefs about the active player's value of a mouse, for spectators who have a higher value themselves (OLS; $p < 0.001$), and a significantly stronger effect of the mouse dying on punishment for spectators who value a mouse more (Table A6; $p < 0.001$). One implication of these findings is that moral choices involving bad outcomes that are more emotionally upsetting may be more likely to generate moral luck. Another is that heterogeneity in moral luck may be partly explained by heterogeneity in how much punishers care about a given type of outcome.

3.5 Prevalence of moral luck

Our within-subject design and use of a non-student sample allows us to investigate the prevalence of moral luck as a bias, as well as have meaningful variation in demographics and other correlates to explore whether the bias varies systematically across different segments of society. We find that exhibiting moral luck, defined as punishing more on average when the mouse dies than when the mouse lives, is the modal choice pattern in Treatment Main. Specifically, if we eliminate the 9 percent of spectators who do not exhibit moral luck because they never punish at all, we find roughly 64 percent exhibit moral luck, while 36 percent exhibit zero moral luck. Thus, moral luck is prevalent but not universal. We do not find significant differences in propensity to exhibit moral luck, or magnitude of moral luck, by gender, age, income, education, or political affiliation. Thus, the bias is found for individuals from across society. As noted above, one trait that does predict strength of moral luck is the spectator's own value of a mouse, pointing to caring about the outcome as a key moderator for moral luck in punishment.

4 Conclusion

This paper provides evidence that actors are punished partly for outcomes that are beyond their control, an example of "moral luck." The results shed light on a long-standing question of whether moral luck is a preference or a bias, showing that it is at least partly a bias. The findings point to a role for emotion in generating the bias, with more upsetting outcomes leading to stronger moral luck. The conclusion that moral luck is at least partly a bias has important policy implications, suggesting a value of interventions to reduce moral luck, in the many different spheres where it may play a role, from legal settings, to economic interactions, to politics. Our findings show the effectiveness of two types of interventions – providing information about the character of an actor, and priming deliberation – but also offer the caveat that these may be insufficient to eliminate the bias.

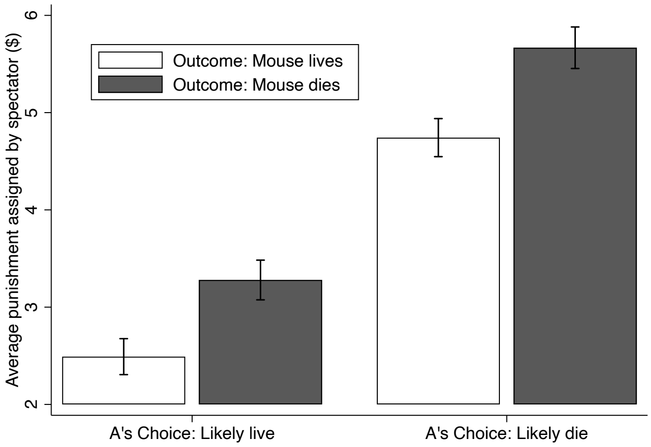
Because we elicited judgments of active players about themselves, we can also explore an intriguing, additional question, which is whether moral luck is to some extent internalized by actors. Adam Smith and others have hypothesized that moral luck is internalized in this way, and one can also find examples from literature with this theme. For example, in ancient Greek tragedy, Oedipus kills his father in a roadside conflict, and marries his mother, without knowing their identities; when he later discovers what he has done, he blinds himself, and goes into exile, even though he would presumably not have done had his vanquished opponent, and his wife, been unrelated to him. If random outcomes influence actors in how they judge themselves, and even potentially punish themselves (psychologically through feelings of guilt, or possibly through costly actions like "penance"), this would be a particularly striking form of moral luck, given that actors presumably have greater certainty about their own characters than external spectators.

