Ashoka University Economics
Discussion Paper 53

Moral Hypocrisy in Social Preferences

February 2021

Abhinash Borah, Ashoka University

https://ashoka.edu.in/economics-discussionpapers

Plot No. 2, Rajiv Gandhi Education City,
National Capital Region,
P.O.Rai, Sonepat, Haryana 131029

Moral Hypocrisy in Social Preferences

Abhinash Borah[*]

February 17, 2021

Abstract

We propose and axiomatize a decision model of social preferences under risk that highlights moral hypocrisy, which we think of as the motivation to appear moral while avoiding the cost of acting morally to the extent possible (Batson et al., 1997). Our model considers a setup with a decision maker (DM) and one other individual. It highlights how the presence of risk enables the DM to exploit the distinction between the other individual's ex post outcome and his ex ante opportunity in a self-serving manner and perceive herself as more moral than what her choices warrant. In turn, this allows her to behave more selfishly in the presence of risk than under certainty and, further, be more risk loving over the other individual's risks than her own. Our axiomatization highlights that the DM acts like a motivated Bayesian when assessing risk faced by the other individual, specifically, she underweights the probabilities of unfavorable outcomes that the other individual may receive in her assessments. We show that our model can explain a wide array of experimental evidence on generous behavior under risk.

JEL Classification: D01, D81, D91
Keywords: social preferences under risk, moral hypocrisy, motivated Bayesian reasoning, ex post outcomes and ex ante opportunities

1 Motivation

In an influential set of experiments involving the two player dictator game, Dana, Weber, and Kuang (2007) [DWK, henceforth] asked lab dictators whether they prefer alternative A that gives the dictator (the decision maker) \$6 and the other player (the recipient) \$1, denote this allocation as $(6, 1)$, or alternative B that gives both \$5 each, denote this allocation as $(5, 5)$. In keeping with the literature, they found that a significant portion (74%) chose B. Then, in a separate treatment, they introduced risk into the environment.

[*] Department of Economics, Ashoka University, Sonepat, Haryana - 131029, India. Email: abhinash.borah@gmail.com. I have greatly benefited from several conversations on the subject of social preferences with Andy Postlewaite, Alvaro Sandroni, Yoram Halevy and Raghvi Garg. Reem Qamar and Nitesh Kumar Singh provided excellent research assistance. Any shortcomings in the paper are, of course, my responsibility.

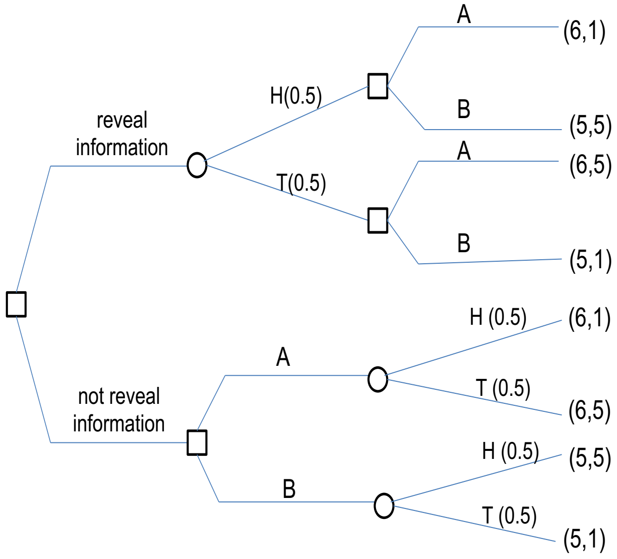

Figure 1: Choices in the DWK experiment. □ denotes a decision node and ○ a chance node. Suppose that the recipient's payoffs are determined by a fair coin-toss in which H(eads) and T(ails) realize with probability 0.5 each. If H realizes, his payoffs are as in the first treatment, i.e., he receives 1 from A and 5 from B. If T realizes, his payoffs are flipped, i.e., he receives 5 from A and 1 from B.

Specifically, whereas the decision maker (DM) still received \$6 from alternative A and \$5 from alternative B, the recipient's payoffs from A and B could, with even chances, be either \$1 and \$5, respectively (as in the earlier treatment), or flipped to be \$5 and \$1, respectively. The key feature of this treatment was that the DM was provided with the option of *privately* and *costlessly* revealing the information about the recipient's true payoffs before making the choice between A and B. That is, it was up to her whether she wanted to make the choice with or without this information. The decision tree faced by a DM in this treatment is illustrated in Figure 1.

Now, consider the following two strategies in the second treatment with the goal of delving into the question of what drives generous behavior for these experimental subjects. The first, call it s_1, involves choosing to reveal information about the recipient's payoffs, followed by the choice of B if the payoffs are as in the baseline treatment [event H], and A if they happen to be flipped [event T].[1] It is straightforward to deduce from Figure 1 that this strategy results in the lottery $[(5,5),0.5;(6,5),0.5]$.[2] The second strategy, call it s_2, involves not revealing information and choosing A. This strategy results in the lottery $[(6,1),0.5;(6,5),0.5]$. As DWK point out, *if* fairness concerns is what drives the DMs that choose alternative B over A in the first treatment, then there should not be a significant portion of them who choose the strategy s_2 over s_1 in the second treatment, i.e., we should not see consistent preference reversals with DMs expressing a preference for $(5,5)$ over $(6,1)$ under certainty but $[(6,1),0.5;(6,5),0.5]$ over $[(5,5),0.5;(6,5),0.5]$ under

[1] Note that if after choosing to reveal information, the DM finds that the payoffs happen to be flipped, then the choice of A and B result in the allocations $(6,5)$ and $(5,1)$, respectively, and any reasonable DM would, arguably, choose A.

[2] Following standard notation, $[(5,5),0.5;(6,5),0.5]$ denotes a lottery in which the outcomes $(5,5)$ and $(6,5)$ realize with probability 0.5 each.

risk. In other words, given that information about the recipient's payoffs can be costlessly acquired, such a fairness motivation on the part of these DMs is inconsistent with them avoiding this information and using the ignorance as an excuse to choose A in the second treatment.

In this regard, the interesting finding that DWK's experimental results throw up is that whereas the proportion that chose A in the first treatment was 26%, in the second, the proportion that chose not to reveal the payoff information and go with this option (strategy s_2) went up to around 40%. Correspondingly, the proportion that opted for strategy s_1 was about 47% and significantly lower than the 74% who chose B in the first treatment. What these experimental results, therefore, suggest is that there may exist a significant proportion of DMs for whom evidence of generous behavior may not reflect a *deep* preference for fairness or altruism. Rather, it may be driven by more self-serving and egoistical ends. Specifically, this may be the behavior of DMs who wish to maintain a self image of being moral while avoiding the cost of behaving morally to the extent possible—often by exploiting situational excuses and wiggle room, like hiding behind uncertainty in the DWK experiment. The psychology literature refers to such motivation, notwithstanding the harsh connotation of the term, as *moral hypocrisy* [Batson et al. (1997), Batson et al. (1999)].[3]

In this paper, we propose and axiomatize a decision model of social preferences under risk that highlights such moral hypocrisy in behavior. Our decision model considers a set-up with a DM and one other individual. The critical insight that it captures is how the presence of risk makes it possible for the DM to exploit the distinction between the other individual's *ex post outcome* and his *ex ante opportunity* in a self-serving manner and perceive herself as more moral than what is warranted by the consequences of her choices for him. This, in turn, allows her to behave more selfishly in environments with risk (than under certainty) without concomitantly hurting her self-image of being moral. At the same time, this also makes her more risk loving when assessing risk faced by the other individual as compared to identical risk faced by her.

To understand the critical insight underlying our decision model, let's apply it to the observed choices in the DWK experiment. First, consider the choice between the allocations $(5, 5)$ and $(6, 1)$ in the first treatment. In it, if the DM chooses the latter allocation, under

[3] The finding of the DWK experiment reported above has been replicated in other studies as well, specifically, in Larson and Capra (2009), Matthey and Regner (2011), Feiler (2014) and Grossman and Van Der Weele (2017). The broader theme that moral behavior may have a more self-serving basis, including drifting into the territory of moral hypocrisy, and that situational excuses are often employed to justify immoral behavior has been reaffirmed in a large body of experimental findings beyond these papers. Without trying to be exhaustive, some papers emphasizing this theme are Pillutla and Murnighan (1995), Schweitzer and Hsee (2002), Mazar, Amir, and Ariely (2008), Wiltermuth (2011), Shalvi et al. (2011), Lewis et al. (2012), Rodriguez-Lara and Moreno-Garrido (2012), Gino, Ayal, and Ariely (2013), Lin, Zlatev, and Miller (2017), Exley (2018), Garcia, Massoni, and Villeval (2020), Falk, Neuber, and Szech (2020), Gneezy et al. (2020), and Exley (2020).

which the other individual receives only \$1, then presumably her self-image of being moral is undermined and this, in part, influences her to choose the allocation $(5,5)$. On the other hand, in the second treatment, where there is risk in the environment, the distinction between the *ex ante opportunity* available to the other individual and his *ex post outcome* provides the DM with an additional mechanism to maintain her self-image as moral. To see this, consider the lottery $[(6,1),0.5;(6,5),0.5]$ that results from not revealing information and choosing A in the second treatment. Whereas in the case of certainty, the choice of the allocation $(6,1)$ may undermine her self-image as moral, under risk, even if this allocation were to realize from the lottery, she may still be able to maintain it, at least partially. She can do so by reasoning that although the other individual ended up receiving only \$1, her choice did provide him with a better ex ante opportunity than that—e.g., she may reason that his expected earning of \$3 was much higher than what he ended up with ex post. In other words, when the other individual receives an unfavorable (ex post) outcome under a lottery, the presence of (ex ante) risk allows the DM to think of herself as more moral than what his outcome warrants by falling back on the excuse that his overall opportunity was much more favorable than his actual outcome—as if saying to herself, *"Well, I intended better, but fate is to be blamed for his unfavorable outcome!"* This, in short, is the moral hypocrisy in behavior that we capture in our decision model.

To explain things a little more formally, in the way of notation, let X and Y, respectively, denote the set of outcomes of the DM and the other individual, so that $X \times Y$ denotes the set of allocations for this two-member society. Let p be a (simple) lottery on the allocation space $X \times Y$, with p_X and p_Y denoting its marginals over X and Y, respectively. Under our proposed *moral hypocrisy (MH)* representation of preferences, the DM's assessment of an allocation-lottery like p is given by:

$$W(p) = \sum_{x \in X} p_X(x)u(x) + \sum_{y \in Y} p_Y(y) \max \left\{ v(y), \sum_{\tilde{y} \in Y} p_Y(\tilde{y})v(\tilde{y}) \right\}$$

Here, the functions $u : X \to \mathbb{R}$ and $v : Y \to \mathbb{R}$ reflect the DM's assessment of her own and the other individual's outcomes, respectively. First, observe that the DM's assessment of a (sure) allocation like $(x,y) \in X \times Y$ is simply given by $u(x) + v(y)$. Next, consider her assessment of a non-degenerate allocation-lottery, p. To understand this assessment, observe that $\sum_{\tilde{y} \in Y} p_Y(\tilde{y})v(\tilde{y})$ represents an expected utility like evaluation, based on the function v, of the overall risk, p_Y, faced by the other individual under p—think of this as the DM's assessment of the other individual's ex ante opportunity under this lottery. Further, let $\overline{Y}_p = \{y \in Y : v(y) \geq \sum_{\tilde{y} \in Y} p_Y(\tilde{y})v(\tilde{y})\}$ denote the set of outcomes for the other individual that the DM considers to be at least as good as his ex ante opportunity under p. Similarly, let $\underline{Y}_p = \{y \in Y : v(y) < \sum_{\tilde{y} \in Y} p_Y(\tilde{y})v(\tilde{y})\}$ denote the set of outcomes for him that she considers worse than his ex ante opportunity. We may, then, rewrite her assessment of p under an MH representation as:

$$W(p) = \overbrace{\sum_{x\in X} p_X(x)u(x) + \sum_{y\in \overline{Y}_p} p_Y(y)v(y)}^{\textit{consequentialist assessment}} + \overbrace{\sum_{y\in \underline{Y}_p} p_Y(y) \sum_{\tilde{y}\in Y} p_Y(\tilde{y})v(\tilde{y})}^{\textit{counterfactual MH assessment}}$$

When it comes to assessing her own risk, p_X, under p, the DM goes by a standard consequentialist expected utility evaluation, $\sum_{x\in X} p_X(x)u(x)$. The same holds true when it comes to her assessment of the other individual's outcomes that are at least as good as his ex ante opportunity under the lottery, i.e., $y \in \overline{Y}_p$. However, when it comes to outcomes, $y \in \underline{Y}_p$, that are worse than his ex ante opportunity, the DM abandons consequentialist reasoning. Instead, she correctly anticipates that in the event of the other individual receiving such an outcome, ex post, she will have the wiggle room to engage in *counterfactual MH reasoning* of the "I–intended–better–but–fate–is–to–be–blamed" type. Specifically, in her evaluation, instead of attributing her assessment of the actual outcome he receives to such events, she attributes her assessment of his ex ante opportunity under the lottery, $\sum_{\tilde{y}\in Y} p_Y(\tilde{y})v(\tilde{y})$. In other words, with respect to these events, the DM's assessment inflates her moral self image beyond what is warranted by the actual consequences experienced by the other individual under her choice.

Going back to DWK's experiment, as per the MH representation, to rationalize the choice of the allocation $(5, 5)$ over $(6, 1)$ requires that $u(5) + v(5) > u(6) + v(1)$; or, equivalently, $u(6) - u(5) < v(5) - v(1)$. On the other hand, to rationalize the choice of the lottery $p = [(6, 1), 0.5; (6, 5), 0.5]$ over the lottery $q = [(5, 5), 0.5; (6, 5), 0.5]$ requires that:

$$W(p) = u(6) + 0.5v(5) + 0.5[0.5v(5) + 0.5v(1)] > 0.5u(5) + 0.5u(6) + v(5) = W(q)$$

i.e., $u(6) - u(5) > 0.5[v(5) - v(1)]$. In other words, an MH type DM will make these two choices in the two treatments if $0.5[v(5) - v(1)] < u(6) - u(5) < v(5) - v(1)$.

In this paper, we provide a justification for the MH representation by demonstrating that it can be derived from plausible axioms on behavior. The key idea that our axiomatization builds on is that the DM behaves like a motivated Bayesian (Gino, Norton, and Weber, 2016) when it comes to evaluating risk faced by the other individual. By motivated Bayesian, we refer here to a tendency to subjectively process objective probabilities in a systematically biased manner with the goal of maintaining an elevated moral self image. Specifically, we show that the DM's choice behavior can be understood in terms of a cognitive manipulation wherein she underweights the probabilities of unfavorable outcomes that the other individual may receive and, accordingly, overweights the probabilities of favorable ones. Our key axiom, morally motivated Bayesianism, uses a well known method to decompose lotteries introduced in Gul (1991) to precisely identify the exact nature of this underweighting/overweighting. Here, it is worth noting that there is experimental evidence supporting the viewpoint that DMs may subjectively bias objective probabilities to embellish their self image of being moral. For instance, Exley (2016) reports the findings of an experiment in which subjects had to assess risky prospects that impacted their own payoffs and that of a charity (American Red Cross). The experimental results demonstrate that when comparing prospects that involved tradeoffs between their own payoffs and that of the charity, subjects behaved as if they were systematically distorting objective probabilities so as to provide themselves with an excuse for not giving to the charity.[4] Other studies, e.g., Eil and Rao (2011) and Mobius et al. (2011), document that, like in our model, underweighting negative and overweighting positive outcomes is a prominent channel through which individuals maintain an exaggerated self image.

Besides providing a behavioral foundation for the model, we also emphasize its empirical content by showing that it is consistent with a wide array of existing experimental evidence on generous behavior in risky social environments. Specifically, the model helps us connect evidence from different domains. For instance, we show how the phenomenon of information avoidance in the context of moral decisions, like in the DWK experiment, is connected behaviorally to why DMs may be more risk loving over others' prospects than their own and why they may be much more tolerant to ex post inequalities when they can fall back on the excuse that everyone had a chance or opportunity of doing well.

