3-28-2022

Agency, Benevolence and Justice

Prithvijit Mukherjee
Mount Holyoke College, prithvijit@mtholyooke.edu

J. Dustin Tracy
Chapman University, tracy@chapman.edu

Agency, Benevolence and Justice

Comments

ESI Working Paper 22-03

Agency, Benevolence and Justice

Prithvijit Mukherjee* J Dustin Tracy†

March 28, 2022

Abstract

We test for social norms regarding how Agents should select between risky prospects for Principals, including norms consistent with observations by Adam Smith. We elicit norms from subjects serving as "impartial spectator[s]" about the choice of risky prospects selected by the Agents. We find strong evidence for the existence of norms, consistent with Smith's observations. Furthermore, we find that Agents are more likely to select more normative options. In contrast, we find that Principals' allocation for bonuses depends on the realization of the risky prospect rather than whether the Agents' choice was consistent with the norm.

Key Words: Social norms, Decisions-making for others, Laboratory experiments, Principal-Agent, Decision-making under risk

JEL Codes: C9, D63, D81, D90, G41

1 Introduction

> Our continual observations upon the conduct of others insensibly lead us to form to ourselves certain general rules concerning what is fit and proper either to be done or to be avoided.
>
> -Adam Smith (*The Theory of Moral Sentiments*, p. 140)

In finance and life in general, we must rely on others for making choices. Yet, most, if not all, contracts are incomplete and do not state stipulations for every possible contingency; some things seem so self-evident that there is no need to state them; some situations are never anticipated, and it would be inefficient to try and write a complete contract for every imaginable situation. Can trust bridge this incompleteness, allowing the granting of agency to another despite imperfect contracts? Do societal (perhaps universal) norms buttress this trust? Is there a mutual understanding of certain principles within society that provide an unspoken and unwritten framework for these contracts and traversing situations they do not address? In particular, Smith (1759) suggests this includes rewarding benevolence and punishing malevolence, but not rewarding lack of malevolence nor punishing a lack of benevolence.

There is a large literature that finds that social norms and rules can be pivotal in explaining deviation from "self-utility" maximizing Agents in many economic interactions.[1] There is evidence that moral and social norms systematically influence behavior in dictator games (Krupka and Weber, 2013), ultimatum games (Smith and Wilson, 2018), trust games (Johnson and Mislin, 2011, Smith, 2020), public goods games (Chaudhuri, 2011, Kimbrough and Vostroknutov, 2016), among others. However, the inherent stochasticity

We thank Erik Kimbrough and Erin Krupka for their input and comments on the design; and Kevin James for his help programming the software. Any mistakes are our own.

* Affiliation: Mount Holyoke College. Email: prithvijit@mtholyooke.edu

† Affiliation: Chapman University. Email: tracy@chapman.edu

[1] Please see Kimbrough and Wilson (2021) for an excellent overview of the experimental and theoretical literature on rule-following.

in financial decision-making drives a wedge between actions individuals take on behalf of others and the outcomes the others receive. For instance, making a risky investment decision that yields a higher payoff is viewed differently than choosing a safe option for a lower but more certain payoff. This tension between intention and outcomes complicates the process of understanding which norms are most appropriate to the decision. The literature studying how individuals make financial decisions for others finds that risk attitudes of the decision-makers cannot consistently explain choices made for others (Eriksen et al., 2020).[2]

In this paper, we design a set of experiments to test for the existence of social norms regarding Principal-Agent interactions, identify potential guiding principles to these norms and explore how information about outcomes impacts norms, test how closely the norms predict Agent behavior, and see Principal reward (punish) adherence to (departure from) these norms. The Agent chooses between two risky prospects, A and B. Across the treatments, we vary whether the Agent is making a choice for themselves (Self), for a Principal who has no option to reward or punish the Agent (Other), and for a Principal who has the opportunity to send a reward to the Agent (Consequence). As a preliminary step, we use a group of independent observers or Judges who rate the social appropriateness of the choice and whether the choice deserves a reward or a punishment. We elicit the norms using a coordination game (Krupka and Weber, 2013), where the Judges who guess the most popular option get an additional payoff. These Judges' rating forms the basis for evaluating the decision made in the Principal-Agent game. We conducted the entire study with an online subject pool to avoid existing social norms and bonds between students in a university experimental laboratory.[3]

We find that there is strong coordination among Judges on social norms, which include rewarding benevolence and punishing malevolence, but does not include rewarding lack of malevolence or punishing a lack of benevolence Smith (1759). We find that social norms we elicit from Judges successfully predict the Agent selection of prospects for Principals. We find that Agents tend to make more risk-neutral choices for the Principal than for themselves, but that defaults impact choices; choices were closer to risk-neutral when the starting prospect was more risk-neutral of the pair. However, we find that social norms have little impact on the financial consequences Principals impose on Agents; the value of the realized prospect seems to dominate this decision.

This paper makes several contributions to the literature and helps forge valuable connections between parallel strands. It applies a framework of propositions from Adam Smith, which have been shown to hold within the trust and ultimatum games, to a Principal-Agent setting, featuring uncertainty, which expands the understanding of under what circumstances are these the salient principles. Additionally, by eliciting norms, we can start to connect these behaviors to social norms. It advances the literature on social norms by eliciting norms online, though still in an incentivized manner, but with a population with no clear focal identity, e.g., all students at the same university. Additionally expands our understanding of norms by exploring how norms interact with certainty. The norm literature benefits from connection to the Principal-Agent literature because it provides insight and theory as to why there is convergence on a norm. It expands the understanding of the social underpinning of the literature on making risky choices for others by connecting it to the research on norms.

The next section reviews the relevant literature. Section 3 presents hypotheses drawn from that literature and details how the study seeks to test them. Section 4 details the design of the experiment. Section 5 presents the results. Section 6 discusses the results and implications for further research.

2 Background

There is evidence from early experiments investigating the effect of norms in bargaining to a wide literature studying dictator games, trust games, and public good games demonstrating that when making decisions, individuals consider not only their own payoffs but the payoff of others. Vostroknutov (2020) provides an excellent overview of the literature and presents a theory unifying how both descriptive and injunctive

[2] Polman and Wu (2020) in a meta-analysis only finds a very small indication of more risk-taking behavior when making decisions for others.

[3] We use participants registered on prolific.co.

norms might impact an individual's utility. Additionally, provides evidence that the relative appropriateness of action within the set will influence the likelihood that action is taken. Levitt and List (2007) model moral costs of actions depending upon the scrutiny the action will receive.

The literature studying risk-taking decisions on behalf of others has inconclusive patterns. Some papers find individuals are more likely to take more risks when they make decisions for others than for themselves (Chakravarty et al., 2011, Polman, 2012, Agranov et al., 2014, Pollmann et al., 2014). In contrast, other papers find there is an increase in risk aversion when individuals make decisions for others (Charness and Jackson, 2009, Reynolds et al., 2009, Bolton and Ockenfels, 2010, Eriksen and Kvaløy, 2010, Pahlke et al., 2012). The third group of papers finds there is no difference in risk-taking behavior (Harrison et al., 2013, Luzuriaga et al., 2017, Barrafrem and Hausfeld, 2020). A meta-analysis of the literature finds there is a very small indication of an increase in risk-taking behavior (Polman and Wu, 2020).

Specific to Principal-Agent settings, Lazear (1995) notes that traditionally employee pay is only adjusted negatively when the employee fails to meet some minimal standard, e,g, a factory worker who is late to a shift is docked. And all other adjustments are positive, e,g, employees who exceed productivity expectations are given bonuses. Fehr et al. (1997) conduct Principal-Agent experiments in Principals can state wages and desired effort from Agents. The market functions better for both Principals and Agents when the Principal can punish Agents relative to when no punishment was possible. Yet the market works best when the Principal can punish or reward. Marchegiani et al. (2016) conduct an experiment in which contracts are either lenient or severe and find that neglecting to reward deserving Agents is more detrimental than rewarding undeserving Agents. Rubin and Sheremeta (2016) find that when stochastic shocks are introduced (to the same environment), even with perfect information about shocks, Principals reward based on output rather than effort.

Smith and Wilson (2019, pp 85-90) restate observations Adam Smith made in *The Theory of Moral Sentiments* (1759) as a series of propositions.

Beneficence Proposition 1:	If X does something good (Z^{good}) for Y because she wants to do something good for Y, Z^{good} appears, with nothing further needed, to deserve reward by Y.
Beneficence Proposition 2:	If X does not do something good (Z^{good}) for Y because she does not want to do something good for Y, the lack of Z^{good} does not, solely by itself, to deserve punishment by Y.
Injustice Proposition 1:	If X does something bad (Z^{bad}) to Y because he wants to do something bad to Y, Z^{bad} appears, with nothing further needed, to deserve punishment by Y.
Injustice Proposition 2:	If X does not do something bad (Z^{bad}) to Y because she does not want to do something bad to Y, the lack of Z^{bad} appears, solely by itself, to deserve reward by Y.

The high proportion of second movers who return money in trust games (Berg et al., 1995) yet not in the involuntary trust game (McCabe et al., 2003) provides evidence in support of Beneficence Proposition (BP) 1, and the high proportion of first movers who send money a testament to that there wide recognition that others can be relied upon to apply it (McCabe and Smith, 2000, Cox and Deck, 2005, Gillies and Rigdon, 2017). The Ultimatum Game (Harsanyi, 1961, Güth et al., 1982) provides some evidence for the propositions. However, X fails to do something good or does something bad, and therefore, whether BP2 or Injustice Proposition (IJ) 1 is often a matter of framing. In a binary version of the game in which an 8,2 offer was one option, and the other option was varied, 8-2 was seen as good or bad depending upon the other option; thus, selection and rejection rates differed (Falk et al., 2003). List (2007), Bardsley (2008) and Korenok et al. (2014) all show difference between giving and taking in dictator games.

Kimbrough and Vostroknutov (2016) and Kimbrough and Vostroknutov (2018) provide evidence that there is individual variation in the propensity to follow rules, and provide methods to measure a proxy of this

parameter.Additionally, they show that there is a correlation between the propensities to follow and enforce norms. Akerlof and Kranton (2000) provide extensive theory regarding how characteristics of the individual and the situation impact adherence to and enforcement of norms. Their model covers a much broader scope than Kimbrough and Vostroknutov's and therefore appears very different. However, the models are compatible and may be two sides of the same coin. Akerlof and Kranton's model implies that Kimbrough and Vostroknutov's elicited social norms vary not only by the individual but by role in society and situation. Despite this variation, we would still expect the correlation Kimbrough and Vostroknutov find; within a group with some power, the members that are at the periphery are most likely to conform to the norms of the group (because they are most at risk of losing membership) and most likely to enforce norms particularly to exclude non-members (because they gain the most from membership).

3 Theory and Model

In this section, we extend the models reviewed in the previous session to propose a model for norms regarding prospect selection. Our model keeps the spirit of Kimbrough and Vostroknutov (2016) while dealing with decisions that cannot be mapped to a single axis. Prospects at the very least involve two axes: expected value and variance but could include other axes to account for higher-order risk preferences. Our approach will not attempt to specify which axes are relevant. For simplicity, we limit the number of prospects to two, A and B. Asked to determine a norm about which to choose, the Judge who finds pros and cons to both implicitly references a third (unavailable) prospect, perhaps a portfolio of A and B, which we will call C.

The utility from norm adherence for both option A and B can be calculated as to how close they are to C, in the various dimensions, x_i and those dimensions can also be given weights φ_i to produce an overall score:

$$u_{norms}(\alpha) = \sum_i \varphi_i g_i(x_i^\alpha - x_i^C),\ \alpha = \{A, B\}$$

The utility from norm adherence for both options A and B can be calculated as to how close they are to C in the various dimensions, x_i, and those dimensions can also be given weights φ_i to produce an overall score:

> **Hypothesis 1:** Subjects will be able to coordinate on how socially appropriate and deserving of punishment or requiring reward each prospect is.

In most Principal-Agent situations involving prospects, the Agent makes a decision before the prospect is realized. At the same time, the Principal's income is impacted by that realization, so they evaluate the decision post realization. As such, we want to elicit norms at both points. However, information about the potential realization conflicts with information about the choice, e.g., what was a good decision might have a bad outcome, leading to:

> **Hypothesis 2:** Elicitation of social norms in which the outcome is not known will yield better coordination than elicitation of norms in which only the action is known.

Assuming that subjects converge on norms,[4] the next goal of the project is to examine guiding Principals to those norms. Having elicited $u_{norms}(A)$ and $u_{norms}(B)$, and specifying that inaction results in Prospect A, the sign of $u_{norms}(B) - u_{norms}(A)$ will establish if switching to Prospect B is consistent with, or contrary to norms. The subsequent two hypotheses follow from Smith's Beneficence and Injustice Propositions. Also, note that because these propositions distinguish action from inaction, this necessitates a design that also makes such a distinction.

[4] As stated in our preregistration, we planned to run the Judges experiment first to ensure that there was convergence on norms before proceeding to the next experiment, which serves little purpose if there were no norms. However, they were conceived as a single project.

Hypothesis 3: If the Agent changes the Principal initial investment to a better (worse) investment, the Principal will reward (punish) the Agent.

Hypothesis 4: If the Agent does not change the status-quo, the Principal will not reward or punish the Agent.

Comparing choices Agents make for themselves to choices they make for others will illuminate how norms might drive any difference in choices. This indicates the design should include treatment in which Agents make choices for themselves rather than Agents. Polman and Wu's meta-analysis (2020) leads to:

Hypothesis 5: When making choices for Principals, who cannot affect the Agent's pay, Agent choices will be closer to risk-neutral than when making choices for themselves.

Situations and contracts vary, so they may or may not allow the Principal to impose consequences (positive or negative) on the Agent. This ability to impose consequences has an ambiguous impact on the role of norms; it might increase scrutiny of the action per Levitt and List (2007) and increase the impact. However, it has also been suggested that moral accounting is separate from financial accounting In order to resolve this ambiguity, our experiment will vary the Principals' ability to impose financial consequences.

Hypothesis 6: When Principals can affect Agent pay, Agent choices will be guided by the social appropriateness norms regarding the decision.

As the design will feature action versus inaction, and vary both whether the choice is for the Agent or a Principal, and if for a Principal whether that Principal can impose financial consequences on the Agent, we offer hypotheses, regarding inaction.

Hypothesis 7: Agents who are making decision for themselves rather than a Principal will be more likely to consider acting and explore the action space

Hypothesis 8: Agents who are making a decision for themselves rather than a Principal will be more likely to consider acting and exploring the action space

Finally, as we will allow Principals to impose consequences on Agents, we will explore how norms guide Principals' responses to Agent decisions. To be consistent with the Judges' treatment, we elicit Principals' responses to the social appropriateness of the Agent's decision and whether they deserve a reward or a punishment before and after showing the realization of the prospect.

Hypothesis 9: When Principals can affect Agent pay, Principal choices to reward or punish will be guided by the norms regarding the decision.

We design a set of experiments to test these hypotheses.

4 Design

The core of the experiment is a Principal-Agent game, for which we elicit social norms (from 'Judges') regarding the appropriateness of Agents' actions. All treatments are implemented between subjects. We use the sessions, we used neutral terms, e.g. "another participant". However, throughout the paper, we use the terms Principal, Agent, and Judge. The Principal is endowed with a risky prospect, and the Agent can switch it for another risky prospect. The Principals are paid based upon the Agents' choices and random draws. Table 1 summarizes the experimental design. In the Consequences treatment, the Principals can (at a cost) adjust the Agents' pay to punish or reward them. This primary investigation tests whether social norms guide Principals' pay adjustments and thereby guide the decisions of Agents who anticipate the adjustments.

Table 1: Experimental Design

	Self	**Consequences**	**Other**
Agent	Make 6 choices between pairs of prospects		
Nature	Select one choice and realize value of chosen prospect		
Principal	NA	Adjust Bonus	See Outcome
		See Outcome	Rate Social Appropriateness
		Re-Adjust Bonus	Rate Deserve Punish/Reward
Pay Agnt.	Realized prospect	Adjusted Bonus	Random prospect
Pay Prnc.	NA	Realized prospect	Realized prospect

The experiment also includes a treatment in which the Principals cannot adjust the pay of the Agents; comparing the decisions in the 'Other' treatment to those in the Consequence treatment allows us to test anticipated pay adjustments impact decisions. We also include a treatment in which the Agents are their own Principals; the Self treatment tests if Agents select different prospects when the prospect will determine their own pay rather than other participants' pay. Finally, we vary whether the pool of Judges knew the lot drawn within the prospect or simply the prospect the Agent selected for Principal in order to explore how intent (expectation) versus outcome impact judgment.

