Moral Luck: Mechanisms, Robustness, and Prevalence*

Armin Falk Sven Heuser David Huffman

Abstract

In many types of decisions, individuals can influence the probabilities of good or bad outcomes by their actions, but there is still a role for chance in determining final outcomes. If punishment and rewards are conditioned on such random outcomes, this violates a property of optimal incentives. It has been posited since ancient times that humans do assign punishments and rewards based on factors outside of actors' control, a tendency called "moral luck." This paper provides new evidence on the prevalence and robustness of moral luck, and on a key open question of whether moral luck is a preference or a bias. The results are from controlled experiments that can cleanly identify moral luck, but also involve real, consequential moral choices that are a matter of life and death for a third party (a mouse). We find moral luck in punishment, and show that this is at least partly due to a bias. Our findings support a causal chain in which random outcomes lead to biased judgments and incentivized beliefs about the nature of the actor, even though they contain zero information, and this in turn causes punishments to vary with outcomes. We also show that the bias is strong enough to remain in the face of an intervention that encourages deliberation. The bias is prevalent, but not universal, it is unrelated to most demographics, and is present regardless of high or low cognitive ability or education. We also find evidence that actors exhibit internalized moral luck in how they evaluate themselves based on outcomes.

*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.

In many types of decisions, individuals can influence the probabilities of good or bad outcomes by their actions, but there is still 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. In all of these cases, realized random outcomes do not contain any additional information about the intentions or effort of the actor beyond the observed actions. If punishments and rewards do vary with such outcomes, however, this violates a property of optimal incentives, sometimes called the "informativeness" principle (Holmström, 1979; Bolton and Dewatripont, 2004). Such violations would have profound implications for the functioning of legal systems, employment relationships, democracies, and meritocratic societies, by undermining the motivating and deterrent value of rewards and punishments.

It has been posited since ancient times that there is, in fact, a human tendency to reward or punish actors for outcomes that are beyond their control (Aristotle, 1984), a phenomenon sometimes called "moral luck," but the prevalence and robustness of this phenomenon are still not fully understood, and a key open question continues to be debated, which is whether moral luck is a preference or a bias. Understanding whether moral luck is a preference or a bias is important, because if it is a preference, having punishments vary with outcomes satisfies some notion of what is appropriate or desired, which might offset the costs of providing suboptimal incentives. If the phenomenon is a bias, however, then this raises important questions about the desirability of how punishments and rewards are determined in many areas of society. It also points to a possible value of interventions to de-bias decisions. An early contributor to the debate on mechanisms was Adam Smith, who proposed that sentiments or emotions aroused by outcomes can affect perceptions of the actor, even though they contain no information about intentions, and thereby distort attributions of merit or demerit (Smith, 1790). More modern philosophers, however, have continued to debate whether moral luck could instead be due to a coherent moral preference (e.g., Nagel, 1979; Williams, 1981; for a survey see Dana, 2021), and the question remains a current one for legal scholars (e.g., Enoch and Marmor, 2007).

This paper provides new evidence on the existence and prevalence of moral luck, and provides evidence on the question of mechanisms, indicating that it is at least partly a bias. As a first step we show evidence of moral luck in punishment behavior. Second,

we show that random outcomes influence various judgments about the character of the actor, as well as incentivized beliefs about one aspect of the preferences of the actor, despite the random outcomes containing zero information. These biased judgements and beliefs are in turn correlated with punishment behavior. Third, to complete the causal chain, we exogenously vary whether punishers are provided with information about the actor's character, and show that this significantly reduces the influence of outcomes on punishment. This indicates that impact of random outcomes on beliefs about the actor is a mechanism underlying the variation of punishment with outcomes. We check robustness of the bias to an intervention that encourages deliberative rather than spontaneous decision making, but find that an influence of outcomes on beliefs and punishment remains, indicating that the bias is relatively deep seated and hard to remove. Interestingly, we also find that actors tend to internalize moral luck, in terms blaming themselves differently depending on random outcomes, which indicates that the phenomenon can emerge even when actions and outcomes are unobserved.

