amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Morality Meets Risk: What Makes a Good Excuse for Selfishness

Wanxin Dong
Jiakun Zheng

WP 2025 - Nr 22

Morality Meets Risk: What Makes a Good Excuse for Selfishness

WANXIN DONG
School of Finance, Renmin University of China, Haidian District, 100872, Beijing, China

JIAKUN ZHENG
Aix Marseille Univ, CNRS, AMSE, Centrale Méditerranée, Marseille, France

Prior work finds that individuals are often less prosocial when they can exploit uncertainty as an excuse. In contrast to prior work that largely explores the relevance of excuses in the gain domain, this paper investigates the relevance of excuses in both the loss and gain domains. In our laboratory experiment, participants evaluated risky payoffs for themselves and their partners in either the gain or loss domain, with or without interpersonal trade-offs. We found that participants exhibited excuse-driven risk behaviors in both domains. We also documented significant individual heterogeneity in the degree of excuses, influenced by factors such as individuals' risk preferences, beliefs about others' risk preferences, and the size of the risk. We present a self-signaling model that incorporates self-image concerns to explain our experimental findings. We show that excuse-driven risk behavior arises because people misattribute their selfish behavior to risk preferences rather than a reduced level of altruism.

KEYWORDS. Prosocial behavior, Risk preferences, Self-image, Misattribution, Experiment.

JEL Codes. D71, D80, D91.

Wanxin Dong: dongwanxin@ruc.edu.cn
Jiakun Zheng : jiakun.zheng@outlook.com

1. Introduction

Are people born moral? A substantial body of economic literature documents that people often engage in morally right actions, such as volunteer work (Finkelstien, 2009, Kroll and Vogel, 2018), voluntary contributions to public goods (Dickinson, 1998, Zelmer, 2003), or charitable giving (Agerström et al., 2016, Dannenberg and Martinsson, 2021). Early studies suggest that individuals may experience a sense of subjective well-being when acting morally (Andreoni and Miller, 2002, Gebauer et al., 2008). A natural way to model such behavior is to include a moral argument in the utility function, where agents derive utility from acting morally (Andreoni, 1990, Fehr and Schmidt, 1999, Gibson et al., 2013). However, recent research challenges this perspective, arguing instead that people may not inherently be moral but rather strive to feel moral (e.g., Gino et al., 2016). In particular, decision contexts matter for the observed level of morality. When decisions arise in settings with sufficient flexibility, individuals will take advantage of plausible excuses to justify their selfish behavior and prioritize self-interest over the welfare of others (Gino et al., 2016, Engel and Szech, 2020). A classic example is the moral wiggle room studied by Dana et al. (2007). Agents avoid feedback about adverse consequences of their actions for others, i.e., willful ignorance, to avoid moral condemnation or punishment (Bartling et al., 2014, Grossman and van der Weele, 2017).

More recent literature highlights that uncertainty, either risk or ambiguity, can also serve as an excuse for selfish behavior in charitable contexts while maintaining a sense of moral integrity (Exley, 2016, Garcia et al., 2020). Specifically, Exley (2016) demonstrates that individuals exhibit excuse-driven risk behavior that deters charitable giving. She shows that participants misjudge the likelihood of self risk and charity risk when their decisions involve trade-offs between personal financial gain and charitable contributions. Participants tend to become more averse to charity risk while being less averse to self risk. Such distorted risk preferences bias individuals toward maximizing self-interest, as though they are hindered by cognitive limitations (Exley and Kessler, 2024b).

However, to date, research on excuse-driven risk behavior has predominantly focused on contexts where individuals' decisions could positively impact others' well-being, i.e., how much individuals are willing to benefit others under gain risk. In contrast, many real cases involve how much one is willing to endure self-losses without compromising others' interests, i.e., social decisions in the loss domain. Examples are pervasive, ranging from significant events like the initiation of wars—whether cold or hot—to collective ignorance about catastrophic climate risks imposed on future generations, as well as the negative externalities we impose on others in our daily lives. In such situations involving losses, agents may also use risk to act egoistically, a topic that has received limited attention in the existing literature. In this paper, we replicate Exley (2016)'s study in the gain domain and expand the research to the loss domain. Moreover, we delve into individual heterogeneity in responses to various risks and examine the factors determining the degree of excuses.

Our experimental design follows closely that of Exley (2016). We implement a between-subject design, randomly assigning participants to either the gain or loss treatment. Participants first complete a normalization price list, where they make binary choices between receiving (or losing) a ten-yuan amount themselves and their partners receiving (or losing) a certain amount in the gain (loss) treatment. From these choices, we elicit a certainty-equivalent amount such that participants are indifferent between receiving (or losing) ten yuan themselves and the corresponding amount for their partners. We then use 10 yuan and the certainty-equivalent amount as the nonzero outcomes for the self lotteries and the others lotteries, respectively. In the subsequent valuation price lists, participants make binary decisions between risky lotteries and riskless amounts for either themselves or their partners in each treatment. Two decision contexts can be distinguished, depending on whether a self–others trade-off is present. In the no self-others trade-off context, participants choose between self lotteries and self-certain amounts, or between others lotteries and others-certain amounts. In the self-others trade-off context, participants face trade-offs between self lotteries and others-certain amounts, or between others lotteries and self-certain amounts. In this way, we measure certainty equivalents for self and others lotteries in terms of both self valuations and others valuations. For comparability, lottery valuations are expressed as percentages of the corresponding baseline amounts, which allows us to interpret them as decision weights in lottery evaluation. Differences in these decision weights across contexts then indicate the degree of excuse-driven risk behavior.

Our results provide significant evidence for excuse-driven risk behavior in both the gain and loss domains. In the no self-others trade-off context, excuses for selfishness are irrelevant. Participants have nearly indistinguishable responses to risk in self lotteries and others lotteries for both domains. In addition, in the self-others trade-off context, our findings are consistent with Exley (2016) in the gain domain. Participants become more averse to others' risks and more seeking for their own risks compared to the no self-others trade-off context. That is, they appear to overestimate the probability that others' lotteries yield zero-yuan payoffs and that self lotteries yield nonzero gains, using this as an excuse to pursue their own material interests in the gain domain. However, this pattern is reversed in the loss domain, resulting in a reflection effect in excuse-driven risk behavior. Participants behave as if they are more averse to self-loss risk and more seeking to others-loss risk. They overweight the probability that others lotteries lose nothing and that self lotteries lose non-zero amounts, using it as an excuse to avoid the potential losses for themselves.

Despite the existence of excuse-driven behavior in both the gain and loss domains, we find that the degree of excuse-driven behavior is subject to the size of the risk and who bears the risk. When the risk affects themselves, participants always exhibit stronger excuse-driven behavior under small and moderate probabilities. When the risk falls on others, participants exhibit a marked tendency for excuse-driven risk behavior under high probabilities. We conjecture that it may depend on individuals' risk preferences and their beliefs about others' risk preferences. Additional results verify that participants whose risk preferences are contrary to the direction of excuses exhibit a greater degree of excuses. Specifically, risk-averse participants are more likely to use self-gain risk and others-loss risk as an excuse while risk-seeking ones prefer to take advantage of others-gain risk and self-loss risk as an excuse for selfishness. However, there is a limit to excuse-driven risk behavior. Participants whose risk preferences deviate from common trends even exhibit more altruistic behavior under risk. For instance, participants seeking self-gain risk or avoiding self-loss risk exhibit more altruistic behavior under high probabilities. Similarly, participants who are averse to others-gain risk or loving to others-loss risk may prioritize prosocial actions at the expense of their own interests under small probabilities.

To shed light on the mechanisms driving these behaviors, we propose a self-signaling model in which the agent, motivated by self-image concerns, engages in self-serving actions (Kunda, 1987, 1990). This model builds on the work of Bénabou and Tirole (2006) and extends the framework developed by Grossman and van der Weele (2017) to explain willful ignorance in social decision-making. The agent comprises two selves: a decision-maker self and an observer self. Each agent is characterized by a preference type, including the degree of altruism, self-image concerns and risk preferences. The decision-maker self is aware of her true preferences and makes choices to balance various motives, including self-payoff, intrinsic altruism, and self-image. The observer self lacks knowledge of her preferences and the true motivations underlying the behavior. She can only update her beliefs about her preferences based on the chosen action in a given signaling environment, which affects the utility derived from her self-image.

Aligned with our experimental setup, the agent in our model makes binary choices between returns to themselves or to others across various decision environments, whether risky or risk-free. In a risk-free environment, the observer self only updates her beliefs about the decision-maker's altruism based on her actions. Due to additional self-image concerns, the agent behaves more prosocially compared to scenarios where only pure altruism is at play. However, in a risky environment, two significant changes occur. First, the trade-off between self-image and the utility derived from self-payoff versus others' payoff through altruism is altered. Specifically, it becomes less costly to maintain a high-standard image since payoffs are now uncertain. We refer to this as the pure risk channel, which results in more altruistic behavior in a risky environment compared to a riskless environment. Second, because the observer self lacks precise information about her preference type, she may mistakenly attribute selfish behavior to risk preferences rather than a reduced level of altruism, creating room for excuse-driven risk behavior. However, this misattribution channel is not unlimited and depends on the decision context and the true risk preferences of the decision-maker self. Consider, for instance, the valuation tasks in the self-others trade-off context, where others encounter risks in the domain of gains. It is implausible to attribute selfish behavior to risk aversion in scenarios where ordinary individuals are typically risk-seeking—that is, when the risks are small (e.g., Tversky and Kahneman, 1992). Similarly, individuals who are already quite risk-averse may have less room for selfishness compared to those who are not. Overall, we demonstrate that incorporating self-image concerns into the agent's objective effectively reconciles with our experimental results.

Our contribution to the existing literature is twofold. Firstly, our study adds to the extensive body of work on moral reasoning, with a specific focus on how uncertainty shapes morality in the loss domain. Existing research has shown that individuals are adept at justifying immoral behavior through information processing (e.g., Dana et al., 2006), belief distortions (e.g., Tella et al., 2015), and misattribution to innocuous preferences (e.g. Exley, 2016, Garcia et al., 2020). Among these studies, Exley (2016) and Garcia et al. (2020) are the closest to ours. Both studies found that uncertainty, risk or ambiguity, can serve as an excuse for selfishness in the context of charitable donations. Our results in the gain domain successfully replicate their findings in a Chinese context, which may differ from western contexts in various aspects. Beyond that, we extend the research on excuse-driven risk behavior to the domain of losses, as reflected in issues such as public bads (Pace and van der Weele, 2020), common resource extractions (Hopfensitz et al., 2019), and various negative externalities (Rovira et al., 2000). Our study complements the existing literature by confirming the adverse impact of risk on prosocial behavior in social decisions involving losses. Moreover, using a well-defined notion of the degree of excuses, we find that the effects of risk on morality vary widely depending on the decision environment, individuals' risk preferences, and their beliefs about others' risk preferences. This heterogeneity offers a fresh perspective on existing findings and provides key insights into the mechanisms underlying excuse-driven risk behavior.

Secondly, our study contributes to the literature by presenting a self-signaling theory with self-image concerns to explain the observed heterogeneity in excuse-driven risk behavior. The classical approach to modeling prosocial behavior typically involved adding a preference term that is not belief-based to the objective of the decision-maker, often driven by factors like warm-glow (Andreoni, 1990), social pressure (DellaVigna et al., 2012), or social norms (Elster, 1989, Krueger et al., 2008). However, this approach sometimes fails due to its lack of flexibility in accounting for the information environment in which the decision-maker is situated. In line with Grossman and van der Weele (2017), we assume that the agent is not only inherently altruistic but also cares about her self-image, which is measured by how the observer self believes she is altruistic. Although our model shares similarities with that of Grossman and van der Weele (2017), the underlying mechanism is quite different. To explain the concept of moral wiggle room in Dana et al. (2007), Grossman and van der Weele (2017) demonstrated that by concealing relevant information, the observer self cannot fully discern the true intentions of the decision-maker, thereby introducing noise into belief updating. In contrast, in our model, due to a lack of information about the decision maker's preference type, the observer self must update beliefs about both altruism and risk preferences. This joint learning process creates opportunities for misattributing selfish behavior to innocuous risk preferences rather than to a lower level of altruism. Moreover, this misattribution is constrained by the observer's beliefs about the true preferences of the decision-maker. Our theory, aligned with Grossman and van der Weele (2017), also suggests that greater decision transparency could improve prosocial behavior. In particular, we have shown that if the true risk preferences of either oneself or others, depending on the decision environment under consideration, could be revealed, risk would always enhance prosocial behavior by reducing the costs of maintaining a high-standard self-image.