The experiment consists of six pairs of lotteries. Which lottery, within a pair, is the initial endowment (default) is varied to create flipped pairs and test (norms and payment adjustments of) action versus inaction. Smith's Beneficence and Injustice Propositions predict differing consequences when an Agent "does something" from when the Agent does not do the opposite.[5] Table 2 shows details of the lotteries used. In general, Option B was the more rational (risk-neutral) choice, though to varying degrees. Figure S.1 compares the utility of each option (within the pairs) as risk preference varies. Pair 1 was expected to provide the strongest norm; Option B first-order stochastic dominates (FOSD) Option A. Pair 4 is Pair 2, with every outcome on both options reduced by \$0.30, which creates the possibility that the lower outcome from one option is a loss from the initial \$0.50 endowment, so also expected to provide a strong norm. Figure S.6 shows an example of the Agent decision screen. In order to create a strong sense of inaction versus action, in which Option A represented inaction, the terms of Option B are not initially visible. The Agent had to click on the image with the question mark. Once the image was clicked, it changed to a plot akin to the one for Option A, and the text alongside the image appeared. The Agent could only select Option A before the terms of Option B were revealed. After revealing Option B, either could be selected. An example like this was presented to all participants as part of the instructions. They could not proceed until they had clicked the question mark to ensure they understood. After the experiment, all participants completed a ten-question Big Five Personality Test.[6]

Table 2: Lottery Pair Details

Pair	Option A					Option B					Note
	Pr. 1	Pay 1	Pr. 2	Pay 2	EV	Pr. 1	Pay 1	Pr. 2	Pay 2	EV	
1	0.75	\$0.90	0.25	\$0.10	\$ 0.70	0.75	\$1.15	0.25	\$0.15	\$0.90	FOSD
2	0.75	\$1.10	0.25	\$0.30	\$ 0.90	0.75	\$1.35	0.25	\$0.00	\$1.01	↑ EV & Variance
3	0.75	\$1.00	0.25	\$0.20	\$ 0.80	0.75	\$0.96	0.25	\$0.32	\$0.80	EV Equal, ↓ Var.
4	0.75	\$0.80	0.25	\$0.00	\$ 0.60	0.75	\$1.05	0.25	-\$0.30	\$0.71	↑ EV, $pay_2 < 0$
5	0.75	\$1.10	0.25	\$0.30	\$ 0.90	0.50	\$1.80	0.50	\$0.45	\$1.10	↑ EV & pr(Min)
6	0.75	\$0.90	0.25	\$0.10	\$ 0.70	0.25	\$1.45	0.75	\$0.45	\$0.70	↑ Max & pr(Min)

[5] Author's personal communications with Vernon Smith confirmed that the propositions concerned action rather than cost.
[6] https://www.ocf.berkeley.edu/~johnlab/bfi.htm

The experiment is preregistered on aspredicted.org as #83458, #83631 and #84401. It was programmed in oTree (Chen et al., 2016), and run on prolific.co.

4.1 Judges

We ran the Judge role first to check that there was convergence on social norms before we ran the other roles. Judges rated 24 decisions, both Options for all 6 pairs and 6 flipped pairs. Hot Judges only rated one outcome per decision. They were asked to rate each decision on two scales: one, how socially acceptable it was; two, whether the decision deserved punishment or required reward. Both were on seven-point scales between [-3, 3] with zero as neither. Figure S.7 shows an example of a decision screen. One set of Judges, the 'cold' Judges (N=120), only saw the Agent decision. The other 'hot' set (N=240) saw the Agent decision, as well as the payment the Principal would receive because of the decision and lottery realization. We only asked a Judge to rate one prize per possible decision, so we doubled the number of Judges. Judges could earn up to \$6.00 in bonus payments. One decision was randomly chosen for each Agent; for each scale, if their rating matched were the modal response, they earned \$3.00. We recorded time spent on each page and also checked if participants who finished particularly quickly ratings showed a lack of variation across decisions.[7] We detected no evidence of decision fatigue, so we used all responses.

4.2 Agents

Agents choose one lottery out of each pair for a total of six decisions. Figure S.8 shows an example decision screen after the Agent had clicked on Option B's image and revealed its terms. Agents either saw Pairs or the Flips but not both; thus, crossed with the three treatments, there were six cells of Agents, with 120 per cell and 720 total. Agents were paid \$0.60 for completing the six decisions. Bonus payments varied by treatment. For the Self treatment, in which they made choices for themselves, one pair was selected, and the Agents were paid a realization of their chosen lottery plus an endowment of \$0.50. For the Other treatment, in which they knew one of their decisions would determine a Principal's pay, they were paid the same as a randomly drawn Principal. For the Consequence treatment, the default bonus pay was \$0.60, but it could be adjusted by the Principal and varied from \$0.00 to \$1.20.

4.3 Principals

There were 480 Principals, one for each Agent (aside from the Self treatment). They earned \$0.30 for participating and were paid a bonus of a realization of the lottery selected for them plus an endowment of \$0.50. In the Other treatment, the Principal learned the Agent's choice and their own pay and then made the same ratings of their Agents' decisions as the Judges. In the Consequences treatment, the Principal learned the Agent's decision and then was asked to select a bonus for the Agent for each of the two possible payment outcomes. The default bonus to the Agent was \$0.60. However, in \$0.20 increments, the Principal could decrease the bonus to \$0.00 or increase it to \$1.20, at a cost to the Principal of one-fifth of the (absolute value of the) adjustment. Figure S.9 shows an example of this decision screen. On the next screen, the Principal saw the realization of the lottery and resulting pay and was allowed to change the adjustment to the bonus. Figure S.10 shows an example of this decision screen. The decision from the previous screen for the lottery realization was pre-checked, but any adjustment could be chosen.

5 Results

5.1 Judges

Prolific provides elapsed time between when a participant accepts an invitation until they submit a completion code. The mean completion time for the cold Judges was 646 seconds. Cold Judges were paid a completion fee of \$1.50 and a bonus based on matching modal ratings. The mean bonus payment was

[7] The authors thank Erin Krupka for these design suggestions.

$1.88. The mean completion time for the hot Judges was 1105 seconds. Hot Judges were paid a completion fee of $1.25 and a bonus based on matching modal ratings. The mean bonus payment was $1.74.

Figure 1 displays histograms of how Judges who did not know outcomes rated Agent decisions. The ratings of 'extremely' social inappropriate or deserving of punishment are coded as -3, 'moderately', and 'slightly' as -2 and -1, respectively. Neither is coded as zero. Socially appropriate and requiring rewards are coded with positive values of equivalent intensities. There are clear modes in each panel. The mean proportion of subjects selecting the modal response is 0.30 for social appropriateness and 0.37 for deserving of punishment or requiring reward, with some panels having coordination as high as 0.5.

> **Result 1:** Consistent with Hypothesis 1 Judges are able to coordinate on norms of how socially appropriate and deserving of punishment or requiring reward Agent decisions are.

Within a "Pair" and its "Flip", opting for A in the pair is selecting the same lottery as opting for B in the Flip. The Pair is directly above the Flip. The only difference is that in the former it was the default, or option the Principal started with. If the starting option (default) was not relevant, we would expect the Flip plots to look very similar to the Pair plots above them. However, in many cases switching the default changes how the action is rated. For example: in Pair 1, selecting Option A (even though B is FOSD) has modal ratings of slightly socially inappropriate (the highest blue bar is -1) and neither deserving punishment nor requiring reward (the highest pink bar is 0), supporting BP2; yet, in Flip 1 selecting Option B has the same modal rating for social appropriateness, but now has a modal rating of slightly deserving punishment, supporting IP1. Both rating distributions shift clearly to the left. Wilcoxian test have respective p-values of 0.002 and 0.008. Similarly, in Pair 4, they're opting for Lottery B, which has the potential of a loss, has a modal rating of slightly socially inappropriate and slightly deserving punishment, supporting IP1; yet in Flip 4 Option A, not switching the Principal away from the same lottery, has a modal rating of neither on both dimensions supporting BP2. Wilcoxian test have respective p-values of 0.022 and 0.023. The modal rating of Pair 1 Option B switches the Principal to the better lottery as moderately requiring reward is consistent with BP1. The modal rating of Pair 4 Option A of not switching the Principal to the lottery with a potential loss as neither is consistent with IP2.

> **Result 2:** Consistent with Hypotheses 3 and 4, cold ratings from Judges are consistent with Beneficence and Injustice Propositions.

Figures 2 and 3 are akin to Figure 1 except they are the ratings from Judges who knew what lottery realization (payment) the Principal would get; they saw the text in a red box in Figure S.7. Judges rated both outcomes from each lottery; there are separate histograms for each prize. In all cases, Prize 1 has a greater value than Prize 2. There is a clear pattern of the distributions for Prize 2 shifting to the left (relative to the distributions for Prize 1). Pair 4 Option B provides one of the most extreme divergences in payment and thus has the most divergence in ratings. When the prize is $1.05 the modal rating is that the choice is slightly socially appropriate and moderately requiring reward. However, when the prize is -$0.30, the choice is rated as slightly socially inappropriate and slightly deserving of punishment.

> **Result 3:** Consistent with Hypothesis 2, we find stronger coordination on norms when Judges did not know outcomes than when they did.

To understand how the two sets of ratings relate to each other, Table 3 regresses how Judges, who knew outcomes (of lottery draws) rated Agent decisions on mean rating by Judges who did not know outcomes. We use a panel regression and cluster errors on Judges. The latter 'cold' ratings have strong predictive power ($p < 0.001$) of the former 'hot' ratings once the value of the lottery prize is added as an independent variable. Hot ratings from Judges appear to primarily be based on the intent of the Agent and secondarily upon outcomes; coefficients on the cold rating are .79, whereas those on prizes are lower. Also, note the range of ratings (-3, 3) is greater than that of prizes (-.3, 1.8).

> **Result 4:** Hot ratings from Judges take into account both Agent actions and prize values.