This study uses an approach that addresses some important methodological challenges to studying moral luck. With naturally occurring observational data on punishments and rewards, e.g., sentencing decisions from court cases, a problem is the difficulty of observing the roles of chance versus actions as perceived by those deciding on punishments and rewards. Without knowing how much information people have about these factors, it is difficult to establish if punishments are varying with outcomes in a way that reflects moral luck, as opposed to just inferring hidden action from outcomes.[1] One solution is to study decisions in controlled environments, where the researcher causes outcomes to vary in ways that are plausibly or explicitly due to chance (see Robbenholt, 2020; Martin and Cushman, 2016; Gurdal et al., 2013; Brownback and Kuhn, 2019), but using such artificial environments raises the difficulty of having real, consequential moral decisions. Having consequential decisions may be important for recruiting key mechanisms in a realistic way, e.g., eliciting strong emotions. Such realism is desirable for assessing how strong the phenomenon is, e.g., in the face of interventions designed to mute emotions and de-bias behavior. This study uses a framework that combines both clean identification of moral luck, with consequential moral choices that are a matter of life and death for a third party (a mouse).

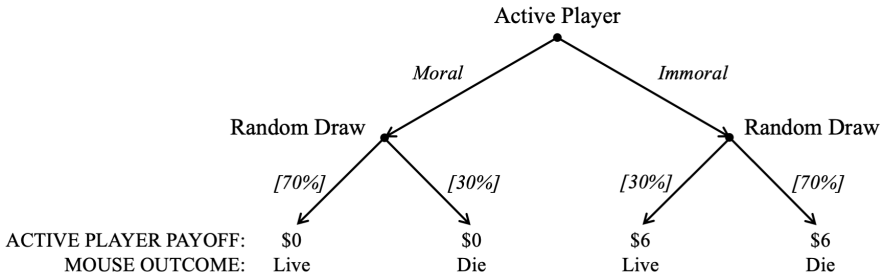
Experiments. In a first stage of our experiments, shown in Figure 1, subjects in the role of active players make a choice between two lotteries, denoted the moral lottery and the immoral lottery, where outcomes are consequential in that they involve life

[1] If actions that influence the probabilities of good or bad outcomes are unobserved, or observed only with noise, then outcomes become informative signals of the intentions and actions of the actor, and thus the informativeness principle entails varying punishments and rewards with the outcomes to some extent. Without knowing exactly what decision makers believe, it is difficult to assess if punishments and rewards vary with outcomes to the optimal degree.

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.

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



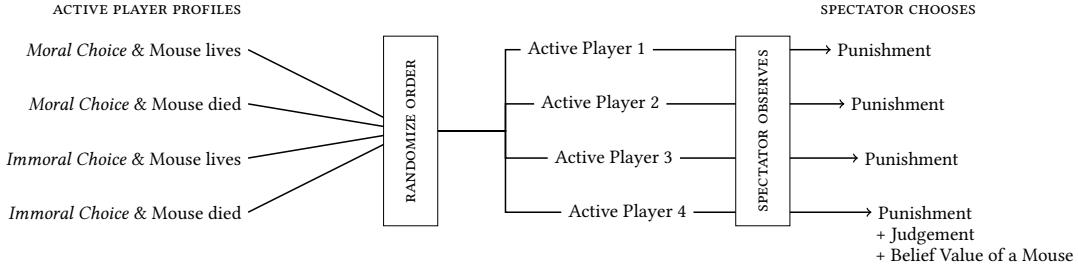
Notes: The active player first chooses one of two options, shown in the figure as *moral* or *immoral*, although more neutral, factual labels "option likely live" and "option likely die" were used with subjects. The moral choice leads to a subsequent random draw with a low probability of death for the 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.

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

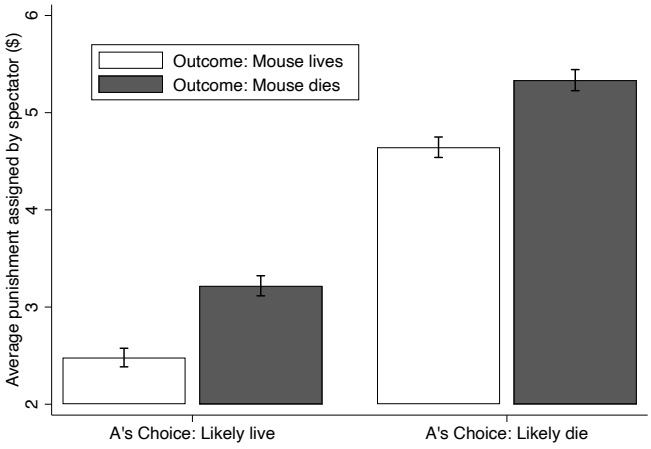
The study also elicited judgements of the spectators about the fourth active player they saw, e.g., in terms of morality of the choice, and whether the active player was a good person. The elicitation asks only about the final active player a spectator saw, to reduce complexity of asking about all previous 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.