This paper is structured as follows. Section 2 outlines the experimental framework and predictions. Section 3 provides details on the experimental design and its implementation. Section 4 presents our main results, while Section 5 introduces a self-signaling model to explain these findings. Finally, Section 6 concludes.

2. Decision Frameworks and Predictions

2.1 *Decision Frameworks*

In this section, we present the decision frameworks that participants encounter in the experiment. Decisions can be made in different domains, either the gain domain or the loss domain, denoted by d with $d \in \{g, l\}$. Specifically, in the gain domain, participants face similar types of decisions as elaborated in Exley (2016) and can earn additional money on top of their initial endowment. Decisions in the loss domain are simply the flip side of those in the gain domain, and participants can only lose money, deducted from their initial endowment. These initial endowments are supposed to be the reference point for the subjects in different domains (e.g., Tversky and Kahneman, 1992). For consistency and comparability, we set different endowments in the two domains such that subjects face the same decisions, i.e., 10 yuan in the gain domain and 20 yuan in the loss domain.[1] This also generates the same expected payoffs across domains.

For ease of comparison, we closely follow the notations and derivations of Exley (2016). Since all amounts used in our experiment are expressed in Yuan, we omit the unit in our descriptions for simplicity. Participants make a series of binary decisions between riskless amounts and risky lotteries for themselves or for their partners, who are randomly and anonymously selected from other participants. For simplicity, these two roles are denoted by i with s for the participants ("self") and o for the partners ("others"), i.e., $i \in \{s, o\}$. A riskless amount produces a non-zero gain or loss with certainty given to the participants or their partners, denoted by C_d^i with $d \in \{g, l\}$ and $i \in \{s, o\}$. It can be a "self-gain-certain amount", a "self-loss-certain amount", an "others-gain-certain amount" or an "others-loss-certain amount". A risky lottery produces a non-zero gain or loss with probability P and 0 with probability $1 - P$, denoted by P_d^i with $d \in \{g, l\}$, $i \in \{s, o\}$ and $P \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$. Participants themselves receive the outcomes of "self-gain lotteries" or "self-loss lotteries"; their partners receive the outcomes of "others-gain lotteries" or "others-loss lotteries".

A "self lottery" generates a gain or loss of 10 with probability P and 0 with probability $1-P$ for a participant in the corresponding domain. An "others lottery" generates a gain or loss of X_d with probability P and 0 with probability $1 - P$ for a partner in the corresponding domain. X_d is defined as the others-domain-specific values such that participants are indifferent between themselves receiving (or losing) a ten-yuan amount and their partners receiving (or losing) an amount denoted by X_d. We will elicit X_d values for each subject to ensure comparability between others lotteries and self lotteries. Intuitively, we can use the allocation ratio $10/X_d$ to measure participants' level of altruism or prosocial motivation in a risk-free context. It normally lies between 0 and 1. In particular, when $10/X_d$ equates to one, it is a fair decision; when it is less than 1, it means that subjects put a higher weight on self utility than on others. Under the assumption of monotonicity in preferences, X_d is unique if it exists.

[1] In particular, when looking at aggregated payoffs, lotteries in the gain domain (i.e., $10 + (0, P; 10, 1 - P)$) are equivalent to those in the loss domain (i.e., $20 + (-10, P; 0, 1 - P)$).

	Self valuation	Others valuation
Self lottery	$C_d^s(P_d^s)$	$C_d^o(P_d^s)$
Others lottery	$C_d^s(P_d^o)$	$C_d^o(P_d^o)$

Table 1. Four Types of Lottery Valuations in the Gain $(d = g)$ and loss $(d = l)$ Domain

Let a bundle (W^s, W^o) denote an allocation where the participant receives W^s and a partner receives W^o. Self and others lotteries are defined as follows:

$$P_g^s = P(10,0) + (1-P)(0,0), \quad P_g^o = P(0,X_g) + (1-P)(0,0); \tag{1}$$

$$P_l^s = P(-10,0) + (1-P)(0,0), \quad P_l^o = P(0,-X_l) + (1-P)(0,0). \tag{2}$$

Here, $P \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$, and X_g and X_l are others-domain-specific values satisfying

$$(10,0) \sim (0,X_g), \quad (-10,0) \sim (0,-X_l). \tag{3}$$

In the experiment, we measure certainty equivalents for each type of lotteries in terms of self valuations and others valuations. More specifically, a self valuation of lotteries, denoted by $C_d^s(P_d^i)$, indicates that subjects are indifferent between the lottery P_d^i and the self-certain amount $C_d^s(P_d^i)$ that they receive. Likewise, an others valuation of lotteries, denoted by $C_d^o(P_d^i)$, indicates that subjects are indifferent between the lottery P_d^i and the others-certain amount $C_d^o(P_d^i)$ that their partners receive.

In short, there are two types of lotteries (self lotteries and others lotteries) and two types of valuations (self valuations and others valuations) in each domain. This results in four types of lottery valuations, {self-gain lottery, others-gain lottery} × {self-gain valuation, others-gain valuation} and {self-loss lottery, others-loss lottery} × {self-loss valuation, others-loss valuation}, as summarized in Table 1. To standardize the denomination for these valuations, we refer to self lotteries in self valuations $(C_d^s(P_d^s))$ and others lotteries in others valuations $(C_d^o(P_d^o))$ as lottery valuations in the no self-others trade-off context. In contrast, we refer to self lotteries in others valuations $(C_d^o(P_d^s))$ and others lotteries in self valuations $(C_d^s(P_d^o))$ as lottery valuations in the self-others trade-off context.

Notice that the valuations $(C_d^j(P_d^i))$ are positive in the gain domain and negative in the loss domain. Moreover, these valuations are elicited in different units for oneself and one's partner. For ease of comparison, we employ a two-step procedure. Firstly, we represent these certainty equivalents with absolute values to ensure that they are always positive numbers. Then, we rescale the certainty equivalents of lotteries into the same unit, expressed as a percentage of the corresponding baseline amounts. Specifically, self valuations $(C_d^s(P_d^i))$ are scaled as percentages of self-baseline amounts of 10 yuan. Others valuations $(C_d^o(P_d^i))$ are scaled as percentages of partners' baseline amounts of X_d yuan. For simplicity, we replace the above ratios with $Y_d^s(P_d^i)$ and $Y_d^o(P_d^i)$. Formally,

$$Y_d^s(P_d^i) = \frac{|C_d^s(P_d^i)|}{10}, \quad Y_d^o(P_d^i) = \frac{|C_d^o(P_d^i)|}{X_d}, \quad \text{with } d \in \{g,l\}, i \in \{s,o\}.$$

Same as in Exley (2016), we rescale the valuations based on the assumption of linear utilities in payoffs. After rescaling these valuations, $Y_d^j(P_d^i)$ indicates the decision weights on payoffs based on objective probabilities, as demonstrated in cumulative prospect theory (Tversky and Kahneman, 1992). Specifically, in the no self-others trade-off context, $Y_d^s(P_d^s)$ captures the risk preferences of participants themselves. $Y_d^o(P_d^o)$ reflects participants' beliefs about the risk preferences of their partners. As such, valuations for different lotteries are influenced by probability weighting instead of utility curvature. Then, we define the degree of excuse-driven risk behavior as the difference in decision weights when evaluating the same lottery, i.e., $Y_d^o(P_d^i) - Y_d^s(P_d^i)$ with $d \in \{g, l\}$ and $i \in \{s, o\}$. A larger deviation indicates stronger excuse-driven risk behavior.

2.2 *Predictions*

According to the independence axiom, Equation (1)-(3) jointly imply that participants should be indifferent between self lotteries and others lotteries. That is,

$$P_d^s \sim P_d^o, \quad d \in \{g, l\} \text{ for the same probability } P. \tag{4}$$

Recall that we use the allocation ratio $10/X_d$ to represent participants' altruism levels in a risk-free context. Similarly, in situations involving risk, we can employ the allocation ratio of self and others' certainty equivalents for the same lotteries to measure altruism levels, i.e., $C_d^s(P_d^i)/C_d^o(P_d^i)$. Assuming linear utilities in payoffs and standard risk preferences, the altruism levels across various risk situations satisfy the equation: $C_d^s(P_d^i)/C_d^o(P_d^i) = 10/X_d$. That is, $Y_d^s(P_d^i)/Y_d^o(P_d^i) = 1$, indicating equal decision weights under risk. Following Exley (2016), the following prediction serves as the null hypothesis that levels of altruism are not influenced by the presence of risk.

PREDICTION 1 (Standard Risk Preferences). *All else equal, if individuals have standard risk preferences for some probability P, then:*

$$Y_d^o(P_d^s) = Y_d^s(P_d^s) = Y_d^o(P_d^o) = Y_d^s(P_d^o), \quad d \in \{g, l\}. \tag{5}$$

However, Exley (2016) rejected this null hypothesis and documented the presence of excuse-driven risk behavior in the gain domain. Individuals may evaluate the same lottery differently depending on whether a trade-off between self and others' payoffs is involved. In the no self–others trade-off context, excuse-driven behavior is not relevant. By contrast, in the self–others trade-off context, participants may underweight the probability that others' lotteries yield positive outcomes as an excuse to favor self-certain amounts, or overweight the probability that self lotteries yield positive outcomes as an excuse to reject others-certain amounts. Participants thus become more risk-seeking toward self risk and more risk-averse toward others' risk. This is summarized in Prediction 2.

PREDICTION 2 (Excuse-Driven Risk Behavior in the Gain Domain). *All else equal, if individuals exhibit excuse-driven risk behavior for some probability P in the gain domain, then:*

$$Y_g^o(P_g^s) > Y_g^s(P_g^s) = Y_g^o(P_g^o) > Y_g^s(P_g^o). \tag{6}$$

Building on Predictions 1 and 2, we extend the analysis to the loss domain. Since many experimental studies have documented that individuals are typically risk-averse for gains but risk-seeking for losses, i.e., the reflection effect (e.g. Tversky and Kahneman, 1992, Bleichrodt and van Bruggen, 2022), we conjecture that excuse-driven risk behavior will reverse when gains are replaced by losses. Specifically, when participants choose between others-loss lotteries and self-loss certain amounts, they may display greater risk-seeking toward others' lotteries (i.e., assigning a lower decision weight to the negative outcome) to avoid losses for themselves. Conversely, when choosing between self-loss lotteries and others-loss-certain amounts, they may exhibit greater risk aversion toward self lotteries (i.e., assigning a higher decision weight to the negative outcome), using the potential losses to themselves as a justification for their choices. This leads to Prediction 3.

PREDICTION 3 (Excuse-Driven Risk Behavior in the Loss Domain). *All else equal, if individuals exhibit excuse-driven risk behavior for some probability P in the loss domain, then:*

$$Y_l^o(P_l^s) > Y_l^s(P_l^s) = Y_l^o(P_l^o) > Y_l^s(P_l^o). \tag{7}$$

3. EXPERIMENTAL DESIGN

In this section, we describe the experimental design of our main experiment. The experiment utilized a between-subject design consisting of two treatment conditions, gain and loss. Each treatment included 30 multiple price lists, comprising a normalization price list, a buffer price list, and 28 valuation price lists. Each price list involves a series of binary decisions. The complete instructions can be found in Appendix A. In the following subsections, we describe each part of the experiment in detail.