Figure 1: Judges Cold Ratings of Agent Decisions

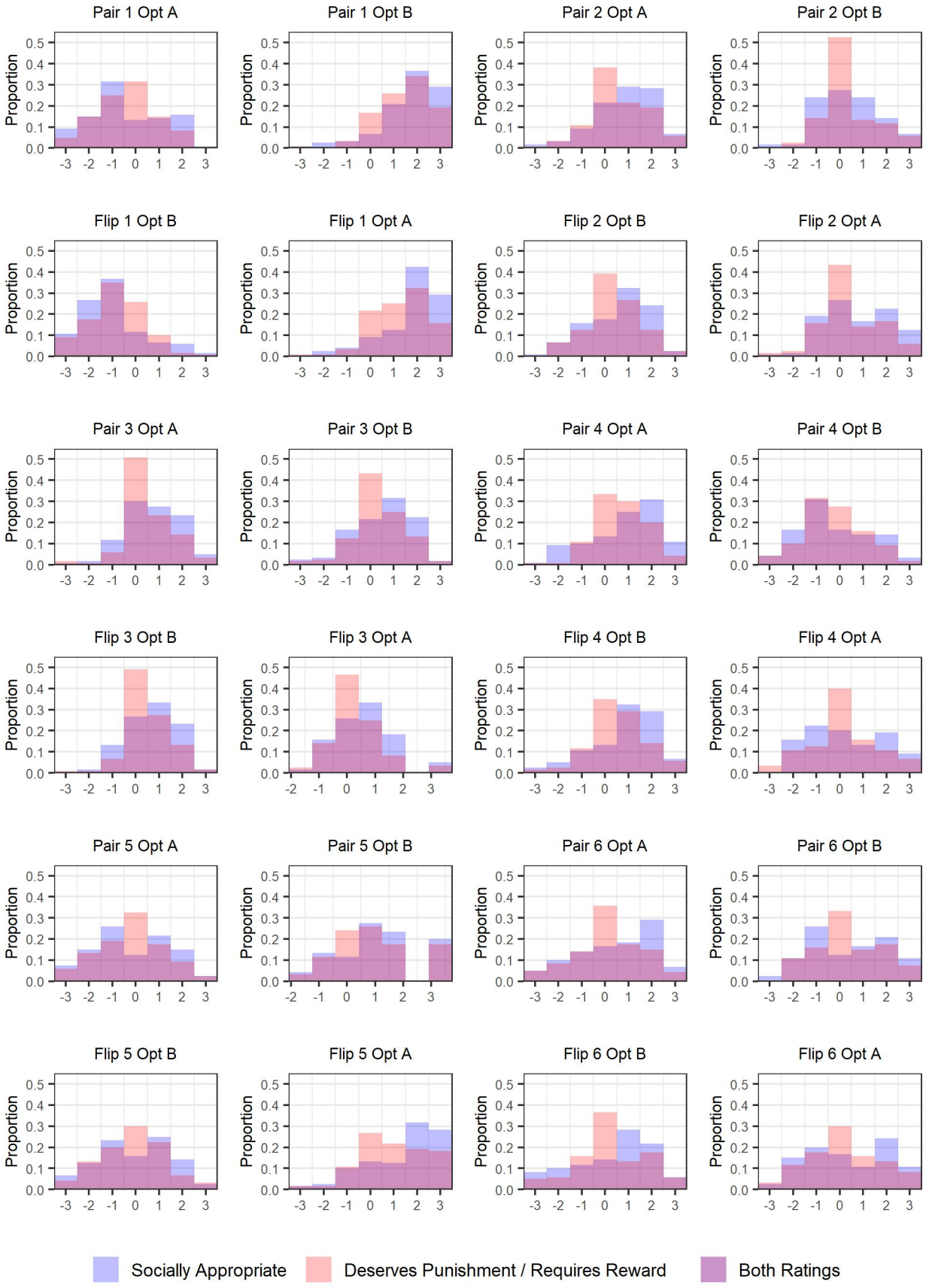


Figure 2: Judges Hot Ratings of Agent Decisions

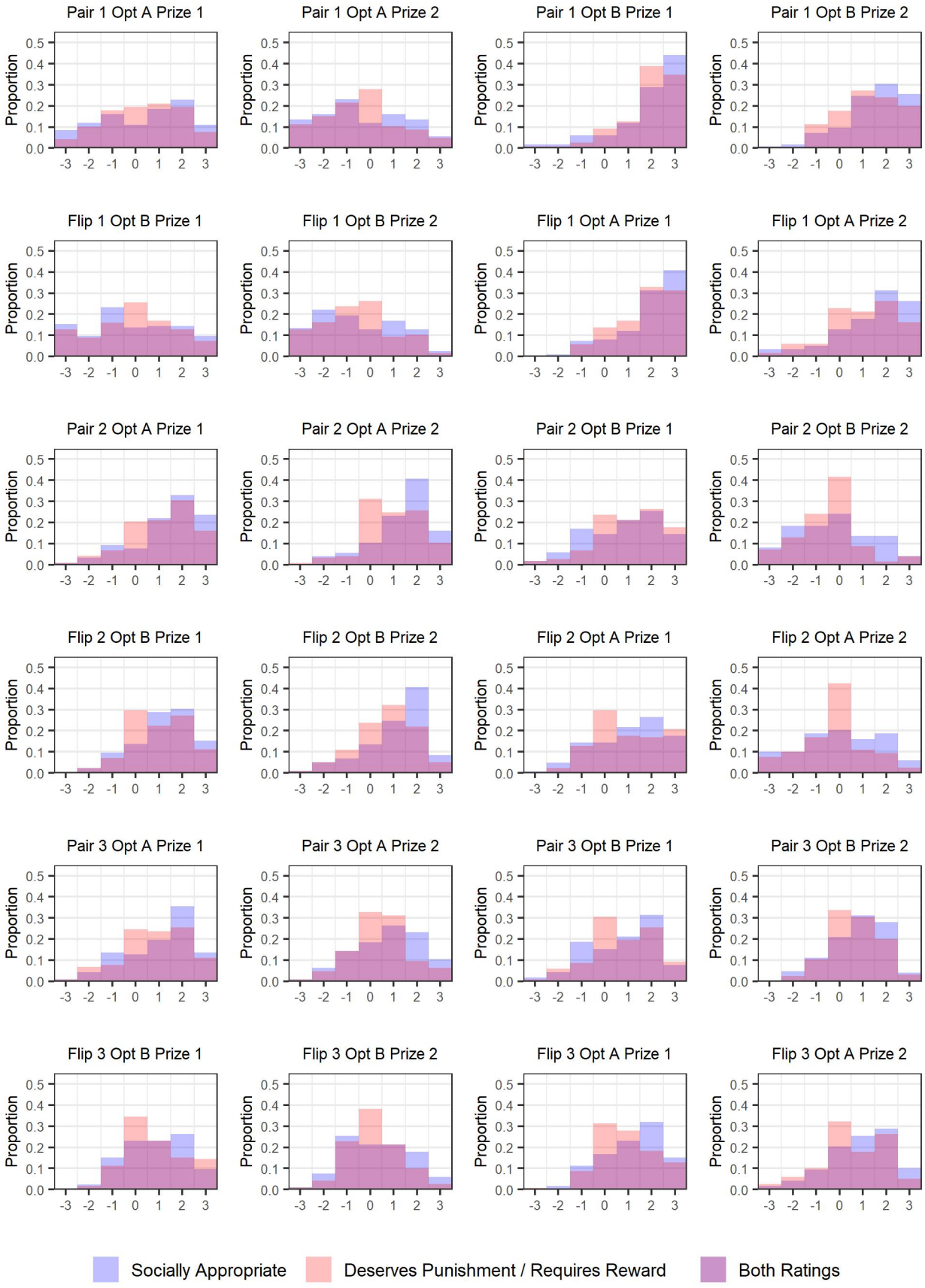


Figure 3: Judges Hot Ratings of Agent Decisions, cont

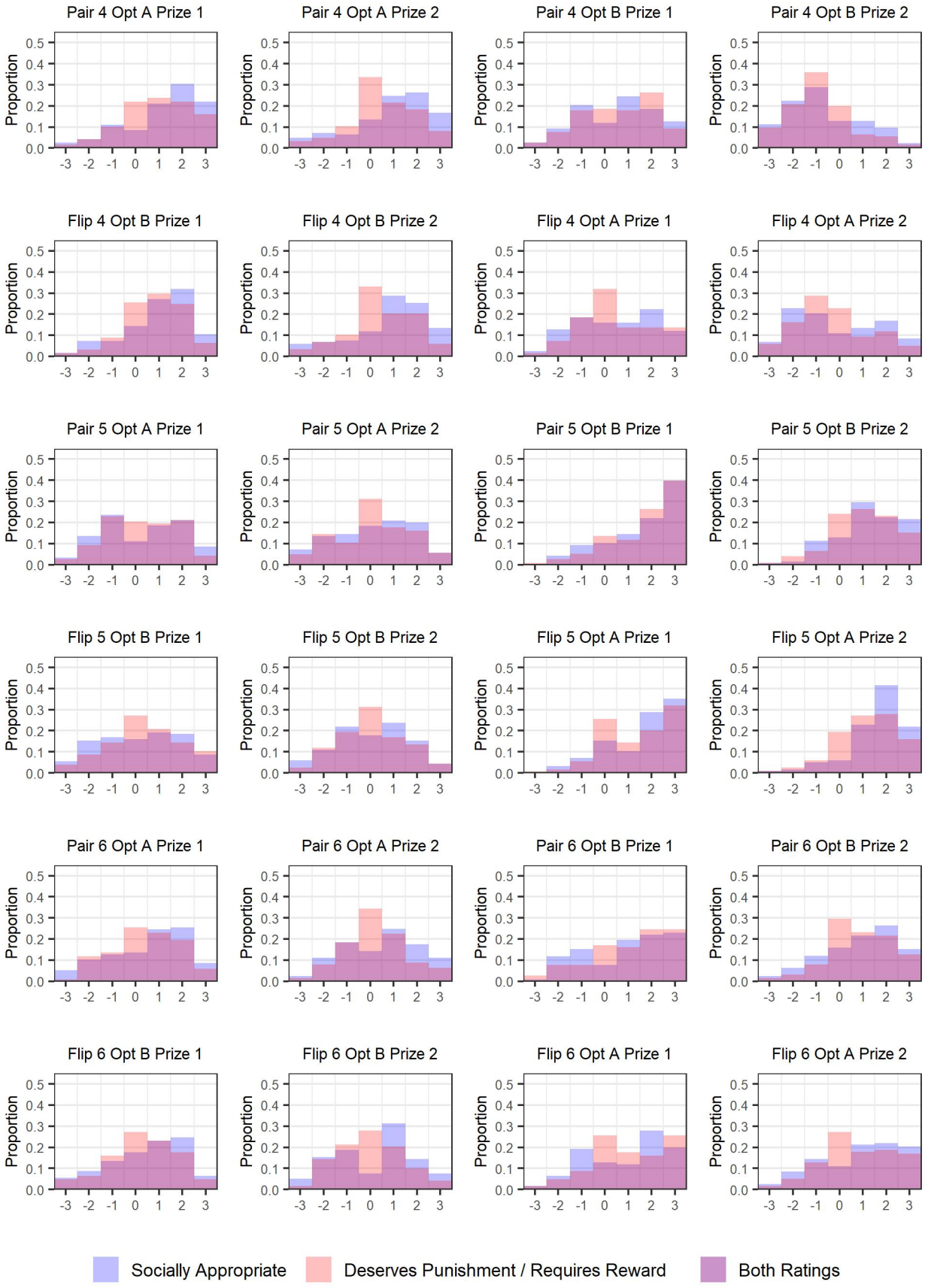


Table 3: Regression of Ratings from Hot Judges on Ratings from Cold Judges

	(1) Soc. App. Opt. A	(2) Soc. App. Opt. B	(3) Dev. Bonus Opt. A	(4) Dev. Bonus Opt. B
Soc. App. Opt. A Cold	0.794***			
	(0.0588)			
Soc. App. Opt. B Cold		0.775***		
		(0.0549)		
Dev. Bonus Opt. A Cold			0.793***	
			(0.0683)	
Dev. Bonus Opt. B Cold				0.805***
				(0.0589)
Prize A	0.371***		0.527***	
	(0.0666)		(0.0663)	
Prize B		0.466***		0.687***
		(0.0671)		(0.0664)
Constant	0.0887	0.000463	-0.0158	-0.143**
	(0.0711)	(0.0661)	(0.0593)	(0.0539)
Observations	2916	2916	2916	2916
Number of IDs	243	243	243	243
Chi Squared	232.1	274.4	228.1	340.9

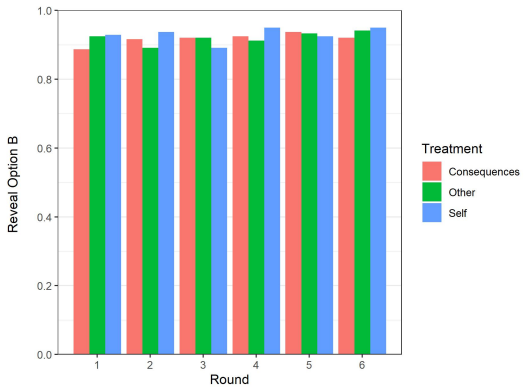
Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.2 Agents

Figure 4 graphs by treatment and across round, the proportion of Agents who clicked on the stand-in image of Option B to reveal the true image and lottery terms. There is no decline across rounds, nor is there any difference by treatment; Agents made just as much effort across all treatments. Prolific provides elapsed time between when a participant accepts an invitation until they submit a completion code. The mean time for the Self Treatment is 187 seconds, whereas mean time for the Consequences and Other Treatments are 231 and 226 seconds, suggesting Agents spent more time deliberating decisions that impacted Principals pay than ones that impacted their own pay. Wilcoxon rank-sum (Mann-Whitney) tests indicate that all three differences are statistically significant, p-values < 0.001.

Figure 4: Proportion of Agents Revealing Option B by Treatment



Result 5: Contrary to Hypothesis 7 and 8 Agents exert as much effort and greater time in making selections for others as for themselves. Neither substantial increase with the Principals' ability to impose financial consequences.

Figure 5 graphs the proportion of Agents selecting Option B. The Pairs are plotted in the left panel, and the Flips of the Pairs in the right. Option B is the more risk-neutral choice for the left and the risk-averse choice for the right. The Y-axis is reversed in the right panel to reflect that selecting Option B in the Flip is equivalent to selecting Option A in the Pair. The reversed axis also illustrates that in most cases the bars fail to meet. For example: in Pair 2 Consequences, the selection rate for Option B is about 0.25; in Flip 2 Consequences, it's about 0.55. If the selection was driven entirely by the preference for one lottery over another, the expected sum is one. While any two bars among the 18 in each panel may vary across all 18, the expectation should hold. There is a clear pattern of a sum less than one. The implication is that many Agents stuck with the default because it was the default. Some of these Agents simply never revealed Option B (compare the gap between the two bars to the gap between the bar and the top of the axis in Figure 4). However, for Pair 2, a larger proportion revealed Option B. A Wilcoxon rank-sum (p-value = < 0.001) indicates the default is more likely to be chosen.

Figure 5: Proportion of Agents Selection Option B by Treatment

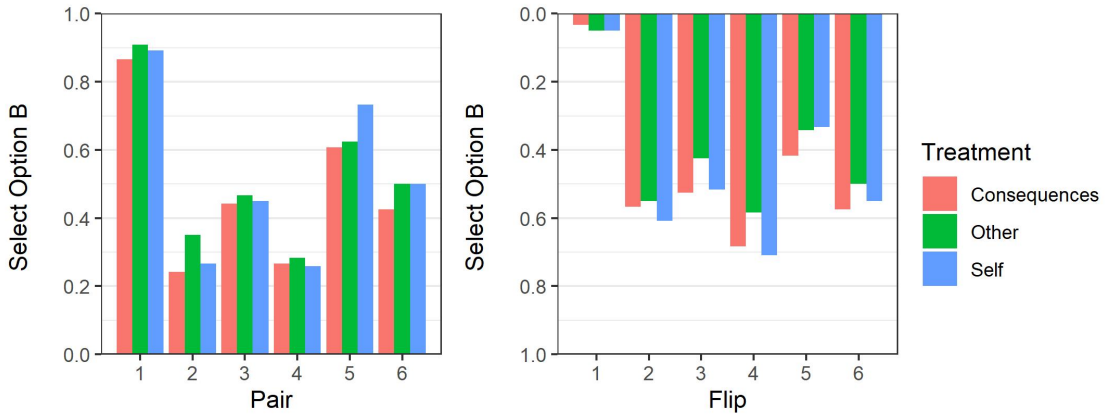


Table 4 reports estimated marginal effects for the likelihood that an Agent selects Option B. Column 1 is restricted to Pairs, and Column 2 is restricted to Flips. In each, Pair (Flip) 1 is the reference decision. Results indicate relative to Pair (Flip) 1, selecting Option B in other Pairs (Flips) is statistically significantly less (more) likely. The estimates indicate there are statistically significant differences across Pairs (Flips) and confirm the visual difference seen in Figure 5. However, treatment is only statistically significant for the Flips in the Other treatment, in which Option B is 6% less likely to be chosen.

> **Result 6:** Consistent with Hypothesis 5, Agents make more risk-neutral choices for Principals than for themselves. However, if the Principal can impose financial consequences choices are much closer to Agent's risk-averse self choices.

Table 5 reports marginal effects from a probit regression testing how social norms impact the likelihood that an Agent selects Option B. Selections from the Self treatment are not included. Ratings (from Judges who did not know the lottery realizations) as to how socially appropriate the selection and whether making the selection deserves punishment or requires reward are both good predictors of selection. Higher ratings of Option A make selecting B less likely. Higher ratings of Option B make selecting B more likely. Specifications 2 and 5 difference the ratings of A and B to create a single variable. However, when an indicator for the Consequence treatment is interacted with the differences, in specifications 3 and 6, the variables are not statistically significant, indicating that a Principal's ability to impose financial consequences does impact the importance of norms. There is also some evidence that the Consequence treatment is associated with more risk-averse selections. The estimate for the indicator is negative, meaning Option B is less likely to be chosen in the Pairs and fairly consistent across all the specifications though only statistically significant in the final specification. The variable interacting indicators for the treatment and Flips is positive, meaning in the Flips, it is more likely to be chosen. The sum of the first three variables is the estimate of the likelihood.

Table 4: Marginal Effects from Probit Regression on Agents Likelihood to Pick Option B by Treatment

	(1) Pairs	(2) Flips
Pair=2	-0.603***	0.530***
	(0.0291)	(0.0289)
Pair=3	-0.437***	0.444***
	(0.0287)	(0.0285)
Pair=4	-0.620***	0.614***
	(0.0287)	(0.0271)
Pair=5	-0.234***	0.319***
	(0.0274)	(0.0281)
Pair=6	-0.415***	0.497***
	(0.0306)	(0.0281)
Consequences	-0.0489	0.00510
	(0.0309)	(0.0275)
Other	0.00717	-0.0597*
	(0.0291)	(0.0263)
Observations	2160	2160
ChiSquared	340.6	284.0
NumberIDs	360	360

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Result 7: Consistent with Hypothesis 6, Agent lottery choices are strongly influenced by social norms regarding both the appropriateness of action and whether it deserves punishment or requires reward.

Table 6 is akin to Table 5, but tests if Agent decisions are loss averse regarding the potential impact on bonuses, so it only includes selections from the Consequences treatment. It uses ratings from Judges who know the realizations of the lotteries. Each lottery had two possible outcomes; the ratings of the selection when the realization was the higher (lower) value prize are labeled 'High' ('Low'). The regression shows that the ratings associated with the higher value prize are more predictive of selection than the rating associated with the low value. This could be explained by the fact that the higher value prizes were generally more likely; therefore, they should be given greater weight. However, across all Pairs and Flips, the probability of High was 0.6875, only 2.2 times the probability of Low, and the ratios of coefficients are much higher, suggesting upside bias in Agents' choices rather than loss aversion which would give less weight to high outcomes.

Mean prizes (realizations of the prospect) for the Consequence, Other and Self treatments were \$0.818, \$0.828 and \$0.824, respectively. A Wilcoxon rank-sum tests for differences were not statistically significant; p-values for Consequences Treatment versus Other Treatment is 0.506, for Other versus Self is 0.694, and for Consequence versus Self is 0.798.

Table 5: Marginal Effects from Probit Regression on Agents Likelihood to Pick Option B

	(1)	(2)	(3)	(4)	(5)	(6)
Consequences	-0.0469	-0.0468	-0.0491	-0.0472	-0.0470	-0.0506*
	(0.0286)	(0.379)	(0.0256)	(0.0250)	(0.0249)	(0.0250)
Flip Opt A & B	0.0178	0.0255	0.0218	0.0354	0.0338	0.0302
	(0.0237)	(0.203)	(0.0236)	(0.0234)	(0.0233)	(0.0238)
Consequence * Flip	0.104*	0.104	0.112**	0.104**	0.104**	0.111**
	(0.0440)	(0.847)	(0.0373)	(0.0339)	(0.0339)	(0.0350)
Social Appropriate, Opt. A	-0.105*					
	(0.0448)					
Social Appropriate, Opt. B	0.233***					
	(0.0693)					
Diff in Soc. App. (A - B)		-0.173	-0.166***			
		(1.418)	(0.0220)			
Cons * Diff in Soc. App.			-0.0129			
			(0.0158)			
Deserves Bonus, Opt. A				-0.153**		
				(0.0492)		
Deserves Bonus, Opt. B				0.265***		
				(0.0413)		
Diff in Des. Bonus (A - B)					-0.214***	-0.206***
					(0.00737)	(0.0127)
Cons * Diff in Des. Bonus						-0.0140
						(0.0193)
Observations	2880	2880	2880	2880	2880	2880
Chi Squared	482.3	492.3	506.3	479.7	484.2	494.0
Number of IDs	480	480	480	480	480	480

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Marginal Effects from Probit Regression on Agents Likelihood to Pick Option B

	(1)	(2)
Flip Opt A & B	0.177***	0.214***
	(0.0320)	(0.0349)
Social Appropriate Opt. A, High	-0.252***	
	(0.0598)	
Social Appropriate Opt. B, High	0.217***	
	(0.0554)	
Social Appropriate Opt. A, Low	0.109	
	(0.0607)	
Social Appropriate Opt. B, Low	0.0654	
	(0.0439)	
Deserve Punish / Require Reward Opt. A, High		-0.268***
		(0.0621)
Deserve Punish / Require Reward Opt. B, High		0.143**
		(0.0472)
Deserve Punish / Require Reward Opt. A, Low		0.0135
		(0.0531)
Deserve Punish / Require Reward Opt. B, Low		0.0762
		(0.0392)
Observations	1440	1440
Number of IDs	240	240
Log Likelihood	-877.0	-883.1

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.3 Principals

Figure 6 displays histograms of how Principals report they will adjust their Agent's bonus based on the Agent's lottery choice. Principals reported two values, one for the lottery's high realization (green) and a second for the lottery's low realization (red). Zero indicates no adjustment; the Agent's bonus is the default \$0.60. Each level of adjustment is a \$0.20 change to the Agent's bonus and costs the Principal \$0.04. At the limits, at the cost of \$0.12, the Principal could reduce the Agent's bonus to \$0.00 or double it to \$1.20. The most notable feature of the plots is the lack of overlap between the two bonuses.