Results. 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. 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. 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 in a between-subject comparison, using only first choices of spectators. 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.

Figures 4 and 5 explore one possible explanation, 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 about the active player. Figure 4 shows that spectators are more likely to judge the

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

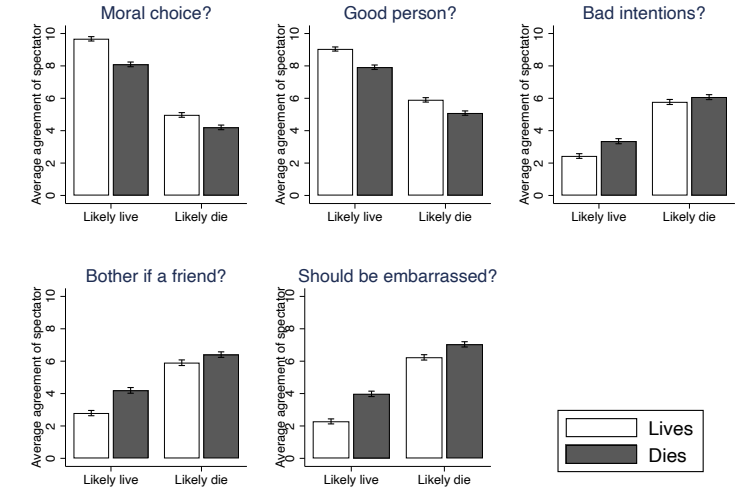


Notes: Average punishment levels for each of the four cases. "A's choice" refers to Active Player's choice. Each spectator chooses punishment for all four cases (within-subject comparison). Figure shows standard error bars clustering on spectator.

active player's choice as immoral, the active player as less of a good person, and the active player as having bad intentions, if the mouse died, whether the active player choose the good lottery or the bad lottery. If the mouse dies they also agree more that the active player should be embarrassed, and that it would bother them if the active player were their friend. Isolating more precisely an impact of random outcomes on a belief about the nature (preferences) of the active player, Figure 5 shows that beliefs about the active player's value of a mouse are also influenced by the random outcome, even though these beliefs were incentivized. The effect on beliefs is smaller and not statistically significant for the choice likely die, which can be understood as reflecting a ceiling effect, as the immoral choice is consistent with the active player having only a relatively narrow range of values.

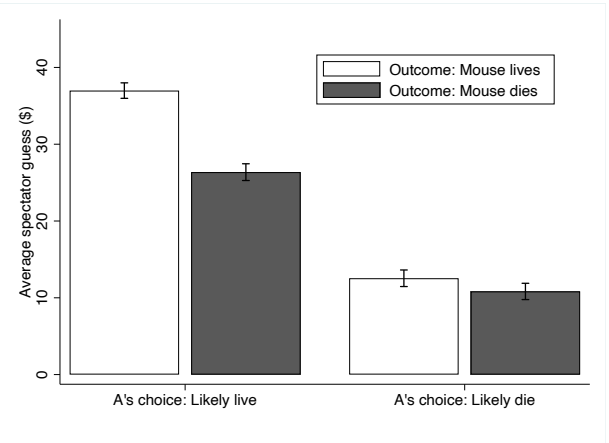
As shown in Table 1, punishment behavior is significantly correlated with judgements and beliefs. Columns (1) and (2) show that spectators who agree more, e.g., that the active player had bad intentions, punish significantly stronger. The judgement about being a good person is an exception, as it is not significant, but this reflects a very high correlation with perceived morality of the choice (Spearman correlation = 0.82; $p<0.001$); jointly these two judgements are highly significant. Columns (2) and (4) Table 1 show that 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 the value. These findings are consistent with a bias in perceptions of the active player, due to the random outcome, driving the different punishment levels, but the evidence is correlational.

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 (between-subject comparison). Figure shows standard error bars.

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



Notes: Average incentivized guess about the active player's value of the life of a mouse. Each spectator makes a guess for one case (between-subject comparison). Figure shows standard error bars.

