Exploiting moral wriggle room: Behavior inconsistent with a preference for fair outcomes

Jason Dana
Social and Decision Sciences, Carnegie Mellon University

Roberto A. Weber
Social and Decision Sciences, Carnegie Mellon University

Jason Xi Kuang
Katz School of Business, University of Pittsburgh

June 24, 2003[*]

[*] We thank George Loewenstein, Robyn Dawes, Colin Camerer, John Patty, Charlie Plott, Matthew Rabin, Ernst Fehr, Iris Bohnet and seminar participants at Carnegie Mellon and participants at the 2003 Public Choice / Economic Science meetings in Nashville and at the 2003 Economic Science meetings in Pittsburgh for helpful comments and suggestions. We greatly appreciate the access to resources at the Pittsburgh Experimental Economics Laboratory (PEEL) at the University of Pittsburgh. This research was funded by a Carnegie Mellon Berkman Faculty Development Grant to Weber.

Abstract

Subjects in economic experiments regularly appear to reveal concern for the payoffs of others. Economic models typically assume that this behavioral regularity reflects a preference for fair or equitable outcomes. This implies, among other things, that decision makers will backward induct along a game or decision tree to take the action expected to best satisfy these preferences. The present research demonstrates allocation choices among selves and others that are inconsistent with such a preference.

Our experiments reveal that eliminating the direct link between actions and harmful outcomes to others results in significantly more self-interested behavior. This is true even when the link can be easily re-established or when the link between actions and positive social outcomes is clear. Specifically, we show that in a binary dictator game, subjects typically choose the more equitable outcome. However, when making the same allocation decision, subjects behave more selfishly when the others' payoffs are uncertain, even though this uncertainty can be costlessly resolved. We also find that when two subjects, either of whom can guarantee the equitable outcome, make the allocation decision, they tend to favor the "selfish" outcome.

We argue that this behavior is consistent with a model in which most people do not want to take actions that indicate (to themselves) that they are willing to harm others out of self-interest. By removing the direct link between own actions and others' outcomes, we allow people to behave self-interestedly without negatively affecting their beliefs about themselves.

Introduction

In contrast with the assumptions of traditional neoclassical economics, subjects across a wide variety of experiments appear to exhibit concern for other people's welfare (see Thaler, 1992; Camerer, 2003). This phenomenon is perhaps clearest in dictator games, where a "proposer" chooses to divide an endowment between herself and a passive "receiver" who must accept the division. In this situation, a truly self-interested proposer seeking only to maximize her monetary payment should simply keep the entire endowment. However, Forsythe, et al., (1994) found that proposers gave a mean of 20 percent of the endowment to the receiver. Further, Kahneman, Knetsch and Thaler (1986) found that in a binary choice between an inequitable outcome ($18, $2) that monetarily favored the proposer and an even split ($10, $10), 76% of proposers chose the even split. Even when elaborate steps are taken to ensure double-blind anonymity (possibly creating an experimental demand for selfishness), the amount given to an unknown partner is still often greater than zero (Hoffman, et al., 1994).

The mere demonstration of altruistic behavior in experiments, however, need not pose a threat to economic models that presume rational self-interest. It could simply be the case that people voluntarily give because they gain utility from the basic act – a sort of "warm glow" (Andreoni 1990, 1995). For instance, Andreoni and Miller (2000) found that subjects' repeated social allocation choices obeyed revealed preference axioms with "giving" treated as simply another good from which decision makers derive welfare. Thus, others' payoffs could be treated as a consumption good, and costly benevolent allocations as rational.

Several recent economic theories model other-regarding payoffs by including concern for others' welfare or for relative outcomes in the utility function, incorporating various psychological insights into formal theories about the way people make decisions. For example, Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) present models that propose that decision makers are averse to inequitable outcomes. Thus, they will pay a premium in order to bring payoffs closer to equity. Charness and Rabin's quasi-maximin theory (1999) posits that people are partly motivated to maximize the total surplus accruing to all players in a game.