Table 7 reports results from regressing the bonuses plotted in Figure 6, on Judges' ratings of whether the choice deserved punishment or required reward. Each Principal made two decisions, and the regressions cluster errors on Principals. Columns 1 and 2 use ratings from Judges who did not know the realization of the lottery, and Columns 3 to 5 use ratings from Judges who knew. Columns 2 and 4 use differences in the rating of selecting Option A and selecting Option B, rather than each rating as a separate variable. Column 5 repeats Column 4 but adds the value of the lottery as an independent variable. Lottery realizations have a more profound impact on bonuses to Agents than the Agents' lottery choice. Bonus adjustments are not a continuous variable; Table S.5 reports all the regressions from Table 7 run as ordered probits. The inferences are equivalent.

Figure 7 plots the adjustments to the bonus the Principals reported they would give their Agent after learning which lottery the Agent chose for them, but before learning the realization of that lottery on the horizontal axis, and the adjustments they actually made after learning the realization on the vertical axis. Principals reported adjustments for both possible realizations; only the adjustment for the realization the Principal eventually received is plotted. There is a limited number of possible responses; circle size is the number of participants choosing the combination. The 45-degree line represents no change to the adjustment after learning the realization. Most participants are on the line. Points above (below) represent more

Figure 6: Principal Cold Bonuses to Agents

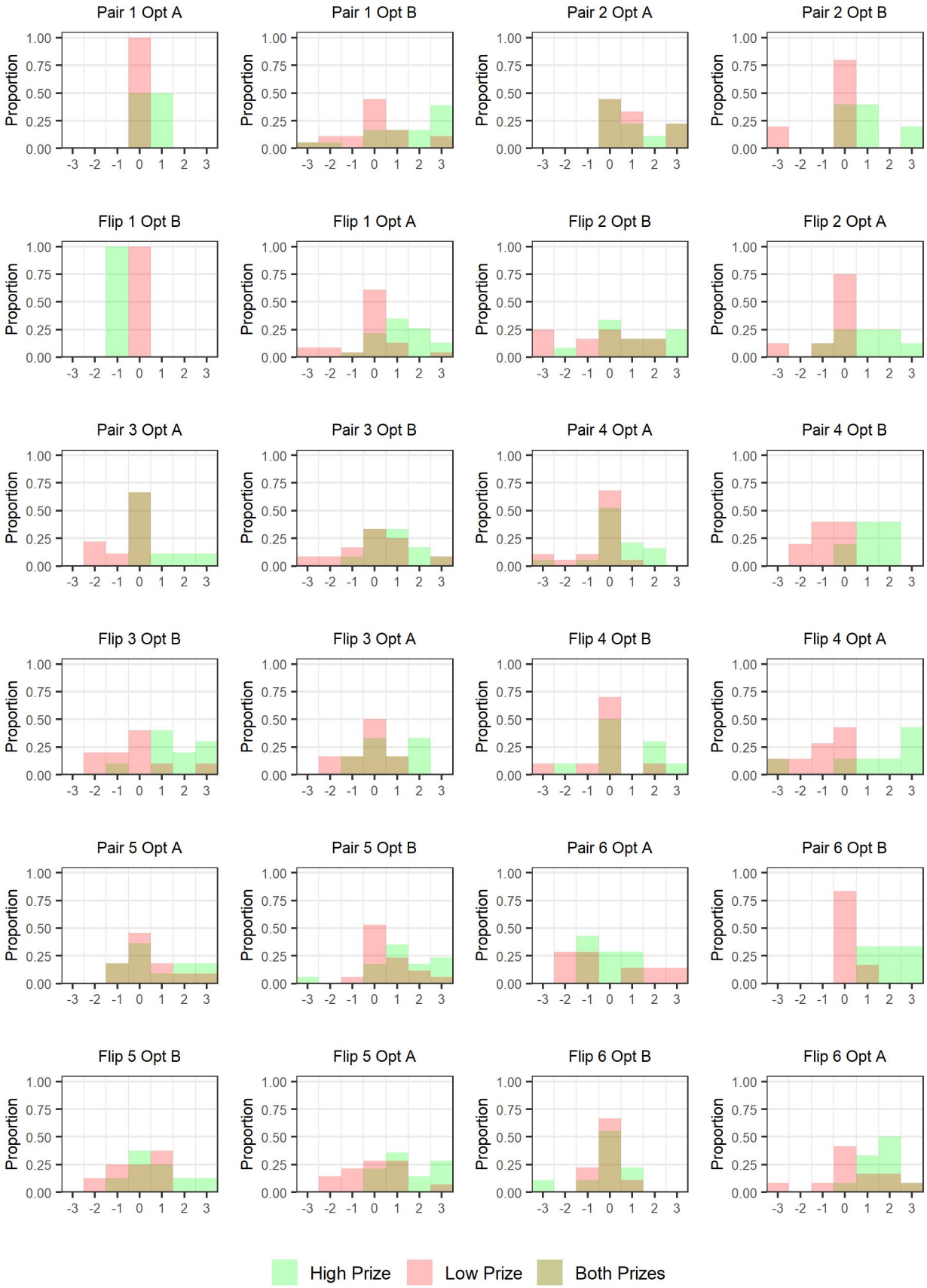


Table 7: Regression on Principals Cold Bonuses to Agents

	(1) Bonus, Cold	(2) Bonus, Cold	(3) Bonus, Cold	(4) Bonus, Cold	(5) Bonus, Cold
Deserves Bonus of Pick	0.662				
	(0.367)				
Deserves Bonus of Not Pick	0.478				
	(0.333)				
Diff in DB (Picked - Not)		0.0911			
		(0.0673)			
Deserves Bonus of Pick, Hot			0.242		
			(0.158)		
Deserves Bonus of Not, Hot			-0.0876		
			(0.132)		
Diff in DB (Picked - Not) Hot				0.162*	0.0808
				(0.0730)	(0.0716)
Lottery Prize, Cold					1.111***
					(0.0970)
Constant	0.0166	0.413***	0.285	0.393***	-0.332***
	(0.249)	(0.0730)	(0.194)	(0.0727)	(0.0927)
Observations	480	480	480	480	480
Number of IDs	240	240	240	240	240
Chi Squared	3.693	1.832	5.049	4.918	135.5

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(less) generous adjustments after learning. There are more instances of increased generosity than decreased generosity.

> **Result 8:** Principals' decisions regarding bonuses to Agents do not vary much between the cold and hot states. There is more variation in the direction of increased generosity than decreased generosity.

Table 8 reports regressions results for the Principal's implemented bonus decisions. Column 1 shows that Judges' cold ratings still are not statistically significant predictors of bonuses. The remaining columns use Judges' hot ratings. Column 2 tests for leniency in punishments. The first two variables are based on Judges' hot ratings of selecting that option. Deserves Punishment includes only negative ratings (not positive ratings were coded 0), and Requires Reward includes only positive ratings. If Principals were lenient on Agents whose actions deserved punishment but delivered all required rewards. The estimate latter would be larger than the former. Neither estimate is statistically significant. Column 3 indicates hot ratings from Judges have predicted both Principal's hot bonus decisions. However, once lottery prize values are included (Column 4), they are no longer statistically significant. Bonus adjustments are not a continuous variable; Table S.6 reports all the regressions from Table 8 run as ordered probits. The inferences are equivalent.

> **Result 9:** Contrary to Hypothesis 9, lottery realizations have a more profound impact on bonuses to Agents than social norms regarding Agents' actions.

Figure 7: Principals Bonuses to Agents before and after Learning Lottery Realization

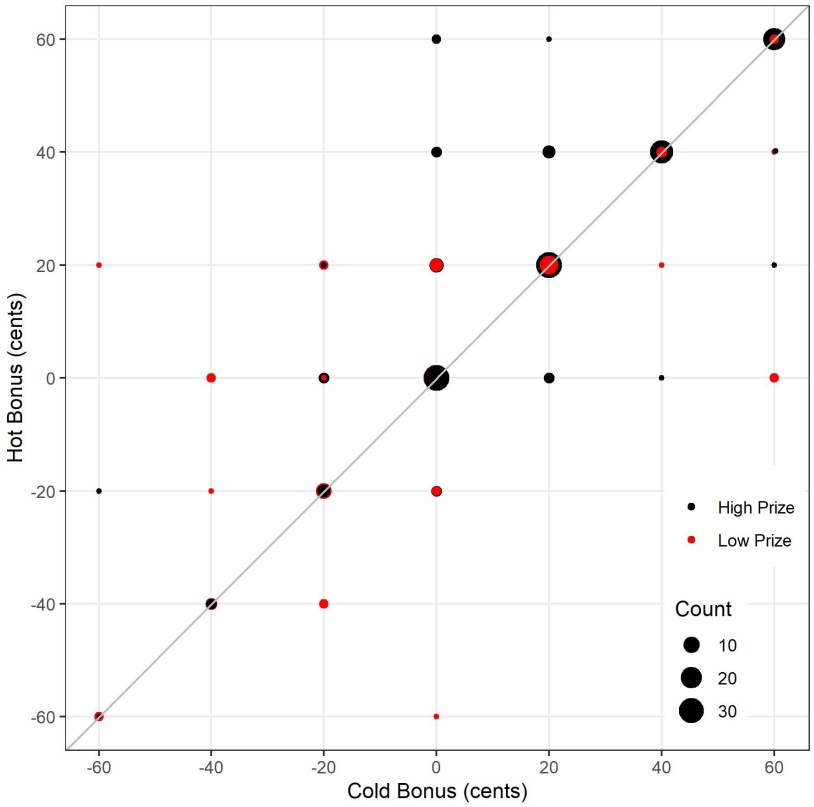


Table 8: Regression on Principals Hot Bonuses to Agents

	(1) Bonus, Hot	(2) Bonus, Hot	(3) Bonus, Hot	(4) Bonus, Hot
Deserves Bonus of Pick	0.502			
	(0.435)			
Deserves Bonus of Not Pick	0.154			
	(0.414)			
Deserves Punishment, Pick Hot		1.281		
		(0.833)		
Requires Reward, Pick Hot		0.191		
		(0.186)		
Lottery Prize		1.071***		1.204***
		(0.182)		(0.304)
Deserves Bonus of Pick, Hot			1.089***	0.134
			(0.172)	(0.298)
Deserves Bonus of Not Pick, Hot			0.648***	-0.114
			(0.147)	(0.240)
Constant	0.410	-0.295	-0.566**	-0.307
	(0.300)	(0.181)	(0.206)	(0.212)
Observations	240	240	240	240
R squared	0.0177	0.205	0.153	0.202

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Discussion and Conclusion

Our results provide support that Beneficence and Injustice Propositions from *The Theory of Moral Sentiments* hold in Principal Agents settings, in which there is stochasticity in the realization of the payoff separating intention from the outcome. Neutral observers coordinate on norms, and Agents make selections for Principals based on those norms. While prospect selections are not much different than selections made for one's self–there may be little daylight between "Do unto others as you would have them do unto you" and the Beneficence and Injustice Propositions–selections for Principals are somewhat more risk-neutral, particularly when the default choice provides cover for the Agent. This effect is consistent with the Beneficence and Injustice Propositions; failure to do good does not deserve punishment, nor does resisting doing bad require a reward. We find this effect despite eliciting norms that accounted for which option was the default. However, when the Principal can impose consequences on the Agent, choices are closer to those made for self. In contrast, these norms seem to have little impact on Principals' bonuses to Agents, which are dominated by the realized value of the prospect.

However, the result is not inconsistent with Adam Smith's writings. Smith (1759, p 96) observes that, "The thief, whose hand has been caught in his neighbor's pocket before he had taken any thing out of it, is punished with ignominy alone. If he had got time to take away a handkerchief, he would have been put to death." What is fair or just also depends on perspective. Hot ratings (from the Judges who were told the outcomes) still placed some weight on intent as measured by cold ratings, while the Principal's bonuses seemed to give intent virtually no weight.

Additionally, other norms might be invoked. The purpose of norms is to cohere a group of people around an idea of fairness. The arbitrariness of lotteries–gains that come through luck rather than work–requires a different standard of fairness, one which recognizes that fortune might have favored someone else, balancing the desire to maintain endowments against the condemnation of greedy defense of arbitrary gains (of others). There are stronger norms for sharing gains from high-variance hunting than from low-variance gathering or hunting (Kaplan et al., 2012).

One limitation of our study is the neutral framing and lack of context for the Agent's decision. We did not use the term 'Agent' in the participant's interface. We simply said one participant makes a decision for another. Using the term Agent, stating that the Agent was hired by the Principal, or offering further specifications about the Agent's duties could impact participant perceptions. Thus our results may not generalize to all financial interactions. Norms might depend on small details such as these, so changing them could alter the norms, thus the ratings by Judges, selections by Agents, and bonuses from Principals. It is also possible the details could decrease (or increase) the ability to converge on a norm; if there were a decrease (increase) in the proportion of Judges selecting the modal rating, we also expected those ratings would have less (more) of an impact on Agent selections and a great less (more) impact on Principals' bonus awards. These are important questions for future research.

Our study also featured perfect and public information that did not have any variation from *ex-ante* Agent selection to *ex-post* Principal bonus selection, for very simple prospects. These conditions give norms their best shot. However, they are rarely, if ever, found in financial interactions. Generally, Agents have richer and more nuanced information than Principals, who lack the expertise or inclination to digest that level of detail. Even unbiased attempts to aggregate from one level to another might lead to divergent perspectives. During a bubble, the probability of bust seems minuscule; yet, after a burst, it seems like it was inevitable. However, a Principal or third party might not account for this shift in perspective when judging an Agent's actions or determining the consequences of those actions. Further research is needed to explore how aspects of identity such as sex and gender interact with norms. Section S.6 contains an exploratory analysis of how sex impacted actions in our experiment. We find no conclusive evidence, but it was not part of our set of hypotheses, and we cannot rule out effects.

In our experiment, aside from the effort to reveal Option B there were no costs to the Agent; thus, the Agent could comply with the norm with no opportunity cost relative to not complying. If there were opportunity costs and the Judges were aware of the costs, different norms may have emerged. Agents would likely make decisions that balanced norm compliance and opportunity costs. We also expect Principals would

consider opportunity costs when adjusting bonuses. However, outside of experiments, information about and saliency of the opportunity costs is likely to vary by role, as would how much weight is given to the opportunity costs.

In line with the findings of the literature, we observe a small increase in risk-taking behavior when making decisions for others Polman and Wu (2020). The lack of statistical significance can be due to small stake sizes. As our study was done online and did not take participants much time, our prospects had fairly small expected values. Additionally, we used an intuitive way to represent lotteries, which graphically depicted both the prize value and the probability. Many experiments use Hey and Orme (1994) style pie charts that only use graphical representation of probability.

Our results help connect two strands of literature, one evidencing how principles or meta-rules guide actions and reactions, the second eliciting norms about how appropriate various actions are in particular situations, which we assert is the product of subjects deciding principle(s) are most salient to the situation and what is their implication. We hope future research further tests this assertion and leverages the insights from both strands to further our understanding of how norms are formed and impact economic decision-making. The most immediate extension would be to see how Judges would rate Principals' bonus decisions, e.g., was the bonus too small, the right size, or too big.