Our central theme that generous behavior may not be driven by a deep concern for fairness or altruism but rather by self-serving, egoistical emotions, has been developed in other decision theoretic papers as well. Dillenberger and Sadowski (2012) build on the framework of Gul and Pesendorfer (2001) and develop a model that formalizes the idea that the underlying motivation behind generous behavior may be the desire to avoid shame. They consider an environment with a DM and one other individual in which the DM's choice of an allocation from a set of allocations is observable to the other individual but the choice of the set itself, made in an earlier stage, is not. Choosing a selfish option when a more prosocial one is available in the choice set may inflict shame on the DM and generate generous behavior. At the same time, when provided with the opportunity to choose the choice set itself at the earlier stage, such a DM, who is assumed to be forward-looking, may choose it in a way that optimally solves the trade-off between her desire to behave selfishly but in a way that avoids shame. Their model formalizes this line of reasoning and behaviorally characterizes it. Evren and Minardi (2017) work with a set-up that is identical to the one in Dillenberger and Sadowski (2012) and behaviorally characterize a decision model in which the DM experiences a warm glow from taking a prosocial action, but only when it is publicly observed. That is, the DM's warm glow comes from the social acclaim she receives from behaving prosocially when her choice is observed by the other individual. The key to experiencing this warm glow rests in the ability to choose a privately costly prosocial alternative when a more selfish alternative is available in the menu. Because of this, an important behavioral distinction emerges between the two models—whereas in Dillenberger and Sadowski (2012), anticipating the experience of shame, the DM has

[4] These results were replicated when the role of the charity was taken by another individual.

a preference for smaller menus in the first stage, in Evren and Minardi (2017), the need to have selfish alternatives in the menu to enjoy social acclaim results in a preference for larger menus. Working within the same framework, Saito (2015) behaviorally characterizes the phenomena of impure altruism, defined as the tendency to make prosocial choices in order to feel pride in acting altruistically and to avoid the shame of acting selfishly; and impure selfishness, defined as the temptation to act selfishly faced by a DM who otherwise wants to act altruistically. Saito's representation allows us to better understand how these forces of pride, shame and temptation to act selfishly may interact in conflicting ways and how a DM's choices in social settings is shaped by this interaction.

Our work also relates to the literature that has looked into the foundations of social preferences under risk. The first generation of social preference models [e.g., Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Charness and Rabin (2002)] were proposed for risk-free environments. The literature soon discovered, though, that these models cannot always be readily extended to environments of risk using standard approaches like expected utility or the available non-expected utility theories. The reason for this is that these standard models of decision making under risk are all outcome-based. As such, they fail to capture concerns for opportunities that DMs with social preferences very often exhibit in environments featuring risk.[5] Hence, the quest in the literature has been to develop appropriate models of opportunity-sensitive social preferences under risk. In this regard, the dominant approach in the literature has been that of procedural fairness. Specifically, in environments featuring risk, the case has been made that concerns for fairness translate not just to a concern for equality of ex post outcomes but also for equality of ex ante opportunities, i.e., procedural fairness. A particularly compelling way of implementing this viewpoint has been proposed by Fudenberg and Levine (2012) and Saito (2013). Their proposal, formalized in the expected inequality aversion (EIA) model of Saito (2013), involves using the expected outcome that different individuals receive under an allocation-lottery as a proxy for the ex ante opportunities available to them under it and using the Fehr-Schmidt functional form to assess not just the distribution of outcomes but also that of opportunities. The assessment of an allocation-lottery under this model is determined by taking a weighted average of these two separate Fehr-Schmidt assessments of outcomes and opportunities. As should be evident, a key difference between our paper and this work is the very different perspective we take on the role that opportunity concerns play in the context of social preferences. Specifically, the difference in emphasis between the MH and EIA models regarding what drives generous behavior—a deep concern for equality or a more self serving pursuit to appear moral—translates into two very different theories of why opportunity concerns may matter to DMs in risky social environments. We will show that this difference can indeed be empirically validated.

[5] Formally speaking, concern for opportunities often results in the preferences of such decision makers violating the property of stochastic dominance, which is shared by all the standard models of decision making under risk. For instance, in the DWK experiment, the preference for $(5,5)$ over $(6,1)$ along with that for the lottery $[(6,1),0.5;(6,5),0.5]$ over the lottery $[(5,5),0.5;(6,5),0.5]$ violates stochastic dominance.

The rest of the paper is organized as follows. Section 2 lays out the framework and formally defines an MH representation. Section 3 provides a behavioral foundation for the representation and contains our representation result. Section 4 further highlights the empirical content of the MH model by relating it to a wide array of experimental evidence on generous behavior in risky environments. The proof of the representation result appears in the Appendix.

2 A Decision Model of Moral Hypocrisy

2.1 Preliminaries

We consider a set-up with a decision maker (DM) and one other individual. Associated with each individual is a well-defined set of outcomes. We denote the set of outcomes of the DM by X and that of the other individual by Y. We take these sets to be connected, separable metric spaces.[6] Accordingly, $X \times Y$ denotes the set of allocations for this two-member society. We denote generic elements of X by x, x' etc., that of Y by y, y' etc., and that of $X \times Y$ by (x, y), (x', y') etc. We denote the set of simple probability measures (lotteries, for short) on the sets $X \times Y$, X and Y by Δ, $\Delta(X)$ and $\Delta(Y)$, respectively. These sets are endowed with the topology of weak convergence. We refer to elements of Δ as *allocation-lotteries* and denote generic elements of this set by p, q etc. For any allocation-lottery $p \in \Delta$, we denote the marginal probability measure of p on X and Y by $p_X \in \Delta(X)$ and $p_Y \in \Delta(Y)$, respectively. Since any lottery in $\Delta(X)$ is the marginal probability measure on X of some allocation-lottery in Δ, to economize on notation, we also denote generic elements of $\Delta(X)$ by p_X, q_X etc. Analogously, we denote generic elements of $\Delta(Y)$ by p_Y, q_Y etc. For any $p \in \Delta$, $p(x, y)$ denotes the probability that p assigns to the outcome $(x, y) \in X \times Y$. Similarly, $p_X(x)$ and $p_Y(y)$ denote the probabilities that p_X and p_Y assign to the outcomes x and y, respectively. For any $p_X \in \Delta(X)$ and $p_Y \in \Delta(Y)$, denote by $p_X \circ p_Y$ the product measure in Δ given by $(p_X \circ p_Y)(x, y) = p_X(x) \times p_Y(y)$. We abuse notation and do not distinguish between an outcome and the degenerate lottery that gives that outcome with unit probability. Thus, $(x, y) \in X \times Y$ is also understood as $(x, y) \in \Delta$, $x \in X$ as $x \in \Delta(X)$, etc. We assume reduction of compound lotteries all along and any convex combination of lotteries, $\sum_{k=1}^{K} \alpha^k p^k$, $p^k \in \Delta, \alpha^k \in [0, 1]$, $k = 1, \ldots, K$, and $\sum_{k=1}^{K} \alpha^k = 1$, denotes an element in Δ that gives the outcome (x, y) with probability $\sum_{k=1}^{K} \alpha^k p^k(x, y)$.

[6] The set-up of course allows for the case where X and Y are the same set. For example, X and Y may be some interval of $\mathbb{R}$ in the case where outcomes are monetary ones.

2.2 Preferences and Representation

The DM has preferences over the set Δ of allocation-lotteries that is specified by a binary relation $\succcurlyeq \subseteq \Delta \times \Delta$. The symmetric and asymmetric components of $\succcurlyeq$ are defined in the usual way and denoted by $\sim$ and $\succ$, respectively. We now formally define a moral hypocrisy (MH) representation of $\succcurlyeq$.

Definition 1. *An MH representation of $\succcurlyeq$ consists of a pair of continuous functions $u : X \to \mathbb{R}$ and $v : Y \to \mathbb{R}$ such that the function $W : \Delta \to \mathbb{R}$, given by*

$$W(p) = \sum_{x \in X} p_X(x)u(x) + \sum_{y \in Y} p_Y(y) \max \left\{ v(y), \sum_{\tilde{y} \in Y} p_Y(\tilde{y})v(\tilde{y}) \right\},$$

represents $\succcurlyeq$. That is, $p \succcurlyeq q$ iff $W(p) \geq W(q)$.

As suggested in the Introduction, when it comes to assessing the risk, p_X, that she faces under an allocation-lottery p, the DM goes by a standard expected utility criterion. That is, her assessment of the own-lottery p_X under p is given by $\sum_{x \in X} p_X(x)u(x)$, where $u : X \to \mathbb{R}$ represents her assessment of her own outcomes. On the other hand, when it comes to assessing the risk, p_Y, faced by the other individual under p, what is "non-standard" about her assessment is that it incorporates counterfactual moral hypocrisy reasoning. Such reasoning comes to the fore in relation to those events in which the DM considers the outcome, y, that the other individual receives to be worse than her assessment of his ex ante opportunity under the lottery, i.e., when $v(y) < \sum_{\tilde{y} \in Y} p_Y(\tilde{y})v(\tilde{y})$, where $v : Y \to \mathbb{R}$ represents her assessment of the other individual's outcomes. With respect to such events, the DM's assessment internalizes the fact that, ex post, she will end up protecting her self-image of being moral on the pretext that the other individual had a better ex ante opportunity under this lottery than the outcome he actually received. The exact way she incorporates this reasoning in her evaluation is by attributing her assessment of his ex ante opportunity under this lottery, $\sum_{\tilde{y} \in Y} p_Y(\tilde{y})v(\tilde{y})$, to such events.

Example: Avoiding Information

As a further illustration of how the MH representation works, we now consider an extension of the DWK experiment conducted by Feiler (2014). Recall that in the hidden information treatment of DWK, the probability that the two alternatives, A and B, result in aligned payoffs for the DM and the other individual, i.e., $(6, 5)$ and $(5, 1)$, respectively, is the same as the probability that they result in non-aligned payoffs, i.e., $(6, 1)$ and $(5, 5)$, respectively. Feiler (2014) considers the case where these probabilities need not be the same. Specifically, five different probability values of the aligned payoffs realizing were considered, $\alpha = 0, 0.2, 0.5, 0.8, 1$, with the non-aligned payoffs realizing with complementary probability. Their experimental results show that participants were less likely to reveal information higher was the probability of the aligned payoffs realizing. In their probit estimates, the probability that a participant would reveal information decreased by 0.11 when the probability of the aligned payoffs increased from 0.5 to 0.8 and by 0.13 when this probability increased from 0.2 to 0.8. Further, as one would expect, participants who in the choice problem between the allocations $(6, 1)$ and $(5, 5)$ [the case of $\alpha = 0$] chose the latter were much more likely to reveal information than the ones who chose the former.

We will now provide a rationalization for this evidence in the context of the MH model. To that end, let the aligned and non-aligned payoffs realize with probabilities α and $1-\alpha$, respectively, where $\alpha > 0$. Then the choice of A without revealing information results in the lottery $p = [(6, 5), \alpha; (6, 1), 1-\alpha]$. On the other hand, the choice to reveal information followed by the choice of A if the payoffs are aligned and that of B if they are not results in the lottery $q = [(6, 5), \alpha; (5, 5), 1-\alpha]$. According to the MH representation, the assessment of these lotteries is given by:

$$\begin{aligned}
W(p) &= u(6) + \alpha v(5) + (1-\alpha)[\alpha v(5) + (1-\alpha)v(1)] \\
&= u(6) + v(5) - (1-\alpha)^2[v(5) - v(1)] \\
W(q) &= \alpha u(6) + (1-\alpha)u(5) + v(5) \\
\Rightarrow W(p) - W(q) &= (1-\alpha)[u(6) - u(5)] - (1-\alpha)^2[v(5) - v(1)] \\
\Rightarrow W(p) \geq W(q) &\Leftrightarrow \alpha \geq 1 - \frac{u(6) - u(5)}{v(5) - v(1)}
\end{aligned}$$

Recall that in the choice between the allocations $(6, 1)$ and $(5, 5)$ an MH type DM will choose the former if $u(6)-u(5) > v(5)-v(1)$, or, $\frac{u(6)-u(5)}{v(5)-v(1)} > 1$. Hence, what our model says is that any DM who, in the treatment without uncertainty, chooses A over B will choose not to reveal information for any value of α. Therefore, if the proportion of DMs who choose to reveal information decreases as α increases, then this change in behavior must come from the ones who choose B over A under certainty. For such a DM, $u(6) - u(5) \leq v(5) - v(1)$ and she will choose to not reveal information and choose A if the probability α of the payoffs being aligned is at least as large as the cutoff $1 - \frac{u(6)-u(5)}{v(5)-v(1)} \in (0, 1)$. Further, amongst these DMs, the ratio $\frac{u(6)-u(5)}{v(5)-v(1)}$ will, presumably, be smaller for a relatively more prosocial DM compared to a less prosocial one. As such, this cutoff probability will be higher for the former compared to the latter. This means that as α increases relatively more prosocial ones from this group will start switching to not revealing information.

Example: Risk loving attitude over others' risks

Consider the choice between the lottery $[\$200, 0.5; \$0, 0.5]$ vs. \$100 for sure; or that between $[\$100, 0.7;\ \$0, 0.3]$ vs. \$70 for sure. Since in each of these comparisons the sure amount is the expected value of the lottery, the logic of risk aversion suggests that most DMs in these comparisons would choose the sure amount if they were choosing for themselves. Is the same true if DMs were making these choices for another individual? Evidence suggests that this may not always be the case. E.g., Mengarelli et al. (2014) gave choices like the ones above to experimental subjects with the outcomes of these choices being received by a matched participant in the experiment. They found that in many of these choices a very significant number of subjects, sometimes more than a majority, chose the lottery instead of its expected value for sure when choosing for others, suggesting risk loving attitudes over others' risks. Such risk loving behavior over others' prospects is something that MH-type DMs may engage in. To see this, consider such a DM whose choices are represented by the pair (u, v), both of which, say, are from the CRRA family, i.e., $u(x) = \frac{x^{1-\lambda}}{1-\lambda}$ and $v(y) = \frac{y^{1-\hat{\lambda}}}{1-\hat{\lambda}}$, $\lambda, \hat{\lambda} \in (0, 1)$. This DM, for instance, will choose the lottery $[200, 0.5; 0, 0.5]$ over the sure amount of 100 when choosing for the other individual if:[7]

$$0.5 \times \frac{200^{1-\hat{\lambda}}}{1-\hat{\lambda}} + 0.5 \left[0.5 \times \frac{200^{1-\hat{\lambda}}}{1-\hat{\lambda}} + 0.5 \times 0 \right] \geq \frac{100^{1-\hat{\lambda}}}{1-\hat{\lambda}} \Leftrightarrow \hat{\lambda} \leq 1 - \frac{\ln(4/3)}{\ln(2)} = 0.585$$

To further illustrate this theme that MH-type DMs may be risk loving when it comes to risk faced by others, consider the following piece of experimental evidence from Cettolin, Riedl, and Tran (2017). Specifically, consider two of the tasks that were given to subjects in this experiment. The first of these was a standard dictator game. In it, out of an endowment of M experimental currency units, the DM had to decide what amount $y \in [0, M]$ to give to the other individual, so that the resulting allocation was $(M - y, y)$. In the second, the DM still had to decide on the amount y to give from M, but whereas her payoff continued to be $M - y$, the other individual's was decided by the lottery $[1.25y, 0.8; 0, 0.2]$. The experimental goal was to compare the value of y that subjects chose in these tasks. Observe that for any value of y, the other individual's prospects in the second task is a mean preserving spread of that under the first. In that sense, the second task involves more risk for the other individual. Another way to see this is to take a standard utility function from, say, the CRRA or CARA family and note that for a given y, the expected utility of the lottery faced by the other individual in the second task is smaller than the utility from the sure amount under the first. Therefore, if the DM were to assess the risk faced by the other individual according to an expected utility criterion, her choice of y would be higher in the first task than in the second. However, this is not what the experimental results showed. On average, experimental subjects chose to give about 15% more in the second

[7] Assume that DM's own earnings across these two options are the same.

task compared to the first. This evidence is however consistent with the MH model for a wide class of specifications of the utility functions. For instance, take u and v from the CRRA family like above. Assuming an interior solution, the amount $\hat{y}$ given in the first task solves:

$$u'(M-\hat{y}) = (M-\hat{y})^{-\lambda} = \hat{y}^{-\hat{\lambda}} = v'(\hat{y})$$

On the other hand, the MH assessment of the lottery in the second task is given by:

$$u(M-y) + 0.8v(1.25y) + 0.2[0.8v(1.25y) + 0.2v(0)] = u(M-y) + 0.96v(1.25y)$$

and the amount $\tilde{y}$ given in this task solves:

$$(M-\tilde{y})^{-\lambda} = 0.96 \times 1.25(1.25\tilde{y})^{-\hat{\lambda}} = 1.2 \times 1.25^{-\hat{\lambda}}\tilde{y}^{-\hat{\lambda}}$$

Observe that for $\hat{\lambda} < \frac{\ln 1.2}{\ln 1.25} = 0.817$, we have $1.2 \times 1.25^{-\hat{\lambda}} > 1$. Accordingly, $(M-\tilde{y})^{-\lambda} > \tilde{y}^{-\hat{\lambda}}$ and, it follows that $\tilde{y} > \hat{y}$. That is, for $\hat{\lambda} < 0.817$, this DM gives more in the second task compared to the first in line with the experimental evidence.