3.1 *Normalization Price Lists*

The first price list serves as a normalization task that determines others-domain-specific X_d values such that participants are indifferent between themselves receiving (or losing) a ten-yuan amount and their partners receiving (or losing) an amount of X_d. Then, we use these 10 self values and their X_d others' values as the non-zero baseline amounts in the subsequent valuation price lists. Note that participants are uninformed about this arrangement. Instead, they are just asked to make 16 binary decisions between option A and option B. In the gain domain, option A columns always pay 10 yuan to the participant. As proceeding down the rows of the list, option B columns always pay an increasing amount from 0 to 30 yuan by increments of 2 yuan to the participant's partner. By contrast, in the loss domain, option A columns always involve the partner losing increasing amounts from 0 to 30 yuan by increments of 2 yuan. Option B columns always involve the participant losing 10 yuan with certainty. This opposite design of options in the two domains ensures that most people start by selecting option A and then switch to option B.

Following Exley (2016), we evaluate others-domain-specific X_d values as follows. Suppose that a participant switches from option A to option B on the i^{th} row in the gain domain, and their partners receive corresponding B_i. It reveals that others-domain-specific X_g values fall into the range from B_{i-1} to B_i (i.e., $B_{i-1} < X_g < B_i$). We estimate X_g values as its upper bound of B_i. Similarly, assume a participant switches from option A to option B on the j^{th} row in the loss domain. It implies that the participant is willing to endure 10 yuan losses themselves to prevent at least A_j losses for the partner (i.e., $A_{j-1} < X_l < A_j$). We also estimate X_l values as its upper bound of A_j.[2]

In our data, there is only one case where X_d cannot be accurately estimated, i.e., censored X_d values. It occurs if some selfish participants never choose option B in either treatment. Our main results exclude those participants with censored X_d values, while our robustness analysis includes this group by assuming their X_d values as the lower bound of 30. Unlike Exley (2016), the inaccurately estimated case does not involve multiple switching points. We utilize one switching-point elicitation method in all price-list tasks rather than binary comparisons for each choice. Participants do not need to click on all options separately in each price list. Instead, they determine the switching point where they intend to alter their choices. The software then automatically generates the option results based on their switching points.[3]

After finishing the normalization price list, participants should complete the second price list served as a buffer between the normalization task and valuation tasks. The buffer task is similar to the normalization one. The only difference is that the self-baseline amount changes from 10 yuan to 5 yuan.

3.2 *Valuation Price Lists*

After the initial two price lists, participants in each treatment encounter 28 valuation price lists to elicit their lottery valuations. As described in Section 2.1, there are four blocks, each with seven price lists. The price lists within each block only differ in probabilities and are ordered sequentially by increasing probabilities P, i.e., $P \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$. Four blocks in each treatment are presented in random order at the individual level. Participants complete all price lists in one block, then another block.

[2] We select the upper bound because overestimating X_d biases our results against finding evidence for excuse-driven risk behavior. Others valuations ($C^o_d(P^i_d)$) are scaled as percentages of partners' baseline amounts of X_d. Therefore, overestimations of X_d lead to underestimations of rescaled others valuations. Excuse-driven risk behavior implies that participants overweight the probability P of self lotteries or underweight the probability P of others lotteries in the self-others trade-off context. Thus, excuse-driven risk behavior would be underestimated.

[3] In our experiment, we enforce one single switching point with a monotonic preference restriction. Although participants may exhibit more than one switching point, Exley (2016) suggests that only 1% of participants have demonstrated multiple switching points, which is significantly lower than the 11% observed in Garcia et al. (2020). One single switching point ensures the consistency of our main results. Moreover, if participants can choose multiple switching points, they may employ various excuses to rationalize their selfish choices. Thus, one single switching point can avoid the influence of additional channels and allow us to focus on excuse-driven risk behavior.

Participants complete 21 binary decisions between option A and option B in each price list. In the gain domain, option A columns always involve a fixed self-gain lottery or others-gain lottery. Option B columns always involve a self-certain amount or an others-certain amount. Self-certain amounts range from 0 to 10 yuan by increments of 0.5, while others-certain amounts range from 0 to X_d by increments of $X_d/20$. In the loss domain, we swap the lotteries in option A with certain amounts in option B, and change the expressions from gains to losses. This setting ensures that participants always begin by preferring option A and then switch to option B.

We estimate participants' lottery valuations as follows. Suppose participants always switch from option A to option B on the i^{th} row. Since the certain amount always increases as proceeding down the rows, the certainty equivalent must lie in the interval of B_{i-1} and B_i in the gain domain, or the interval of A_{i-1} and A_i in the loss domain. We evaluate their valuations as the middle point of the interval, i.e., $(B_{i-1} + B_i)/2$ in the gain domain and $(A_{i-1} + A_i)/2$ in the loss domain.[4] As with the normalization price lists, there are also censoring problems. We estimate those selfish individuals' lottery valuations as the lower bound of 10 or X_d.

3.3 *Experimental Implementation*

The experiment was preregistered in the AEA-RCT (ID: AEARCTR-0008697) and conducted with the o-Tree program in March 2022 at the School of Finance, Renmin University of China. We recruited 86 full-time students for the gain treatment and 92 subjects for the loss treatment via the Yanzhonglab platform. Subjects were not informed about the tasks that they would be asked to complete before the experimental session. Each session took an average of 30 minutes. The average payoffs were 29.35 yuan in the gain treatment and 31.27 yuan in the loss treatment.

Upon arrival, subjects were randomly assigned a seat in the lab. In the preparation phase, we read aloud and explained the experimental instructions as simply as possible. Subjects were informed to participate in a two-person matched experiment where we paired each subject randomly with another. Throughout the experiment, both partners remained anonymous without any communication. After reading the instructions, participants were asked to answer some comprehension questions to ensure their understanding of the tasks. Once they answered all the questions correctly, participants began completing the main experiment, which included 30 price lists, three moral wiggle room questions (Dana et al., 2007),[5] and an anonymous questionnaire containing standard socio-demographic questions.

When both partners completed all the tasks, the o-Tree program calculated their payoffs based on the following rules. In the price lists phase, each participant received

[4] For our previous estimation of X_d, we choose the upper bound rather than the midpoint since participants face the estimated values of X_d in the study. However, when it comes to lottery valuations, different estimations are feasible. Specially, the midpoint estimations would yield more accurate results.

[5] Three moral wiggle room questions include a choice-to-reveal question, a revealed-unaligned-state question, and a revealed-aligned-state question. The average payoff of this part is 5 yuan in both treatments.

an initial endowment of 10 yuan in the gain treatment and 20 yuan in the loss treatment. The program randomly selected one of the two matching partners and one of the decisions from the chosen partner to implement extra payments given to both partners. In the moral wiggle room phase, each participant received an endowment of 0 yuan in the gain treatment and 8 yuan in the loss treatment. Similarly, the program randomly selected one of the questions from the chosen partner to implement the additional payment. In the final questionnaire phase, participants were paid 10 yuan. The average total payoff amounted to 30 yuan in both treatments.

4. Results

Section 4.1 presents the results on participants' levels of altruism in the normalization tasks without risk. Section 4.2 presents evidence of excuse-driven risk behavior in both treatments, implying a reflection effect. Section 4.3 considers the heterogeneous degree of excuses across different risk levels at the individual level.

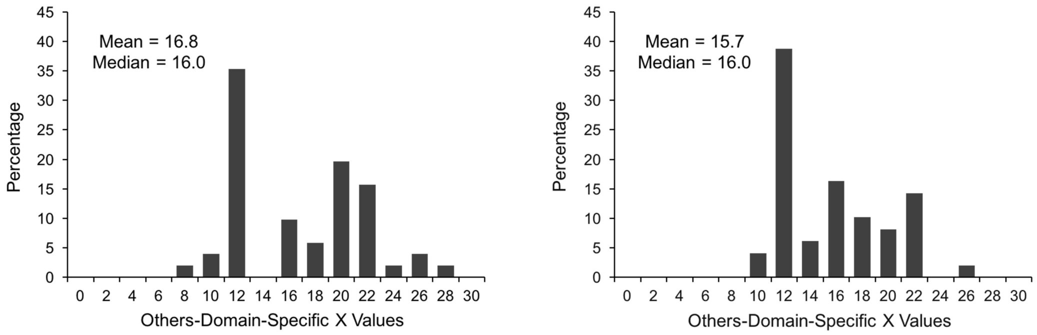
4.1 *Altruistic Behavior in the Absence of Risk*

In our normalization price list tasks, approximately 40% of subjects have censored X_d values. Specifically, among the 86 subjects in the gain domain, 34 were unwilling to give up 10 yuan for themselves to earn at least 30 yuan for their partners. Similarly, among the 92 subjects in the loss domain, 37 were reluctant to accept a loss of 10 yuan for themselves to prevent their partners from a loss of at least 30 yuan. Our main results exclude those with censored X_d values to ensure that the results are still relevant among an altruistic group less prone to making self-excusing decisions.[6] It is therefore a more conservative test of excuse-driven risk behavior. We further exclude participants who failed the internal consistency check: one in the gain domain and two in the loss domain. Consistent participants typically showed monotonic or near-monotonic changes in decision weights as probabilities increased from 5% to 95%, whereas the excluded participant exhibited multiple reversals or near-random responses, suggesting random choices or misunderstanding. In the loss treatment, we also exclude four participants with $X_l = 2$ yuan. The latter displayed extreme prosociality, preferring to incur a loss of 10 to prevent others from losing at most 2 yuan. Including these subjects could underestimate our results. This leaves 51 participants in the gain treatment and 49 in the loss treatment. Unless otherwise noted, our analysis focuses on this remaining sample.

Figure 1a and Figure 1b present the distribution of X_g and X_l for the remaining sample, respectively. In the gain domain, the average X_g value is 16.8 and lower than the average of 18.5 reported in Exley (2016), indicating that our participants are more generous in other-regarding decisions.[7] More than 94% of these participants evaluate X_g values exceeding 10 yuan, i.e., they are only willing to sacrifice 10 yuan for themselves in exchange for their partners receiving more than 10 yuan. In the loss domain, the average

[6] The magnitude of X_d values reflects the extent of individuals' selfishness. Those with censored X_d values are thus more likely to exhibit self-centered behavior.

[7] All the numbers we set in our experiment are the same as Exley's but in different units.



(a) X_g Values in the Gain Domain

(b) X_l Values in the Loss Domain

Figure 1. Distribution of X_d Values for the Remaining Sample in the Two Domains

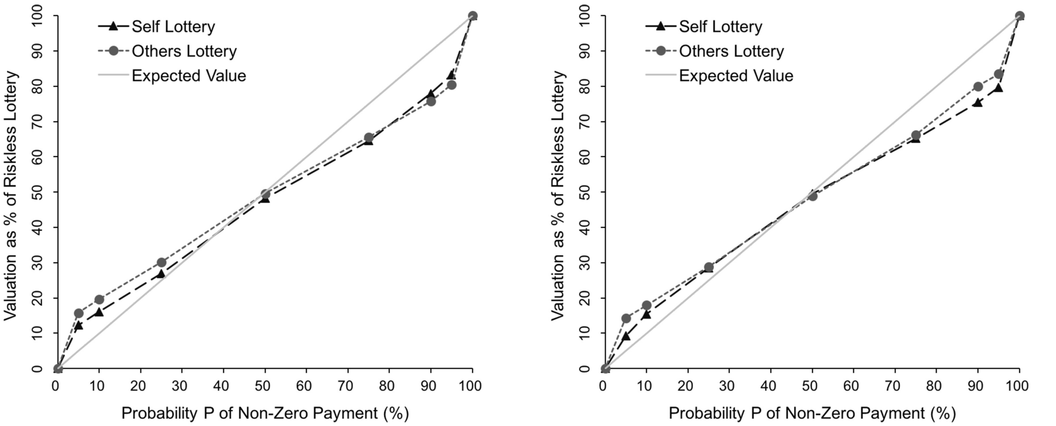
X_l value is 15.7 and slightly lower compared to the gain domain, but there is no significant difference between the two domains. 96% of the participants in the loss treatment are willing to accept a loss of 10 yuan for themselves only when it saves their partners from a loss larger than 10 yuan. Both domains have a median value of 16 yuan.