References

Agranov, Marina, Alberto Bisin, and Andrew Schotter, "An experimental study of the impact of competition for other people's money: the portfolio manager market," *Experimental Economics*, 2014, *17* (4), 564–585. 3

Akerlof, George A. and Rachel E. Kranton, "Economics and Identity," *The Quarterly Journal of Economics*, 2000, p. 715. 3. 4

Bardsley, Nicholas, "Dictator game giving: altruism or artefact?," *Experimental Economics*, June 2008, *11* (2), 122–133. 3

Barrafrem, Kinga and Jan Hausfeld, "Tracing risky decisions for oneself and others: The role of intuition and deliberation," *Journal of Economic Psychology*, 2020, *77*, 102188. 3

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, reciprocity, and social history," *Games and economic behavior*, 1995, *10* (1), 122–142. Publisher: Elsevier. 3

Bolton, Gary E and Axel Ockenfels, "Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states: Comment," *American Economic Review*, 2010, *100* (1), 628–33. 3

Chakravarty, Sujoy, Glenn W Harrison, Ernan E Haruvy, and E Elisabet Rutström, "Are you risk averse over other people's money?," *Southern Economic Journal*, 2011, *77* (4), 901–913. 3

Charness, Gary and Matthew O Jackson, "The role of responsibility in strategic risk-taking," *Journal of Economic Behavior & Organization*, 2009, *69* (3), 241–247. 3

Chaudhuri, Ananish, "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature," *Experimental economics*, 2011, *14* (1), 47–83. 1

Chen, Daniel L., Martin Schonger, and Chris Wickens, "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, March 2016, *9*, 88–97. 7

Cox, James C. and Cary A. Deck, "On the nature of reciprocal motives," *Economic Inquiry*, 2005, *43* (3), 623–635. Publisher: Wiley Online Library. 3

Eriksen, Kristoffer W and Ola Kvaløy, "Myopic investment management," *Review of Finance*, 2010, *14* (3), 521–542. 3

_ , _ , and Miguel Luzuriaga, "Risk-taking on behalf of others," *Journal of Behavioral and Experimental Finance*, 2020, *26*, 100283. 2

Falk, Armin, Ernst Fehr, and Urs Fischbacher, "On the Nature of Fair Behavior," *ECONOMIC INQUIRY*, 2003, *41* (1), 7. 3

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger, "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica*, 1997, *65* (4), 833. 3

Gillies, Anthony and Mary L. Rigdon, "Plausible Deniability and Cooperation in Trust Games," *Available at SSRN 3030482*, 2017. 3

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze, "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, December 1982, *3* (4), 367–388. 3

Harrison, Glenn W, Morten I Lau, E Elisabet Rutström, and Marcela Tarazona-Gómez, "Preferences over social risk," *Oxford Economic Papers*, 2013, *65* (1), 25–46. 3

Harsanyi, John C., "On the rationality postulates underlying the theory of cooperative games," *Journal of Conflict Resolution*, June 1961, *5* (2), 179–196. Publisher: SAGE Publications Inc. 3

Hey, John D. and Chris Orme, "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica*, 1994, *62* (6), 1291–1326. Publisher: [Wiley, Econometric Society]. 21

Johnson, Noel D and Alexandra A Mislin, "Trust games: A meta-analysis," *Journal of economic psychology*, 2011, *32* (5), 865–889. 1

Kaplan, Hillard S., Eric Schniter, Vernon L. Smith, and Bart J. Wilson, "Risk and the evolution of human exchange," *Proceedings of the Royal Society B: Biological Sciences*, August 2012, *279* (1740), 2930–2935. 20

Kimbrough, Erik O and Alexander Vostroknutov, "Norms make preferences social," *Journal of the European Economic Association*, 2016, *14* (3), 608–638. 1, 3, 4

Kimbrough, Erik O. and Alexander Vostroknutov, "A portable method of eliciting respect for social norms," *Economics Letters*, July 2018, *168*, 147–150. 3

Kimbrough, Erik O and Bart J Wilson, "Rule-following," *Available at SSRN 3838374*, 2021. 1

Korenok, Oleg, Edward L. Millner, and Laura Razzolini, "Taking, giving, and impure altruism in dictator games," *Experimental Economics*, September 2014, *17* (3), 488–500. 3

Krupka, Erin L and Roberto A Weber, "Identifying social norms using coordination games: Why does dictator game sharing vary?," *Journal of the European Economic Association*, 2013, *11* (3), 495–524. 1, 2

Lazear, Edward P., *Personnel economics [computer file] / Edward P. Lazear* The Wicksell lectures: 1993, Cambridge, Mass. : MIT Press, c1995., 1995. 3

Levitt, Steven D. and John A. List, "What do laboratory experiments measuring social preferences reveal about the real world?," *The Journal of Economic Perspectives*, 2007, *21* (2), 153–174. 3, 5

List, John A., "On the Interpretation of Giving in Dictator Games," *Journal of Political Economy*, June 2007, *115* (3), 482–493. Publisher: The University of Chicago Press. 3

Luzuriaga, Miguel et al., "Taking Risk with Other People's Money: Does Information about the Others Matter?," *Review of Behavioral Economics*, 2017, *4* (2), 107–133. 3

Marchegiani, Lucia, Tommaso Reggiani, and Matteo Rizzolli, "Loss averse agents and lenient supervisors in performance appraisal," *Journal of Economic Behavior & Organization*, November 2016, *131*, 183–197. 3

McCabe, Kevin A. and Vernon L. Smith, "A comparison of naive and sophisticated subject behavior with game theoretic predictions," *Proceedings of the National Academy of Sciences*, 2000, *97* (7), 3777–3781. Publisher: National Acad Sciences. 3

_ , Mary L. Rigdon, and Vernon L. Smith, "Positive reciprocity and intentions in trust games," *Journal of Economic Behavior & Organization*, 2003, *52* (2), 267–275. Publisher: Elsevier. 3

Pahlke, Julius, Sebastian Strasser, and Ferdinand M Vieider, "Risk-taking for others under accountability," *Economics Letters*, 2012, *114* (1), 102–105. 3

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

Polman, Evan, "Self–other decision making and loss aversion," *Organizational Behavior and Human Decision Processes*, 2012, *119* (2), 141–150. 3

_ and Kaiyang Wu, "Decision making for others involving risk: A review and meta-analysis," *Journal of Economic Psychology*, 2020, *77*, 102184. 2, 3, 5, 21

Reynolds, Douglas B, Jacob Joseph, Reuben Sherwood et al., "Risky shift versus cautious shift: determining differences in risk taking between private and public management decision-making," *Journal of business & economics research (JBER)*, 2009, *7* (1). 3

Rubin, Jared and Roman Sheremeta, "Principal–Agent Settings with Random Shocks," *Management Science*, April 2016, *62* (4), 985–999. 3

Smith, Adam, *The Theory of Moral Sentiments*, H. G. Bohn, 1759. Google-Books-ID: 96Wpjtx4H4wC. 1, 2, 3, 20

Smith, Vernon L, "Trust, reciprocity, and social history: New pathways of learning when max U (own reward) fails decisively," 2020. 1

_ **and Bart J Wilson**, "Equilibrium play in voluntary ultimatum games: Beneficence cannot be extorted," *Games and Economic Behavior*, 2018, *109*, 452–464. 1

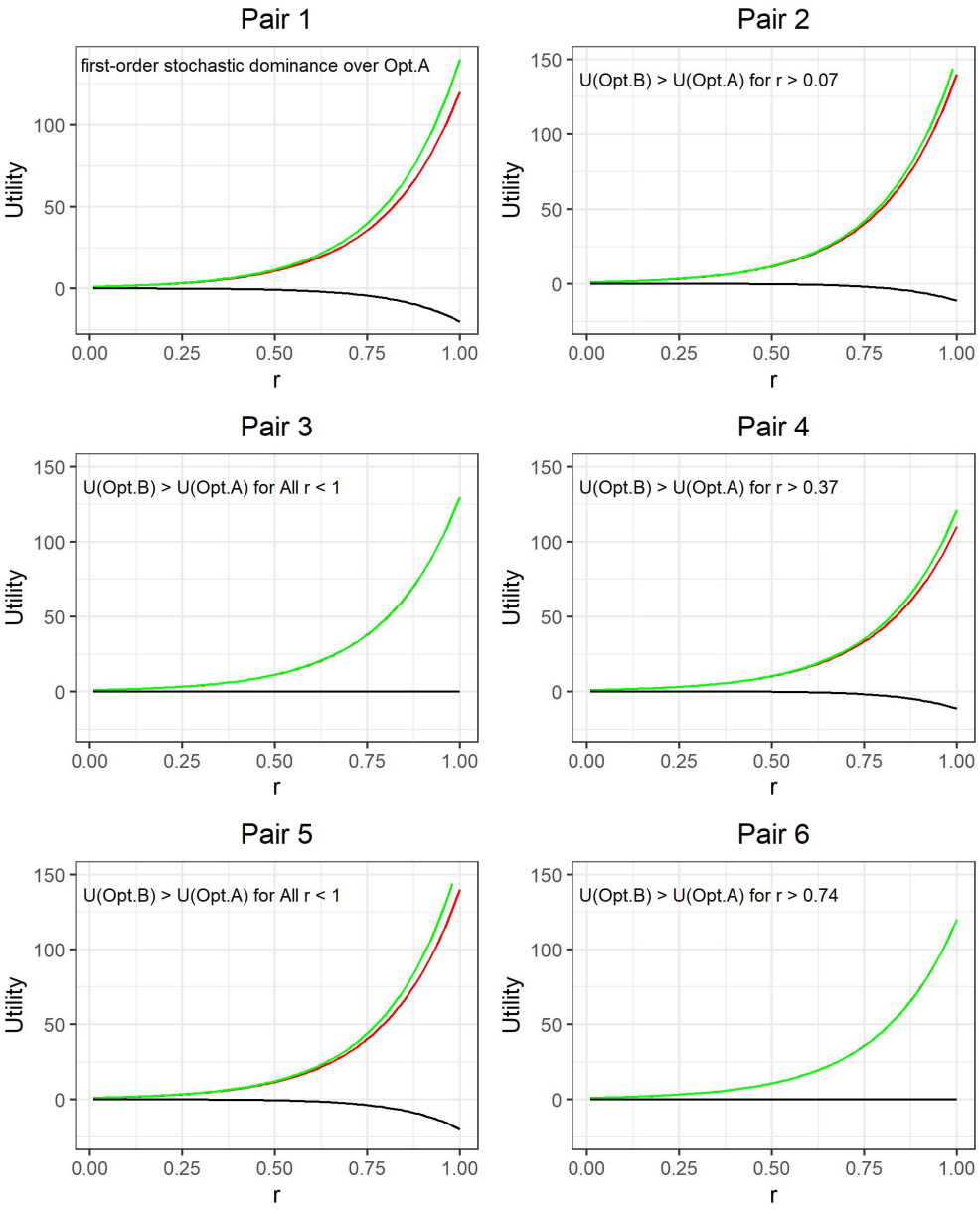
Smith, Vernon L. and Bart J. Wilson, *Humanomics: Moral sentiments and the wealth of nations for the twenty-first century*, Cambridge University Press, 2019. 3

Vostroknutov, Alexander, "Social norms in experimental economics: Towards a unified theory of normative decision making," *Analyse & Kritik*, 2020, *42* (1), 3–40. 2