We do find evidence of internalized moral luck for active players, although it differs in an interesting way from that of spectators. Specifically, active players judge their own immoral choice as significantly less immoral if the mouse lives than if the mouse dies (Table A7; $p < 0.03$). There is also suggestive evidence that actors who make the immoral choice change their view about being a good person based on the outcome, relative to a baseline assessment before their choice; the reduction in self-esteem if the mouse dies is marginally significant for individuals who have above median baseline self-image and therefore do not have a floor effect working against a reduction (Table A7; $p < 0.08$). Interestingly, however, we find an asymmetry, in that for active players there is little internalized moral luck for the moral choice. Active players view the moral choice as highly moral, regardless of the outcome, and also do not adjust their views of themselves as a good person (Table A7; $p < 0.9$, $p < 0.16$). These findings suggest that actors have a conviction that the moral action clearly indicates a good character, which cannot be shaken by having the mouse die, whereas they have more malleable views about the immoral action. This could potentially be motivated, if actors want to believe they are a good person; it may be possible to convince themselves of this in all cases, except for the immoral choice with the mouse dying. At the same time, we see that actors' feelings of embarrassment vary significantly with the outcome, for the moral as well as the immoral choice (Table A7; $p = 0.01$, $p < 0.01$). This suggests that actors anticipate that others may evaluate them based on outcomes for the moral choice, even if they themselves do not do so. This asymmetry in external versus internal moral luck that we find is in line with the type of tension hypothesized to arise in meritocracies, by Sandel (2019) and others, such that individuals who have had bad luck feel unfairly judged by others. Also, good or bad luck may have lasting influences on how individuals view themselves.

Online Appendix

A Additional results

Figure A1: Punishment levels in Treatment Main using only first choices for between-subject comparison



Notes: Average punishment levels for each of the four cases, with separate groups of subjects in each category. Figure shows standard error bars. Punishment is significantly higher when the mouse dies than when the mouse lives, for both the moral and immoral choice (OLS; $p < 0.004$, $p < 0.001$).

Table A1: Moral luck in judgments in Treatment Main

	Moral	Good	Embarrassing	Bother if friend	Bad intent.
	(1)	(2)	(3)	(4)	(5)
Immoral choice	-4.70***	-3.14***	3.34***	3.11***	3.95***
	(0.20)	(0.19)	(0.21)	(0.24)	(0.22)
Die	-1.58***	-1.13***	0.91***	1.40***	1.70***
	(0.18)	(0.17)	(0.20)	(0.24)	(0.22)
Immoral*Die	0.81***	0.31	-0.62**	-0.90***	-0.90***
	(0.29)	(0.27)	(0.31)	(0.34)	(0.33)
Constant	9.67***	9.04***	2.44***	2.79***	2.28***
	(0.10)	(0.11)	(0.12)	(0.16)	(0.13)
Observations	1441	1446	1443	1446	1446
Adjusted R^2	0.416	0.281	0.227	0.167	0.275

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) Choice was moral; (2) active player is a good person; (3) active player should be embarrassed; (4) would bother if active player were a friend; (5) active player had bad intentions. Independent variables are indicators for the immoral choice, outcome of mouse dying, and the interaction of these. Self-reported judgements refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Robust standard errors in parentheses.

Table A2: Moral luck in judgments in Treatment Main with controls

	Moral (1)	Good (2)	Embarrassing (3)	Bother if friend (4)	Bad intent. (5)
Immoral choice	-4.70***	-3.13***	3.33***	3.12***	3.92***
	(0.20)	(0.19)	(0.21)	(0.24)	(0.22)
Die	-1.62***	-1.16***	0.96***	1.44***	1.73***
	(0.19)	(0.18)	(0.20)	(0.24)	(0.22)
Immoral*Die	0.82***	0.32	-0.65**	-0.94***	-0.89***
	(0.29)	(0.27)	(0.31)	(0.34)	(0.33)
Constant	9.86***	9.31***	2.87***	2.92***	1.75***
	(0.46)	(0.44)	(0.50)	(0.58)	(0.54)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1441	1446	1443	1446	1446
Adjusted R^2	0.422	0.290	0.244	0.176	0.284

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) Choice was moral; (2) active player is a good person; (3) active player should be embarrassed; (4) would bother if active player were a friend; (5) active player had bad intentions. Independent variables are indicator for the immoral choice, outcome of mouse dying, and the interaction of these. Self-reported judgements refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Controls consist of: The spectator's own value of a mouse; gender; age; income range; educational attainment. Robust standard errors in parentheses.