These examples illustrate the fact that MH type DMs can demonstrate more risk loving attitudes over others' risks than their own. We will talk more about this feature of the model in Section 4 when we further discuss its empirical content.

3 Axiomatic Foundations of the MH Model

We now introduce a set of axioms on the DM's preferences that characterizes an MH representation. The first two axioms are standard.

Axiom (Weak Order). $\succcurlyeq$ is complete and transitive.

Axiom (Continuity). For any $p \in \Delta$, the sets $\{q \in \Delta : q \succcurlyeq p\}$ and $\{q \in \Delta : p \succcurlyeq q\}$ are closed (in the topology of weak convergence).

Our next axiom weakens the standard Independence condition of vNM expected utility theory.

Axiom (MH Independence). For any $p, p', q, q' \in \Delta$ with $p_Y = p'_Y$, $q_Y = q'_Y$ and $p' \sim q'$,

$$p \succcurlyeq q \Longleftrightarrow \alpha p + (1-\alpha)p' \succcurlyeq \alpha q + (1-\alpha)q'$$

The reason the standard independence axiom may fail to hold in our set-up is because when we take the probability mixture of two allocation-lotteries to form a compound lottery, the risk faced by the other individual under the compound lottery and those under the component sub-lotteries may not be the same. Accordingly, the scope for the DM to engage in counterfactual MH reasoning may differ across these allocation-lotteries and, in turn, may result in preference reversals and violations of independence. For instance, in the DWK experiment, consider an MH type DM who prefers the allocation $(5,5)$ to $(6,1)$. Given that this is a comparison between sure allocations and no risk is involved, there is no scope for MH reasoning. But in the comparison between the 50:50 mixtures, $\frac{1}{2}(5,5)+\frac{1}{2}(6,5)$ and $\frac{1}{2}(6,1)+\frac{1}{2}(6,5)$, this is not the case. Whereas in the mixture $\frac{1}{2}(5,5)+\frac{1}{2}(6,5)$, the other individual faces no risk and there is still no scope for MH reasoning, in the mixture $\frac{1}{2}(6,1)+\frac{1}{2}(6,5)$, as we have seen, the DM can indeed profitably engage in such reasoning. This is what may produce a preference reversal and violation of independence: $(5,5) \succ (6,1)$ but $\frac{1}{2}(6,1)+\frac{1}{2}(6,5) \succ \frac{1}{2}(5,5)+\frac{1}{2}(6,5)$. However, like in the statement of the axiom, when we consider allocation-lotteries like p and p' (respectively, q and q') under which the other individual faces the same risk, then the risk faced by him under the compound lottery $\alpha p + (1-\alpha)p'$ (respectively, $\alpha q + (1-\alpha)q'$) is also the same. As such, the DM has the same scope for counterfactual MH reasoning under the respective mixtures as under their component sub-lotteries. Hence, in comparisons between such mixtures, the DM, presumably, should not have a reason to violate the logic of independence. Note that if the DM's preferences respect MH Independence, then her assessment of risks over her own outcomes follows the logic of Bayesian expected utility maximization.

We now introduce the key axiom of the paper that delineates the scope of MH reasoning. The axiom identifies the departure from Bayesian rationality involved in the DM's assessment of the risk faced by the other individual, benchmarking this departure by the fact that when it comes to assessing the risk she faces, she very much behaves like a Bayesian. This departure can be thought of as a form of *motivated* Bayesian reasoning with the DM subjectively processing objective probabilities in a way that underweights probabilities of unfavorable outcomes that the other individual may receive. The axiom identifies the exact nature and extent of this distortion. To state it, we need to introduce two definitions.

The first of these definitions provides a way to think about risk faced by the other individual in terms of a similar risk faced by the DM.[8]

Definition 2. $p_X \circ y = [(x_1, y), \alpha_1; \ldots; (x_n, y), \alpha_n] \in \Delta$ *is a risk translation of* $x \circ p_Y = [(x, y_1), \alpha_1; \ldots; (x, y_n), \alpha_n] \in \Delta$ *if for each* $i = 1, \ldots, n$, $(x_i, y) \sim (x, y_i)$.

Observe that under the allocation lottery $x \circ p_Y = [(x, y_1), \alpha_1; \ldots; (x, y_n), \alpha_n]$, the DM doesn't face any risk and all the risk is borne by the other individual. If one were to think of this risk faced by the other individual in terms of an equivalent risk faced by the DM, what would that risk be? The notion of a risk translation provides a natural way of answering this question.

[8] Recall that for any $p_X \in \Delta(X)$ and $p_Y \in \Delta(Y)$, we denote by $p_X \circ p_Y$ the product measure in Δ given by $(p_X \circ p_Y)(x, y) = p_X(x) \times p_Y(y)$.

Our next definition refers to a way of decomposing lotteries that was proposed in Gul (1991). Although this decomposition can be applied to any lottery, for our purpose, we restrict attention to allocation-lotteries in which all the risk is borne by the DM and the other individual doesn't face any risk.

Definition 3. *$(\alpha, q_X \circ y, r_X \circ y) \in [0,1] \times \Delta \times \Delta$ is a Gul decomposition of $p_X \circ y \in \Delta$ if:*

1. *$p_X \circ y = \alpha(q_X \circ y) + (1-\alpha)(r_X \circ y)$*
2. *(x, y) in the support of $q_X \circ y$ implies $p_X \circ y \succsim (x, y)$*
3. *(x, y) in the support of $r_X \circ y$ implies $(x, y) \succsim p_X \circ y$*

A Gul decomposition of the lottery $p_X \circ y$ is arrived at in the following way. First, the support of the lottery is decomposed into two parts, one comprising those outcomes which are less preferred to $p_X \circ y$ ("unfavorable outcomes"), and the other comprising those outcomes which are preferred to $p_X \circ y$ ("favorable outcomes"). Then, the probabilities of all unfavorable outcomes are normalized by dividing them by α, the sum of all unfavorable outcome probabilities, to arrive at the lottery $q_X \circ y$. Similarly, the probabilities of all favorable outcomes are normalized by dividing them by $1-\alpha$, the the sum of all favorable outcome probabilities, to arrive at the lottery $r_X \circ y$. Hence, $p_X \circ y = \alpha(q_X \circ y) + (1-\alpha)(r_X \circ y)$. Observe that as long as any certainty equivalent of $p_X \circ y$ is not in its support, this decomposition is unique. Otherwise, there will be an infinity of such decompositions.[9]

We can now state our key axiom that clarifies the exact nature of motivated Bayesian reasoning involved in the DM's assessments of risk faced by the other individual.

Axiom (Morally Motivated Bayesianism). If $p_X \circ y \in \Delta$ is a risk translation of $x \circ p_Y \in \Delta$ and $(\alpha, q_X \circ y, r_X \circ y)$ is a Gul decomposition of $p_X \circ y$, then

$$x \circ p_Y \sim \alpha^2(q_X \circ y) + (1-\alpha^2)(r_X \circ y)$$

To understand the axiom, first, note that if the DM behaved like a Bayesian expected utility maximizer, she would be indifferent between $x \circ p_Y$ and $p_X \circ y = \alpha(q_X \circ y) + (1-\alpha)(r_X \circ y)$. On the other hand, according to the axiom, the DM that we are modeling is indifferent

[9] In Gul (1991), such a decomposition is referred to as an elation/disappointment decomposition (EDD), with the elation and disappointment outcomes corresponding to what we are referring here as favorable and unfavorable outcomes, respectively. We avoid the EDD terminology as, in our context, the emotions at work when using this concept do not correspond to elation and disappointment. Further, when it comes to notation, note that in a triple specifying an EDD in Gul (1991), the first entry refers to the total probability of elation (favorable) outcomes, the second to the elation (favorable) lottery and the third to the disappointment (unfavorable) lottery. In terms of that notation, the decomposition in Definition 3 would be written as $(1-\alpha, r_X \circ y, q_X \circ y)$. We prefer the slightly different notation here as it is more efficacious when it comes to stating our axiom.

between $x \circ p_Y$ and $\alpha^2(q_X \circ y) + (1-\alpha^2)(r_X \circ y)$. Note that $q_X \circ y$ and $r_X \circ y$ are essentially the decomposed risk translations of the unfavorable and favorable outcomes, respectively, that the other individual receives under the lottery $x \circ p_Y$. Therefore, given that $\alpha^2 < \alpha$ for $\alpha \in (0,1)$, this means that the DM's assessment of $x \circ p_Y$ is akin to underweighting (respectively, overweighting) the likelihood of unfavorable (respectively, favorable) outcomes that the other individual receives under it, compared to how a Bayesian would weight these outcomes. The axiom provides a precise specification of the extent of this underweighting/overweighting. In other words, the axiom captures the exact sense in which the DM can be thought of as subjectively distorting objective probabilities in a self-serving manner when taking account of the consequences of her choices for the other individual.

In an MH representation, the DM's assessment of the risk she faces and that faced by the other individual are separable and, in particular, correlations between the outcomes of the two do not matter. Our next axiom provides the foundation for this. To understand what it says, consider allocation lotteries like $p, p', q, q' \in \Delta$ for which $p_Y = p'_Y$ and $q_Y = q'_Y$, i.e., the other individual faces the same risk under p and p', as well as under q and q'. Further, suppose $p \sim q$ and $p' \sim q'$. This means that the DM considers the preference or "utility" difference between p and p' to be the same as that between q and q'. If correlations do not matter, then the difference between q and q' is the same as that between $q_X \circ q_Y$ and $q'_X \circ q'_Y = q'_X \circ q_Y$. Additionally, if assessments of the risk faced by the two are separable, then this difference should not change if the common lottery, q_Y, faced by the other individual under both is replaced by any other lottery, say, the lottery p_Y. Putting everything together, this means that for a DM whose assessment of the risks that she and the other individual face are separable, the difference between p and p' should be the same as that between $q_X \circ p_Y$ and $q'_X \circ p_Y$. The axiom below formalizes this basic idea.

Axiom (Separability). For all $p, p', q, q' \in \Delta$ with $p_Y = p'_Y$ and $q_Y = q'_Y$,

$$[p \sim q, p' \sim q'] \Longrightarrow \frac{1}{2}p + \frac{1}{2}(q'_X \circ p_Y) \sim \frac{1}{2}p' + \frac{1}{2}(q_X \circ p_Y)$$

The indifference condition, $\frac{1}{2}p + \frac{1}{2}(q'_X \circ p_Y) \sim \frac{1}{2}p' + \frac{1}{2}(q_X \circ p_Y)$, is nothing but a tradeoff condition establishing that the DM considers the difference between p and p' the same as that between $q_X \circ p_Y$ and $q'_X \circ p_Y$. To see why, first note that the other individual faces the same risk under the four allocation lotteries, p, p', $q_X \circ p_Y$ and $q'_X \circ p_Y$. Accordingly, following MH Independence, the DM's comparison of the two lotteries, $\frac{1}{2}p + \frac{1}{2}(q'_X \circ p_Y)$ and $\frac{1}{2}p' + \frac{1}{2}(q_X \circ p_Y)$, can be viewed separably across the two 50:50 events. The indifference between these two lotteries, therefore, reveals that an increase, say, in the DM's utility from replacing p' with p must be exactly compensated by a decrease in her utility from replacing $q_X \circ p_Y$ with $q'_X \circ p_Y$; hence, the conclusion stated above.

Finally, for our representation result, we need a richness condition on the domain of preferences. Essentially, it requires the domain to be rich enough to ensure that the DM's own outcomes have a relatively stronger bearing on her preference assessments than those of the other individual, in the sense that any change in her "utility" achieved by varying the other individual's outcomes can always be achieved by varying her own. From a behavioral perspective, a condition of this kind seems a natural one for the type of DM we are modeling for whom moral concerns are largely driven by self-serving impulses.

Condition (**Richness**)**.** For all $x \in X$, $y, y' \in Y$, if $(x, y) \succ (x, y')$, then there exists $x' \in X$ such that either *(i)* $(x', y') \succ (x, y)$ or *(ii)* $(x, y') \succ (x', y)$.

The preference $(x, y) \succ (x, y')$ reveals that holding her own outcome fixed at x and changing that of the other individual from y' to y makes the DM strictly better off. Now, if x' exists such that $(x', y') \succ (x, y) \succ (x, y')$, then the extent of this improvement is smaller than when her own outcome changes from x to x' with the other individual's held fixed at y'. On the other hand, if x' exists such that $(x, y) \succ (x, y') \succ (x', y)$, then too, the extent of the improvement is smaller than when her own outcome changes from x' to x with the other individual's held fixed at y.

The axioms listed above together constitute a choice-theoretic foundation for the MH representation as the following theorem establishes.

Theorem. *Suppose the Richness condition holds. Then $\succcurlyeq$ has an MH representation if and only if it satisfies Weak Order, Continuity, MH Independence, Morally Motivated Bayesianism and Separability. Further, if (u, v) and (u', v') are both MH representations of $\succcurlyeq$, then there exists constants $\alpha > 0, \beta, \beta'$ such that $u' = \alpha u + \beta$ and $v' = \alpha v + \beta'$.*

Proof. Please refer to Sections A.1 and A.2 in the Appendix. □

The Theorem also establishes that the two utility functions representing the DM's assessment of her own outcomes and those of the other individual are unique up to a common positive affine transformation.

4 Empirical Content of the MH Model: Further Comments

In this Section, we further highlight the empirical content of the MH model and its connection to observed patterns of behavior seen in experiments. As a first step in this exercise and to provide it with greater context, it is instructive to formally distinguish the moral hypocrisy motivation underlying generous behavior as outlined in our model from a fairness motivation, understood as an aversion to inequality. We specifically consider inequality aversion as it is not only the predominant paradigm within which social preferences have been studied in the literature but also because the behavioral motivation underlying it is in sharp contrast to that underlying decision making in our model. In particular, we want to highlight how our theory and the inequality aversion paradigm produce two very different rationales for why opportunity concerns may matter in social preferences; and we show below that this difference can be substantiated based on observed behavior.