To provide evidence on whether the outcome influences punishment by influencing perceptions of the active player, 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. For a given choice, they were matched with two active players who had the same value of a mouse, but who had different outcomes for the mouse.[2] If part of the reason why punishment

[2]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

Table 1: Relationships of punishment choices to judgements and beliefs

	(1)	(2)	(3)	(4)
Moral choice	-0.64***	-0.60***		
	(0.20)	(0.22)		
Good person	-0.11	-0.12		
	(0.18)	(0.19)		
Bad intentions	0.45***	0.45***		
	(0.16)	(0.16)		
Bother if a friend	0.43***	0.45***		
	(0.15)	(0.15)		
Embarrassing	0.34**	0.36**		
	(0.15)	(0.15)		
Belief about active player			-0.77***	-0.36***
			(0.10)	(0.12)
Constant	3.72***	3.61***	4.02***	2.43***
	(0.10)	(0.69)	(0.11)	(0.71)
Controls	No	Yes	No	Yes
Observations	1439	1439	1446	1446
Adjusted R^2	0.162	0.173	0.037	0.084

Notes: OLS regressions. Dependent variable is punishment of the fourth active player seen by the spectator. Independent variables include self-reported judgements 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. Columns (2) and (4) include controls: Dummy variables for choice of the fourth active player and outcome for the mouse, with moral_live as the omitted category; the spectator's own value of a mouse; gender; age; income range; educational attainment. Robust standard errors in parentheses.

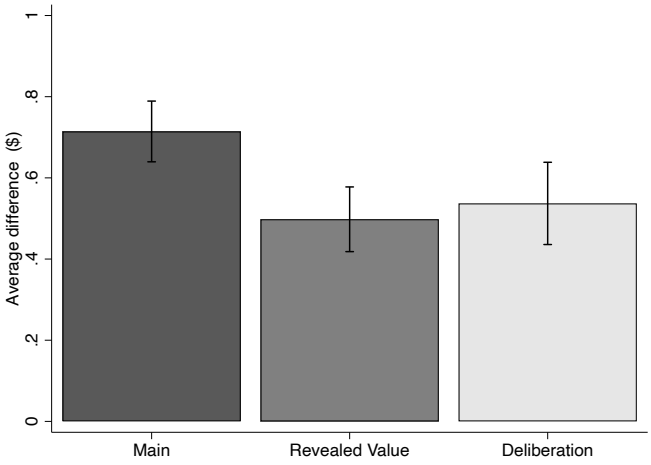
varies with outcomes is the bias in beliefs about the active player value of a mouse, we would expect moral luck to be weaker in Treatment Revealed Value, since the value of mouse is known to the spectator and does not vary with the outcome conditional on choice.

As shown in Figure 6, we find that punishment does, indeed, vary significantly less with the random outcome in Treatment Revealed Value, compared to Treatment Main. Moral luck is still present and significant in Treatment Revealed Value, however, which could reflect the fact that the treatment only shuts down one type of inference about the active player, among many that appear 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.

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

for the matching.

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



Notes: Average difference in punishment of die minus punishment of live, using each spectator's choices for all four cases (within-subject comparison). Figure shows standard error bars clustering on spectator.

et al., 2012; Gino et al., 2008). If so, this 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. Moral luck in punishment is still highly significant within Treatment Deliberation, and we also find that moral judgements and incentivized beliefs are all still influenced by random outcomes in Treatment Deliberation, in ways that are not significantly different from in Treatment Main. This indicates that the mechanism involving bias in judgements and beliefs about the active player is not eliminated by simply taking more time to deliberate and thus appears relatively robust and deeply rooted.

Additional analysis. In additional analysis to be reported in an online appendix (TBA) we explore some alternative explanations for why punishment might vary with outcomes, besides the bias that we identify, but find little support for these. One explanation could be that some individuals disagree with the control principle, and have in mind some alternative, consequentialist moral principle. In a survey question about the control principle, however, the median individual agrees strongly with the principle, and we find strong moral luck even among the sub-sample who indicate complete agreement. Another explanation for why punishment varies with outcomes could be an imperfect understanding of the role of chance in our study, due to limited cognitive ability, or inattention to information provided, especially if this inattention is skewed

towards noticing outcome information more than choice information. Working against such explanations, however, is the fact that spectators were required to correctly answer comprehension questions before making their choices. We also find that exhibiting moral luck is unrelated to measures of cognitive ability (cognitive reflection test, and Raven's progressive matrixes), or to educational attainment. Higher cognitive ability does predict a lower propensity to exhibit anti moral luck, suggesting that such behavior may be a cognitive mistake. We also find that spectators were attentive to the information provided. In an incentivized question at the very end of the study asked spectators to recall the choice and outcome for the final active player they saw, and accuracy rates are quite high, about 85 percent, and essentially identical for choices and outcomes. Thus, inattention to information does not appear to explain moral luck.