RESULT 1. *On average, participants value 10 yuan for themselves as equivalent to 16.8 yuan for partners in gains and 15.7 yuan in losses. There are no significant domain differences in altruism under riskless conditions.*

4.2 *Excuse-Driven Risk Behavior*

We now examine the existence of excuse-driven risk behavior in the gain domain and loss domain. Figure 2 and Figure 3 plot the mean values of self and others lottery valuations for each probability. The gray diagonal line represents a risk-neutral agent's lottery valuations under expected utility theory. Figure 2 shows that participants have nearly indistinguishable responses to risk in self lotteries and others lotteries for both the gain and loss domains in the no self-others trade-off context.[8] Concerning the participants' valuations of self risk and others' risk, it reveals an inverse S pattern of lottery valuations that is consistent with the probability weighting function in cumulative prospect theory (Tversky and Kahneman, 1992). Specifically, participants tend to overestimate small probabilities while underestimating moderate and high ones. This, in turn, results in risk-seeking behavior associated with high-risk gains and low-risk losses, as well as risk-averse behavior associated with low-risk gains and high-risk losses.

[8] These equivalent responses are partly due to the normalization design described in Section 2.1, since others-domain-specific values capture the differences in individuals' utility curvatures for self and others' money. When corresponding payoffs are normalized, excuses for selfishness solely come from the deviation of decision weights on risk, implying changes in altruism levels under different risk conditions. Figure B.1 in Appendix B also confirms that the differences between valuations for self lotteries and others lotteries are not statistically different at the 5% level in most cases.

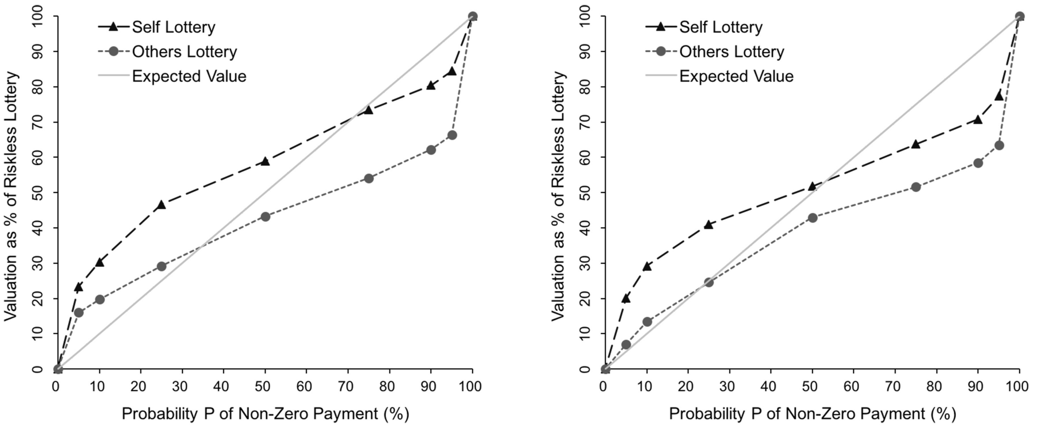


(a) Lottery Valuations in the Gain Domain

(b) Lottery Valuations in the Loss Domain

Figure 2. Lottery Valuations in the No Self-Others Trade-off Context

Figure notes: The estimates are valuations in the no self-others trade-off context. The self lottery indicates the valuations of $Y_d^s(P_d^s)$, while the others lottery indicates the valuations of $Y_d^o(P_d^o)$. Valuations in self-money are scaled as a percentage of 10 yuan. Valuations in others-money are scaled as a percentage of X_d yuan. The expected value indicates the expected lottery valuations scaled as a percentage of 10 yuan, i.e., a 45-degree line against probability P.



(a) Lottery Valuations in the Gain Domain

(b) Lottery Valuations in the Loss Domain

Figure 3. Lottery Valuations in the Self-Others Trade-off Context

Figure notes: The estimates are valuations in the self-others trade-off context. The self lottery indicates the valuations of $Y_d^o(P_d^s)$, while the others lottery indicates the valuations of $Y_d^s(P_d^o)$. Valuations in self-money are scaled as a percentage of 10 yuan. Valuations in others-money are scaled as a percentage of X_d yuan. The expected value indicates the expected lottery valuations scaled as a percentage of 10 yuan, i.e., a 45-degree line against probability P.

In contrast, Figure 3 reveals that there is a significant difference between participants' responses to self risk and others' risk in the self-others trade-off context. Consistent with Prediction 2 on excuse-driven risk behavior in the gain domain, participants act more averse to others' risk and less averse to self risk, as documented in Exley (2016). From $P = 0.95$ to $P = 0.05$, decision weights for self risk and others' risk diverge by 15 percentage points on average. In response to the 50% risk, participants reduce their valuations to 41% for self lotteries but increase them to 57% for others lotteries. As described in Prediction 3, we observe an opposite pattern in the loss domain. Participants become more seeking to others' risk and less seeking to self risk. Their responses to others' risk and self risk now diverge by only 13 percentage points on average. In response to the 50% risk, the difference between valuations of self and others lotteries is 8.8 percentage points, lower than 15.6 percentage points in the gain domain. The starkest differences occur with the introduction of high probabilities. From $P = 1$ to $P = 0.95$, decision weights for others' risk decrease by 36.5 percentage points, compared to only a 16.5 percentage point reduction in the no self-others trade-off context. Participants underweight the possibility that others lose non-zero payoffs, using it as an excuse to avoid self-losses.[9]

To provide more empirical evidence on excuse-driven risk behavior, consider the linear regressions with standard errors clustered at the individual level in Equation (8). The dependent variable Y_{pi} indicates an individual i's valuation of a particular lottery with probability p in the gain or loss domain (as a percentage of the corresponding baseline amount, i.e., 10 or X_d). The independent variables include two dummy variables and one interaction term of the two: $others_{pi}$ is equal to 1 for others lotteries, $tradeoff_{pi}$ is equal to 1 for valuations elicited in the self-others trade-off context, and $others * tradeoff_{pi}$ is the interaction term. We also include probability fixed effects (λ_p) or individual fixed effects (μ_i) in our regression model.

$$Y_{pi} = \beta_0 + \beta_1 others_{pi} + \beta_2 tradeoff_{pi} + \beta_3 others * tradeoff_{pi} + \Sigma_p \lambda_p + \Sigma_i \mu_i + \epsilon_{pi} \quad (8)$$

The coefficient of $others_{pi}$ captures the difference between self and others lottery valuations in the no self-others trade-off context. The coefficient of $tradeoff_{pi}$ captures the difference between self lottery valuations in the no self-others trade-off context and those in the self-others trade-off context. The sum of the coefficients of $tradeoff_{pi}$ and $others * tradeoff_{pi}$ thus indicates the difference between others lottery valuations in the no self-others trade-off context and those in the self-others trade-off context. Recall that the degree of excuse-driven risk behavior is defined as the gap between others valuation and self valuation of the same lottery (see more details in Section 2.1). Hence, the coefficients of $tradeoff_{pi}$ and $others * tradeoff_{pi}$ also indicate the degree of excuse-driven risk behavior.

Column 1 of Table 2 displays the corresponding regression results with no control variables in the gain domain. Note that the coefficient of $others_{pi}$ is not statistically different from zero, suggesting no significant differences in preferences toward

[9] Figures B.2–B.5 in Appendix B present excuse-driven risk behavior for each type of lottery valuation in both domains.

Regression:	Ordinary Least Squares			Tobit	
Dependent Variable:			Y_{pi}		
	1	2	3	4	5
others	1.11	1.11	1.11	5.18**	4.99***
	(0.65)	(0.63)	(0.64)	(2.76)	(2.74)
tradeoff	9.76***	9.76***	9.76***	25.70***	24.45***
	(4.09)	(4.02)	(4.09)	(6.82)	(6.69)
*others*tradeoff*	-16.36***	-16.36***	-16.36***	-44.07***	-42.66***
	(-3.71)	(-3.64)	(-3.70)	(-7.03)	(-6.96)
I(P=0.10)			4.61***		4.52**
			(4.15)		(4.96)
I(P=0.25)			16.40***		14.06***
			(11.14)		(12.04)
I(P=0.50)			33.19***		29.30***
			(15.16)		(17.39)
I(P=0.75)			47.61***		43.63***
			(16.36)		(19.68)
I(P=0.90)			57.23***		51.62***
			(17.77)		(20.26)
I(P=0.95)			61.81***		56.00***
			(18.70)		(21.44)
Constant	47.03***	62.50***	15.48***	46.80***	18.30***
	(33.86)	(65.32)	(6.50)	(41.26)	(10.14)
Ind FE	no	yes	no	no	no
Censored X	no	no	no	yes	yes
Observations	1428	1428	1428	2408	2408
Subjects	51	51	51	86	86

Table notes: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Standard errors are clustered at the individual level and shown in parentheses. The dependent variables Y_{pi} are lottery valuations in the gain domain, i.e., $Y^s_g(P^s_g)$, $Y^s_g(P^o_g)$, $Y^o_g(P^o_g)$, $Y^o_g(P^s_g)$. Probability fixed effects and individual fixed effects (Ind FE) are shown when included. "Censored X" indicates whether participants with censored X_g values are included or not. Tobit regressions include participants with censored values of X_g. We set their others-domain-specific values X_g with a right-censoring limit equal to 30.

Table 2. Regression Analysis on Excuse-Driven Risk Behavior in the Gain Domain

self risk and others' risk in the no self-others trade-off context. By contrast, the coefficient of $tradeoff_{pi}$ shows that participants' decision weights for the winning state of self lotteries are about 10 percentage points higher on average in the self-others trade-off context. Participants use self-gain risk as an excuse to choose more self lotteries over others-certain amounts. Additionally, the sum of the coefficients of $tradeoff_{pi}$ and $others * tradeoff_{pi}$ indicates that others-gain lotteries are valued lower by about 6.6 percentage points on average in the self-others trade-off context. Participants appear to overweight the possibility that others lotteries yield zero payoffs, using it as an excuse to choose more self-certain amounts. Columns 2-3 confirm the robustness of the results to the inclusion of individual fixed effects and probability fixed effects. Columns 4-5 include participants with censored values who are considered the most selfish in this study. The results are much more significant when considering Tobit regressions as a robustness check. Similar results have been documented in Exley (2016, 2020). This larger finding could, in part, result from the most selfish individuals being more excuse-driven.[10]

Column 1 of Table 3 displays the regression results with no control variables in the loss domain. Also, the coefficient of $others_{pi}$ is not significantly different from zero. Namely, decision-making for oneself and others is quite similar in the no self-others trade-off context. The coefficient of $tradeoff_{pi}$ shows that the average decision weights for the losing state of self lotteries increase by 4.4 percentage points when measured in terms of others valuations. The estimates of the degree of excuses using self risk in the loss domain are in the same direction as in the gain domain but with a much smaller effect (9.8 percentage points in the gain domain). In contrast, the sum of the coefficients of $tradeoff_{pi}$ and $others * tradeoff_{pi}$ shows that participants underweight the probability that others lotteries result in non-zero losses by about 11 percentage points on average, using it as an excuse to avoid self losses. That is, the degree of excuses using others' risk in the loss domain is twice as high as that in the gain domain. Columns 2-3 also confirm the robustness of our results to the inclusion of individual fixed effects and probability fixed effects. Columns 4-5 include participants who are excluded in our main results. The results remain significant, and the documented effects become larger when considering Tobit regressions as a robustness check.