While these and similar models of "fairness preferences" do a good job of explaining behavior in economics experiments, it is worth noting that these experiments typically place subjects in situations where their decisions unambiguously and directly help or harm others. For example, the standard dictator experiment presents subjects with a simple decision to give or not give. However, in many situations outside the laboratory, the connection between our behavior and others' outcomes is often not so clear or direct. For instance, we often make choices that might or might not result in harmful consequences for others, depending on the actions of a third party or of nature. Moreover, we often make choices (such as whether to acquire information about the consequences of our actions) that may or may not put us in a situation similar to the kind of decision captured by the dictator game.

This discrepancy between the kinds of choices made in and out of the laboratory raises the possibility that there are important features of real-world giving that the models, which are based largely on experimental data, do not explain. Specifically, will people pursue opportunities to behave kindly to others, even when not doing so will not

ensure inequity? Models that include preferences for altruism or fairness often predict that they should. Someone who reveals a preference for fairness in a dictator game should try to implement similar fair outcomes in a game that is fundamentally identical to the dictator game. The presence of additional steps that do not change the payoffs – such as the option to acquire new information – should not affect outcomes. However, to see why this might not be true, consider the following example: A man is faced with the opportunity of being on a donor registry. It occurs to him that if he is on the registry, then he may be contacted as a match for someone else. He knows that if he finds out that he is a match, he will go through the donor process in the hope of saving a life. However, thinking that this process would be difficult and painful he decides he would rather not place himself on the list, thereby removing the possibility that he might be contacted as a match and face a decision where he knows he would choose to donate.

Following the logic of revealed preference, the above scenario seems inherently contradictory. The man will donate given the choice, revealing a preference to save a life at the cost of a painful procedure. Yet he avoids being faced with such a choice by not placing himself on the registry, revealing the opposite preference. Despite the apparently contradictory nature of these preferences, we suspect that many readers will relate with the above scenario.[1]

The problem for economic models of fairness that is highlighted in this example is that giving – whether it produces a warm glow by the mere act, by equalizing

[1] Three other examples of similar kinds of decisions are: 1) A man potentially carrying a sexually transmitted disease might show concern for the welfare of prospective sexual partners if he knows he has the disease, but may choose to avoid getting tested in order to not have to do so, 2) The developer of a profitable new product might choose to avoid finding out about the potential harmful effects of the product in order to not feel guilty about selling it, and 3) A doctor might not seek to know the costs of a particular drug in order to not feel guilt for prescribing it and accepting gifts from the drug manufacturer's representatives.

outcomes, or by social welfare maximization – may not in itself provide utility in the manner of a consumption good. It might be the case that a decision-maker will give when faced with an unambiguous choice between implementing a "fair" outcome and a selfish one (which is consistent with the existing experimental literature). However, there may be no preference for simply incrementing other's payoffs (or equitable outcomes), and in situations where the "selfish" choice does not immediately or directly produce an unfair outcome, the decision-maker will forego opportunities to help others in order to behave more self-interestedly. That is, people may exploit moral "wriggle room," when it is available, in order to behave selfishly. Evidence of such behavior would pose a problem for economic models of fairness preferences.

In this paper, we present experimental results that provide this evidence. In one experiment, subjects pay a small premium to allocate equitably rather than leave a much smaller payoff to another party. However, when the impact of the choice on the other party is unknown because payoff information is hidden, subjects choose not to (costlessly) reveal this information and instead maximize their own payoffs. Thus, in a situation where choosing not to acquire information does not directly result in an unequal outcome, subjects essentially avoid an opportunity to behave altruistically.