The bias we identify raises questions about what might be the deeper mechanisms underlying 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 no significant relationship between belief in a just world and moral luck. 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. Lastly, we consider whether the bias might be stronger for individuals who have stronger emotional reactions, suggesting a mechanism based on emotion. We elicited a survey measure of emotions about the active player, and find that random outcomes significantly influence emotions about the active player. We also find that moral luck in punishment is stronger for individuals who have stronger emotional reactions. As a proxy for caring more about the outcomes, we use the spectator's own value of a mouse, and find that such spectators have stronger emotional reactions, stronger biases in beliefs, and stronger moral luck in punishment. One implication of such a mechanism is that moral choices involv-

ing bad outcomes that are more emotionally upsetting may be more likely to generate moral luck, and heterogeneity in moral luck may be partly explained by heterogeneity in how emotionally upsetting spectators find a given bad outcome.

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 43 percent exhibit moral luck, 36 percent zero moral luck, and 21 percent anti-moral luck. Thus, moral luck is prevalent but not universal. As noted above, anti-moral luck is less likely when individuals have higher cognitive ability, and it is also smaller in magnitude than moral luck, suggesting that this pattern reflects noise. 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.

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 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. 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 assess-

ment before their choice; the reduction in self-esteem if the mouse dies is statistically significant for individuals who have above median baseline self-image and therefore do not have a floor effect working against a reduction. 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. 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. 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.

Discussion.

Our findings have important implications for theories of human punishment behavior. Models of reciprocity theorize that individuals will engage in costly punishment of actions that cause harm. This can reflect a strategy of deterrence in repeated interactions, or it can arise as a heuristic or a preference for punishing those who would create harm by their actions, and manifest even in one-shot interactions (e.g., Trivers, 1971; Axelrod and Hamilton, 1981; Rabin, 1993; Levine, 1998; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). A complicating factor in reality, however, is that both actions and chance often play a role in determining whether harm is caused. Our findings show that punishment behavior is influenced partly by actions, consistent with reciprocity theories, but also partly by random outcomes, something that cannot be explained by traditional models of reciprocity. Furthermore, we show that a key reason that random outcomes influence punishments is by biasing judgements and beliefs about the intentions of the actor. This implies lasting reputational effects of random outcomes, which could in turn lead to longer-run consequences for punishments that are also not explained by reciprocity models, e.g., in the form of future avoidance or ostracism of actors who are viewed as bad types due to previous bad luck with outcomes. Another novel prediction from our findings is that punishments may be particularly sensitive to outcomes when there is more limited information about intentions or

characters of actors. Strangers, therefore, might be more subject to moral luck in how they are evaluated, compared to individuals' whose characters and reputations are well-known to evaluators (good or bad). Our findings call for modifying traditional models of reciprocity to allow for a bias in which evaluators wrongly infer about intentions and character from random outcomes.

The existence of moral luck in punishment also offers a new angle from which to theorize about how and why human punishment behaviors may have evolved. Evolutionary theories posit that human punishment behaviors may have played a crucial role in allowing the human species to sustain large-scale cooperation, by deterring actions that lead to harmful outcomes (e.g., Trivers, 1971; Axelrod and Hamilton, 1981; Henrich and Boyd, 2001; Fehr and Gaechter, 2002). Such theories, however, abstract away from the role of both actions and chance in determining harmful outcomes. One explanation for our finding that humans have a deep rooted tendency to condition punishment on outcomes, could be that it evolved to solve a problem of deterrence that arises in such cases, if actions are hard to observe (Gurdal et al., 2013). When actions influence the probability of harm, but are unobserved, outcomes are signals of actions, and optimal incentives involve conditioning punishment on the occurrence of harm. Our findings suggest that punishments are likely to be overly sensitive to outcomes, since they respond to outcomes even when actions are perfectly observable. But as a fast and frugal heuristic (Gigerenzer, 2004), moral luck could have been adaptive, if conditions with hard to observe actions were sufficiently frequent. One factor that may have also minimized the scope for moral luck to cause distortions in early societies, in cases when actions were observed, is the high frequency of repeated interactions. The result that having more information about actors reduces moral luck suggest that in early societies, with dense social networks and well-established reputations, distortions due to moral luck could potentially have been small, whereas in modern societies, where social networks are less dense, and there is less information about others' characters, moral luck has more scope to distort behavior.