Table A3: Punishment choices in Treatment Main as a function of judgements and beliefs with controls

			Punishment ($)			
	(1)	(2)	(3)	(4)	(5)	(6)
Moral choice	-1.42***					
	(0.10)					
Good person		-1.28***				
		(0.10)				
Bad intentions			1.40***			
			(0.10)			
Bother if a friend				1.31***		
				(0.11)		
Embarassing					1.34***	
					(0.10)	
Belief about active player						-0.82***
						(0.11)
Constant	3.73***	3.86***	3.43***	3.61***	3.86***	3.38***
	(0.66)	(0.67)	(0.67)	(0.68)	(0.67)	(0.70)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1437	1442	1439	1442	1442	1442
Adjusted R^2	0.135	0.107	0.123	0.106	0.115	0.044

Notes: OLS regressions. The dependent variable is punishment in dollars of the fourth active player seen by the spectator. Independent variables include self-reported judgements about the active player given information about the fourth active player's choice and outcome for the mouse: Morality of active player's choice; active player is a good person; it would bother the spectator if active player was a friend; active player had bad intentions. Another independent variable is the spectator's incentivized guess about the active player's value of the life of a mouse, in dollars. All independent variables are standardized, so coefficients give the impact of a one standard deviation increase in the independent variable. Controls consist of: Dummy variables for choice of the observed (fourth) active player and outcome for the mouse, with moral choice and lives as the omitted category; the spectator's own value of a mouse; gender; age; income range; educational attainment. Self-reported judgements and beliefs refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Robust standard errors in parentheses.

Table A4: Treatment comparisons of moral luck in punishment

	Punishment ($)	
	(1)	(2)
Die	0.71***	0.71***
	(0.07)	(0.07)
T. Revealed Value	0.30**	0.29**
	(0.13)	(0.13)
T. Deliberation	-0.23*	-0.23*
	(0.14)	(0.13)
Die*T. Revealed Value	-0.22**	-0.22**
	(0.11)	(0.11)
Die*T. Deliberation	-0.18	-0.18
	(0.13)	(0.13)
Constant	3.56***	3.65***
	(0.08)	(0.34)
Controls	No	Yes
Observations	12120	12116
Adjusted R^2	0.007	0.010

Notes: OLS regressions. Dependent variable is punishment in dollars. Die is an indicator for the outcome of the mouse dying, measuring the differential punishment when the mouse dies versus when the mouse lives, averaged across the cases of moral and immoral choice. T. Revealed Value and T. Deliberation are treatment dummies, respectively. Controls include: The spectator's own value of a mouse; gender; age; income range; educational attainment. Each spectator chooses punishment for all four cases of active player choice and outcome so there are four observations per spectator (within-subject comparison). Robust standard errors in parentheses, clustering on spectator.