Supplementary Materials

S.1 Lottery Characteristics

Figure S.1: Utility Comparisons of Lotteries by Risk Parameter



S.2 Software Screen Shots

Figure S.2: Instructions for Judges

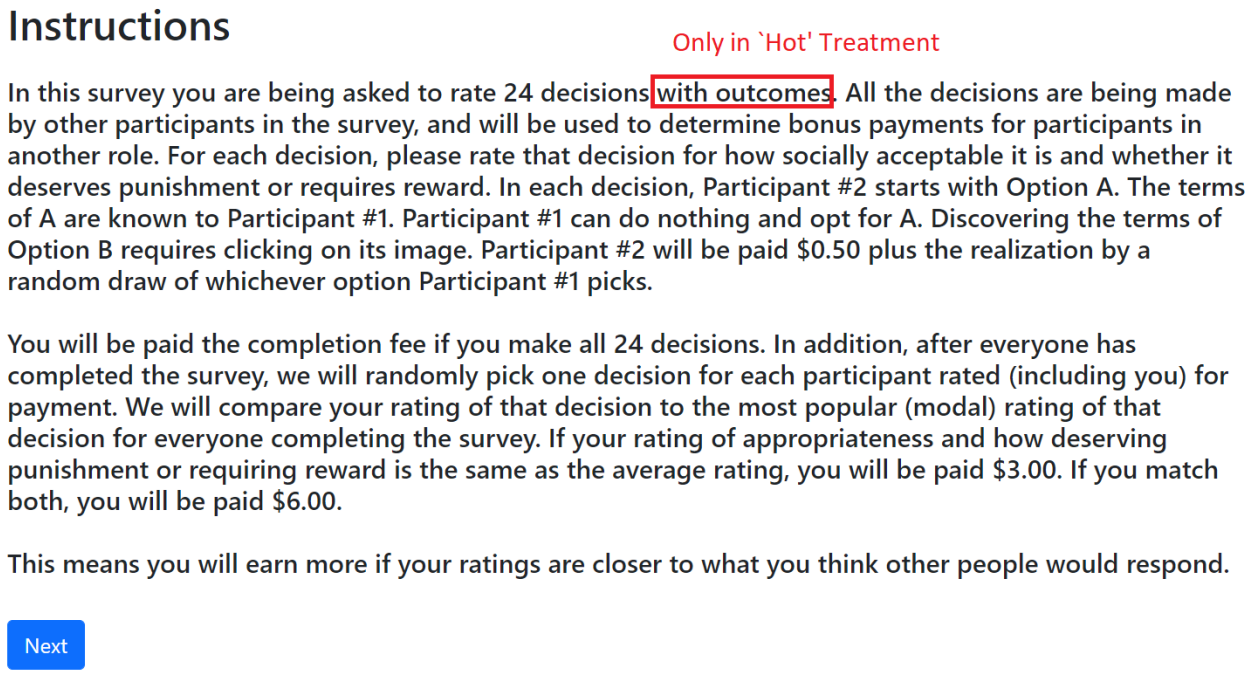


Figure S.3: Instructions for Agents

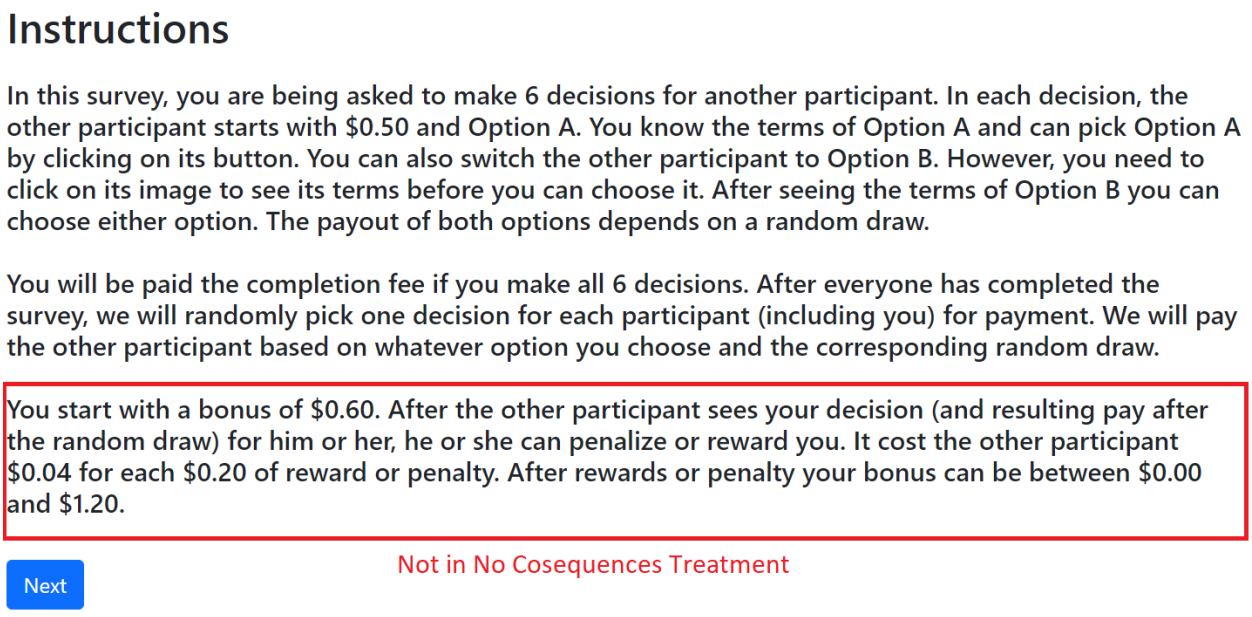


Figure S.4: Instructions for Principals

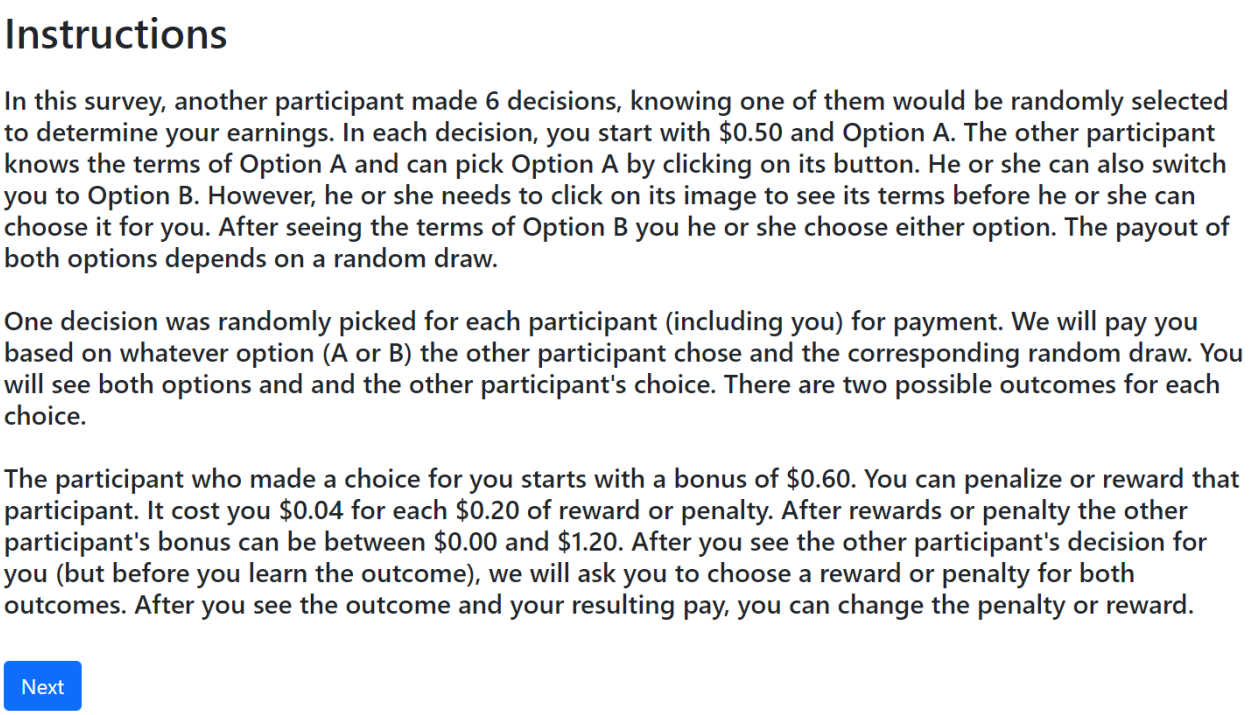


Figure S.5: Instructions for Principals

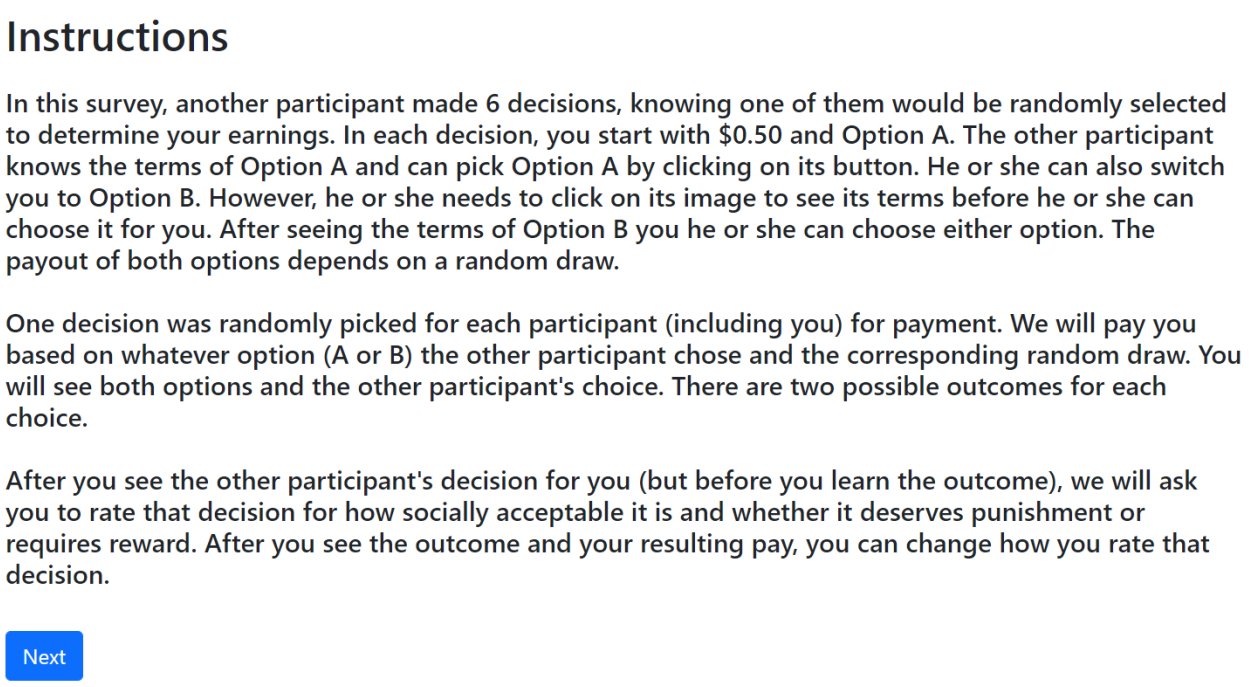


Figure S.6: Example of Strong Default

Example decision screen

This an example of Participant 1's decision screen.

Participant 1 can select Option A immediately, but cannot select Option B without clicking on the question mark. After revealing Option B, either option can be selected.

Please click on the question mark.

Figure S.7: Screen for Judges to Rate Agent Decision

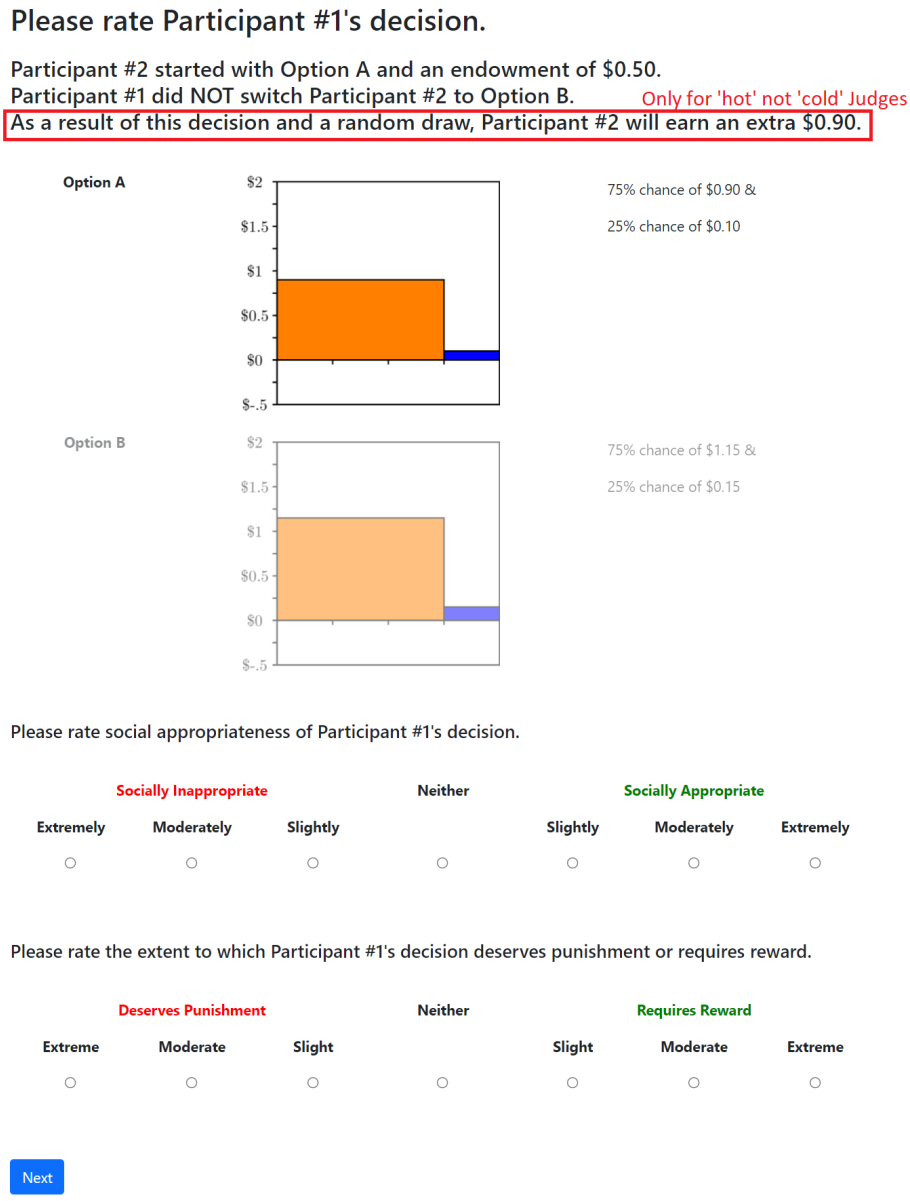


Figure S.8: Agent Decision Screen, Pair 5

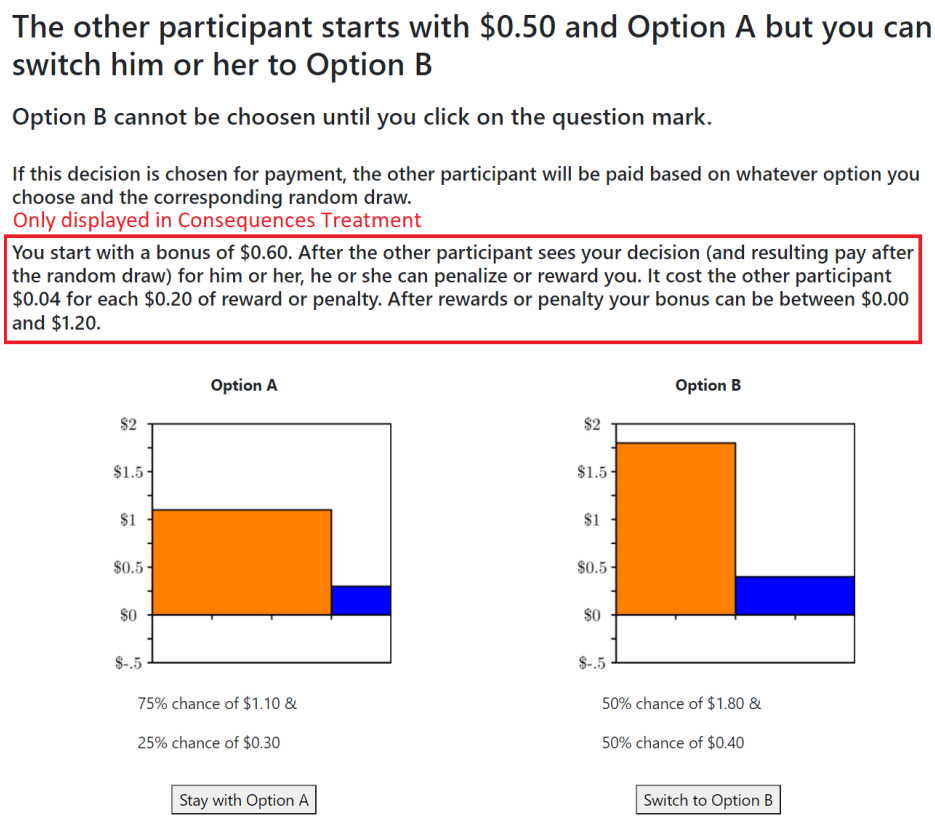


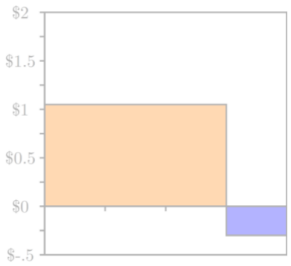
Figure S.9: Principal Cold Bonus Amount Decision Screen

Results

You started with Option A and $0.50

The other person choose Option B, therefore you will be paid based on Option B.

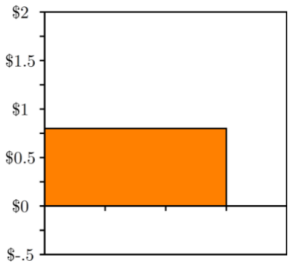
Option A



75% chance of $1.05 &

25% chance of $-0.30

Option B



75% chance of $0.80 &

25% chance of $0.00

If your pay from the Option is $0.80,
how much would you reward or penalize the person who choose which Option (A or B) you would get?

	Penalty					Reward	
to other Person	-60¢	-40¢	-20¢	0¢	+20¢	+40¢	+60¢
	○	○	○	○	○	○	○
Cost to you	-12¢	-8¢	-4¢	0¢	-4¢	-8¢	-12¢

If your pay from the Option is $0.00,
how much would you reward or penalize the person who choose which Option (A or B) you would get?

	Penalty					Reward	
to other Person	-60¢	-40¢	-20¢	0¢	+20¢	+40¢	+60¢
	○	○	○	○	○	○	○
Cost to you	-12¢	-8¢	-4¢	0¢	-4¢	-8¢	-12¢

Next

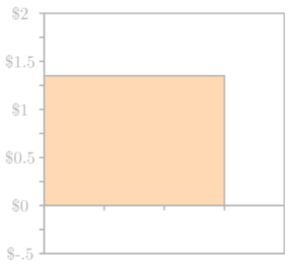
Figure S.10: Principal Hot Bonus Amount Decision Screen

Results

You started with Option A and $0.50

The other person choose Option B, therefore you will be paid based on Option B.

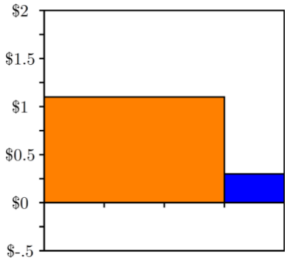
Option A



75% chance of $1.35 &

25% chance of $0.00

Option B



75% chance of $1.10 &

25% chance of $0.30

The random number drawn for you was 30 (out of 100), so your pay from the Option is $1.10.

Your choice for this outcome is below. You can adjust your choice to reward or penalize the person who choose which Option (A or B) you would get.