As is well known, the leading model in economics that captures the idea of fairness as inequality aversion is due to Fehr and Schmidt (1999). The decision-theoretics underlying the original formulation of the model was primarily geared towards an environment of certainty and in it a DM's assessment of social allocations is allowed to be sensitive to inequality in outcomes. Subsequent research has highlighted that when there is risk in the environment, DMs may care not just about inequality of ex post outcomes but also about inequality of ex ante opportunities, i.e., care about procedural fairness. To accommodate this concern about inequality of both outcomes and opportunities, Fudenberg and Levine (2012) and Saito (2013) have proposed an extension of the Fehr-Schmidt model that, following the latter, we refer to as the *expected inequality aversion (EIA)* model.

The EIA Model

We now formally define the EIA model in the context of our primitive set up. To do so, in this section, we will assume that X and Y are intervals of the real line. Further, in the way of notation, for any probability measures $p \in \Delta$, $p_X \in \Delta(X)$ and $p_Y \in \Delta(Y)$, we shall denote by $\mathbb{E}_p[.], \mathbb{E}_{p_X}[.]$ and $\mathbb{E}_{p_Y}[.]$, respectively, the expectations operator w.r.t. these measures.

Definition 4. *An EIA representation of $\succcurlyeq$ on Δ is a triple $(\beta, \delta, \gamma) \in \mathbb{R}^2_+ \times [0,1]$ such that the function $W^{EIA} : \Delta \to \mathbb{R}$, given by*

$$W^{EIA}(p) = \gamma \mathbb{E}_p[w^{FS}(x,y)] + (1-\gamma) w^{FS}(\mathbb{E}_{p_X}[x], \mathbb{E}_{p_Y}[y])$$

represents $\succcurlyeq$, where $w^{FS} : X \times Y \to \mathbb{R}$ is the Fehr-Schmidt functional form given by

$$w^{FS}(x,y) = x - \beta \max\{x-y, 0\} - \delta \max\{y-x, 0\}, \text{ with } \delta > \beta.$$

Under an EIA assessment of the allocation-lottery p, the term $\mathbb{E}_p[w^{FS}(x,y)]$ incorporates the DM's aversion to inequality of ex post outcomes. To see this, observe that this term is nothing but the expected utility of the lottery, p, evaluated with respect to the Fehr-Schmidt utility function w^{FS}. Under the function w^{FS}, in assessing any allocation (x,y), the term $\beta \max\{x-y,0\}$ captures the DM's disutility from advantageous inequality, whereas the term $\delta \max\{y-x,0\}$ captures her disutility from disadvantageous inequality. The condition $\delta > \beta$ implies that the DM is more sensitive to disadvantageous than advantageous inequality. On the other hand, the term $w^{FS}(\mathbb{E}_{p_X}[x], \mathbb{E}_{p_Y}[y])$ incorporates the DM's aversion to inequality of ex ante opportunities. Observe that $\mathbb{E}_{p_X}[x]$ and $\mathbb{E}_{p_Y}[y]$ specify, respectively, the DM's and the other individual's expected outcomes under the lottery, p. Hence, thinking of these expected outcomes as indicative of the ex ante opportunities available to the two individuals under p and using them as arguments of w^{FS} captures aversion to inequality of opportunities. Finally, γ and $1-\gamma$ serve as weights that the DM puts on the ex post and ex ante concerns, respectively. In the subsequent discussion, we will think of the logic of inequality aversion as it applies to risky social environments in the context of the EIA model.

Moral hypocrisy and inequality aversion are observationally distinct

We have already shown that the MH model can accommodate the evidence of the DWK experiment. That is, a DM whose preferences have an MH representation, (u, v), simultaneously chooses the allocation $(5, 5)$ over $(6, 1)$ as well as the lottery $[(6, 1), 0.5; (6, 5), 0.5]$ over $[(5, 5), 0.5; (6, 5), 0.5]$ if $0.5[v(5) - v(1)] < u(6) - u(5) < v(5) - v(1)$. As a first step towards establishing that the MH model is observationally distinct from the EIA model, we show that these choices are not consistent with the latter model. To see this, note that for an EIA representation, (β, γ, δ), to accommodate the preference $(5, 5) \succ (6, 1)$ requires that $5 > 6 - 5\beta$. That is, it requires that $\beta > 0.2$. On the other hand, for it to accommodate the preference $[(6, 1), 0.5; (6, 5), 0.5] \succ [(5, 5), 0.5; (6, 5), 0.5]$ requires that:

$$\begin{aligned} &\gamma[0.5(6-5\beta)+0.5(6-\beta)]+(1-\gamma)[6-3\beta] \\ &\qquad > \gamma[0.5(5)+0.5(6-\beta)]+(1-\gamma)[5.5-0.5\beta] \\ \Leftrightarrow \quad &\gamma[6-3\beta]+(1-\gamma)[6-3\beta] > \gamma[5.5-0.5\beta]+(1-\gamma)[5.5-0.5\beta] \end{aligned}$$

That is, it requires that $6-3\beta > 5.5-0.5\beta$, or, $\beta < 0.2$. Hence, the pattern of choices seen in the DWK experiment cannot be accommodated by the EIA model, thus confirming, in the context of this formal model, that for many DMs generous behavior may not be driven by a deep preference for fairness but rather by more self-serving emotions.

Procedural fairness?

We next show that choice behavior that is often interpreted as resulting from a concern for procedural fairness or aversion to inequality of ex ante opportunities may also be rationalizable within the MH paradigm, thus suggesting an alternative interpretation of this data. To see this, consider the two player probabilistic dictator (PD) game. In such a game, the dictator (the DM) is endowed with a fixed amount of money. However, unlike the standard dictator game, she is not allowed to divide the money between herself and the other individual. Rather, she is given the option, if she so chooses, to share *chances* of getting the money with him. In particular, she can assign him any probability of getting the entire amount while retaining the amount herself with complementary probability. For example, if the fixed amount is $20 and the DM assigns to the other person a probability $\alpha \in [0, 1]$, then the allocation $(0, 20)$ in which the other person gets the 20 dollars (and the DM gets 0) results with probability α and the allocation $(20, 0)$ in which the DM gets the 20 dollars (and the other person gets 0) results with probability $1 - \alpha$. Experimental evidence (Krawczyk and Le Lec, 2010; Brock, Lange, and Ozbay, 2013) indicates that a significant portion of lab subjects do give the other individual a positive probability of getting the money. The reason they share ex ante opportunities or chances with the other individual, it is often argued, is to compensate for the inequality of ex post outcomes that is inevitable in this setting. As such, positive giving in the PD game is often suggested as a leading example of a concern for procedural fairness amongst decision makers.[10] This may well be true for many DMs and such choices on their part may indeed reflect a deep concern for procedural fairness. But, a degree of caution is warranted as it need not necessarily be true that choices of this nature always reflect concerns for procedural fairness. To highlight this point, we next show that the MH paradigm can also accommodate this evidence.

To that end, consider a DM's problem of deciding what probability $\alpha \in [0, 1]$ she wants to assign to the other individual of getting the 20 dollars. Any choice of α generates an allocation-lottery, $p(\alpha) = [(0, 20),\ \alpha;\ (20, 0),\ 1-\alpha]$. If this DM's preferences have an MH representation, (u, v), then her assessment of any such lottery, $p(\alpha)$, is given by:

$$W(p(\alpha)) = \alpha u(0) + (1 - \alpha)u(20) + \alpha v(20) + (1 - \alpha)[\alpha v(20) + (1 - \alpha)v(0)]$$

Note that $\partial W/\partial\alpha = 2(1 - \alpha)(v(20) - v(0)) - (u(20) - u(0))$. So, for α close to zero, $\partial W/\partial\alpha > 0$ as long as $v(20)-v(0) > 0.5(u(20)-u(0))$. Therefore, if this condition is satisfied, the DM chooses a positive α, i.e., chooses to give the other individual a positive probability of getting the money. This illustrates that the moral hypocrisy paradigm may also be able to account for choice behavior that a priori appears to be motivated by a concern for procedural fairness. Therefore, we need to be careful when we attempt to map back such evidence of generous behavior to underlying motivations.

Generous behavior under certainty and risk: a comparison

Next, we turn to a set of dictator game experiments reported in Brock, Lange, and Ozbay (2013) that further highlights how the distinction between certainty and risk may have an important bearing on the contours of generous behavior. Once again, to understand the

[10] Given that the EIA model is premised on accommodating a concern for procedural fairness, it can rationalize the evidence of sharing chances in the PD game.

scope of our model in elucidating this distinction, it will be instructive to contrast it with the EIA model in the context of these experimental findings. Drawing this contrast should also help further clarify the very differing roles that opportunity concerns play in these two paradigms. In the experiment, lab dictators (DMs) were given several tasks that involved allocating 100 tokens between themselves and their matched recipients. Tokens translated to monetary payments with the exact nature of this translation varying from one task to the other. Here we focus on two of the tasks in this experiment. The first replicated the standard dictator game under certainty. In it, if the DM gave $\theta \geq 0$ tokens to the other individual, then the resulting allocation was $(100-\theta, \theta)$. In the second, the tokens that the DM gave translated to a lottery for the other individual. More precisely, if the DM gave θ tokens to the other individual, then her own payoff in the task was $100-\theta$ like in the first. On the other hand, the recipient's payoff was determined by the lottery $[50, \frac{\theta}{50}; 0, 1-\frac{\theta}{50}]$, where the number of tokens θ that could be given in this task was capped at 50. In both these tasks a significant proportion of decision makers gave a positive number of tokens to the recipient. The question that interests us here is in which task were more tokens given by non-selfish dictators.[11] Brock, Lange, and Ozbay (2013) report that when attention is restricted to these non-selfish dictators, on average, more tokens were given in the first task than in the second, with this difference being statistically significant at the 1% level.

We first show that this evidence can be rationalized by the MH model. To see this, first, note that for any choice of θ in the first task, the DM's assessment in the MH model of the resulting allocation $(100-\theta, \theta)$ is given by: $u(100-\theta) + v(\theta)$. Assume that both u and v are differentiable, increasing and strictly concave. Accordingly, assuming an interior solution and that tokens are divisible, the number of tokens that she optimally gives to the other individual in this task solves:

$$u'(100-\tilde{\theta}) = v'(\tilde{\theta}),$$

On the other hand, in the second task, the DM's assessment of the lottery $q = [(100-\theta, 50), \frac{\theta}{50}; (100-\theta, 0), 1-\frac{\theta}{50}]$ generated by the choice to allocate θ tokens to the other individual is given by:

$$W(q) = u(100-\theta) + \frac{\theta}{50}v(50) + \left(1-\frac{\theta}{50}\right)\left[\frac{\theta}{50}v(50) + \left(1-\frac{\theta}{50}\right)v(0)\right]$$

In this case, the optimal choice of tokens solves:

$$u'(100-\hat{\theta}) = \frac{v(50)-v(0)}{50} \times \frac{50-\hat{\theta}}{25}$$

We can find standard specifications for the utility functions u and v for which, from the

[11] Brock et al. consider a dictator to be non-selfish if she gave her matched recipient a positive number of tokens in at least one of their tasks.

two equations above, we get $\tilde{\theta} > \hat{\theta}$.[12]

In contrast, as Brock, Lange, and Ozbay (2013) point out, the EIA model predicts that the number of tokens given in both these tasks are the same. It is instructive to go through the calculations as to why the EIA model predicts this as it helps to further clarify the difference in the motivations underlying behavior in the two models, particularly, the role that opportunity concerns play in their, respective, decision making processes. Observe that for a DM whose preferences have an EIA representation, her assessment of the allocation $(100-\theta,\theta)$ is given by:

$$w^{FS}(100-\theta,\theta) = 100-\theta-\beta\max\{100-2\theta,0\}-\delta\max\{2\theta-100,0\}$$

It is straightforward to verify that the optimal θ chosen can never be greater than 50, so that $w^{FS}(100-\theta,\theta) = 100-\theta-\beta(100-2\theta)$ and the optimal allocation rule is specified by:

$$\theta^* = \begin{cases} 0, & \beta < 0.5 \\ \in [0,50], & \beta = 0.5 \\ 50, & \beta > 0.5 \end{cases}$$

Now consider this DM's assessment of the lottery $q = [(100-\theta,50),\frac{\theta}{50};(100-\theta,0),1-\frac{\theta}{50}]$ generated by the choice to allocate $\theta \in [0,50]$ tokens to the other individual in the second task:

$$\begin{aligned} W^{EIA}(q) &= \gamma\left[\frac{\theta}{50}w^{FS}(100-\theta,50)+\left(1-\frac{\theta}{50}\right)w^{FS}(100-\theta,0)\right] \\ &\quad +(1-\gamma)w^{FS}(100-\theta,\theta) \\ &= \gamma\left[100-\theta-\frac{\theta}{50}\beta(100-\theta-50)-\left(1-\frac{\theta}{50}\right)\beta(100-\theta)\right] \\ &\quad +(1-\gamma)w^{FS}(100-\theta,\theta) \\ &= \gamma[100-\theta-\beta(100-2\theta)]+(1-\gamma)w^{FS}(100-\theta,\theta) \\ &= \gamma w^{FS}(100-\theta,\theta)+(1-\gamma)w^{FS}(100-\theta,\theta) = w^{FS}(100-\theta,\theta) \end{aligned}$$

Accordingly, this DM allocates the same number of tokens to the other individual in both these tasks. As the above calculations clarify, what drives this conclusion is the fact that in the EIA model, ex post concerns for outcomes and ex ante concerns for opportunities serve as perfect substitutes. This is consistent with the interpretation under this model that the DM has a deep preference for fairness and reducing inequality; and whether this objective

[12] For instance, suppose u and v are from the CARA family, i.e., $u(x) = k - \frac{1}{\lambda}e^{-\lambda x}$ and $v(y) = -\frac{1}{\hat{\lambda}}e^{-\hat{\lambda}y}$, $k > 0$, $\lambda, \hat{\lambda} > 0$. Here, k is a scaling parameter to ensure that $u(z) > v(z)$ for all z. Further, given that for a typical DM, marginal utility from an extra dollar to herself is, presumably, greater than that to the other individual, it is reasonable to assume that $\hat{\lambda} > \lambda$. For values of $\lambda = 0.005$ and $\hat{\lambda} = 0.01$, which are reasonable values for prospects of the size at play in the experiment (Babcock, Choi, and Feinerman, 1993), we can calculate that $\tilde{\theta} = 33$ and $\hat{\theta} = 27$. These values are along the lines of the average number of token given in the two tasks in the experiment.

is achieved in terms of reducing inequality in ex post outcomes or ex ante opportunities plays a symmetric role. On the other hand, in the MH model, the DM uses the ex post and ex ante considerations in an opportunistic and self serving way. Specifically, when it comes to assessing the other individual's prospects, she goes by the ex post consideration whenever he does better by it, and by the ex ante consideration otherwise. That is why the second task affords her the leeway to offer a fewer number of tokens than the first, for even if the other individual ends up receiving zero, the DM can protect her self-image of being moral by reasoning that the ex ante opportunity that her choice afforded him was much better than what he received ex post—after all, he had a chance of getting 50.

Risk Attitudes

A key question in the literature on social preferences is whether DMs are more or less risk averse when it comes to making choices for others as compared to making these choices for themselves. The literature provides evidence for both. Whereas some studies show a risky shift in choices made for others, others show a more cautious shift or no shift at all. In a meta-analysis done with 71 papers (totaling 14,443 observations), Polman and Wu (2019) find that, on average, there is a significant, though small, effect in favor of a risky shift when DMs choose for others. Based on such evidence in the literature, it will be fair to say that both types of DMs exist—ones who are less risk averse and may be even risk loving when it comes to choosing for others and ones who are more risk averse. Therefore, it becomes important to identify these two types of decision makers in terms of deeper motivations underlying their social behavior. In this context, our model, as we have seen above, can contribute by identifying one set of motivating behavioral factors that drive a class of DMs to exhibit a risky shift when choosing for others.