Table A5: Treatment comparisons of moral luck in judgements

	Moral	Good	Embarrassing	Bother if friend	Bad intent.
	(1)	(2)	(3)	(4)	(5)
Die	-1.38***	-1.12***	1.43***	1.09***	0.75***
	(0.18)	(0.16)	(0.19)	(0.18)	(0.17)
T. Revealed Value	-0.33	-0.28	0.16	0.28	0.21
	(0.22)	(0.18)	(0.22)	(0.22)	(0.20)
T. Deliberation	-0.21	-0.17	0.23	0.12	0.18
	(0.22)	(0.19)	(0.22)	(0.22)	(0.21)
Die*T. Revealed Value	0.49	0.44*	-0.23	0.04	-0.23
	(0.31)	(0.26)	(0.31)	(0.31)	(0.29)
Die*T. Deliberation	0.55*	0.52**	-0.59*	-0.18	-0.44
	(0.31)	(0.26)	(0.32)	(0.32)	(0.29)
Constant	7.47***	7.58***	4.13***	4.25***	4.00***
	(0.13)	(0.11)	(0.13)	(0.13)	(0.12)
Observations	3008	3030	3030	3030	3027
Adjusted R^2	0.025	0.021	0.028	0.021	0.007

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) Choice was moral; (2) active player is a good person; (3) active player should be embarrassed; (4) would bother if active player were a friend; (5) active player had bad intentions. Independent variable Die is an indicator for the outcome of mouse dying, T. Revealed Value and T. Deliberate are treatment dummies, respectively. Self-reported judgements refer to the fourth active player choice and outcome combination seen by the spectator, so there is one observation per spectator (between-subjects). Robust standard errors in parentheses.

Table A6: Moral luck in punishment choices in Treatment Main as a function of potential mechanisms

	Punishment ($)					
	punish	punish	punish	punish	punish	punish
Die	0.71***	0.72***	0.71***	0.72***	0.72***	0.51***
	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)	(0.08)
Agreement control principle	-0.11					
	(0.08)					
Die*Agreement control principle	0.04					
	(0.08)					
CRT score		-0.29***				
		(0.08)				
Die*CRT score		0.02				
		(0.07)				
Raven's IQ score			-0.33***			
			(0.09)			
Die*Raven's IQ score			-0.01			
			(0.07)			
Educational attainment				-0.01		
				(0.08)		
Die*Educational attainment				0.05		
				(0.07)		
Belief in a just world					-0.11**	
					(0.05)	
Die*Belief in a just world					-0.01	
					(0.05)	
Spectator's value of a mouse						-0.01
						(0.01)
Die*Spectator's value of a mouse						0.38***
						(0.09)
Constant	3.56***	3.54***	3.57***	3.56***	3.59***	3.71***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.21)
Observations	5784	5784	5784	5784	5772	5784
Adjusted R^2	0.008	0.011	0.014	0.007	0.009	0.009

Notes: OLS regressions. The dependent variable is punishment in dollars of the fourth active player seen by the spectator. Independent variables in include: Self-reported agreement with the control principle; CRT test score; Raven's IQ test score; spectator's own value of the life of a mouse. All independent variables are standardized, so coefficients give the impact of a one standard deviation increase in the independent variable. Each spectator makes a choice for all four cases so there are four observations per spectator (within-subjects). Robust standard errors in parentheses, clustering on spectator.

Table A7: Moral luck for active players

	Immoral Choice			Moral Choice		
	Moral	Good	Embarrassed	Moral	Good	Embarrassed
	(1)	(2)	(3)	(4)	(5)	(6)
Die	-0.72**	-0.32*	0.94***	-0.05	0.15	1.78***
	(0.32)	(0.18)	(0.36)	(0.34)	(0.10)	(0.46)
Constant	4.23***	-0.43***	2.73***	8.38***	-0.12*	0.80***
	(0.28)	(0.14)	(0.29)	(0.21)	(0.06)	(0.21)
Observations	257.00	133.00	257.00	170.00	103.00	170.00
Adjusted R^2	0.02	0.01	0.02	-0.01	0.01	0.09

Notes: OLS regressions. Dependent variables in Columns (1) to (5) are measured on scales from 0 to 10, indicating levels of agreement with: (1) and (4) Choice was moral; (2) and (5) I am a good person (difference after-before choice); (3) and (6) I would be embarrassed if a friend learned my choice. Columns (2) and (5) consider only those subjects with above median self-esteem to avoid floor effect. Die indicates whether the mouse died. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$