	Penalty				Reward		
to other Person	-60¢	-40¢	-20¢	0¢	+20¢	+40¢	+60¢
	○	○	◉	○	○	○	○
Cost to you	-12¢	-8¢	-4¢	0¢	-4¢	-8¢	-12¢

Next

S.3 Participant Demographics

Table S.1: Participant Characteristics

Age	Mean	SD	Obs
	31.63	12.72	1564
Sex	Female	Male	Prefer not to say
	5	1040	509

S.4 Agents

Table S.2: Probit Regression on Agents Likelihood to Pick Option B by Treatment

	(1)	(2)	(3)	(4)
	Self vs. Other		Self vs. Consequences	
	Pairs	Flips	Pairs	Flips
Pair=2	-1.888***	-1.888***	1.924***	1.924***
	(0.210)	(0.210)	(0.226)	(0.226)
Pair=3	-1.384***	-1.384***	1.691***	1.691***
	(0.179)	(0.179)	(0.230)	(0.230)
Pair=4	-1.915***	-1.915***	2.199***	2.199***
	(0.211)	(0.211)	(0.225)	(0.225)
Pair=5	-0.621***	-0.621***	1.217***	1.217***
	(0.184)	(0.184)	(0.236)	(0.236)
Pair=6	-1.256***	-1.256***	1.775***	1.775***
	(0.187)	(0.187)	(0.223)	(0.223)
Other & Pair=1	0.0972	0.0972		
	(0.225)	(0.225)		
Other & Pair=2	0.242	0.242		
	(0.173)	(0.173)		
Other & Pair=3	0.0430	0.0430		
	(0.165)	(0.165)		
Other & Pair=4	0.0762	0.0762		
	(0.176)	(0.176)		
Other & Pair=5	-0.312	-0.312		
	(0.173)	(0.173)		
Other & Pair=6	0.0000207	0.0000207		
	(0.165)	(0.165)		
Consequences & Pair=1			-0.190	-0.190
			(0.294)	(0.294)
Consequences & Pair=2			-0.107	-0.107
			(0.164)	(0.164)
Consequences & Pair=3			0.0210	0.0210
			(0.163)	(0.163)
Consequences & Pair=4			-0.0716	-0.0716
			(0.171)	(0.171)
Consequences & Pair=5			0.221	0.221
			(0.166)	(0.166)
Consequences & Pair=6			0.0636	0.0636
			(0.163)	(0.163)
Constant	1.256***	1.256***	-1.649***	-1.649***
	(0.155)	(0.155)	(0.196)	(0.196)
lnsig2u	-3.453**	-3.453**	-5.355	-5.355
	(1.154)	(1.154)	(7.085)	(7.085)
Observations	1440	1440	1440	1440
Chi Squared	232.8	232.8	227.3	227.3
Number IDs	240	240	240	240

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S.3: Probit Regression on Agents Likelihood to Pick Option B, Other vs. Consequences

	(1) Pairs	(2) Flips
Pair=2	-1.846***	2.002***
	(0.183)	(0.241)
Pair=3	-1.283***	1.897***
	(0.166)	(0.239)
Pair=4	-1.766***	2.311***
	(0.191)	(0.236)
Pair=5	-0.854***	1.623***
	(0.164)	(0.255)
Pair=6	-1.326***	2.023***
	(0.181)	(0.251)
Other & Pair=1	0.220	0.189
	(0.220)	(0.292)
Other & Pair=2	0.322	-0.0422
	(0.174)	(0.163)
Other & Pair=3	0.0646	-0.252
	(0.165)	(0.163)
Other & Pair=4	0.0499	-0.267
	(0.176)	(0.166)
Other & Pair=5	0.0433	-0.197
	(0.168)	(0.165)
Other & Pair=6	0.192	-0.189
	(0.165)	(0.163)
Constant	1.134***	-1.834***
	(0.144)	(0.220)
lnsig2u	-3.408**	-14.43
	(1.112)	(65802.5)
Observations	1440	1440
Chi Squared	240.1	197.3
Number IDs	240	240

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S.4: Probit Regression on Agents Likelihood to Pick Option B by Treatment

	(1) Other vs. Self		(2) Cons. vs. Self		(3) Cons. vs. Other	
player.PairD=2	-1.621***	(0.194)	-1.981***	(0.203)	-2.040***	(0.204)
player.PairD=3	-1.319***	(0.191)	-1.416***	(0.196)	-1.484***	(0.197)
player.PairD=4	-1.809***	(0.196)	-1.901***	(0.202)	-1.962***	(0.203)
player.PairD=5	-0.917***	(0.192)	-0.983***	(0.197)	-1.060***	(0.199)
player.PairD=6	-1.236***	(0.191)	-1.459***	(0.196)	-1.526***	(0.197)
player.PairD=7	-2.881***	(0.249)	-3.143***	(0.275)	-3.178***	(0.272)
player.PairD=8	-1.110***	(0.191)	-1.094***	(0.196)	-1.168***	(0.197)
player.PairD=9	-1.425***	(0.191)	-1.201***	(0.195)	-1.274***	(0.197)
player.PairD=10	-1.025***	(0.192)	-0.778***	(0.198)	-0.858***	(0.200)
player.PairD=11	-1.644***	(0.193)	-1.481***	(0.196)	-1.548***	(0.198)
player.PairD=12	-1.236***	(0.191)	-1.072***	(0.196)	-1.147***	(0.198)
FlipPairTx=11	0.0952	(0.221)				
FlipPairTx=20	-0.238	(0.170)	0.0801	(0.179)		
FlipPairTx=21	0	(.)			0.317	(0.172)
FlipPairTx=30	-0.0420	(0.162)	0.0217	(0.166)		
FlipPairTx=31	0	(.)			0.0635	(0.163)
FlipPairTx=40	-0.0755	(0.174)	-0.0265	(0.178)		
FlipPairTx=41	0	(.)			0.0499	(0.174)
FlipPairTx=50	0.304	(0.170)	0.358*	(0.174)		
FlipPairTx=51	0	(.)			0.0436	(0.165)
FlipPairTx=60	-8.57e-09	(0.162)	0.193	(0.166)		
FlipPairTx=61	0	(.)			0.190	(0.163)
FlipPairTx=110	-0.000107	(0.273)	0.197	(0.300)		
FlipPairTx=111	0	(.)			0.192	(0.294)
FlipPairTx=120	0.149	(0.163)	0.109	(0.167)		
FlipPairTx=121	0	(.)			-0.0425	(0.163)
FlipPairTx=130	0.231	(0.163)	-0.0211	(0.165)		
FlipPairTx=131	0	(.)			-0.253	(0.163)
FlipPairTx=140	0.338*	(0.168)	0.0725	(0.174)		
FlipPairTx=141	0	(.)			-0.268	(0.167)
FlipPairTx=150	-0.0228	(0.167)	-0.224	(0.169)		
FlipPairTx=151	0	(.)			-0.198	(0.166)
FlipPairTx=160	0.126	(0.162)	-0.0648	(0.166)		
FlipPairTx=161	0	(.)			-0.190	(0.163)
FlipPairTx=12			-0.123	(0.215)	-0.220	(0.217)
Constant	1.236***	(0.152)	1.265***	(0.156)	1.337***	(0.160)
lnsig2u	-7.538	(42.76)	-3.076***	(0.607)	-4.843	(3.008)
Observations	2880		2880		2880	
Chi Squared	429.7		489.7		442.2	
Number IDs	480		480		480	

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

S.5 Principals

Figure S.11 Plots the bonuses Principals awarded Agents against the realization from the prospect the Agent choose for the Principal. Color indicates the difference between the chosen and not chosen option in ratings of deserving punishment or requiring reward. Dot size indicates the number of Principals with a given combination. There is clear relationship between the realized value and bonus amount. Because rating drove selection, there are more positive ratings than negative ratings. However, there is no clear effect of rating on bonus.