Although it may be obvious, it is worth reiterating that in an MH representation, the DM's assessment of any allocation-lottery, p, can be decomposed in terms of the individual risks, p_X and p_Y, faced by her and the other individual, respectively. As such, correlations between their outcomes don't matter and we can derive risk attitudes for each that are independent of the risk faced by the other. The key to understanding why an MH-type DM may be less risk averse when it comes to others' prospects compared to her own lies in the morally motivated Bayesianism axiom. Recall how that axiom postulates that the DM, in assessing risk faced by the other individual, overweights the probabilities of the favorable outcomes and underweights those of the unfavorable ones. It is precisely this departure from the Bayesian calculus that makes her less risk averse over his risks. To understand this better, consider the following investment problem from Gneezy and Potters (1997), which has been used in experiments to study whether DMs are more or less risk averse when it comes to choosing for others [e.g., Pollmann, Potters, and Trautmann (2014)]. In this problem, a DM has to decide how to invest an endowment of M monetary units between a risky and a safe asset. The amount invested in the risky asset, call it m, has a return of 250% with probability 1/3, and a return of -100% with probability 2/3. The amount invested in the safe asset, $M - m$, has a return of 0%. Consider two variants of this investement problem: one in which the outcome of the investment accrues to the DM and another in which it accrues to another individual. Suppose these problems are faced by an MH type DM characterized by functions (u, v). Assume that the two functions are differentiable, increasing and strictly concave. Further, to keep matters simple, assume that the individual whose outcome is not determined by the realization of this risky investment receives a fixed amount k. Then, in the first case, when the DM invests for herself, the amount invested in the risky asset solves,

$$\max_m \left\{ \frac{1}{3} u(M + 2.5m) + \frac{2}{3} u(M - m) + v(k) \right\},$$

with any interior solution characterized by the first order condition:

$$\frac{u'(M + 2.5m^*)}{u'(M - m^*)} = \frac{4}{5}$$

On the other hand, when the DM takes the investment decision for another individual, the amount invested in the risky asset solves,

$$\begin{aligned} & \max_m \left\{ u(k) + \frac{1}{3} v(M + 2.5m) + \frac{2}{3} \left[\frac{1}{3} v(M + 2.5m) + \frac{2}{3} v(M - m) \right] \right\} \\ = \; & \max_m \left\{ u(k) + \frac{5}{9} v(M + 2.5m) + \frac{4}{9} v(M - m) \right\} \end{aligned}$$

with the solution characterized by the first order condition:

$$\frac{v'(M + 2.5m^{**})}{v'(M - m^{**})} = \frac{8}{25} \left(< \frac{4}{5} = \frac{u'(M + 2.5m^*)}{u'(M - m^*)} \right)$$

It is reasonable to hypothesize that for a typical DM, marginal utility from an extra dollar to herself is greater than that to the other individual at any level of wealth. Even though this may mean that v is more concave than u, it is possible for such a DM under the MH model to choose $m^{**} > m^*$, i.e., she invests more in the risky asset when making this choice for the other individual than when making it for herself.[13] The reason why this is the

[13] To see this concretely, suppose u and v are from the CRRA family, i.e., $u(x) = s + \frac{x^{1-\lambda}}{1-\lambda}$ and $v(y) = \frac{y^{1-\hat{\lambda}}}{1-\hat{\lambda}}$, with $k > 0$ and $0 < \lambda < \hat{\lambda} < 1$. Here, s is a scaling parameter to ensure that $u(z) > v(z)$ for all $z > 0$. Further, $\lambda < \hat{\lambda}$ ensures that DM's marginal utility for her own money is greater than that for the other individual's (for monetary levels greater than one). Accordingly, $\frac{u'(M+2.5m^*)}{u'(M-m^*)} = \left(\frac{M-m^*}{M+2.5m^*}\right)^{\lambda} = \frac{4}{5}$. Since, $\frac{M-m^*}{M+2.5m^*} < 1$, there exists $\bar{\lambda} > \lambda$ s.t. $\left(\frac{M-m^*}{M+2.5m^*}\right)^{\bar{\lambda}} = \frac{8}{25}$. Finally, note that for any $\hat{\lambda} \in (\lambda, \min\{\bar{\lambda}, 1\})$, the function $f : [0, M] \to \mathbb{R}$ given by $f(m) = \left(\frac{M-m}{M+2.5m}\right)^{\hat{\lambda}}$ is continuous and strictly decreasing with $f(m^*) > \frac{8}{25}$ and $f(M) = 0$. Accordingly, there exists $m^{**} > m^*$, such that $f(m^{**}) = \frac{8}{25}$, i.e., $\frac{v'(M+2.5m^{**})}{v'(M-m^{**})} = \left(\frac{M-m^{**}}{M+2.5m^{**}}\right)^{\hat{\lambda}} = \frac{8}{25}$. In other words, for $\lambda < \hat{\lambda} < \min\{\bar{\lambda}, 1\}$, we have $m^{**} > m^*$.

case under the MH model can be seen from the second maximization problem capturing the DM's investment decision for the other individual. Observe that while evaluating the lottery over the other individual's outcomes corresponding to a choice of m, the MH reasoning that the DM engages in makes her overweight the favorable outcome of $M+2.5m$ (subjective weight of $\frac{5}{9}$ instead of $\frac{1}{3}$) and, correspondingly, underweight the unfavorable outcome of $M-m$ (subjective weight of $\frac{4}{9}$ instead of $\frac{2}{3}$). This is nothing but the logic of the morally motivated Bayesianism axiom at work and a natural corollary of it is that the DM behaves in a less risk averse way when it comes to assessing others' risks compared to her own.

How much does ex post inequality matter in the presence of risk?

Another question that has generated interest in the social preferences literature is whether ex post inequalities matter in the presence of risk. That is, if people can justify to themselves that ex ante opportunities for everyone involved are more or less fair, do they still care about ex post inequalities? The question has relevance for public policy issues like support for the welfare state or redistributive politics. Specifically, individuals may be less likely to support such policies if they can fall back on the excuse that everyone got a fair shake. There is a body of experimental evidence that indeed suggests that many DMs whose choices may otherwise be responsive to the goal of reducing ex post inequalities, may care far less about it if they can rationalize that everyone had a chance or opportunity of doing well.

For instance, building on earlier findings from Bohnet et al. (2008), Bolton and Ockenfels (2010) provide experimental evidence suggesting that for decisions with social comparisons, risk taking may not depend on whether the risky option yields unequal payoffs. In their experiment, they considered a set of allocations, $A=\{(7,7),(7,16),(7,0),(9,9),(9,16),(9,0)\}$. For each allocation, $(x,y)\in A$, they gave DMs in their experiment the following two choices: (i) (x,y) vs. $[(16,16),0.5;(0,0),0.5]$ and (ii) (x,y) vs. $[(16,0),0.5;(0,16),0.5]$. If ex post inequality does not matter in the presence of risk, especially when ex ante opportunities are deemed to be equal, then it should be the case that DMs' choices do not vary across (i) and (ii) in terms of whether the safe or risky alternative is chosen. This is indeed what the aggregate choice behavior in their subject pool seem to suggest. In turn, this is indicative of the fact that for many DMs, the assessment of the two allocation-lotteries $[(16,16),0.5;(0,0),0.5]$ and $[(16,0),0.5;(0,16),0.5]$ is roughly identical. Observe that this is true in the MH model, where a DM whose preferences can be represented thus is indifferent between these two lotteries.

A similar insight as the above experiment emerges more comprehensively in the important paper by Cappelen et al. (2013). They too are interested in the question about whether in situations marked by equality of ex ante opportunities, DMs still care about ex post inequalities. They find that a large proportion of DMs in their experiment do not. But, at the same time, there was also a significant proportion who did. This shows the diversity amongst DMs when it comes to this facet of behavior in social situations of risk. In this regard, the particular psychology behind choices that our model develops may be one reason underlying the behavior of DMs who do not care much about ex post inequalities when they can reason that everyone had a fair ex ante chance.

A Appendix

A.1 Preliminaries

We first present a few Lemmas that follow from our axioms and the Richness condition. We use these Lemmas to prove the representation Theorem.

Lemma 1. *For any $p \in \Delta$, $p \sim p_X \circ p_Y$.*

Proof. By Completeness, $p \sim p$ and $p_X \circ p_Y \sim p_X \circ p_Y$. Applying Separability, it follows that $\frac{1}{2}p + \frac{1}{2}(p_X \circ p_Y) \sim \frac{1}{2}(p_X \circ p_Y) + \frac{1}{2}(p_X \circ p_Y)$ Accordingly, MH Independence implies that $p \sim p_X \circ p_Y$. □

Lemma 2. *For any $p_X \circ p_Y \in \Delta$, there exists $\tilde{x} \in X$, such that $p_X \circ p_Y \sim \tilde{x} \circ p_Y$.*

Proof. By MH Independence, it follows that there exists $\overline{x}$ and $\underline{x}$ such that $\overline{x} \circ p_Y \succcurlyeq p_X \circ p_Y \succcurlyeq \underline{x} \circ p_Y$. Let $X_1 = \{x \in X : x \circ p_Y \succcurlyeq p_X \circ p_Y\}$ and $X_2 = \{x \in X : p_X \circ p_Y \succcurlyeq x \circ p_Y\}$. Note that: (i) $X_1 \neq \emptyset$ since $\overline{x} \in X_1$, $X_2 \neq \emptyset$ since $\underline{x} \in X_2$; (ii) X_1 and X_2 are closed in X by virtue of the Continuity axiom; and (iii) $X_1 \cup X_2 = X$. Therefore, since X is connected, it follows that $X_1 \cap X_2 \neq \emptyset$. That is, there exists $\tilde{x} \in X_1 \cap X_2$ and, accordingly, $p_X \circ p_Y \sim \tilde{x} \circ p_Y$. □

Lemma 3. *Any $x \circ p_Y \in \Delta$ has a risk translation $p_X \circ y$.*

Proof. Let $x \circ p_Y = [(x, y_1), \alpha_1; \ldots; (x, y_n), \alpha_n]$ and assume wlog that $(x, y_1) \succcurlyeq (x, y_2) \succcurlyeq \ldots \succcurlyeq (x, y_n)$. If $(x, y_1) \sim (x, y_2) \sim \cdots \sim (x, y_n)$, then the conclusion is immediate as any $x \circ y_i$, $i = 1, \ldots, n$, is a risk translation of $x \circ p_Y$. So, assume $(x, y_1) \succ (x, y_n)$. By the Richness condition, it follows that there exists $x' \in X$ such that either $(x', y_n) \succ (x, y_1)$, or $(x, y_n) \succ (x', y_1)$, i.e., either $(x', y_n) \succ (x, y_1) \succcurlyeq (x, y_2) \succcurlyeq \ldots \succcurlyeq (x, y_n)$, or $(x, y_1) \succcurlyeq (x, y_2) \succcurlyeq \ldots \succcurlyeq (x, y_n) \succ (x', y_1)$. In the first case, by virtue of the Continuity axiom and the fact that X is a connected space, it follows that there exists $x_1, \ldots, x_n$ such that $(x, y_i) \sim (x_i, y_n)$ for all $i = 1, \ldots, n$. Accordingly, $p_X \circ y_n = [(x_1, y_n), \alpha_1; \ldots; (x_n, y_n), \alpha_n]$

is a risk translation of $x \circ p_Y$. A similar argument establishes the conclusion in the second case as well. $\square$

Lemma 4. *For any* $x \circ p_Y \in \Delta$, *there exists* $\overline{y}$, $\underline{y} \in Y$ *such that* $(x, \overline{y}) \succcurlyeq x \circ p_Y \succcurlyeq (x, \underline{y})$.

Proof. Let $x \circ p_Y = [(x, y_1), \alpha_1; \ldots; (x, y_n), \alpha_n]$ and assume wlog that $(x, y_1) \succcurlyeq (x, y_2) \succcurlyeq \ldots \succcurlyeq (x, y_n)$. Further, let $p_X \circ y = [(x_1, y), \alpha_1; \ldots; (x_n, y), \alpha_n]$ be a risk translation of $x \circ p_Y$ and $(\alpha, q_X \circ y, r_X \circ y)$ a Gul decomposition of $p_X \circ y$. MH Independence and Morally Motivated Bayesianism together imply that:

$$(x_1, y) \succcurlyeq r_X \circ y \succcurlyeq \alpha^2(q_X \circ y) + (1 - \alpha^2)(r_X \circ y) \sim x \circ p_Y \succcurlyeq q_X \circ y \succcurlyeq (x_n, y)$$

The conclusion, therefore, follows since $(x, y_1) \sim (x_1, y)$ and $(x, y_n) \sim (x_n, y)$. $\square$

Lemma 5. *For any* $x, x' \in X, y, y' \in Y$, $(x, y) \succcurlyeq (x', y)$ *iff* $(x, y') \succcurlyeq (x', y')$.

Proof. Towards a contradiction, suppose $(x, y) \succcurlyeq (x', y)$ but $(x', y') \succ (x, y')$, for some $x, x' \in X, y, y' \in Y$. There are four cases that exhaust all possibilities and we show that a contradiction appears in all four.

Case I: $(x, y) \succcurlyeq (x', y') \succ (x, y') \succcurlyeq (x', y)$
In this case, it follows from the Continuity and MH Independence axioms that there exists $p_X, p'_X \in \Delta(X)$ such that $p_X \circ y \sim (x, y')$ and $p'_X \circ y \sim (x', y')$. Accordingly, $p'_X \circ y \succ p_X \circ y$. The Separability axiom then implies that:

$$\tfrac{1}{2}(p_X \circ y) + \tfrac{1}{2}(x', y) \sim \tfrac{1}{2}(p'_X \circ y) + \tfrac{1}{2}(x, y).$$

Finally, note that since $p'_X \circ y \succ p_X \circ y$, it follows from MH Independence that $(x', y) \succ (x, y)$, which brings us to our desired contradiction.

Case II: $(x', y') \succcurlyeq (x, y) \succcurlyeq (x', y) \succcurlyeq (x, y')$
Arguing along similar lines as in *Case I*, we can show that a contradiction arises in this case. We avoid repeating the argument here.

Case III: $(x, y) \succ (x', y')$ and $(x', y) \succ (x, y')$
In this case, since $(x, y) \succ (x, y')$, it follows from the Richness condition that there exists $x'' \in X$ such that either (i) $(x'', y') \succ (x, y)$ or (ii) $(x, y') \succ (x'', y)$.
In case of (i), since $(x'', y') \succ (x, y) \succcurlyeq (x', y) \succ (x, y')$, it follows from the Continuity and MH Independence axioms that there exists $p_X, p'_X \in \Delta(X)$ such that $p_X \circ y' \sim (x, y)$ and $p'_X \circ y' \sim (x', y)$. The Separability axiom then implies that:

$$\tfrac{1}{2}(p_X \circ y') + \tfrac{1}{2}(x', y') \sim \tfrac{1}{2}(p'_X \circ y') + \tfrac{1}{2}(x, y').$$

Now, since $(x', y') \succ (x, y')$, it follows as a consequence of MH Independence that $p'_X \circ y' \succ p_X \circ y'$, in turn, implying that $(x', y) \succ (x, y)$, which brings us to our desired contradiction. On the other hand, in case of *(ii)*, since $(x, y) \succ (x', y') \succ (x, y') \succ (x'', y)$, once again by virtue of Continuity and MH Independence axioms, it follows that there exists $p_X, p'_X \in \Delta(X)$ such that $p_X \circ y \sim (x, y')$ and $p'_X \circ y \sim (x', y')$. We once again arrive at a contradiction following a similar argument as in *Case I*.