In a second experiment, we employ a game with two strategic players and a third strategically irrelevant dummy player where the payoffs are similar to those in the first experiment. We find that players choose strategies consistent with maximizing their own payoffs at a cost to the third "player" similar to the selfish allocation in the first experiment. This happens in spite of the fact that either player can guarantee an equal allocation to all players. Therefore, when subjects are faced with a choice that guarantees

a fair outcome versus a choice that might produce either a fair outcome or a selfish one (depending on what another player does), they tend to pick the latter.

Based on the results of these experiments, we question the general validity of theories that model fairness and altruism as preferences for others' welfare or for equitable outcomes. Behavior in our experiments is inconsistent with these models. We suggest that a better description of "altruistic" or "fair" behavior – which is consistent with the above results – is that people dislike taking actions that are inconsistent with their beliefs about themselves. A key part of this interpretation is that people care not only about outcomes that result from their own choices, but also about the choices themselves and what they imply for these "self-beliefs." Specifically, people have beliefs about what kind of person they are and about what this implies for what "should" or "ought" to be done, and they dislike taking actions that are inconsistent with these beliefs, independently of the consequences of these actions. For most people placed in experimental situations such as the dictator game, the relevant self-beliefs are "I am a fair person" and "fair people do not harm others out of self-interest." Therefore, they are averse to keeping all of the money because it implies selfishness, which is inconsistent with most people's self-beliefs. However, in our experiments, we blur the connection between "selfish" behavior and harm to others, making self-interested behavior more acceptable, even though the consequence is the same as in the dictator game. We develop this theory of "self-belief consistency" – including how to model it – after presenting the experiments.

Experiment 1

Subjects were 102 undergraduates at the University of Pittsburgh who participated voluntarily in response to advertising for paid decision experiments. All experimental sessions were run with at least 12 subjects. The experimental material was presented via computer interface, and all interaction occurred via the computers. Exactly half of all subjects were assigned to the role of allocator (henceforth proposer) while the rest were in the role of another party (henceforth receiver) whose payoff was determined by a proposer. The roles of proposer and receiver were presented to subjects as Player X and Player Y, respectively.

Subjects were randomly assigned to one of three conditions. In the *Known* condition, proposers chose one of two allocations: ($6, $1) or ($5, $5) for themselves and a matched anonymous receiver. These allocation choices were labeled with a context-neutral "A" and "B" respectively. All payoffs represented dollars that the subjects received at the conclusion of the experiment in addition to a $5 participation bonus.

Note that this decision differs from the standard dictator game where a proposer divides an endowment with another party. In this treatment, there is a binary allocation choice and the self-interested option also represents a loss of social welfare. We chose this payoff structure for two reasons. First, we wanted to prompt a large proportion of benevolent choices in the Known condition, thus making a demonstration of self-interest under minor situational manipulations more striking. Second, the binary choice allows for a simple classification of actions as either selfish or generous**.**