We therefore document significant evidence for excuse-driven risk behavior in both the gain domain and the loss domain. These results are summarized as follows:

RESULT 2. *Participants demonstrated significant evidence of excuse-driven risk behavior in both the gain and loss domains. Specifically, they became more risk-averse to others' risks and more risk-seeking for their own risks in the gain domain when self-others trade-offs were involved. However, this pattern reversed in the loss domain, resulting in a reflection effect in excuse-driven risk behavior.*

4.3 *Heterogeneity in Risk Effects on the Level of Altruism*

We now turn to the examination of potential heterogeneous excuse-driven risk behavior. We define the degree of excuse-driven risk behavior as the gap between others' valuation and self-valuation of the same lottery,[11] i.e., $Y_d^o(P_d^i) - Y_d^s(P_d^i)$ with $d \in \{g, l\}$ and $i \in \{s, o\}$.

[10] Tables B.1 and B.2 in Appendix B examine the heterogeneous effects of baseline selfishness on excuse-driven behavior in both the gain and loss domains. Our results show that, when excluding subjects with censored values, individuals who exhibit higher selfishness in a riskless environment do not display significantly greater excuse-driven risk behavior. This contrasts with the findings reported in Exley (2016). However, when censored subjects are included in the analysis, the estimated coefficients become significant. These patterns suggest a non-monotonic relationship between baseline selfishness in a riskless environment and excuse-driven risk behavior.

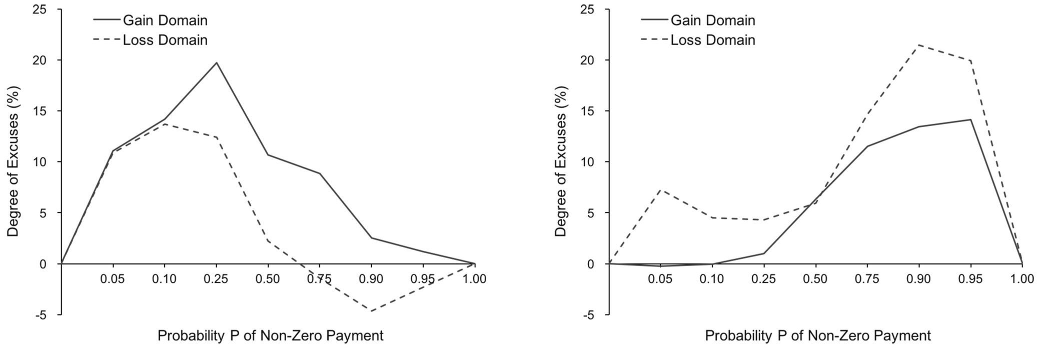
[11] While individuals may assign different utilities to equivalent levels of gains and losses, Figure B.6 in Appendix B shows that differences in decision weights across domains for self-lotteries without self–other trade-offs are not statistically significant at the 5% level for any probability level.

Regression:	Ordinary Least Squares			Tobit	
Dependent Variable:			Y_{pi}		
	1	2	3	4	5
others	2.38	2.38	2.38	1.82	1.67
	(1.36)	(1.34)	(1.36)	(0.99)	(0.92)
tradeoff	4.40*	4.40*	4.40*	21.17***	20.22***
	(1.73)	(1.70)	(1.73)	(6.02)	(5.89)
*others*tradeoff*	-15.56***	-15.56***	-15.56***	-40.59***	-39.54***
	(-3.51)	(-3.45)	(-3.50)	(-7.59)	(-7.58)
I(P=0.10)			6.34***		5.65***
			(6.51)		(8.64)
I(P=0.25)			18.10***		15.33***
			(13.64)		(15.96)
I(P=0.50)			35.61***		30.60***
			(18.95)		(18.19)
I(P=0.75)			49.06***		42.50***
			(22.32)		(22.05)
I(P=0.90)			58.52***		50.06***
			(23.29)		(22.11)
I(P=0.95)			63.36***		55.41***
			(23.86)		(23.01)
Constant	46.15***	42.82***	13.15***	44.42***	15.90***
	(30.69)	(35.45)	(7.39)	(34.28)	(10.75)
Ind FE	no	yes	no	no	no
Censored X	no	no	no	yes	yes
Observations	1372	1372	1372	2576	2576
Subjects	49	49	49	92	92

Table notes: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Standard errors are clustered at the individual level and shown in parentheses. The dependent variables Y_{pi} are lottery valuations in the loss domain, i.e., $Y_l^s(P_l^s)$, $Y_l^s(P_l^o)$, $Y_l^o(P_l^o)$, $Y_l^o(P_l^s)$. Probability fixed effects and individual fixed effects (Ind FE) are shown when included. “Censored X" indicates whether participants with censored X_l values are included or not. Tobit regressions include participants with censored values of X_l. We set their others-domain-specific values X_l with a right-censoring limit equal to 30.

Table 3. Regression Analysis on Excuse-Driven Risk Behavior in the Loss Domain

Figure 4 illustrates that the degree of excuses depends on both who bears the risk and the size of the risk. In particular, participants display significantly stronger excuse-driven risk behavior at small and moderate probabilities when the risk is borne by themselves. By contrast, when the risk falls on others, they show a pronounced tendency toward excuse-driven risk behavior at high probabilities. From $P = 0.05$ to $P = 0.95$, the degree of excuse-driven risk behavior for the self risk drops from 10% to about 0% and even becomes negative. In contrast, the degree of excuses can reach up to 20% when others



(a) Degree of Excuses for Self Risk

(b) Degree of Excuses for Others' Risk

Figure 4. The Degree of Excuse-Driven Risk Behavior

Figure notes: The estimates are the degree of excuse-driven risk behavior for self risk and others' risk in the gain and loss domains. The degree of excuses for risk i in the domain d is given by $Y_d^o(P_d^i) - Y_d^s(P_d^i)$ with $d \in \{g, l\}$ and $i \in \{s, o\}$. The solid lines estimate the degree of excuses in the gain domain. The dashed lines estimate the degree of excuses in the loss domain.

face the risk at a 95% probability level, but it is only about 5% when the probabilities are small or moderate.[12]

Therefore, there exists significant heterogeneity in excuse-driven risk behavior across different risk levels. We conjecture that it may depend on individuals' risk preferences toward self risk and others'. To investigate the potential mechanism, we categorize participants into two groups based on their risk preferences: risk-averse and risk-seeking. Notably, in the no self-others trade-off context, decision weights for self risk and others' risk reflect two distinct types of risk preferences as mentioned in Section 2.1. Decision weights for self risk ($Y_d^s(P_d^s)$) represent the risk preferences of participants themselves, while decision weights for others' risk ($Y_d^o(P_d^o)$) capture participants' beliefs about the risk preferences of their partners. We select the appropriate risk preferences to classify participants in the corresponding situations. Recall that Figure 2 in Section 4.2 indicates that participants' own risk preferences and their beliefs about others' risk preferences are nearly indistinguishable. We also observe a significant positive correlation between the risk preferences for oneself and others, with correlation coefficients of $\rho_g = 0.80$ in the gain domain and $\rho_l = 0.83$ in the loss domain. Such findings suggest that participants tend to project their own risk preferences onto others. Seemingly, we can categorize participants solely based on their own risk preferences. While it might appear straightforward to adopt a single risk preference, further analysis suggests a more nuanced view.

[12] Figure B.7 and B.8 in Appendix B provide t-test results for the degree of excuses in any given probability. Figure B.9 in Appendix B presents a comparison of excuse-driven risk behavior at different risk levels between the loss and gain domains. Overall, when the risk affects themselves, participants exhibit slightly stronger excuse-driven risk behavior in the gain domain than in the loss domain. Conversely, when the risk affects others, they show slightly stronger excuse-driven risk behavior in the loss domain than in the gain domain.

Considering the impact of the two risk preferences on the degree of excuses across different risk levels, we construct a linear regression with standard errors clustered at the individual level based on Equation (9). The dependent variable Y_{pi}^{j} indicates an individual i's degree of excuses for risk j with probability p in the gain or loss domain, where risk j represents self risk or others' risk, i.e., $j \in \{s, o\}$. The independent variables include two dummy variables for risk preferences for oneself and others: $rpself_{pi}$ and $rpothers_{pi}$, which are equal to 1 for risk aversion. In particular, when decision weights for states with non-zero outcomes surpass their objective probabilities, participants are classified as risk-seeking in the gain domain but risk-averse in the loss domain, and vice versa. We also include probability fixed effects (λ_p) in the regression model. Results in Table 4 demonstrate that only participants' beliefs about others' risk preferences can significantly explain the degree of excuse-driven risk behavior for others' risk, while their own risk preferences have limited explanatory power. Although participants' beliefs about others' risk preferences slightly significantly influence the degree of excuses for self risk in the gain domain, participants' risk preferences have higher explanatory power for excuse-driven risk behavior across both domains. Therefore, a more reasonable approach is to categorize participants as follows: when the risk is on the self-side, we use participants' risk preferences; when the risk is on the others-side, we use participants' beliefs about others' risk preferences.

$$Y_{pi}^{j} = \beta_0 + \beta_1 rpself_{pi} + \beta_2 rpothers_{pi} + \Sigma_p \lambda_p + \epsilon_{pi} \tag{9}$$

After classifying participants based on their preferences toward self and others' risk, Figure 5 and 6 plot individual heterogeneity in the degree of excuse-driven risk behavior in the gain and loss domains.[13] The findings reveal that risk-averse individuals exhibit nearly three times stronger excuse-driven behavior than risk-seeking individuals when faced with self-gain risk and others-loss risk. In contrast, the opposite is true in cases of others-gain risk and self-loss risk. Alternatively speaking, participants whose risk preferences are contrary to the direction of excuses exhibit a greater degree of excuses. Notice that this effect is somewhat mechanical and easy to understand, given how the degree of excuses is defined. For instance, participants who are risk-averse in self-gain risk in the no self-others trade-off context would be considered more selfish than participants who are risk-loving, even if they made the same choice in the self-others trade-off context.

Furthermore, the degree of excuses varies with the probabilities of risks. As the likelihood of self risk increases, individuals exhibit less excuse-driven risk behavior in both

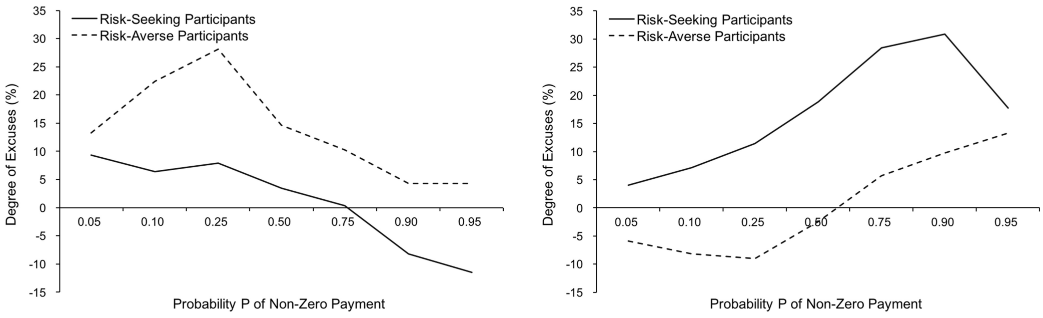
[13] Figure B.10 and B.11 in Appendix B provide results on t-tests for comparing differences in the degree of excuses between risk-averse and risk-seeking individuals. In our main results, we adopt the absolute degree of excuses, measured by $Y_d^o(P_d^i) - Y_d^s(P_d^i)$. Considering that the scale effect of probabilities may influence the robustness of results for individual heterogeneity, we introduce the relative degree of excuses similar to the definition of altruism levels mentioned previously in Section 2.2. The relative one is a better measurement for the degree of excuses since it can transform probabilistic effects into endogenous factors. Define the relative degree of excuses as $Y_d^o(P_d^i)/Y_d^s(P_d^i) - 1$ with $i \in \{s, o\}$ and $d \in \{g, l\}$. We plot individual heterogeneity for the relative degree of excuse-driven risk behavior in Figure B.12 and B.13 in Appendix B. The two figures draw similar conclusions as the findings in Figure 5 and 6, but their visibility is poor. To address this limitation, we ultimately favor the absolute degree of excuses for its good comparison and interpretability of the results.