Case IV: $(x', y') \succ (x, y)$ and $(x, y') \succ (x', y)$
Arguing along similar lines as in *Case III*, we can show that a contradiction arises in this case as well. We do not repeat the details here. □

Lemma 6. *For any $p_X \in \Delta(X)$, there exists $x(p_X) \in X$, such that for any $p_Y \in \Delta(Y)$, $p_X \circ p_Y \sim x(p_X) \circ p_Y$.*[14]

Proof. We first establish the result for the case that the lottery p_Y is a degenerate one. That is, we show that for any $p_X \in \Delta(X)$, there exists $x(p_X) \in X$, such that for any $y \in Y$, $p_X \circ y \sim x(p_X) \circ y \equiv (x(p_X), y)$. To that end, suppose otherwise. Say, for some $p_X \in \Delta(X)$, $p_X \circ y \sim (x, y)$ and $p_X \circ y' \sim (x', y')$.[15] But, $\neg[(x', y') \sim (x, y')]$. If $(x, y') \succ (x', y')$, then by Lemma 5, $(x, y) \succ (x', y)$; and we have $p_X \circ y \sim (x, y) \succ (x', y)$ and $(x, y') \succ (x', y') \sim p_X \circ y'$. On the other hand if $(x', y') \succ (x, y')$ then $(x', y) \succ (x, y)$; and we have $p_X \circ y' \sim (x', y') \succ (x, y')$ and $(x', y) \succ (x, y) \sim p_X \circ y$. The two cases are symmetric; hence, assume wlog that $(x, y') \succ (x', y')$. There are two possibilities: either $p_X \circ y \succ p_X \circ y'$ or $p_X \circ y' \succcurlyeq p_X \circ y$.

Case I: $p_X \circ y \succ p_X \circ y'$
For this case, we consider each of the possibilities, *(i)* $(x', y) \succ (x', y')$ and *(ii)* $(x', y') \succcurlyeq (x', y)$, and show that a contradiction emerges under each.

- In case of *(i)*, we have that $(x, y) \sim p_X \circ y \succ (x', y) \succ (x', y') \sim p_X \circ y'$. Now, since $(x', y) \succ (x', y')$, the Richness condition implies that there exists $x'' \in X$ such that either $(x'', y') \succ (x', y)$ or $(x', y') \succ (x'', y)$. In the first case, it follows that there exists $\alpha \in (0, 1]$ such that $(x'', y') \succ \alpha(p_X \circ y) + (1 - \alpha)(x', y) \equiv (\alpha p_X + (1 - \alpha)x') \circ y$.[16] It therefore follows from the Continuity and MH Independence axioms that there exists $r_X, r'_X \in \Delta(X)$ such that $r_X \circ y' \sim (\alpha p_X + (1 - \alpha)x') \circ y$ and $r'_X \circ y' \sim (x', y)$. Note that since $(\alpha p_X + (1 - \alpha)x') \circ y \succ (x', y)$, we have $r_X \circ y' \succ r'_X \circ y'$. The Separability axiom then implies that:

[14] Observe that this result strengthens the conclusion of Lemma 2.

[15] We know that such (x, y) and (x', y') exist from Lemma 2

[16] Note that if $(x'', y') \succcurlyeq p_X \circ y$, then any $\alpha \in (0, 1]$ works. On the other hand, if $p_X \circ y \succ (x'', y')$, then α will have to be appropriately small.

$$\tfrac{1}{2}(r_X \circ y') + \tfrac{1}{2}(x', y') \sim \tfrac{1}{2}(r'_X \circ y') + \tfrac{1}{2}(\alpha p_X + (1-\alpha)x') \circ y'$$

Now, since $r_X \circ y' \succ r'_X \circ y'$, it follows as a consequence of MH Independence that $(\alpha p_X + (1-\alpha)x') \circ y' \equiv \alpha(p_X \circ y') + (1-\alpha)(x', y') \succ (x', y')$. But, this implies that $p_X \circ y' \succ (x', y')$, which contradicts the maintained assumption that $p_X \circ y' \sim (x', y')$. A similar contradiction also arises in case $(x', y') \succ (x'', y)$.

- In case of (ii), we have $(x, y) \sim p_X \circ y \succ p_X \circ y' \sim (x', y') \succcurlyeq (x', y)$. By the Continuity and MH Independence axioms, it follows that there exists $r_X, r'_X \in \Delta(X)$ such that $r_X \circ y \sim p_X \circ y'$ and $r'_X \circ y \sim (x', y')$. Accordingly, $r_X \circ y \sim r'_X \circ y$. Applying Separability implies that:

$$\tfrac{1}{2}(r_X \circ y) + \tfrac{1}{2}(x', y) \sim \tfrac{1}{2}(r'_X \circ y) + \tfrac{1}{2}(p_X \circ y)$$

Finally, since $r_X \circ y \sim r'_X \circ y$, MH Independence implies that $p_X \circ y \sim (x', y)$, which brings us to our desired contradiction since $p_X \circ y \sim (x, y)$ and $(x, y) \succ (x', y)$.

Case II: $p_X \circ y' \succcurlyeq p_X \circ y$
In this case, we have $(x', y') \sim p_X \circ y' \succcurlyeq p_X \circ y \sim (x, y) \succ (x', y)$. Since, $(x', y') \succ (x', y)$, by the Richness condition, we know that there exist $x'' \in X$ such that either $(x'', y) \succ (x', y')$ or $(x', y) \succ (x'', y')$. For both these cases, we can make similar arguments as the ones above to arrive at our desired contradiction. We do not repeat the details.

Therefore, we have established that for any $p_X \in \Delta(X)$, there exists $x(p_X) \in X$, such that for any $y \in Y$, $p_X \circ y \sim x(p_X) \circ y \equiv (x(p_X), y)$. We now use this fact to prove the general result. Specifically, we show below that $p_X \circ p_Y \sim x(p_X) \circ p_Y$, for any $p_Y \in \Delta(Y)$. To that end, suppose otherwise. We consider here the case $p_X \circ p_Y \succ x(p_X) \circ p_Y$ (the analysis for the other case $x(p_X) \circ p_Y \succ p_X \circ p_Y$ can be carried out along similar lines). Now, we know from Lemma 4 that there exists $\overline{y}, \underline{y} \in Y$ such that $(x(p_X), \overline{y}) \succcurlyeq x(p_X) \circ p_Y \succcurlyeq (x(p_X), \underline{y})$. There are two possibilities.

- $\underline{p_X \circ p_Y \succ (x(p_X), \overline{y})}$: In this case, $p_X \circ p_Y \succ (x(p_X), \overline{y}) \sim p_X \circ \overline{y} \succcurlyeq x(p_X) \circ p_Y$, where the indifference, $(x(p_X), \overline{y}) \sim p_X \circ \overline{y}$, follows from the first part of the proof above. Continuity and MH Independence then implies that there exists $r_X, r'_X \in \Delta(X)$ such that $r_X \circ p_Y \sim (x(p_X), \overline{y})$ and $r'_X \circ p_Y \sim p_X \circ \overline{y}$. Transitivity implies that $r_X \circ p_Y \sim r'_X \circ p_Y$. Further, by Separability it follows that $\tfrac{1}{2}(r_X \circ p_Y) + \tfrac{1}{2}(p_X \circ p_Y) \sim \tfrac{1}{2}(r'_X \circ p_Y) + \tfrac{1}{2}(x(p_X) \circ p_Y)$. Finally, since $r_X \circ p_Y \sim r'_X \circ p_Y$, MH Independence implies that $p_X \circ p_Y \sim x(p_X) \circ p_Y$, which contradicts our maintained assumption of $p_X \circ p_Y \succ x(p_X) \circ p_Y$

- $\underline{(x(p_X), \overline{y}) \succcurlyeq p_X \circ p_Y}$: In this case, $(x(p_X), \overline{y}) \succcurlyeq p_X \circ p_Y \succ x(p_X) \circ p_Y \succcurlyeq (x(p_X), \underline{y})$. Similar to arguments made above, by using the Richness condition along with Continuity and MH Independence, we can conclude that there exists $y' = \overline{y}$ or $\underline{y}$

> and $r_X, r'_X \in \Delta(X)$ such that $r_X \circ y' \sim p_X \circ p_Y$ and $r'_X \circ y' \sim x(p_X) \circ p_Y$. By transitivity, it follows that $r_X \circ y' \succ r'_X \circ y'$. Separability then implies that $\frac{1}{2}(r_X \circ y') + \frac{1}{2}(x(p_X), y') \sim \frac{1}{2}(r'_X \circ y') + \frac{1}{2}(p_X \circ y')$. Since, $r_X \circ y' \succ r'_X \circ y'$, by MH-Independence it follows that, $p_X \circ y' \succ (x(p_X), y')$, which contradicts the conclusion established in the first part of the proof that $p_X \circ y' \sim (x(p_X), y')$.

Hence, for any $p_X \in \Delta(X)$, there exists $x(p_X) \in X$, such that for any $p_Y \in \Delta(Y)$, $p_X \circ p_Y \sim x(p_X) \circ p_Y$. □

Lemma 7. *For any $p_X, q_X \in \Delta(X), y, y' \in Y$, $p_X \circ y \succcurlyeq q_X \circ y$ iff $p_X \circ y' \succcurlyeq q_X \circ y'$.*[17]

Proof. The conclusion follows immediately from Lemmas 5 and 6 since:

$$\begin{aligned} p_X \circ y \succcurlyeq q_X \circ y &\Leftrightarrow (x(p_X), y) \succcurlyeq (x(q_X), y) \\ &\Leftrightarrow (x(p_X), y') \succcurlyeq (x(q_X), y') \\ &\Leftrightarrow p_X \circ y' \succcurlyeq q_X \circ y' \end{aligned}$$

□

Lemma 8. *If $\succ \neq \emptyset$ then:*

(1) $\succ$ restricted to $X \times Y$ is non-empty

(2) for any $y \in Y$, $\succ$ restricted to the set $\Delta_y = \{p_X \circ y : p_X \in \Delta(X)\}$ is non-empty.

Proof. To prove *(1)*, suppose towards a contradiction that $\succ \neq \emptyset$ but $(x, y) \sim (x', y')$, for all (x, y), $(x', y') \in X \times Y$. Specifically, this means that for any $y \in Y$, $(x, y) \sim (x', y)$, for all $x, x' \in X$. Further, MH Independence implies that that $p_X \circ y \sim p'_X \circ y \sim (x, y)$, for all $p_X, p'_X \in \Delta(X)$ and $x \in X$. Putting all of this together, we have that $p_X \circ y \sim p'_X \circ y'$, for all $p_X, p'_X \in \Delta(X)$ and $y, y' \in Y$. Now consider any $x \circ p_Y$ and $x' \circ p'_Y \in \Delta$. Let $p_X \circ y$ and $p'_X \circ y'$ be, respectively, their risk translations.[18] Further, let $(\alpha, q_X \circ y, r_X \circ y)$ and $(\alpha', q'_X \circ y', r'_X \circ y')$ be, respectively, Gul decompositions of $p_X \circ y$ and $p'_X \circ y'$. Then, by Morally Motivated Bayesianism, it follows that

$$\begin{aligned} x \circ p_Y &\sim \alpha^2(q_X \circ y) + (1 - \alpha^2)(r_X \circ y) \\ &\sim \alpha'^2(q'_X \circ y') + (1 - \alpha'^2)(r'_X \circ y') \sim x' \circ p'_Y \end{aligned}$$

[17] It is worth pointing out that the following stronger result which extends the conclusion of this Lemma is implied by our axioms and can be proven along similar lines as the proof of Lemma 5: For any $p_X, q_X \in \Delta(X), p_Y, p'_Y \in \Delta(Y)$, $p_X \circ p_Y \succcurlyeq q_X \circ p_Y$ iff $p_X \circ p'_Y \succcurlyeq q_X \circ p'_Y$. Since we do not require this stronger result to prove our representation result, we do not present its proof here.

[18] By Lemma 3, we know that these risk translations exist.

Finally, consider any $p, q \in \Delta$. By Lemma 1, $p \sim p_X \circ p_Y$ and $q \sim q_X \circ q_Y$. Further, by Lemma 2, $p_X \circ p_Y \sim x \circ p_Y$ and $q_X \circ q_Y \sim x' \circ q_Y$, for some $x, x' \in X$. Moreover, as shown above, $x \circ p_Y \sim x' \circ q_Y$. Therefore, by transitivity of $\succcurlyeq$, it follows that $p \sim q$, which implies that $\succ = \emptyset$!

To prove *(2)*, we know from *(1)* that there exists $(x, y), (x', y') \in X \times Y$ such that $(x, y) \succ (x', y')$. If $y' = y$, then $\succ$ restricted to $\Delta_{y'}$ is non-empty and it follows from Lemma 5 that for, any $y \in Y$, $\succ$ restricted to Δ_y is non-empty. So, consider the case that $y \neq y'$ and note that either $(x, y') \sim (x, y)$ or $\neg[(x, y') \sim (x, y)]$. If it is the former, then $(x, y') \sim (x, y) \succ (x', y')$ and, hence, $\succ$ restricted to $\Delta_{y'}$ is non-empty establishing the claim. On the other hand, if it is the latter, say, $(x, y) \succ (x, y')$, then, by Richness, there exists $x'' \in X$ such that either $(x'', y') \succ (x, y)$, implying $\succ$ restricted to $\Delta_{y'}$ is non-empty; or $(x, y') \succ (x'', y)$ implying $\succ$ restricted to Δ_y is non-empty. Thus, our desired conclusion follows. □

Lemma 9. *If $\succ \neq \emptyset$, then for any $y, y' \in Y$, there exists $\widetilde{p}_X, \widetilde{p}'_X, \widetilde{q}_X, \widetilde{q}'_X \in \Delta(X)$ such that $\widetilde{p}_X \circ y \sim \widetilde{q}_X \circ y' \succ \widetilde{p}'_X \circ y \sim \widetilde{q}'_X \circ y'$.*

Proof. Pick any $p_X \in \Delta(X)$ such that $p_X \circ y$ and $p_X \circ y'$ are not indifferent—if no such p_X exists, then our desired conclusion follows immediately given that $\succ$ restricted to the sets Δ_y and $\Delta_{y'}$ is non-empty. Assume wlog that $p_X \circ y \succ p_X \circ y'$; or, $(x(p_X), y) \succ (x(p_X), y')$. By the Richness condition, it follows that there exists $x' \in X$ such that either $(x', y') \succ (x(p_X), y)$ or $(x(p_X), y') \succ (x', y)$. Under both these cases, it follows by virtue of the two axioms of Continuity and MH Independence that $\widetilde{p}_X, \widetilde{p}'_X, \widetilde{q}_X, \widetilde{q}'_X$ exist as desired. □

A.2 Proof of Theorem

Establishing the representation for the case when $\succ = \emptyset$ is immediate: simply let $u : X \to \mathbb{R}$ and $v : Y \to \mathbb{R}$ be some constant functions. We consider here the proof of sufficiency of the axioms for the representation for the case when $\succ \neq \emptyset$. We break the proof up into several steps. In the way of notation, note that we define the following sets:

- for any $x \in X$, $\Delta_x = \{x \circ p_Y : p_Y \in \Delta(Y)\}$
- for any $y \in Y$, $\Delta_y = \{p_X \circ y : p_X \in \Delta(X)\}$

Step 1: Show that there exists a continuous function $w : X \times Y \to \mathbb{R}$ such that the function $W : \cup_{y \in Y} \Delta_y \to \mathbb{R}$ given by $W(p_X \circ y) = \sum_{x \in X} p_X(x) w(x, y)$ represents $\succcurlyeq$ restricted to $\cup_{y \in Y} \Delta_y$.