The results in this paper complement previous empirical research on moral luck, and related concept of "outcome bias." The most common methodology has been hypothetical vignettes that vary whether an action is described as leading to more or less severe harm, and elicit moral judgements about the actor and views on appropriate punishment. Previous results are mixed, potentially due to issues of subjectivity in how subjects interpret scenario descriptions, especially interpretations of what different outcomes may signal about probabilities of harm, given that probabilities are typically implicit (for a survey and metaanlysis see, e.g., Robbenholt, 2000). Hypothetical measures also potentially encourage intuitive decision making and inattention, and likely

attenuate emotional reactions, which may explain why asking subjects to decide rationally and deliberately has been found to significantly reduce moral luck in hypothetical vignettes (Gino et al., 2008), whereas in our setting with real outcomes and incentives we find persistent moral luck even with a relatively heavy-handed intervention. Previous research has found that incentivized beliefs about an actor can be influenced by random outcomes (Brownback and Kuhn, 2019), like we do, but we complement this finding by providing the first causal evidence that belief distortions due to random outcomes can cause moral luck in punishment behavior.

The findings of our study also add to the debate on whether moral luck is explained by a preference or moral principle, or is instead a bias. Theories of preferences over outcomes posit that individuals can care about outcomes per se, e.g., disliking inequality (Fehr and Schmidt, 1999). If the harm that results from an action leads to more inequality between the actor and another individual, punishment could be motivated by a desire to reduce inequality between these individuals. In our setting, however, it is unclear that inequality aversion applies, since when harm is caused the mouse is dead. Furthermore, we show that moral luck in punishment is driven by an impact of outcomes on perceptions of the actor, so the mechanism appears to be a form of reciprocity with biased beliefs, rather than inequality-averse preferences. It is also hard to explain our findings with adherence to a moral principle, since moral luck in punishment is driven by judgements and beliefs responding to outcomes that contain zero information. Instead, moral luck appears to be a bias. This does not mean, of course, that philosophical inquiry cannot make progress on seeking a moral principle that can justify conditioning punishment on random outcomes. Our findings do not answer the question of how people should make punishment decisions from a normative point of view, rather they shed light on the positive question of how people are making such decisions.

The fact that moral luck appears to reflect a bias, and has distortionary effects on deterrence, also suggests a potential value of interventions to reduce the bias. Our results suggest that effective interventions may include providing additional information about an actor, or encouraging deliberation, although this will probably not eliminate, the bias. One challenge for such interventions, however, is that evaluators may themselves be evaluated by the wider public, if the incidents they evaluate are in the public view. To change behavior might therefore require an intervention to influence the public, not just the evaluator, which may be challenging. A recent example of such a situation could be the very public disqualification of the tennis player, Novak Djokovich from the 2020 U.S. Open. Djokovich hit a tennis ball in frustration towards the back of the court, and hurt a linesperson by hitting her in the throat. Video evidence shows

that he was not looking where he was hitting, and if the path of the ball had been slightly different, no harm would have occurred. The rules of the tennis association call for disqualification for sufficiently severe recklessness, but leave it to officials to judge severity. In subsequent interviews, tennis officials agreed that Djokovich was not trying to hurt someone, and that if harm had not been caused, their decision would likely have been different. Since harm was caused, he was disqualified, and lost the $250,000 that he had earned for reaching the fourth round of the tournament. Tennis officials might have personally felt that the occurrence of harm was relevant for the decision, or they might have had doubts, but decided that the public would not be satisfied by anything less than disqualification.