Regression:	Ordinary Least Squares			
Dependent Variable:	Y_g^s	Y_g^o	Y_l^s	Y_l^o
	1	2	3	4
rpself	14.939***	6.642	-14.756***	-6.406
	(3.68)	(1.26)	(-3.06)	(-1.39)
rpothers	-9.079**	-17.988***	3.048	22.265***
	(-2.08)	(-4.12)	(0.62)	(3.78)
I(P=0.10)	2.907	0.641	2.723	-5.856**
	(0.90)	(0.27)	(1.03)	(-2.18)
I(P=0.25)	7.338*	1.725	0.918	-5.413*
	(1.98)	(0.61)	(0.29)	(-1.90)
I(P=0.50)	-1.897	8.088**	-10.066**	-1.577
	(-0.41)	(2.13)	(-2.51)	(-0.40)
I(P=0.75)	-5.509	14.673***	-15.517***	12.838**
	(-1.25)	(2.85)	(-3.55)	(2.63)
I(P=0.90)	-11.120**	18.045***	-20.122***	17.737***
	(-2.11)	(3.07)	(-4.12)	(3.23)
I(P=0.95)	-11.792**	18.769***	-18.065***	16.530***
	(-2.22)	(3.18)	(-4.10)	(2.84)
Constant	8.258	4.519	17.496***	-2.220
	(1.54)	(0.79)	(3.57)	(-0.52)
Observations	357	357	343	343
Subjects	51	51	49	49

Table notes: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Standard errors are clustered at the individual level and shown in parentheses. The dependent variables Y_d^j are the degree of excuses for risk j in the domain d, $j \in \{s, o\}$ and $d \in \{g, l\}$, i.e., $Y_d^s = Y_d^o(P_d^s) - Y_d^s(P_d^s)$ and $Y_d^o = Y_d^o(P_d^o) - Y_d^s(P_d^o)$. Participants are classified as risk-averse and risk-seeking based on participants' own risk preferences and beliefs about others' risk preferences for the corresponding risk. When decision weights for states with non-zero outcomes surpass their objective probabilities, participants are classified as risk-seeking in the gain domain but risk-averse in the loss domain, and vice versa.

Table 4. Impact of Risk Preferences on the Degree of Excuses

the gain and loss domains. Individuals seeking gain risk or avoiding loss risk even become more altruistic and sacrifice their own interests under higher probabilities. On the contrary, individuals' degree of excuses for others' risks increases as the probability of those risks increases. Under small probabilities, risk-averse individuals in the gain domain or risk-loving individuals in the loss domain tend to exhibit more altruistic behavior, resulting in a negative value in the degree of excuses. It is worth noting that existing experiments have demonstrated that most people exhibit four patterns of risk preferences. Specifically, they tend to be risk-seeking in high-risk gains and low-risk losses, while being risk-averse in low-risk gains and high-risk losses (e.g., Tversky and Kahneman, 1992). This remains true with our sample as already discussed in the no self-others trade-off context in Section 4.2. Thus, participants who display more altruistic behavior

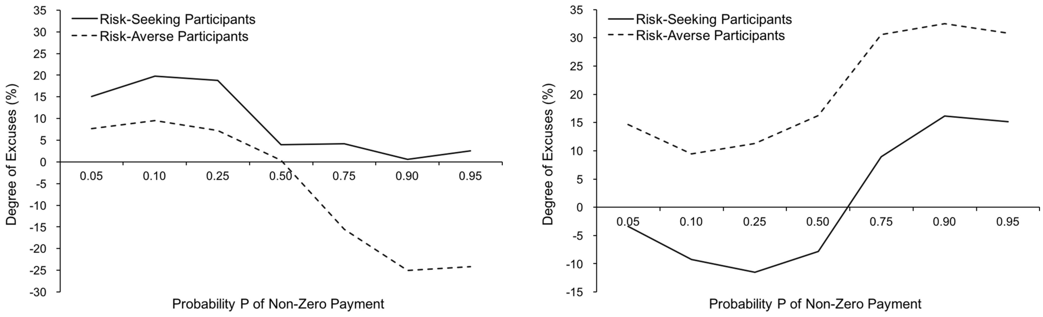


(a) Degree of Excuses for Self Risk

(b) Degree of Excuses for Others' Risk

Figure 5. Individual Heterogeneity in the Degree of Excuses in the Gain Domain

Figure notes: The estimates are the absolute degree of excuses for different types of participants in the gain domain, i.e., $Y_g^o(P_g^i) - Y_g^s(P_g^i)$ with $i \in \{s, o\}$. Participants are classified as risk-averse and risk-seeking based on participants' own risk preferences and beliefs about others' risk preferences for the corresponding risk. The solid lines represent risk-seeking participants while the dashed lines represent risk-averse participants.



(a) Degree of Excuses for Self Risk

(b) Degree of Excuses for Others' Risk

Figure 6. Individual Heterogeneity in the Degree of Excuses in the Loss Domain

Figure notes: The estimates are the absolute degree of excuses for different types of participants in the loss domain, i.e., $Y_l^o(P_l^i) - Y_l^s(P_l^i)$ with $i \in \{s, o\}$. Participants are classified as risk-averse and risk-seeking based on participants' own risk preferences and beliefs about others' risk preferences for the corresponding risk. The solid lines represent risk-seeking participants while the dashed lines represent risk-averse participants.

under risk are those whose risk preferences deviate from commonly observed patterns. This suggests that there is a limit to excuse-driven risk behavior. People cannot simply create excuses out of thin air; they must adhere to certain standards or common knowledge within given decision environments. We summarize these results as follows.

RESULT 3.

1. *Participants with different risk attitudes exhibit significantly varying degrees of excuses. Specifically, risk-averse participants demonstrate nearly three times stronger excuse-driven behavior compared to risk-seeking individuals when faced with self-gain risk and others-loss risk. In contrast, the reverse pattern emerges in situations involving others-gain risk and self-loss risk.*

2. *Participants whose risk preferences deviate from common trends exhibit more altruistic behavior under risk.*

5. THE SELF-SIGNALING MODEL

In Section 5.1, we introduce our self-signaling model, built on Bénabou and Tirole (2006) and Grossman and van der Weele (2017), and provide a more detailed explanation of our experimental findings on excuse-driven risk behavior in Section 5.2.

5.1 *Basic Framework*

In the self-signaling model, the agent possesses two selves: the decision-maker self, who is aware of the true preference type, and the observer self, who lacks this knowledge but can update beliefs about it by observing the actions taken by the decision-maker self in a given decision environment. Each agent has a preference type denoted as (θ, μ, γ). Here, θ represents the social-preference parameter, which reflects the degree of altruism or prosocial motivation of the agent. Agents with low values of θ exhibit less concern for the well-being of others. The parameter μ captures the psychological benefits associated with maintaining a positive self-image and is assumed to be strictly positive, i.e., $\mu > 0$. Individuals prefer to be viewed favorably by themselves and by others (Falk, 2021). Consequently, they may incur moral costs due to a diminished self-image. We further assume that μ is homogeneous and known to the observer self. γ captures risk preferences and determines the shape of probability weighting.

We design our model to mirror the structure of the experimental setup. In our analysis, we assume that subjects evaluate each decision on the choice list in isolation, consistent with narrow bracketing.[14] A similar approach is taken by Grossman and van der Weele (2017), who interpret Dana et al. (2006)'s wiggle-room experiment through a self-signaling framework. In particular, the choice not to reveal information prior to making a decision is treated as irrelevant for belief updating (see also Fehr and Charness (2025), p. 479). Empirically, Bénabou et al. (2024) find that prosociality varies only at the 6–7 percent level when comparing binary choice and multiple price list tasks, which may reflect the prevalence of narrow bracketing among subjects. It is important to note that the same subjects completed all versions of the choice lists—both with and without risk, and with and without self–other tradeoffs—so, in principle, the observer self had access to the full structure of the decision maker's preferences. Nevertheless, our findings, together with those of Exley (2020) and Garcia et al. (2020), indicate that this

[14] We greatly appreciate one anonymous referee for bringing up this discussion.

observer self does not fully exploit the available information; otherwise, excuse-driven risk-taking would not be observed. Interestingly, Exley and Kessler (2024a) show that choice bracketing itself can be strategically motivated in social contexts.

An alternative approach would be to assume that subjects evaluate the entire choice list holistically, treating it as a unified decision problem. In this case, a subject's switching point directly reveals her type, leading to a separating equilibrium. As Bénabou et al. (2024) discuss, this framework introduces additional effects arising from the contingent nature of bids on price lists, which we isolate in the present analysis. For instance, in the riskless environment within the gain domain, increasing X_g allows the decision-maker to raise his expected payoff; however, the effect on the expected utility derived from others-certain amount is negative. These effects vary across riskless and risky environments in self–other tradeoff contexts and complicate equilibrium comparisons. Nevertheless, the central mechanism remains: in a risky environment, individuals can attribute reduced altruism to risk preferences, with the degree of excuse-driven behavior shaped by beliefs about their own and others' risk preferences. With this in mind, we argue that our modeling approach offers greater parsimony without sacrificing realism.

In each row, agents choose between allocating a payoff to themselves ($a = 0$) or to another person ($a = 1$), under varying environments characterized by σ. When an action a is taken, the agent receives $W^s(a)$ as a material payoff, while her partner receives $W^o(a)$. The nature of either payoff can be safe or risky, depending on the decision environment. We assume that the decision-maker self is aware of and acts upon her true preferences (Bodner and Prelec, 2003). The objective function of the decision-maker self can be expressed as follows:

$$\begin{aligned}\max_{a\in\{0,1\}} V(a\mid(\theta,\mu,\gamma),\sigma) = \int w\, d\Pi\big(F_{W^s(a)}\big)(w) + \theta\int w\, d\Pi\big(F_{W^o(a)}\big)(w)\\ + \mu E\big[\theta\mid a,\sigma\big].\end{aligned} \tag{10}$$

The first two terms represent the utility of material payoffs from the agent's actions for both the self and others. Notably, the utility for others is weighted by the altruism parameter, θ. To incorporate non-risk neutrality, probabilities are subject to non-linear weighting, denoted as Π (Kahneman and Tversky, 1979). The final term represents the utility derived from self-image concerns, where the expectation term reflects the observer's posterior beliefs regarding the agent's level of altruism. Given her inability to achieve perfect introspection, the observer self focuses solely on the actual action rather than counterfactual alternatives.

The fundamental mechanism underlying our model is quite straightforward. Since the observer self lacks precise information regarding the preference parameters γ and θ, she may mistakenly attribute selfish behavior to risk preferences instead of a diminished level of altruism, thereby maintaining an unchanged self-image. However, this misattribution is not limitless; it is constrained by her beliefs about the true risk preferences. As will be clarified shortly, changes in the level of altruism from riskless to risky environments can occur through two distinct channels, which we refer to as the pure risk channel and the misattribution channel. To isolate the pure risk channel, we will consider a hypothetical scenario in which the observer self possesses complete information about risk preferences, thus eliminating the possibility of misattributing a reduction in altruism to risk factors.

5.2 *Understanding Experimental Findings through Image Concerns*

To illustrate our experimental results on excuse-driven risk behavior, we consider the following three signaling environments, denoted as σ_j for $j \in \{0,1,2\}$: a riskless environment (σ_0), a risky environment with information on risk preferences (σ_1), and a risky environment without information on risk preferences (σ_2). In each environment, the observer self lacks knowledge of the decision-maker's true level of altruism, θ, but updates her beliefs about it based on the observed actions taken. We assume that the prior beliefs about θ follow a distribution $F(\theta)$ with full support on the interval $[0,1]$. The observer self's posterior beliefs are captured by the behavioral cutoff $\hat{\theta}_j$, where $j \in \{0,1,2\}$, and are derived from the action a and the signal σ_j. We denote the behavioral cutoff from the perspective of the experimenter as $\hat{\theta}_j^*$ in the signaling environment σ_j. Notably, the observer's behavioral cutoff coincides with the experimenter's cutoff in the signaling environments σ_0 and σ_1. However, they diverge in σ_2, as the experimenter is assumed to have full information about the decision-maker's true risk preferences, while the observer self does not. For the sake of clarity, we will focus solely on the domain of gains and scenarios where others encounter risks. However, the same mechanisms are applicable to self risk and the domain of losses.