Since $\succ \neq \emptyset$, we know from Lemma 8 that for any $y \in Y$, $\succ$ restricted to the set Δ_y is non-empty. Further, our three axioms of Weak Order, Continuity and MH Independence imply that for any such set Δ_y, there exists a continuous, bounded function $\tilde{w}_y : X \to \mathbb{R}$ such that the function $W_y : \Delta_y \to \mathbb{R}$, given by $W_y(p_X \circ y) = \sum_{x \in X} p_X(x)\tilde{w}_y(x)$ represents $\succcurlyeq$ restricted to Δ_y.[19] Further, the function $\tilde{w}_y$ and, hence, W_y is unique up to a positive affine transformation. We will next piece together the family of $\tilde{w}_y$ and W_y functions to arrive at the functions $w : X \times Y \to \mathbb{R}$ and $W : \cup_{y \in Y} \Delta_y \to \mathbb{R}$ as hypothesized above. To that end, pick any $y^* \in Y$. Begin by defining the function W on the set Δ_{y^*} by setting $W(p_X \circ y^*) = W_{y^*}(p_X \circ y^*)$, for all $p_X \circ y^* \in \Delta_{y^*}$. Next, we define the function W on the sets Δ_y for $y \neq y^*$. For any such y, we know from Lemma 9 that there exists $\widetilde{p}_X, \widetilde{p}'_X, \widetilde{q}_X, \widetilde{q}'_X \in \Delta(X)$ such that $\widetilde{p}_X \circ y \sim \widetilde{q}_X \circ y^* \succ \widetilde{p}'_X \circ y \sim \widetilde{q}'_X \circ y^*$. As mentioned above, the function W_y is defined uniquely up to a positive affine transformation, that is, we have two degrees of freedom in specifying it. Accordingly, we can *redefine* it by setting $W_y(\widetilde{p}_X \circ y) = W(\widetilde{q}_X \circ y^*)$ and $W_y(\widetilde{p}'_X \circ y) = W(\widetilde{q}'_X \circ y^*)$. Clearly, in the process, we redefine the function $\tilde{w}_y$ as well. We can, then, extend the function W to the set of lotteries in Δ_y by defining $W(p_X \circ y) = W_y(p_X \circ y)$ for all $p_X \circ y \in \Delta_y$. Observe that the MH Independence and Continuity axioms imply that for any $p_X \circ y \in \Delta_y$ and $q_X \circ y^* \in \Delta_{y^*}$, $p_X \circ y \sim q_X \circ y^*$ iff $W(p_X \circ y) = W(q_X \circ y^*)$.[20] Next, define the function $w : X \times Y \to \mathbb{R}$ by $w(x, y) = \tilde{w}_y(x)$. Accordingly, we have defined a function $W : \cup_{y \in Y} \Delta_y \to \mathbb{R}$, given by $W(p_X \circ y) = \sum_{x \in X} p_X(x) w(x, y)$. Using Lemma 9 and MH Independence, it is straightforward to verify that the function W represents $\succcurlyeq$ restricted to the product measures in $\cup_{y \in Y} \Delta_y$, i.e., for all $p_X \circ y, p'_X \circ y'$, $p_X \circ y \succcurlyeq p'_X \circ y'$ iff $W(p_X \circ y) \geq W(p'_X \circ y')$.[21]

The final thing we need to show to establish the conclusion of Step 1 is that the function w is continuous. This is accomplished by showing that the function W is continuous. To that end, let $p^k_X \circ y^k$ be a sequence in $\cup_{y \in Y} \Delta_y$ that converges to some $\hat{p}_X \circ \hat{y} \in \cup_{y \in Y} \Delta_y$. First, consider the case where $\hat{p}_X \circ \hat{y}$ is neither the best nor the worst lottery in Δ. In this

[19] The proof establishing that such a continuous function $\tilde{w}_y : X \to \mathbb{R}$ exists can be shown along similar lines as in Grandmont (1972).

[20] Consider, for instance, the case $\widetilde{p}_X \circ y \sim \widetilde{q}_X \circ y^* \succcurlyeq p_X \circ y \sim q_X \circ y^* \succcurlyeq \widetilde{p}'_X \circ y \sim \widetilde{q}'_X \circ y^*$, where $\widetilde{p}_X \circ y, \widetilde{q}_X \circ y^*, \widetilde{p}'_X \circ y, \widetilde{q}'_X \circ y^*$ are as above. In this case, it follows from the axioms of Continuity and MH Independence that there exists a unique $\alpha \in [0, 1]$ such that $p_X \circ y \sim \alpha(\widetilde{p}_X \circ y) + (1 - \alpha)(\widetilde{p}'_X \circ y)$ and $q_X \circ y^* \sim \alpha(\widetilde{q}_X \circ y^*) + (1 - \alpha)(\widetilde{q}'_X \circ y^*)$. Accordingly, $W(p_X \circ y) = W(\alpha(\widetilde{p}_X \circ y) + (1 - \alpha)(\widetilde{p}'_X \circ y)) = \alpha W(\widetilde{p}_X \circ y) + (1 - \alpha) W(\widetilde{p}'_X \circ y)$ and $W(q_X \circ y^*) = W(\alpha(\widetilde{q}_X \circ y^*) + (1 - \alpha)(\widetilde{q}'_X \circ y^*)) = \alpha W(\widetilde{q}_X \circ y^*) + (1 - \alpha) W(\widetilde{q}'_X \circ y^*)$. Accordingly, $W(p_X \circ y) = W(q_X \circ y^*)$, since $W(\widetilde{p}_X \circ y) = W(\widetilde{q}_X \circ y^*)$ and $W(\widetilde{p}'_X \circ y) = W(\widetilde{q}'_X \circ y^*)$. We arrive at a similar conclusion for the other two cases of $p_X \circ y \sim q_X \circ y^* \succ \widetilde{p}_X \circ y \sim \widetilde{q}_X \circ y^*$ and $\widetilde{p}'_X \circ y \sim \widetilde{q}'_X \circ y^* \succ p_X \circ y \sim q_X \circ y^*$ as well. Similar arguments establish that the implication also goes in the opposite direction, i.e., if $W(p_X \circ y) = W(q_X \circ y^*)$ then $p_X \circ y \sim q_X \circ y^*$.

[21] To see this, first, note that Lemma 9 together with the axioms of Continuity and MH Independence guarantees that there exists $\hat{p}_X, \hat{p}'_X$, $\hat{q}_X, \hat{q}'_X$, $\hat{r}_X, \hat{r}'_X \in \Delta(X)$ such that $\hat{p}_X \circ y \sim \hat{q}_X \circ y^* \sim \hat{r}_X \circ y' \succ \hat{p}'_X \circ y \sim \hat{q}'_X \circ y^* \sim \hat{r}'_X \circ y'$. Based on the conclusion established in the last footnote, it therefore follows that $W(\hat{p}_X \circ y) = W(\hat{q}_X \circ y^*) = W(\hat{r}_X \circ y') > W(\hat{p}'_X \circ y) = W(\hat{q}'_X \circ y^*) = W(\hat{r}'_X \circ y')$. Following similar arguments as in the last footnote, we can then show drawing on MH Independence that for all $p_X \circ y, p'_X \circ y'$, $p_X \circ y \sim p'_X \circ y'$ iff $W(p_X \circ y) = W(p'_X \circ y')$. It is also straightforward to establish that $p_X \circ y \succ p'_X \circ y'$ iff $W(p_X \circ y) > W(p'_X \circ y')$.

case, there exists $\overline{y} \in Y$ and $\overline{q}_X, \underline{q}_X \in \Delta(X)$ such that $\overline{q}_X \circ \overline{y} \succ \hat{p}_X \circ \hat{y} \succ \underline{q}_X \circ \overline{y}$.[22] By the Continuity and MH Independence axioms, it follows that there exists $\hat{\alpha} \in (0,1)$ such that $\hat{p}_X \circ \hat{y} \sim \hat{\alpha}(\overline{q}_X \circ \overline{y}) + (1-\hat{\alpha})(\underline{q}_X \circ \overline{y})$. Further, since $p_X^k \circ y^k$ converges to $\hat{p}_X \circ \hat{y}$, by continuity of $\succsim$, it follows that for all k large enough, $\overline{q}_X \circ \overline{y} \succ p_X^k \circ y^k \succ \underline{q}_X \circ \overline{y}$. Therefore, for all such k, there exists $\alpha^k \in (0,1)$ such that $p_X^k \circ y^k \sim \alpha^k(\overline{q}_X \circ \overline{y}) + (1-\alpha^k)(\underline{q}_X \circ \overline{y})$. It is fairly straightforward to see that $\alpha^k \to \hat{\alpha}$. This implies that $(\alpha^k \overline{q}_X + (1-\alpha^k)\underline{q}_X) \circ \overline{y} = \alpha^k(\overline{q}_X \circ \overline{y}) + (1-\alpha^k)(\underline{q}_X \circ \overline{y})$ converges to $(\hat{\alpha}\overline{q}_X + (1-\hat{\alpha})\underline{q}_X) \circ \overline{y} = \hat{\alpha}(\overline{q}_X \circ \overline{y}) + (1-\hat{\alpha})(\underline{q}_X \circ \overline{y})$. We know from above that the function $W_{\overline{y}}$ is continuous, which implies that $W_{\overline{y}}((\alpha^k \overline{q}_X + (1-\alpha^k)\underline{q}_X) \circ \overline{y}) \to W_{\overline{y}}((\hat{\alpha}\overline{q}_X + (1-\hat{\alpha})\underline{q}_X) \circ \overline{y})$. But, by construction, $W_{\overline{y}} = W$ restricted to the set $\Delta_{\overline{y}}$. Therefore, $W((\alpha^k \overline{q}_X + (1-\alpha^k)\underline{q}_X) \circ \overline{y}) \to W((\hat{\alpha}\overline{q}_X + (1-\hat{\alpha})\underline{q}_X) \circ \overline{y})$ and, accordingly, $W(p_X^k \circ y^k) \to W(\hat{p}_X \circ \hat{y})$. Now, consider the case where $\hat{p}_X \circ \hat{y}$ is the best lottery in Δ. In this case, there exists $\hat{p}'_X \circ \hat{y}$ such that for all k large enough, $\hat{p}_X \circ \hat{y} \succsim p_X^k \circ y^k \succ \hat{p}'_X \circ \hat{y}$ and we can establish that $W(p_X^k \circ y^k) \to W(\hat{p}_X \circ \hat{y})$ along similar lines as above by drawing on the continuity of $W_{\hat{y}}$. Likewise for the case when $\hat{p}_X \circ \hat{y}$ is the worst lottery in Δ. This establishes that the function W is continuous and, accordingly, so is w.

Step 2: Show that there exist continuous functions $u : X \to \mathbb{R}$ and $v : Y \to \mathbb{R}$ such that the function $W : \cup_{y \in Y} \Delta_y \to \mathbb{R}$ given by $W(p_X \circ y) = \sum_{x \in X} p_X(x)u(x) + v(y)$ represents $\succsim$ restricted to $\cup_{y \in Y} \Delta_y$. That is, establish that $\succsim$ restricted to the set $\cup_{y \in Y} \Delta_y$ has an MH representation.

We define the functions u and v using the function w defined in Step 1. To that end, recall from Lemma 7 that for any $p_X, q_X \in \Delta(X)$ and $y, y' \in Y$, $p_X \circ y \succsim q_X \circ y$ iff $p_X \circ y' \succsim q_X \circ y'$. This along with the fact that preferences satisfy the axioms of Continuity and MH Independence implies that the function w derived in Step 1 has the following property:

- For any $y, y' \in Y$, $w(.,y) : X \to \mathbb{R}$ is a positive affine transformation of $w(.,y') : X \to \mathbb{R}$.[23]

[22] We can take $\overline{y} = \hat{y}$, unless $\hat{p}_X \circ \hat{y} \succsim p_X \circ \hat{y}$ for all $p_X \in \Delta(X)$ or, $p_X \circ \hat{y} \succsim \hat{p}_X \circ \hat{y}$ for all $p_X \in \Delta(X)$. In the first case, $\overline{p} \succ \hat{p}_X \circ \hat{y}$ for some $\overline{p} \in \Delta$. In this case, there exists $\overline{y} \in Y$ s.t. $(x(\overline{p}_X), \overline{y}) \succsim x(\overline{p}_X) \circ \overline{p}_Y \sim \overline{p}_X \circ \overline{p}_Y \sim \overline{p} \succ \hat{p}_X \circ \hat{y} \sim (x(\hat{p}_X), \hat{y})$, where the first preference follows from Lemma 4, the second and fifth from Lemma 6, and the third from Lemma 1. Further, $(x(\hat{p}_X), \hat{y}) \succsim (x(\overline{p}_X), \hat{y})$. Hence, it follows from the Richness condition that there exists x' s.t. $(x(\overline{p}_X), \hat{y}) \succ (x', \overline{y})$. Accordingly, $(x(\overline{p}_X), \overline{y}) \succ \hat{p}_X \circ \hat{y} \succ (x', \overline{y})$. Working along similar lines, the same conclusion can also be established for the second case in which $p_X \circ \hat{y} \succsim \hat{p}_X \circ \hat{y}$ for all $p_X \in \Delta(X)$.

[23] This is straightforward to establish. Consider $p_X, q_X \in \Delta(X)$ such that $p_X \circ y \succ q_X \circ y$. This implies that $p_X \circ y' \succ q_X \circ y'$. Now consider any r_X such that $p_X \circ y \succsim r_X \circ y \succsim q_X \circ y$ and, accordingly, $p_X \circ y' \succsim r_X \circ y' \succsim q_X \circ y'$. Lemma 7 along with the axioms of Continuity and MH Independence implies that there exists a unique $\alpha \in [0,1]$ s.t. $r_X \circ y \sim \alpha(p_X \circ y) + (1-\alpha)(q_X \circ y)$ and $r_X \circ y' \sim \alpha(p_X \circ y') + (1-\alpha)(q_X \circ y')$, i.e., $W(r_X \circ y) = \alpha W(p_X \circ y) + (1-\alpha)W(q_X \circ y)$ and $W(r_X \circ y') = \alpha W(p_X \circ y') + (1-\alpha)W(q_X \circ y')$. It follows from these two equalities that:

$$\frac{W(r_X \circ y) - W(q_X \circ y)}{W(p_X \circ y) - W(q_X \circ y)} = \alpha = \frac{W(r_X \circ y') - W(q_X \circ y')}{W(p_X \circ y') - W(q_X \circ y')}$$

Now, pick any $y^* \in Y$ and define the function $u : X \to \mathbb{R}$ by letting $u(x) = w(x, y^*)$. Clearly, the function u is continuous given that w is. Since each of the functions $w(., y)$, $y \in Y$, is a positive affine transformation of $w(., y^*)$, there exist constants $f(y) > 0$ and $v(y)$ such that $w(x, y) = f(y)u(x) + v(y)$. Of course, $f(y^*) = 1$ and $v(y^*) = 0$. This allows us to define the function $v : Y \to \mathbb{R}$. We next show that $f(y) = 1$, for all $y \in Y$, which in turn establishes that v is continuous, given that w is. To do so, recall from Lemma 9 that for any such $y \neq y^*$ we can find $\widetilde{p}_X, \widetilde{p}'_X, \widetilde{q}_X, \widetilde{q}'_X \in \Delta(X)$ such that $\widetilde{p}_X \circ y \sim \widetilde{q}_X \circ y^* \succ \widetilde{p}'_X \circ y \sim \widetilde{q}'_X \circ y^*$. By the Separability axiom, it then follows that $(\frac{1}{2}\widetilde{p}_X + \frac{1}{2}\widetilde{q}'_X) \circ y \sim (\frac{1}{2}\widetilde{p}'_X + \frac{1}{2}\widetilde{q}_X) \circ y$. This implies that:

$$\begin{aligned} & f(y) \sum_{x \in X} (\frac{1}{2}\widetilde{p}_X + \frac{1}{2}\widetilde{q}'_X)(x)u(x) + v(y) \\ & \quad = f(y) \sum_{x \in X} (\frac{1}{2}\widetilde{p}'_X + \frac{1}{2}\widetilde{q}_X)(x)u(x) + v(y) \\ \implies & \sum_{x \in X} \widetilde{p}_X(x)u(x) - \sum_{x \in X} \widetilde{p}'_X(x)u(x) = \sum_{x \in X} \widetilde{q}_X(x)u(x) - \sum_{x \in X} \widetilde{q}'_X(x)u(x) > 0 \end{aligned}$$

At the same time $\widetilde{p}_X \circ y \sim \widetilde{q}_X \circ y^*$ and $\widetilde{p}'_X \circ y \sim \widetilde{q}'_X \circ y^*$ implies that:

$$\begin{aligned} f(y) \sum_{x \in X} \widetilde{p}_X(x)u(x) + v(y) &= \sum_{x \in X} \widetilde{q}_X(x)u(x) \\ f(y) \sum_{x \in X} \widetilde{p}'_X(x)u(x) + v(y) &= \sum_{x \in X} \widetilde{q}'_X(x)u(x) \end{aligned}$$

Subtracting the second equation from the first, then, gives us that:

$$f(y) \left[\sum_{x \in X} \widetilde{p}_X(x)u(x) - \sum_{x \in X} \widetilde{p}'_X(x)u(x) \right] = \sum_{x \in X} \widetilde{q}_X(x)u(x) - \sum_{x \in X} \widetilde{q}'_X(x)u(x)$$

Putting everything together, it follows that $f(y) = 1$. We have, therefore, established that there exist continuous functions $u : X \to \mathbb{R}$ and $v : Y \to \mathbb{R}$ such that the function $W : \cup_{y \in Y} \Delta_y \to \mathbb{R}$, given by $W(p_X \circ y) = \sum_{x \in X} p_X(x)u(x) + v(y)$, represents $\succcurlyeq$ restricted to the set $\cup_{y \in Y} \Delta_y$.