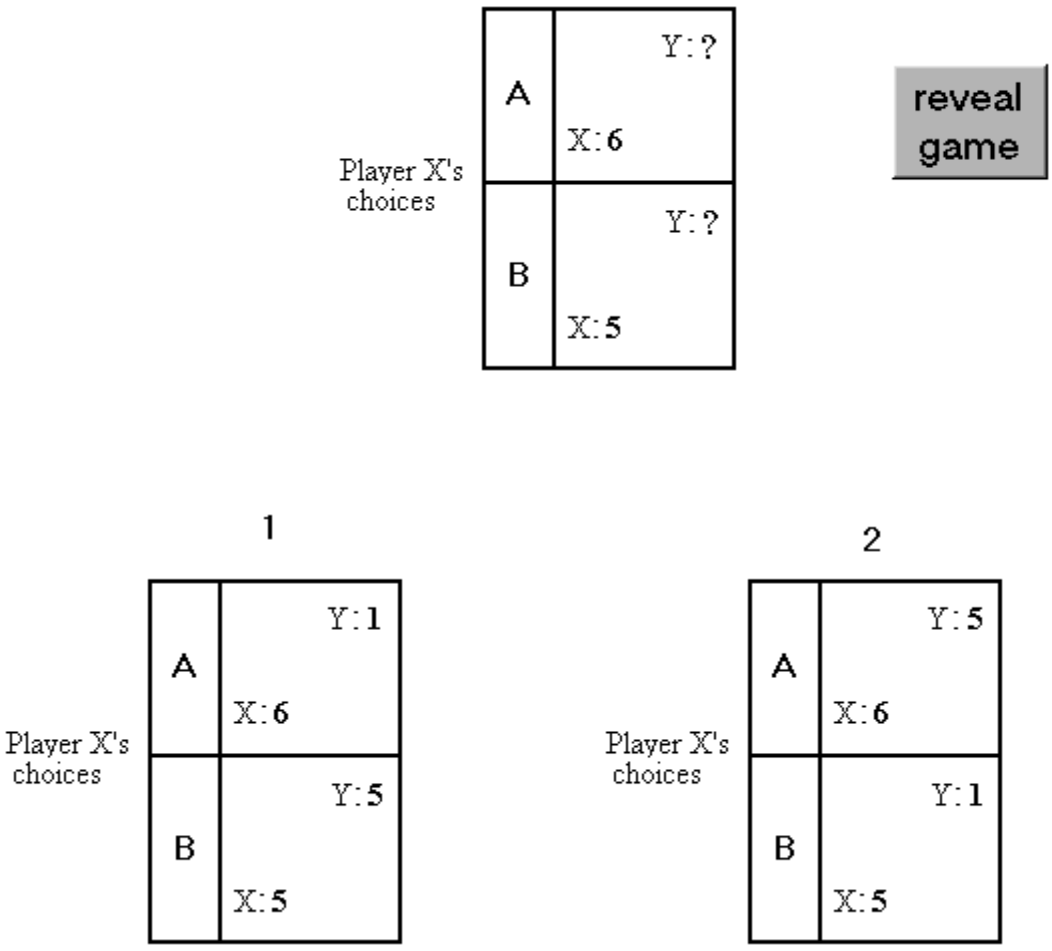


Figure 1. Unrevealed Conditions

In the *Unrevealed 1* and *Unrevealed 2* conditions, proposers also received $6 by choosing "A" and $5 by choosing "B". However, they did not initially know whether their choice of "A" or "B" would cause the receiver to get $5 or $1. This is depicted in Figure 1, which shows part of the interface seen by subjects in these conditions. In the

Unrevealed 1 condition, matrix 1 (which presents the payoffs for the Known condition) was the true payoff table for all pairs, while in the Unrevealed 2 condition, matrix 2 was the true payoff table. Subjects were made aware that the actual payoff table was determined by a coin flip prior to the session. In both unrevealed conditions, subjects were instructed that Player X could privately find out which game was being played by clicking a button, but that clicking the button was not compulsory.

Upon entering a large room with several computer terminals, subjects were randomly assigned identification numbers that they entered into their interfaces. They were instructed that they would be playing a simple game with one other person in the room with whom they would be matched randomly. Subjects were instructed that both players would be paid according to the choice that Player X made. After receiving instructions describing a generic payoff table (see appendix), a short quiz was given to ensure that the task and the payoff representation were understood. Subjects were then informed of the actual payoffs for the experiment. Role assignment and matching were done automatically according to the subject ID numbers after all instructions were completed. Matching and other subjects' roles and choices were anonymous to all participants. While proposers made their choices, receivers answered a question about what they would choose if they were in the role of Player X (in the Unrevealed conditions, they were instructed to assume the game being played was game 1). This protected the anonymity of the roles by ensuring that everyone was clicking on something when proposers made their choices. Upon completion of the game, subjects were paid in private as they exited the room.