5.2.1 *Riskless environment (σ_0)*

We begin with the normalization tasks of the experiment, where choices are made under conditions of certainty. Specifically, participants were asked to choose between receiving 10 for themselves while allowing others to receive 0, or receiving 0 for themselves while permitting others to receive a certain amount as indicated in the normalized price lists. It is intuitive to define the allocation ratio $10/X_g$ as the behavioral cutoff for the experimenter and the observer self in a risk-free context, i.e., $\hat{\theta}_0^* = \hat{\theta}_0 = 10/X_g$. A higher ratio indicates a greater level of altruism exhibited by the agent. When $a = 0$, the observer self understands that the decision-maker's true level of altruism is below $\hat{\theta}_0$; conversely, when $a = 1$, the observer self recognizes that the decision-maker's true level of altruism is above $\hat{\theta}_0$. Therefore, the image utility can be expressed as $E\left[\theta \mid \theta < \hat{\theta}_0\right]$ when $a = 0$, and $E\left[\theta \mid \theta \geq \hat{\theta}_0\right]$ when $a = 1$.

By definition, X_g is a certain amount for others so that the agent is indifferent between bundles $(10,0)$ and $(0,X_g)$, i.e.,[15]

$$\begin{aligned} &V(0 \mid (\theta,\mu,\gamma),\sigma_0) = V(1 \mid (\theta,\mu,\gamma),\sigma_0) \Leftrightarrow \\ &10 + \mu E\left[\theta \mid \theta < \hat{\theta}_0\right] = \theta X_g + \mu E\left[\theta \mid \theta \geq \hat{\theta}_0\right] \end{aligned} \tag{11}$$

[15] Throughout this section, we assume the uniqueness of equilibrium in all three signaling environments. The exact conditions for this uniqueness are derived in Appendix C. Specifically, uniqueness requires that the distribution of θ is not overly concentrated in particular regions of the type space, thereby preventing drastic updates in beliefs about the level of altruism (see also Bénabou and Tirole (2006); Grossman and van der Weele (2017)). Previous studies have shown that multiple switching points occur in fewer than 10% of price lists (Andreoni and Sprenger, 2011, Exley, 2016, Garcia et al., 2020), which supports the reasonableness and realism of our assumption.

It can be simplified to:

$$\nu_0(\hat{\theta}_0 \mid \{\theta, \mu\}) = \mu\delta(\hat{\theta}_0) + 10\frac{\theta}{\hat{\theta}_0} - 10 = 0. \tag{12}$$

where $\delta(\hat{\theta}_0) := E\big[\theta \mid \theta \geq \hat{\theta}_0\big] - E\big[\theta \mid \theta < \hat{\theta}_0\big]$ represents the image reward of prosocial behavior.

PROPOSITION 1. *In our normalization tasks, where choices are made under conditions of certainty,*

1. *The agent behaves more altruistically if she cares more about others' welfare; that is, $\hat{\theta}_0^*$ increases as θ increases.*
2. *The agent also exhibits more altruistic behavior when she is more concerned about her self-image; that is, $\hat{\theta}_0^*$ increases as μ increases.*
3. *The agent with self-image concerns tends to exhibit more altruistic behavior, i.e., $\hat{\theta}_0^* > \theta$.*

PROOF. See the proof in Appendix C.1. □

Intuitively, Proposition 1 illustrates that the agent behaves more altruistically as her inherent altruism and self-image concerns increase. Furthermore, due to the self-image benefits associated with pro-social behavior, the agent's exhibited altruism exceeds her intrinsic level in a risk-free context.

5.2.2 *Risky environment with "complete information on risk preferences" (σ_1)*

Now, we will investigate how risk influences prosocial behavior while assuming complete information regarding risk preferences, thereby examining the pure risk channel. When others are faced with risk, the agent must choose between receiving a certain amount for themselves and allowing others to receive nothing, or receiving nothing themselves while enabling others to receive a lottery payout of P_g^o. At equilibrium, the agent is indifferent between the bundles $(C_g^s(P_g^o), 0)$ and $(0, P_g^o)$.

Suppose that the probability weighting function takes the following form: $\Pi(p) = p^\gamma$, where $\gamma > 0$. When $\gamma = 1$, the agent is considered risk-neutral; when $\gamma < 1$, the agent is risk-seeking; and when $\gamma > 1$, the agent is risk-averse. Furthermore, a higher value of γ indicates a greater degree of risk aversion. In the current decision-making environment, the only preference parameter that the observer self does not know is θ. Similar to a riskless environment, the observer self can utilize the allocation ratio $\hat{\theta}_1 = C_g^s(P_g^o)/C_g^o(P_g^o \mid \gamma)$, which represents the behavioral cutoff to update her beliefs about it. In particular, the image utility is given by $E\big[\theta \mid \theta < \hat{\theta}_1\big]$ when $a = 0$, and $E\big[\theta \mid \theta \geq \hat{\theta}_1\big]$ when $a = 1$. Notice that, under the assumption of complete information on risk preferences, the behavioral cutoff for the observer self is equivalent to the behavioral cutoff for the experimenter, i.e., $\hat{\theta}_1 = \hat{\theta}_1^*$. By definition, $C_g^s(P_g^o)$ is the root of the following equation:

$$\begin{aligned} &V(0 \mid (\theta, \mu, \gamma), \sigma_1) = V(1 \mid (\theta, \mu, \gamma), \sigma_1) \Leftrightarrow \\ &C_g^s(P_g^o) + \mu E\big[\theta \mid \theta < \hat{\theta}_1\big] = \theta X_g p^\gamma + \mu E\big[\theta \mid \theta \geq \hat{\theta}_1\big] \end{aligned} \tag{13}$$

Equation (13) can be rewritten as follows

$$\nu_1(\hat{\theta}_1 \mid \{\theta, \gamma, \mu\}) = \mu\delta(\hat{\theta}_1) + 10p^{\gamma}\frac{\theta - \hat{\theta}_1}{\hat{\theta}_0} = 0. \tag{14}$$

Comparing the agent's altruistic behavior in riskless and risky environments, we obtain the following result.

PROPOSITION 2. *Under conditions of information symmetry regarding true risk preferences γ, the agent exhibits more altruistic behavior in a risky environment compared to a riskless environment (i.e., $\hat{\theta}_1^* > \hat{\theta}_0^*$).*

PROOF. See the proof in Appendix C.2. □

There are two main takeaways from this result. On the one hand, Proposition 2 indicates that, in the absence of an excuse for selfishness, the agent behaves even more altruistically when faced with risk. When the agent exhibits the same level of altruism, $\hat{\theta}_0$, in both contexts and consequently receives the same image reward, the utility cost in a risky context—calculated as $10p^{\gamma}(1-\theta/\hat{\theta}_0)$—is lower than the corresponding cost in a risk-free context, which is $10(1-\theta/\hat{\theta}_0)$. Therefore, the presence of risk encourages the agent to adopt a more altruistic stance, as she can derive greater utility from a positive self-image at a lower cost compared to a risk-free environment. On the other hand, this result suggests that the independence axiom—and, consequently, the predictions regarding standard risk preferences derived by Exley (2016)—do not hold when self-image concerns are present, as these concerns enter the agent's objective function in a nonlinear manner.

As our study and existing research (Exley, 2016, Garcia et al., 2020) all point out, individuals exhibit more selfish behavior in the presence of uncertainty at the aggregate level. This implies that the reduction in altruistic behavior between no-risk and risk environments is purely driven by the misattribution channel, whose effect may be potentially underestimated.

5.2.3 *Risky environment with incomplete information on risk preferences (σ_2)*

In this scenario, the observer self lacks precise knowledge of both true risk preferences (i.e., γ) and the level of altruism (i.e., θ). Consequently, there is a possibility of misattributing selfish behavior to risk preferences rather than altruism, while maintaining the same self-image. However, this misattribution is not limitless; it is constrained by her beliefs regarding true risk preferences. In particular, we will assume that there are both upper and lower limits on γ, denoted as $\overline{\gamma}$ and $\underline{\gamma}$, the formalization of which depends on the decision context. Beyond these limits, a risk preference would be considered implausible by the observer. For example, individuals have been documented to be risk-averse under high probabilities ($\gamma > 1$) and risk-seeking ($0 < \gamma < 1$) under low probabilities in the gain domain (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). Therefore, it may be implausible to assume that displaying extreme risk aversion when others encounter minor risks, or exhibiting extreme risk-seeking behavior when others confront significant risks, is acceptable. Here, we disregard the construction of these value limits by individuals, which may be influenced by their beliefs regarding the distribution of risk preferences across the population.[16] Moreover, we will assume that the observer self interprets evidence in a somewhat favorable manner concerning the decision-maker's self-image. She utilizes $\overline{\gamma}$ ($\underline{\gamma}$) to evaluate the actions taken by the decision-maker when an upper (lower) limit supports her image.

In the scenario under consideration, the agent must choose between two options: one where others receive the lottery amount P_g^o while the agent receives nothing, and another where others receive nothing while the agent receives a guaranteed positive amount. The observer self thus utilizes her behavioral cutoff $\hat{\theta}_2 = C_g^s(P_g^o)/C_g^o(P_g^o \mid \overline{\gamma})$ to update her beliefs about the true level of altruism. Yet, the behavioral cutoff for the experimenter is given by $\hat{\theta}_2^* = C_g^s(P_g^o)/C_g^o(P_g^o \mid \gamma)$.[17]. Then, $\hat{\theta}_2$ is equal to $\hat{\theta}_2^* p^\gamma / p^{\overline{\gamma}}$. Consequently, the image utility is given by $E\big[\theta \mid \theta < \hat{\theta}_2^* p^\gamma / p^{\overline{\gamma}}\big]$ when $a = 0$, and $E\big[\theta \mid \theta \geq \hat{\theta}_2^* p^\gamma / p^{\overline{\gamma}}\big]$ when $a = 1$. By definition, $C_g^s(P_g^o)$ is the root of the following equation:

$$\begin{aligned} & V(0 \mid (\theta, \mu, \gamma), \sigma_2) = V(1 \mid (\theta, \mu, \gamma), \sigma_2) \Leftrightarrow \\ & C_g^s(P_g^o) + \mu E\big[\theta \mid \theta < \hat{\theta}_2^* \frac{p^\gamma}{p^{\overline{\gamma}}}\big] = \theta X_g p^\gamma + \mu E\big[\theta \mid \theta \geq \hat{\theta}_2^* \frac{p^\gamma}{p^{\overline{\gamma}}}\big]. \end{aligned} \tag{15}$$

Equation (15) can be reformulated as

$$\nu_2(\hat{\theta}_2^* \mid \{\theta, \gamma, \mu, \overline{\gamma}\}) = \mu \delta(\frac{p^\gamma}{p^{\overline{\gamma}}} \hat{\theta}_2^*) + 10 p^\gamma \frac{\theta - \hat{\theta}_2^*}{\hat{\theta}_0} = 0. \tag{16}$$

PROPOSITION 3. *In a risky environment with incomplete information regarding risk preferences, the agent's behavior exhibits the following properties:*

1. *The agent behaves more selfishly when the observer self believes that she is more averse to others-gain risk, i.e., $\hat{\theta}_2^*$ decreases as $\overline{\gamma}$ increases.*

2. *The agent exhibits more selfish behavior when she is more tolerant of risks associated with others' gains, i.e., $\hat{\theta}_2^*$ decreases as γ decreases.*

3. *There exists a threshold γ^*, satisfying $\gamma^* < \overline{\gamma}$, such that: (a) if $\gamma \leq \gamma^*$, then $\hat{\theta}_2^* \leq \hat{\theta}_0$; (b) if $\gamma > \gamma^*$, then $\hat{\theta}_2^* > \hat{\theta}_0$.*

PROOF. See the proof in Appendix C.3. □

These results derived from our signaling model can help us understand the individual heterogeneity in prosocial behavior, as illustrated in Figure 5 in Section 4.3. Proposition 3 indicates that an agent can exhibit excuse-driven risk behavior in the presence

[16] Another approach is to assume that the observer self evaluates the action taken by the decision-maker against each possible γ and aggregates these evaluations linearly. The resulting degree of excuses would be lower under this approach. However, the interpretation of the results under different approaches remains very similar.