In some cases, moral luck is seemingly codified in laws or rules within organizations, requiring evaluators to exhibit moral luck, raising the question whether there is a need for policy reform. An example is differences in sentencing guidelines, or rankings of severity of the crime, for attempted murder versus "successful" murder. Because this rule applies regardless of how hard the individual tried to commit murder, it seems that the key difference is whether, due to circumstances beyond the criminal's control, the murder attempt failed, and thus it exhibits moral luck. To the extent that legal judgements need to concur with notions of justice held by the general populace, and murder is an outcome with a particularly profound emotional impact, reforming legal codes to have the same punishment for attempted and successful murder may be difficult. Another argument against reform could be that it is too costly to determine the role of chance, and simpler to just adjust punishment based on whether outcomes occur, as these can be signals of good or bad intenti. This seems contrary, however, to the notion of deliberation in legal judgements. Furthermore, in other areas of law, which involve civil rather than criminal offenses, the law is clear that severity of outcomes is not relevant for setting punishment. These seemingly contradictory ways of handling the role of chance in outcomes may reflect differences in severity of outcomes, and thus emotional reactions, and a tension between what seems rationally correct, and what feels correct.

References

Axelrod, Robert and William Donald Hamilton, "The evolution of cooperation," *science*, 1981, *211* (4489), 1390–1396.

Bolton, Patrick, Mathias Dewatripont et al., *Contract theory*, MIT press, 2005.

Brownback, Andy and Michael A Kuhn, "Understanding outcome bias," *Games and Economic Behavior*, 2019, *117*, 342–360.

Dufwenberg, Martin and Georg Kirchsteiger, "A theory of sequential reciprocity," *Games and economic behavior*, 2004, *47* (2), 268–298.

Enoch, David and Andrei Marmor, "The case against moral luck," *Law and Philosophy*, 2007, *26* (4), 405–436.

Falk, Armin and Nora Szech, "Morals and markets," *science*, 2013, *340* (6133), 707–711.

_ **and Urs Fischbacher**, "A theory of reciprocity," *Games and economic behavior*, 2006, *54* (2), 293–315.

Fehr, Ernst and Klaus M Schmidt, "A theory of fairness, competition, and cooperation," *The quarterly journal of economics*, 1999, *114* (3), 817–868.

_ **and Simon Gächter**, "Altruistic punishment in humans," *Nature*, 2002, *415* (6868), 137–140.

Gigerenzer, Gerd, "Fast and frugal heuristics: The tools of bounded rationality," *Blackwell handbook of judgment and decision making*, 2004, *62*, 88.

Gino, Francesca, Don A Moore, Max H Bazerman et al., *No harm, no foul: The outcome bias in ethical judgments*, Harvard Business School, 2008.

Gurdal, Mehmet Y, Joshua B Miller, and Aldo Rustichini, "Why blame?," *Journal of Political Economy*, 2013, *121* (6), 1205–1247.

Henrich, Joseph and Robert Boyd, "Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas," *Journal of theoretical biology*, 2001, *208* (1), 79–89.

Holmström, Bengt, "Moral hazard and observability," *The Bell journal of economics*, 1979, pp. 74–91.

Levine, David K, "Modeling altruism and spitefulness in experiments," *Review of economic dynamics*, 1998, *1* (3), 593–622.

Martin, Justin W and Fiery Cushman, "The adaptive logic of moral luck," *The Blackwell companion to experimental philosophy*, 2016, pp. 190–202.

Nelkin, Dana K., "Moral Luck," in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, Summer 2021 ed., Metaphysics Research Lab, Stanford University, 2021.

Rabin, Matthew, "Incorporating fairness into game theory and economics," *The American economic review*, 1993, pp. 1281–1302.

Rand, David G, Joshua D Greene, and Martin A Nowak, "Spontaneous giving and calculated greed," *Nature*, 2012, *489* (7416), 427–430.

Robbennolt, Jennifer K, "Outcome severity and judgments of ÒresponsibilityÓ: A meta-analytic review 1," *Journal of applied social psychology*, 2000, *30* (12), 2575–2609.

Roese, Neal J and Kathleen D Vohs, "Hindsight bias," *Perspectives on psychological science*, 2012, *7* (5), 411–426.

Rubin, Zick and Letitia Anne Peplau, "Who believes in a just world?," *Journal of social issues*, 1975, *31* (3), 65–89.

Sandel, Michael J, *The tyranny of merit: What's become of the common good?*, Penguin UK, 2020.

Smith, Adam, *The theory of moral sentiments*, Penguin, 1790/2010.

Thomas, Nagel, "Mortal questions," *Cambridge: CIP*, 1979.

Trivers, Robert L, "The evolution of reciprocal altruism," *The Quarterly review of biology*, 1971, *46* (1), 35–57.

Williams, Bernard, *Moral luck: philosophical papers 1973-1980*, Cambridge University Press, 1981.