Experiment 1 Hypotheses

Four hypotheses will serve as benchmarks for our results. First, as described earlier, a majority of proposers in previous experiments involving comparable decisions chose the fair allocation. As our payoff matrix tests a conceptually similar choice but with different actual dollar amounts, we wish to confirm this result in our sample:

> HYPOTHESIS 1. Most proposers will choose the equitable "B" allocation in the Known condition.

If this hypothesized generosity occurs, most of the existing economic interpretations and theories would involve some sort of stable preference for generosity. However, if there is such a preference, then a trivial situational manipulation such as covering payoffs that can be easily uncovered should not matter in terms of the ultimate allocation choices. Thus, our null hypothesis for the manipulation:

> HYPOTHESIS 2a. There will be no difference in proposer choices between the Known and Unrevealed 1 conditions.

However, we believe that preferences for giving are more complex. Specifically, not revealing payoffs allows proposers to behave self-interestedly without knowingly harming the recipient (creating a justification for self-interest). Thus, the following hypothesis predicts a change in behavior between the two payoff-identical conditions:

HYPOTHESIS 2b. Significantly more proposers will choose the self-interested "A" allocation in the Unrevealed 1 condition than in the Known condition.

Finally, we predict that the reason for the observed difference in behavior will be due to proposers choosing not to acquire the payoff information in the Unknown condition.

HYPOTHESIS 3. A significant proportion of proposers will opt not to reveal the true state in the Unrevealed conditions.

Experiment 1 Results

The results are summarized in Table 1, which shows the proportion of subjects making each choice by condition. The bottom two rows present the hypothetical choices made by receivers (who had no payoff-relevant choices in the experiment).

Proposers		
condition	Proportion choosing "A"	Proportion revealing
known	5/19 (26%)	--
unrevealed 1	10/16 (63%)	8/16 (50%)
unrevealed 2	13/16 (81%)	10/16 (63%)
Receivers		
known	0/19 (0%)	--
Both unrevealed	13/32 (41%)	--

Table 1. Experiment 1 results

Known condition: Thirty-eight subjects were assigned to the Known condition (19 proposers). Results clearly support hypothesis 1. When payoffs were visible, 14 of 19 proposers (74%) chose B. This closely resembles the proportion of equitable dividers reported by Kahneman, Knetsch and Thaler (1986). Of the 19 receivers in the known condition, all 19 indicated that they would choose B if in the role of the proposer. Since the receivers' choices were inconsequential, they may be viewed as little more than cheap talk. However, we feel that the unanimity of these responses, along with their high concordance with proposers' choices, indicates a strong expectation of what proper behavior is in these situations.[2]

Unrevealed conditions: Sixty-four subjects were assigned to the Unrevealed conditions, 32 in the Unrevealed 1 condition. Of the 32 proposers in either Unrevealed condition, 15 (47%) opted not to reveal the true state, providing support for hypothesis 3.

Two of the proposers revealed the true state in the Unrevealed 1 condition, but still chose option "A". Thus, they chose the ($6, $1) allocation when $6 would have been earned by choosing A regardless of the true state. Therefore, these proposers revealed value for the knowledge even though it did not impact their choices (presumably they would have chosen the ($6, $5) allocation if they found out they were in Unrevealed 2). An obvious interpretation for this behavior is curiosity (and self-interest).[3] Note that this behavior works against one of our main hypotheses (hypothesis 3) since subjects are choosing to view the payoffs even though it does not affect their choice.

---

[2] Indeed, Kahneman, et al., demonstrated the consequences of violating this expectation in a second stage of their dictator experiment. They found that subjects would rather divide $10 with someone who earlier dictated an even split than divide $12 with someone who earlier dictated an $18-$2 split.

[3] However, at least one proposer appeared not to understand the game, choosing the strictly dominated B option in the unrevealed 2 condition after revealing the payoffs.