Figure S.11: Bonuses to Agents by Realization from Prospect with Relative Rating

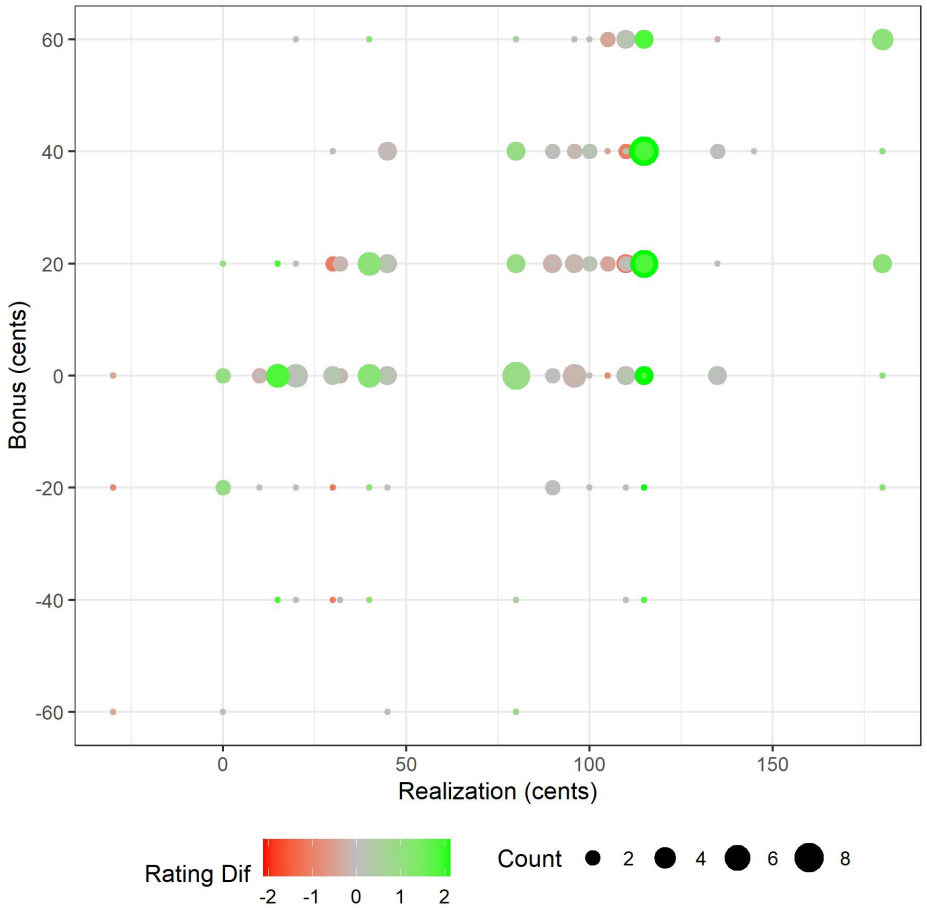


Figure S.12: Principal Ratings of Agent Decisions

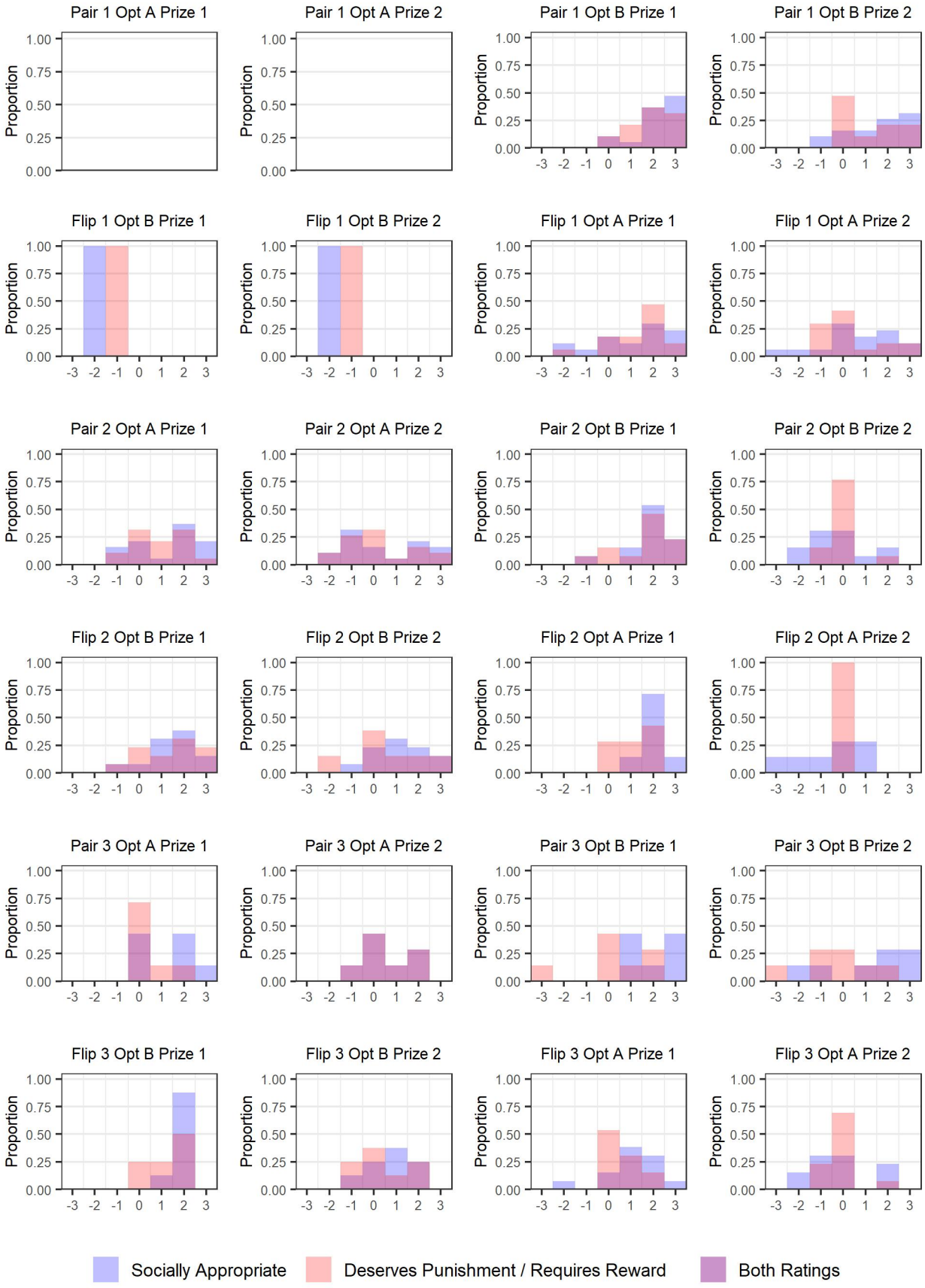


Figure S.13: Principal Ratings of Agent Decisions, cont

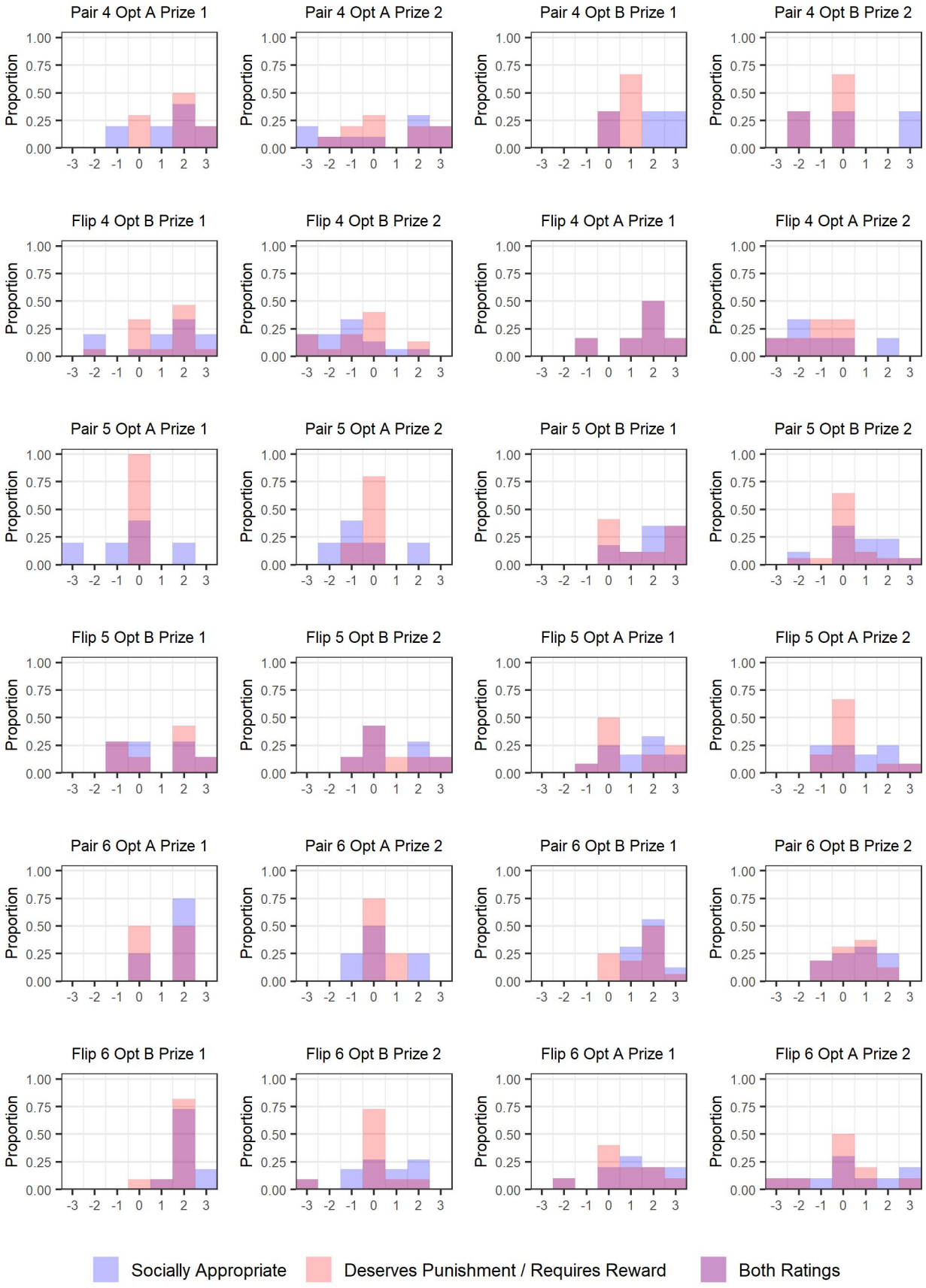


Figure S.14: Principal Cold versus Hot Bonuses to Agents

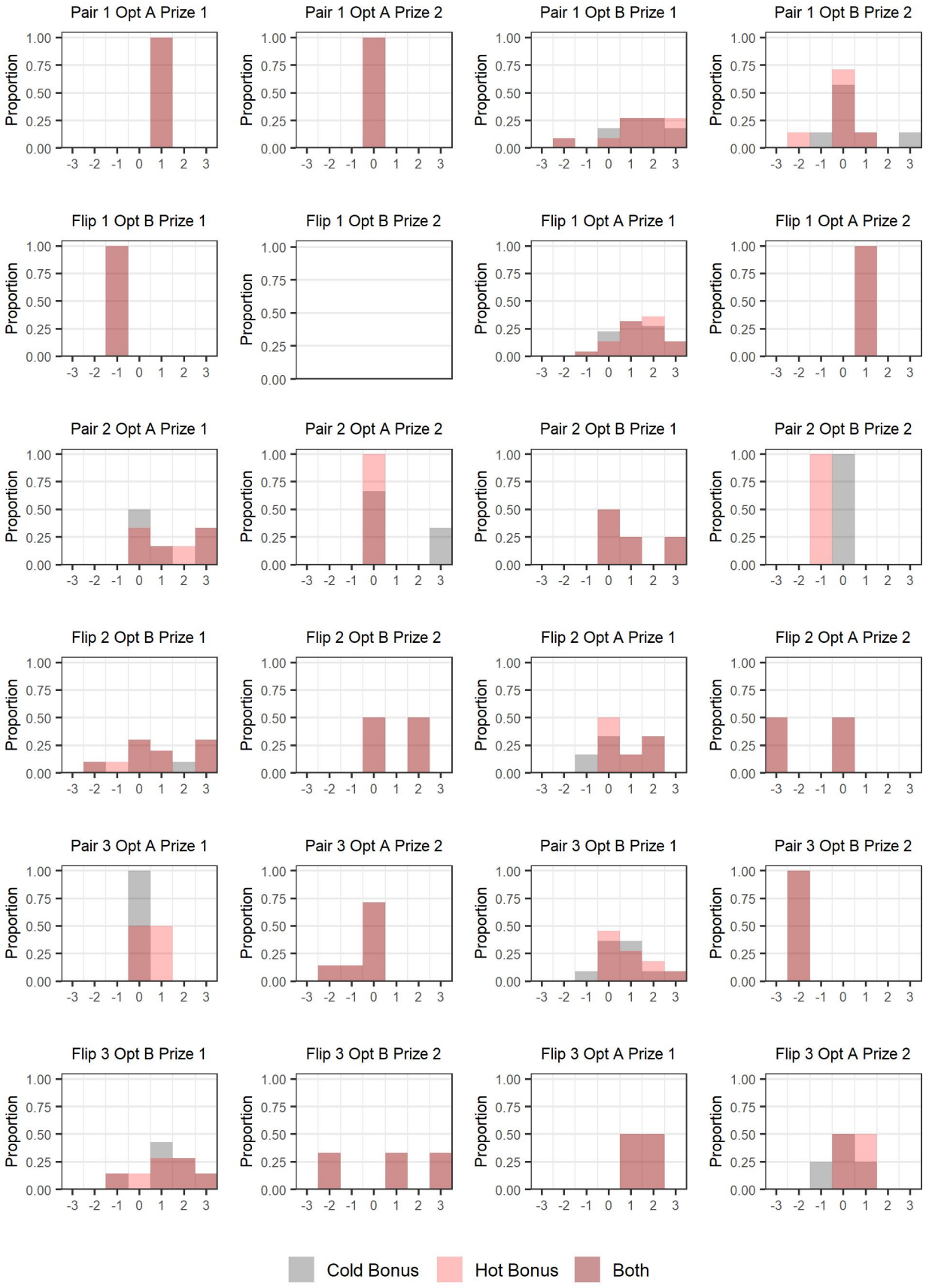


Figure S.15: Principal Cold versus Hot Bonuses to Agents

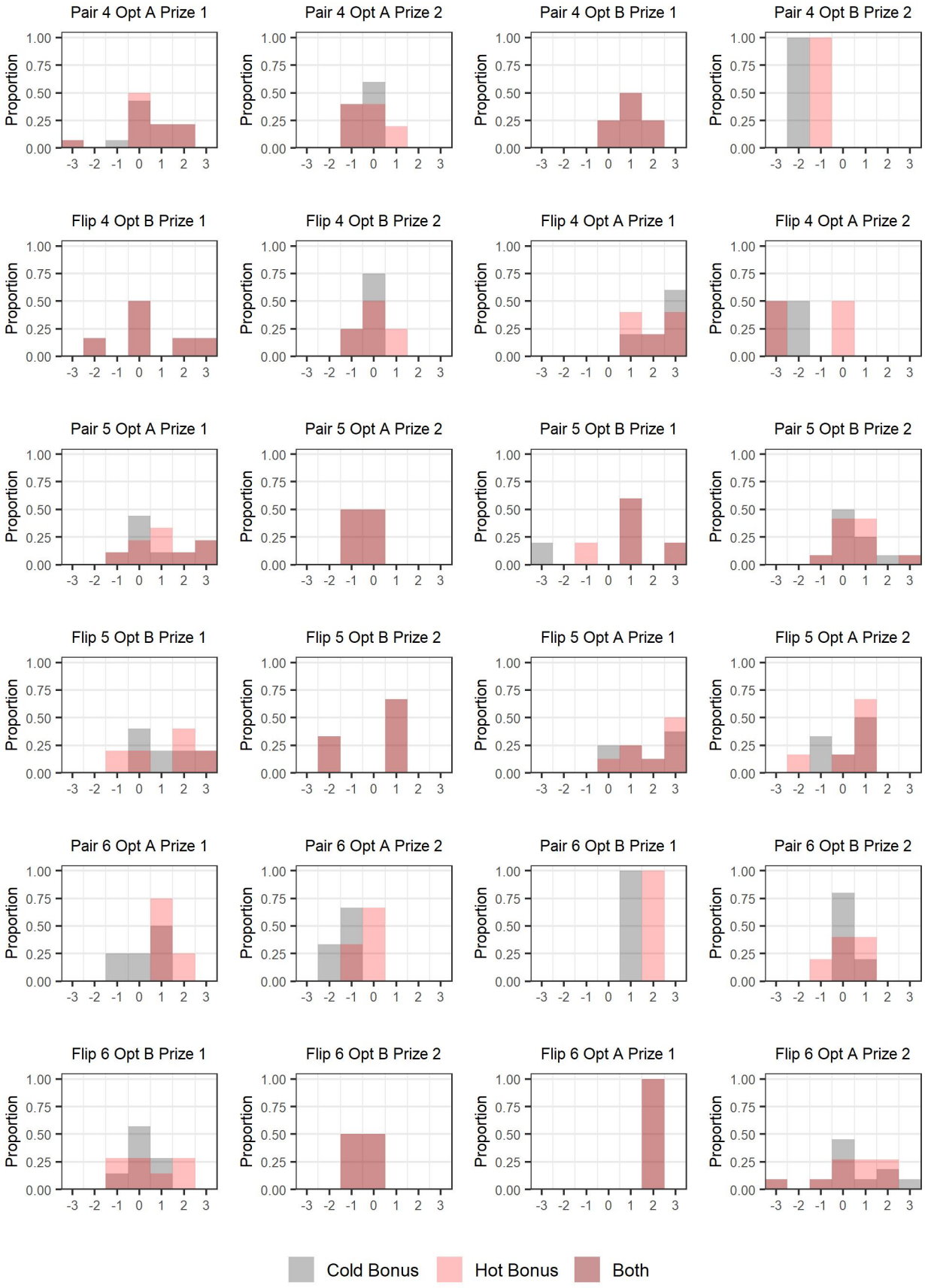


Table S.5: Regression on Principals Cold Bonuses to Agents

	(1) Bonus, Cold	(2) Bonus, Cold	(3) Bonus, Cold	(4) Bonus, Cold	(5) Bonus, Cold
Deserves Bonus of Pick	0.540				
	(0.279)				
Deserves Bonus of Not Pick	0.379				
	(0.251)				
Diff in DB (Picked - Not)		0.0791			
		(0.0510)			
Deserves Bonus of Pick, Hot			0.202		
			(0.118)		
Deserves Bonus of Not, Hot			-0.0619		
			(0.0999)		
Diff in DB (Picked - Not) Hot				0.129*	0.0840
				(0.0558)	(0.0675)
Lottery Prize, Cold					1.064***
					(0.111)
cut1					
Constant	-1.465***	-1.788***	-1.677***	-1.775***	-1.473***
	(0.215)	(0.124)	(0.178)	(0.124)	(0.144)
cut2					
Constant	-1.082***	-1.403***	-1.293***	-1.391***	-1.027***
	(0.193)	(0.102)	(0.156)	(0.102)	(0.120)
cut3					
Constant	-0.603**	-0.922***	-0.812***	-0.910***	-0.455***
	(0.196)	(0.0848)	(0.153)	(0.0843)	(0.105)
cut4					
Constant	0.553**	0.236***	0.347*	0.249***	0.998***
	(0.194)	(0.0670)	(0.152)	(0.0663)	(0.110)
cut5					
Constant	1.142***	0.825***	0.937***	0.839***	1.758***
	(0.201)	(0.0767)	(0.161)	(0.0761)	(0.139)
cut6					
Constant	1.638***	1.319***	1.430***	1.332***	2.374***
	(0.200)	(0.0922)	(0.166)	(0.0938)	(0.168)
sigma2_u					
Constant	0.0681	0.0765	0.0657	0.0671	0.328*
	(0.0857)	(0.0877)	(0.0855)	(0.0860)	(0.129)
Observations	480	480	480	480	480
NumberIDs	240	240	240	240	240
LogLikelihood	-796.8	-798.5	-796.4	-796.8	-741.7

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S.6: Coeficents Estimates from Ordered Probit Regression on Principals Bonuses to Agents - Hot

	(1) Bonus, Hot	(2) Bonus, Hot	(3) Bonus, Hot	(4) Bonus, Hot
Bonus, Hot				
Deserves Bonus of Pick	0.437			
	(0.333)			
Deserves Bonus of Not Pick	0.160			
	(0.318)			
Deserves Punishment, Pick Hot		1.078		
		(0.652)		
Requires Reward, Pick Hot		0.177		
		(0.163)		
Lottery Prize		0.932***		1.048***
		(0.167)		(0.283)
Deserves Bonus of Pick, Hot			0.926***	0.125
			(0.155)	(0.263)
Deserves Bonus of Not Pick, Hot			0.547***	-0.0996
			(0.125)	(0.215)
cut1	-1.879***	-1.508***	-1.230***	-1.491***
	(0.326)	(0.233)	(0.252)	(0.271)
cut2	-1.355***	-0.922***	-0.667***	-0.909***
	(0.238)	(0.185)	(0.193)	(0.194)
cut3	-0.819***	-0.332*	-0.0961	-0.323
	(0.244)	(0.166)	(0.175)	(0.188)
cut4	0.162	0.767***	0.972***	0.770***
	(0.237)	(0.173)	(0.186)	(0.195)
cut5	0.883***	1.585***	1.758***	1.590***
	(0.236)	(0.186)	(0.200)	(0.206)
cut6	1.492***	2.263***	2.415***	2.270***
	(0.230)	(0.212)	(0.218)	(0.223)
Observations	240	240	240	240
LogLikelihood	-396.1	-370.7	-378.1	-371.1

Robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

S.6 Differences by Sex

S.6.1 Judges

Figure S.16: Judges Cold Ratings of Punishment and Reward to Agent by Sex of Judge

S.6.2 Agents

The differences in the proportion of Agents choosing the more risk-neutral prospect in any of the treatments are not statistically significant. The p-values from Wilcoxon rank sum test by Sex for the Consequence, Other and Self treatments are 0.219, 0.203 and 0.185. Coefficients for Male and interactions of Male had confidence intervals which spanned zero.

S.6.3 Principals

Figure S.17 is akin to Figure 7 except color denotes the Principals reported sex. There are no clear sex difference as to how learning the realization changes the Principals awarded bonus.

Figure S.17: Cold versus Hot Bonus Decisions by Sex

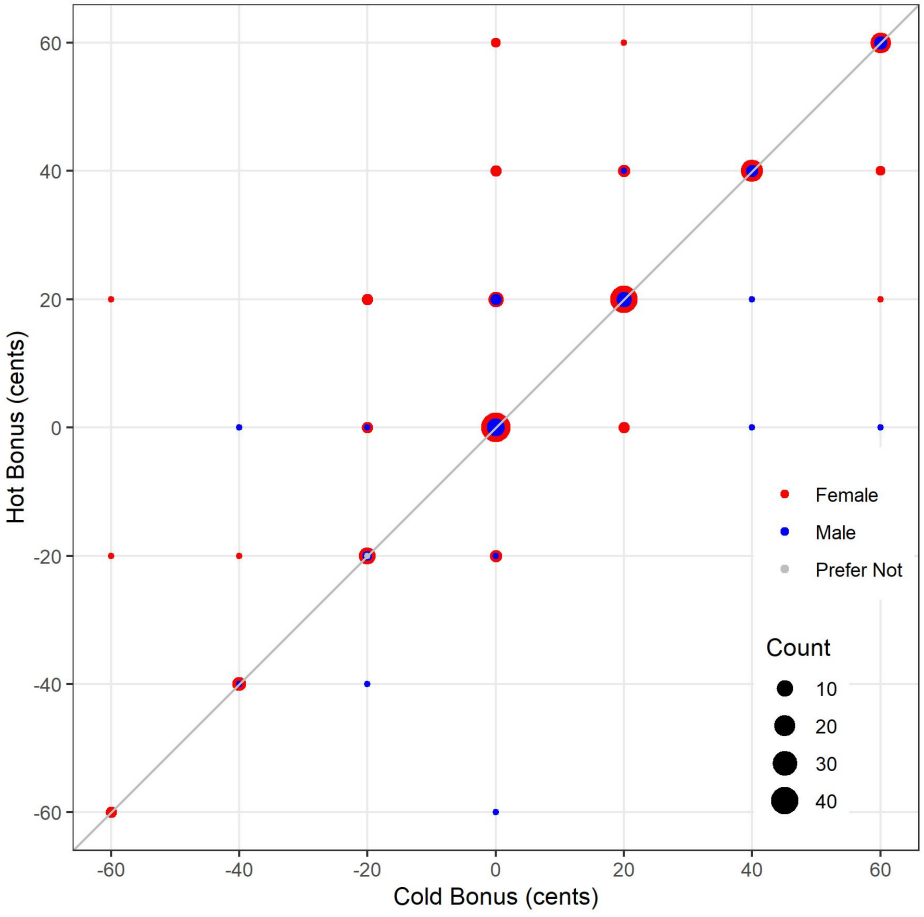


Figure S.18 is a histogram of of the bonuses Principals awarded Agents by sex of the Principal. The distributions show considerable overlap. Females are slightly more positive than males.

Figure S.18: Bonus Decisions by Sex