Before proceeding to the next step of the proof, we note an implication of the conclusion here that we will use in Step 4 below. Recall from Lemma 6 that for any $p_X \in \Delta(X)$, there exists $x(p_X) \in X$ such that for any $y \in Y$, $p_X \circ y \sim x(p_X) \circ y$. Therefore, it follows from our conclusion in this step that $\sum_{x \in X} p_X(x)u(x) = u(x(p_X))$.

which, in turn, implies:

$$W(r_X \circ y) = \frac{W(p_X \circ y) - W(q_X \circ y)}{W(p_X \circ y') - W(q_X \circ y')} W(r_X \circ y') + W(q_X \circ y) - \frac{W(p_X \circ y) - W(q_X \circ y)}{W(p_X \circ y') - W(q_X \circ y')} W(q_X \circ y')$$

It is also straightforward to establish that the same conclusion also follows when $r_X \circ y \succ p_X \circ y$ or $q_X \circ y \succ r_X \circ y$. From this it immediately follows that $w(., y)$ is a positive affine transformation of $w(., y')$.

Step 3: Establish that $\succcurlyeq$ restricted to the set $(\cup_{x\in X}\Delta_x)\cup(\cup_{y\in Y}\Delta_y)$ has an MH representation.

We now extend the MH representation to the set $\cup_{x\in X}\Delta_x$ using the functions $u : X \to \mathbb{R}$ and $v : Y \to \mathbb{R}$ defined in Step 2. Consider any $x \circ p_Y = [(x, y_1), \alpha_1; \ldots ; (x, y_N), \alpha_N] \in \Delta_x$. We know from Lemma 3 that there exists $p_X \circ y = [(x_1, y), \alpha_1; \ldots ; (x_N, y), \alpha_N]$ that is a risk translation of $x \circ p_Y$. Further, let $(\alpha, q_X \circ y, r_X \circ y)$ be a Gul decomposition of $p_X \circ y$ defined by letting (x_n, y) belong to the support of $q_X \circ y$ iff $p_X \circ y \succ (x_n, y)$. It follows from the Morally Motivated Bayesianism axiom that $x \circ p_Y \sim \alpha^2(q_X \circ y) + (1-\alpha^2)(r_X \circ y)$. Let $\underline{\mathcal{N}} = \{n : p_X \circ y \succ (x, y_n)\}$ and $\overline{\mathcal{N}} = \{n : (x, y_n) \succcurlyeq p_X \circ y\}$. It is straightforward to verify that $(x, y_n) \succcurlyeq p_X \circ y$ iff $v(y_n) \geq \sum_{\tilde{n}=1}^{N} \alpha_{\tilde{n}} v(y_{\tilde{n}})$.[24] We can now extend the function W to the set $\cup_{x\in X}\Delta_x$ by defining $W(x \circ p_Y)$ for any $x \circ p_Y$ as follows:[25]

$$\begin{aligned}
W(x \circ p_Y) &= W(\alpha^2(q_X \circ y) + (1-\alpha^2)(r_X \circ y)) = \alpha^2 W(q_X \circ y) + (1-\alpha^2) W(r_X \circ y) \\
&= \alpha^2 \sum_{n\in\underline{\mathcal{N}}} \left(\frac{\alpha_n}{\alpha}\right) W(x_n, y) + (1-\alpha^2) \sum_{n\in\overline{\mathcal{N}}} \left(\frac{\alpha_n}{1-\alpha}\right) W(x_n, y) \\
&= \alpha^2 \sum_{n\in\underline{\mathcal{N}}} \left(\frac{\alpha_n}{\alpha}\right) W(x, y_n) + (1-\alpha^2) \sum_{n\in\overline{\mathcal{N}}} \left(\frac{\alpha_n}{1-\alpha}\right) W(x, y_n) \\
&= \alpha^2 \sum_{n\in\underline{\mathcal{N}}} \left(\frac{\alpha_n}{\alpha}\right) (u(x) + v(y_n)) + (1-\alpha^2) \sum_{n\in\overline{\mathcal{N}}} \left(\frac{\alpha_n}{1-\alpha}\right) (u(x) + v(y_n)) \\
&= u(x) + \alpha^2 \sum_{n\in\underline{\mathcal{N}}} \left(\frac{\alpha_n}{\alpha}\right) v(y_n) + (1-\alpha^2) \sum_{n\in\overline{\mathcal{N}}} \left(\frac{\alpha_n}{1-\alpha}\right) v(y_n) \\
&= u(x) + \alpha \sum_{n\in\underline{\mathcal{N}}} \alpha_n v(y_n) + (1+\alpha) \sum_{n\in\overline{\mathcal{N}}} \alpha_n v(y_n) \\
&= u(x) + \alpha \sum_{n=1}^{N} \alpha_n v(y_n) + \sum_{n\in\overline{\mathcal{N}}} \alpha_n v(y_n) \\
&= u(x) + \sum_{n=1}^{N} \alpha_n \max\left\{ v(y_n), \sum_{\tilde{n}=1}^{N} \alpha_{\tilde{n}} v(y_{\tilde{n}}) \right\}
\end{aligned}$$

It is straightforward to verify that the function $W : (\cup_{y\in Y}\Delta_y)\cup(\cup_{x\in X}\Delta_x) \to \mathbb{R}$ thus defined represents $\succcurlyeq$ restricted to $(\cup_{y\in Y}\Delta_y)\cup(\cup_{x\in X}\Delta_x)$.

Step 4: Establish that $\succcurlyeq$ has an MH representation on the whole of Δ.

Recall from Lemma 1 that for any $p \in \Delta$, $p \sim p_X \circ p_Y$. Further, Lemma 6 establishes that for any $p_X \in \Delta(X)$, there exists $x(p_X) \in X$ such that for any $p_Y \in \Delta(Y)$, $p_X \circ p_Y \sim x(p_X) \circ p_Y$. Hence, by transitivity of $\succcurlyeq$, it follows that $p \sim x(p_X) \circ p_Y$. We now extend

[24] To see this, note that from Step 2, $(x, y_n) \succcurlyeq p_X \circ y$ iff $u(x) + v(y_n) \geq \sum_{\tilde{n}=1}^{N} \alpha_{\tilde{n}}[u(x_n) + v(y)] = \sum_{\tilde{n}=1}^{N} \alpha_{\tilde{n}}[u(x) + v(y_{\tilde{n}})] = u(x) + \sum_{\tilde{n}=1}^{N} \alpha_{\tilde{n}} v(y_{\tilde{n}})$. That is, $(x, y_n) \succcurlyeq p_X \circ y$ iff $v(y_n) \geq \sum_{\tilde{n}=1}^{N} \alpha_{\tilde{n}} v(y_{\tilde{n}})$.

[25] We abuse notation below by writing $W(x, y)$ instead of $W((x, y))$.

the function W to the whole of Δ by letting $W(p) = W(x(p_X) \circ p_Y)$. Accordingly,[26]

$$\begin{aligned} W(p) &= W(x(p_X) \circ p_Y) \\ &= u(x(p_X)) + \sum_{y \in Y} p_Y(y) \max \left\{ v(y), \sum_{\tilde{y} \in Y} p_Y(\tilde{y}) v(\tilde{y}) \right\} \\ &= \sum_{x \in X} p_X(x) u(x) + \sum_{y \in Y} p_Y(y) \max \left\{ v(y), \sum_{\tilde{y} \in Y} p_Y(\tilde{y}) v(\tilde{y}) \right\} \end{aligned}$$

This establishes the sufficiency of the axioms for the representation. The necessity of the axioms for the representation as well as the uniqueness result is straightforward to establish and the details are not included here.

References

Babcock, Bruce A, E Kwan Choi, and Eli Feinerman. 1993. "Risk and probability premiums for CARA utility functions." *Journal of Agricultural and Resource Economics* 18 (1):17–24.

Batson, C Daniel, Diane Kobrynowicz, Jessica L Dinnerstein, Hannah C Kampf, and Angela D Wilson. 1997. "In a very different voice: unmasking moral hypocrisy." *Journal of Personality and Social Psychology* 72 (6):1335–1348.

Batson, C Daniel, Elizabeth R Thompson, Greg Seuferling, Heather Whitney, and Jon A Strongman. 1999. "Moral hypocrisy: appearing moral to oneself without being so." *Journal of Personality and Social Psychology* 77 (3):525–537.

Bolton, Gary E and Axel Ockenfels. 2000. "ERC: A theory of equity, reciprocity, and competition." *American Economic Review* 90 (1):166–193.

Brock, J Michelle, Andreas Lange, and Erkut Y Ozbay. 2013. "Dictating the risk: Experimental evidence on giving in risky environments." *American Economic Review* 103 (1):415–37.

Cappelen, Alexander W, James Konow, Erik Ø Sørensen, and Bertil Tungodden. 2013. "Just luck: An experimental study of risk-taking and fairness." *American Economic Review* 103 (4):1398–1413.

Cettolin, Elena, Arno Riedl, and Giang Tran. 2017. "Giving in the face of risk." *Journal of Risk and Uncertainty* 55 (2-3):95–118.

Charness, Gary and Matthew Rabin. 2002. "Understanding social preferences with simple tests." *The Quarterly Journal of Economics* 117 (3):817–869.

[26] Recall from Step 2 that $u(x(p_X)) = \sum_{x \in X} p_X(x) u(x)$.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory* 33 (1):67–80.

Dillenberger, David and Philipp Sadowski. 2012. "Ashamed to be selfish." *Theoretical Economics* 7 (1):99–124.

Eil, David and Justin M Rao. 2011. "The good news-bad news effect: Asymmetric processing of objective information about yourself." *American Economic Journal: Microeconomics* 3 (2):114–38.

Evren, Özgür and Stefania Minardi. 2017. "Warm-glow Giving and Freedom to be Selfish." *The Economic Journal* 127 (603):1381–1409.

Exley, Christine. 2018. "Incentives for prosocial behavior: The role of reputations." *Management Science* 64 (5):2460–2471.

Exley, Christine L. 2016. "Excusing selfishness in charitable giving: The role of risk." *The Review of Economic Studies* 83 (2):587–628.

———. 2020. "Using charity performance metrics as an excuse not to give." *Management Science* 66 (2):553–563.

Falk, Armin, Thomas Neuber, and Nora Szech. 2020. "Diffusion of being pivotal and immoral outcomes." *The Review of Economic Studies* 87 (5):2205–2229.

Fehr, Ernst and Klaus M Schmidt. 1999. "A theory of fairness, competition, and cooperation." *The Quarterly Journal of Economics* 114 (3):817–868.

Feiler, Lauren. 2014. "Testing models of information avoidance with binary choice dictator games." *Journal of Economic Psychology* 45:253–267.

Fudenberg, Drew and David K Levine. 2012. "Fairness, risk preferences and independence: Impossibility theorems." *Journal of Economic Behavior & Organization* 81 (2):606–612.

Garcia, Thomas, Sébastien Massoni, and Marie Claire Villeval. 2020. "Ambiguity and excuse-driven behavior in charitable giving." *European Economic Review* 124:103412.

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2013. "Self-serving altruism? The lure of unethical actions that benefit others." *Journal of Economic Behavior & Organization* 93:285–292.

Gino, Francesca, Michael I Norton, and Roberto A Weber. 2016. "Motivated Bayesians: Feeling moral while acting egoistically." *Journal of Economic Perspectives* 30 (3):189–212.

Gneezy, Uri and Jan Potters. 1997. "An experiment on risk taking and evaluation periods." *The Quarterly Journal of Economics* 112 (2):631–645.

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. "Bribing the self." *Games and Economic Behavior* 120:311–324.

Grandmont, Jean-Michel. 1972. "Continuity properties of a von Neumann-Morgenstern utility." *Journal of Economic Theory* 4 (1):45–57.

Grossman, Zachary and Joel J Van Der Weele. 2017. "Self-image and willful ignorance in social decisions." *Journal of the European Economic Association* 15 (1):173–217.

Gul, Faruk. 1991. "A theory of disappointment aversion." *Econometrica* 59 (3):667–686.

Gul, Faruk and Wolfgang Pesendorfer. 2001. "Temptation and self-control." *Econometrica* 69 (6):1403–1435.

Krawczyk, Michal and Fabrice Le Lec. 2010. "Give me a chance! An experiment in social decision under risk." *Experimental Economics* 13 (4):500–511.

Larson, Tara and C Monica Capra. 2009. "Exploiting moral wiggle room: Illusory preference for fairness? A comment." *Judgment and Decision Making* 4 (6):467–474.

Lewis, Alan, Alexander Bardis, Chloe Flint, Claire Mason, Natalya Smith, Charlotte Tickle, and Jennifer Zinser. 2012. "Drawing the line somewhere: An experimental study of moral compromise." *Journal of Economic Psychology* 33 (4):718–725.

Lin, Stephanie C, Julian J Zlatev, and Dale T Miller. 2017. "Moral traps: When self-serving attributions backfire in prosocial behavior." *Journal of Experimental Social Psychology* 70:198–203.

Matthey, Astrid and Tobias Regner. 2011. "Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior." *Games* 2 (1):114–135.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The dishonesty of honest people: A theory of self-concept maintenance." *Journal of marketing research* 45 (6):633–644.

Mengarelli, Flavia, Laura Moretti, Valeria Faralla, Philippe Vindras, and Angela Sirigu. 2014. "Economic decisions for others: An exception to loss aversion law." *PLoS One* 9 (1):e85042.

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2011. "Managing self-confidence: Theory and experimental evidence." Tech. rep., National Bureau of Economic Research.

Pillutla, Madan M and J Keith Murnighan. 1995. "Being fair or appearing fair: Strategic behavior in ultimatum bargaining." *Academy of Management Journal* 38 (5):1408–1426.

Pollmann, Monique MH, Jan Potters, and Stefan T Trautmann. 2014. "Risk taking by agents: The role of ex-ante and ex-post accountability." *Economics Letters* 123 (3):387–390.

Polman, Evan and Kaiyang Wu. 2019. "Decision making for others involving risk: A review and meta-analysis." *Journal of Economic Psychology* 47:813–816.

Rodriguez-Lara, Ismael and Luis Moreno-Garrido. 2012. "Self-interest and fairness: self-serving choices of justice principles." *Experimental Economics* 15 (1):158–175.

Saito, Kota. 2013. "Social preferences under risk: Equality of opportunity versus equality of outcome." *American Economic Review* 103 (7):3084–3101.

———. 2015. "Impure altruism and impure selfishness." *Journal of Economic Theory* 158:336–370.

Schweitzer, Maurice E and Christopher K Hsee. 2002. "Stretching the truth: Elastic justification and motivated communication of uncertain information." *Journal of Risk and Uncertainty* 25 (2):185–201.

Shalvi, Shaul, Jason Dana, Michel JJ Handgraaf, and Carsten KW De Dreu. 2011. "Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior." *Organizational Behavior and Human Decision Processes* 115 (2):181–190.

Wiltermuth, Scott S. 2011. "Cheating more when the spoils are split." *Organizational Behavior and Human Decision Processes* 115 (2):157–168.