Perhaps most importantly, hypothesis 2b was also supported by our data, and thus the null hypothesis 2a was rejected. Ten out of 16 proposers (63%) in the Unrevealed 1 condition chose the A option, resulting in the inequitable ($6, $1) allocation. This behavior resulted in spite of the fact that proposers could costlessly reveal that the game was exactly the same as in the Known condition, where a majority (74%) chose the B option. The difference in these proportions is significant (Pearson $\chi^2 = 4.64$, df = 1, $p < 0.05$).

Proposers	
Group	Proportion choosing "A"
unrevealed 1 – chose to reveal (8/16)	2/8 (25%)
unrevealed 1 – chose not to reveal (8/16)	8/8 (100%)
unrevealed 2 – chose to reveal (10/16)	9/10 (90%)
unrevealed 2 – chose not to reveal (6/16)	4/6 (67%)

Table 2. Allocation Choices by Revelation Choices

In hypothesis 3, we suggested that the reason for the difference in proportions of "A" choices across the conditions would be that proposers choose not to reveal payoff information. If we take "B" choices in the Known condition as reflecting a taste for generosity, then we should expect a similar number of generous proposers in the unrevealed conditions. The generous course of action in those conditions is to reveal the true state so that the other-regarding option (which might also be self-interested) can be chosen. Thus, the proportion of proposers in the unrevealed conditions who both reveal

the true state and choose the other-regarding option ("B" in Unrevealed 1 and "A" in Unrevealed 2) should be at least as high as the proportion choosing "B" (74%) in the Known condition.[4] This was not the case, as can be seen in table 2. Only 15 of the 32 proposers (47%) in the Unrevealed conditions revealed the true state and chose the other-regarding option (6 in Unrevealed 1 and 9 in Unrevealed 2). A chi-square analysis revealed this proportion to be significantly different than the proportion choosing B (14/19) in the Known condition (Pearson $\chi^2 = 3.49$, df = 1, $p < 0.07$). This provides evidence that some would-be "B" choosers are opting not to reveal the true state so that they may play "A".

The results of Experiment 1 are inconsistent with a model that incorporates a preference for others' outcomes into the decision maker's utility function. Our subjects appear to have some taste for equitable social allocations when the situation is unambiguous: that is, when they are directly confronted with others' payoffs and there is a direct link between their immediate action and the outcomes for these others. However, this taste for equity is sharply reduced when allocations can be made – if the proposer desires – in ignorance of others' payoffs. If subjects preferred equity, they should have sought the virtually costless information about others' payoffs in order to obtain equity, but they often did not. The subtle change of covering the other party's payoffs breaks the direct link between actions and payoffs: the immediate action of the proposer (to reveal or not reveal payoffs) does not have a direct impact on the welfare of the receiver. By not revealing the true state, subjects then avoid having to make an explicit decision not to give. Like the person who avoids placing himself on the donor registry to avoid opting

[4] In fact, we would expect this proportion to be higher because the choice A in Unrevealed 2 is both selfish (it yields the highest payoff for the proposer) and other-regarding (it yields the highest payoff for the receiver), and because some people – as we find above – are likely to look simply out of curiosity.

for a painful procedure to help someone else, our subjects choose not to know another's payoffs, as this information might only tempt them to forego a small amount of money to help the other person.

Experiment 1 demonstrated that subjects will choose to be selfish if this choice does not directly ensure a harmful consequence for another because the consequences are unknown (and will forego an opportunity to become informed of the consequences). What if, however, the option of ensuring equity was always present? That is, if subjects could ensure equity, but if not doing so did not ensure inequity, would we still observe selfish behavior? In a sense, this would be a stronger test of "fairness preference" models, since subjects could guarantee fair outcomes but not doing so would not guarantee unfair ones. We explored this possibility through the addition of a second strategic player in experiment 2.

Experiment 2

Experiment 2 essentially extends the Known condition from Experiment 1 by adding a second strategic player, as depicted in Figure 2. This changes the decision in that the responsibility for determining the equitable outcome now <u>appears</u> shared, rather than residing solely with one individual. We say "appears" because, in fact, each individual can still individually guarantee the equitable outcome.