[17] In the experiment, although the observer self is unaware of the exact risk preferences, we can observe it according to her choices in the no self-others trade-off context

of risk, depending on her true risk preferences and beliefs about them. The first result states that differences in the degree of excuses at varying risk levels can arise from differing beliefs about risk preferences ($\overline{\gamma}$) at the corresponding risks, even though the decision-maker's risk preferences remain unchanged. As the probability of others' risks increases, the observer self believes that the decision-maker self becomes more risk-averse, as evidence in the no self-others trade-off context has shown. In this scenario, the agent can more readily use risk aversion as an excuse for less altruistic behavior without damaging her self-image. The second result of Proposition 3 elucidates the differences in the degree of excuses between risk-averse and risk-loving individuals at the same level of risk. Risk-loving individuals tend to exhibit more selfish behavior when confronted with the risk of others' gains. They have greater latitude to attribute their altered self-image, resulting from selfishness, to innocuous risk preferences.

As illustrated in Figures 5 and 6 in Section 4.3, not all individuals exhibit more selfish behavior in the presence of risk. The third result of Proposition 3 suggests that the agent may act more altruistically. As shown by the dashed line in Figure 5b, risk-averse participants demonstrate more altruistic behavior under small probabilities than in a risk-free context (i.e., $\hat{\theta}_2 > \hat{\theta}_0$). Notably, while most people are risk-seeking in the face of small risks of gains (Tversky and Kahneman, 1992), these participants' risk preferences deviate from common trends. Their overly altruistic behavior occurs for two reasons. On one hand, it is less costly to behave altruistically under risk, as discussed in Proposition 2; on the other hand, there is less room for excuses for these subjects. These two effects work jointly, leading to more altruistic behavior among these participants.

6. Conclusion

In this paper, we conducted an experiment to study how risk influences prosociality in situations involving either gains or losses. Using a between-subjects design, we replicated the main findings of Exley (2016), which demonstrated that people use risks as excuses for selfishness in the context of charitable donations. We observed that similar patterns persist when decisions result in pure losses. Due to the presence of excuse-driven risk behavior, the risk preferences observed in the context of self-other trade-offs exhibited reflected patterns relative to those observed in contexts without such trade-offs. Beyond complementing the existing literature by providing evidence of excuse-driven risk behavior in the loss domain, we specifically focused on identifying the key elements underlying this behavior. Our subject-level analysis revealed significant heterogeneity in the degree of excuses. Notably, the degree of excuses depends not only on the size of the risk at stake but also on individuals' risk preferences and their beliefs about others' risk preferences.

Drawing on the existing literature on self-image concerns (Bénabou and Tirole, 2006, Grossman and van der Weele, 2017), we developed a signaling model with a dual-self framework to explain our experimental findings, where information asymmetry exists between the decision-maker self and the observer self. Specifically, the observer self lacks knowledge about the decision-maker's type and must simultaneously update her beliefs regarding two preference parameters in a risky environment. This creates room for misattributing selfish behavior to innocuous risk preferences rather than to a lack of altruism. However, as our experimental findings suggest, this misattribution is not unlimited; it is constrained by the observer's beliefs about risk preferences. Generally speaking, when making excuses, people must conform to accepted standards or common knowledge within the given decision-making environment. Otherwise, the excuses they make will lack credibility and cannot be used to justify their selfish behavior, leading to damage to their self-image.

A better understanding of how uncertainty affects prosocial behaviors is particularly important. If we aspire to live in a society where people consistently behave prosocially toward one another, it is crucial to either eliminate these moral obstacles entirely or focus on situations where uncertainty is more likely to erode morality. Our theory suggests that revealing hidden information should be a priority when addressing immoral behavior in the policy sphere. However, information such as preferences may sometimes be difficult to obtain, even for policymakers. An alternative approach would be to rely on risk management tools, such as insurance, to reduce or eliminate uncertainty. For instance, in our experiment, if insurance were available at no cost, we could limit the use of risk as an excuse for selfishness and restore morality. We leave this potential extension for future research.

Acknowledgment

Jiakun Zheng gratefully acknowledges financial support from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University-A*MIDEX.

References

AGERSTRÖM, JENS, RICKARD CARLSSON, LINDA NICKLASSON, AND LINDA GUNTELL (2016): "Using descriptive social norms to increase charitable giving: The power of local norms," *Journal of Economic Psychology*, 52 (C), 147–153. [2]

ANDREONI, J (1990): "Impure altruism and donations to public goods: A theory of warm-glow giving," *Economic Journal*, 100 (401), 464–477. [2, 5]

ANDREONI, JAMES AND JOHN MILLER (2002): "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70 (2), 737–753. [2]

ANDREONI, JAMES AND CHARLES SPRENGER (2011): "Uncertainty equivalents: Testing the limits of the independence axiom," *National Bureau of Economic Research*, 17342. [25]

BARTLING, BJÖRN, FLORIAN ENGL, AND ROBERTO A WEBER (2014): "Does willful ignorance deflect punishment?-An experimental study," *European Economic Review*, 70 (C), 512–524. [2]

BÉNABOU, ROLAND, ARMIN FALK, LUCA HENKEL, AND JEAN TIROLE (2024): "Eliciting Moral Preferences under Image Concerns: Theory and Experiment," . [23, 24]

BÉNABOU, R AND J TIROLE (2006): "Incentives and prosocial behavior," *American Economic Review*, 96 (5), 1652–1678. [4, 23, 25, 29]

BLEICHRODT, HAN AND PAUL VAN BRUGGEN (2022): "The reflection effect for higher-order risk preferences," *Review of Economics and Statistics*, 104 (4), 705–717. [9]

BODNER, RONIT AND DRAZEN PRELEC (2003): "Self-signaling and diagnostic utility in everyday decision making," *Psychology of Economic Decisions*, 1 (105), 26. [24]

DANA, JASON, DAYLIAN M CAIN, AND ROBYN M DAWES (2006): "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and Human Decision Processes*, 100 (2), 193–201. [4, 23]

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. [2, 5, 11]

DANNENBERG, ASTRID AND PETER MARTINSSON (2021): "Responsibility and prosocial behavior-Experimental evidence on charitable donations by individuals and group representatives," *Journal of Behavioral and Experimental Economics*, 90 (C), 101643. [2]

DELLAVIGNA, STEFANO, JOHN A LIST, AND ULRIKE MALMENDIER (2012): "Testing for altruism and social pressure in charitable giving," *Quarterly Journal of Economics*, 127 (1), 1–56. [5]

DICKINSON, DAVID L (1998): "The voluntary contributions mechanism with uncertain group payoffs," *Journal of Economic Behavior and Organization*, 35 (4), 517–533. [2]

ELSTER, JON (1989): "Social norms and economic theory," *Journal of Economic Perspectives*, 3 (4), 99–117. [5]

ENGEL, JANNIS AND NORA SZECH (2020): "A little good is good enough: Ethical consumption, cheap excuses, and moral self-licensing," *PloS One*, 15 (1), e0227036. [2]

EXLEY, CHRISTINE L (2016): "Excusing selfishness in charitable giving: The role of risk," *Review of Economic Studies*, 83 (2), 587–628. [2, 3, 5, 6, 8, 10, 12, 15, 17, 25, 27, 29]

——— (2020): "Using charity performance metrics as an excuse not to give," *Management Science*, 66 (2), 553–563. [17, 23]

EXLEY, CHRISTINE L AND JUDD B KESSLER (2024a): "Equity concerns are narrowly framed," *American Economic Journal: Microeconomics*, 16 (2), 147–179. [24]

——— (2024b): "Motivated errors," *American Economic Review*, 114 (4), 961–987. [2]

FALK, ARMIN (2021): "Facing yourself-a note on self-image," *Journal of Economic Behavior and Organization*, 186, 724–734. [23]

FEHR, ERNST AND GARY CHARNESS (2025): "Social preferences: fundamental characteristics and economic consequences," *Journal of Economic Literature*, 63 (2), 440–514. [23]

FEHR, ERNST AND KLAUS M SCHMIDT (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114 (3), 817–868. [2]

FINKELSTIEN, MARCIA A (2009): "Intrinsic vs. extrinsic motivational orientations and the volunteer process," *Personality and Individual Differences*, 46 (5-6), 653–658. [2]

GARCIA, THOMAS, SEBASTIEN MASSONI, AND MARIE CLAIRE VILLEVAL (2020): "Ambiguity and excuse-driven behavior in charitable giving," *European Economic Review*, 124 (1), 103412. [2, 5, 10, 23, 25, 27]

GEBAUER, JOCHEN E, MICHAEL RIKETTA, PHILIP BROEMER, AND GREGORY R MAIO (2008): "Pleasure and pressure based prosocial motivation: Divergent relations to subjective well-being," *Journal of Research in Personality*, 42 (2), 399–420. [2]

GIBSON, RAJNA, CARMEN TANNER, AND ALEXANDER F WAGNER (2013): "Preferences for truthfulness: Heterogeneity among and within individuals," *American Economic Review*, 103 (1), 532–548. [2]

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. [2]

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. [2, 4, 5, 23, 25, 29]

HOPFENSITZ, ASTRID, CÉSAR MANTILLA, AND JOSEPA MIQUEL-FLORENSA (2019): "Catch uncertainty and reward schemes in a commons dilemma: An experimental study," *Environmental and Resource Economics*, 72, 1121–1153. [5]

KAHNEMAN, DANIEL AND AMOS TVERSKY (1979): "Prospect theory: An analysis of decision under risk," *Econometrica*, 47 (2), 363–391. [24, 27]

KROLL, ALEXANDER AND DOMINIK VOGEL (2018): “Changes in prosocial motivation over time: a cross-sector analysis of effects on volunteering and work behavior,” *International Journal of Public Administration*, 41 (14), 1119–1131. [2]

KRUEGER, JOACHIM I, ADAM L MASSEY, AND THERESA E DIDONATO (2008): “A matter of trust: From social preferences to the strategic adherence to social norms,” *Negotiation and Conflict Management Research*, 1 (1), 31–52. [5]

KUNDA, ZIVA (1987): “Motivated inference: Self-serving generation and evaluation of causal theories,” *Journal of Personality and Social Psychology*, 53 (4), 636–647. [4]

——— (1990): “The case for motivated reasoning,” *Psychological Bulletin*, 108 (3), 480–498. [4]

PACE, DAVIDE AND JOËL VAN DER WEELE (2020): “Curbing carbon: An experiment on uncertainty and information about CO2 emissions,” *Social Science Research Network*, 3693235. [5]

ROVIRA, JOAN, W KIP VISCUSI, FERNANDO ANTOÑANZAS, JOAN COSTA, WARREN HART, AND IRINEU CARVALHO (2000): “Smoking risks in Spain: Part II-Perceptions of environmental tobacco smoke externalities,” *Journal of Risk and Uncertainty*, 21, 187–212. [5]

TELLA, RAFAEL DI, RICARDO PEREZ-TRUGLIA, ANDRES BABINO, AND MARIANO SIGMAN (2015): “Conveniently upset: Avoiding altruism by distorting beliefs about others’ altruism,” *American Economic Review*, 105 (11), 3416–3442. [5]

TVERSKY, AMOS AND DANIEL KAHNEMAN (1992): “Advances in prospect theory: Cumulative representation of uncertainty,” *Journal of Risk and Uncertainty*, 5, 297–323. [4, 6, 8, 9, 13, 21, 27, 29]

ZELMER, JENNIFER (2003): “Linear public goods experiments: A meta-analysis,” *Experimental Economics*, 6, 299–310. [2]