The game employed in Experiment 2 is strategically simple. Two players (Player X and Player Y) choose actions A and B, as they did in Experiment 1, and action A is again each of these players' payoff-maximizing choice. However, as before, choosing A implies (if the other strategic player also chooses A) a loss for a passive recipient.

If subjects have a taste for equity, as proposers revealed in the Known condition of Experiment 1, then we might expect a similarly high frequency of B choices, by which either player can guarantee an equitable payoff. Therefore, if giving reflects a stable fairness preference, then behavior should be comparable to that of the Known condition in Experiment 1.[5]

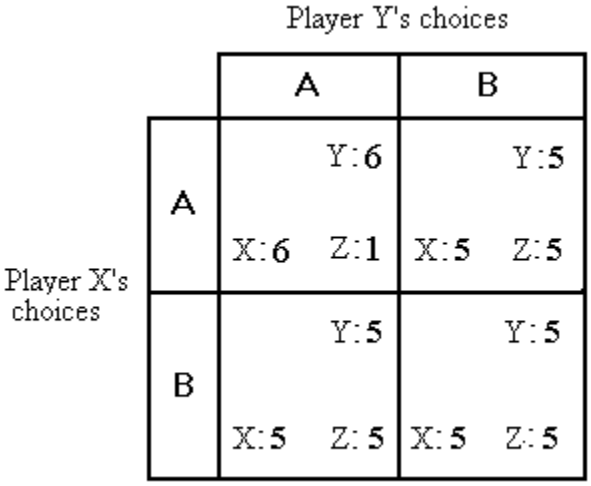


Figure 2. Three player game in Experiment 2.

Subjects were 30 undergraduates at the University of Pittsburgh. Two experimental sessions were each run with 15 subjects present. The procedures were similar to those in Experiment 1, except that subjects were presented with the game depicted in Figure 2. Two thirds of the subjects (20) were assigned to strategic player

[5] Of course, there is a difference between the two experiments in that a choice of B now affects the other strategic player who may have preferred the inequitable outcome. However, the payoff difference for the other player is 1, while this difference is 4 for the receiver, implying that a subject in the role of Player X or Y would have to care about the other strategic player 4 times as much as she cares for the receiver in order for this difference to completely compensate for differences in equity between the two outcomes. If, as is more likely the case, the welfare of both other "players" is valued equally, then the loss of 1 should only matter slightly. Moreover, this loss only negatively affects other players who preferred outcome A to B in the Known condition, which we saw were a minority in Experiment 1.

roles, the rest were passive receivers. The roles were introduced as Player X, Player Y, and Player Z, with Player Z being the strategically irrelevant player. Subjects were instructed that all three players would be paid according to the combined choices of Player X and Player Y. As before, the procedures were introduced using a generic payoff table and subjects answered questions to ensure they understood how payoffs were determined. All roles, groupings, and choices made in the experimental sessions were anonymous to other subjects. While those subjects assigned to the role of X or Y made their choices, those assigned to Z answered a question about which option they thought the majority of players would choose. Upon completion of the game, subjects were paid in private as they exited the room.

Experiment 2 Hypotheses

The addition of a strategic player affects the allocation decision from Experiment 1 in potentially important ways. While strategic players know that they can ensure an equitable division, they also cannot be responsible for an inequitable split. Even if they choose A, they have not ensured that the third party will receive the poorer outcome through their actions alone. The third party will only be hurt if the other player chooses "A" as well. The removal of responsibility for the third party's outcome could elicit a majority of "A" choices, despite the indication from the Known condition of Experiment 1 that players prefer to forego the larger payoff in order to ensure equity.

As the results of the Known condition indicate that most subjects exhibit preferences for equitable outcomes, we might expect subjects to ensure the fair outcome with a similar frequency:

HYPOTHESIS 4a. The proportion of subjects choosing A will be the same as the proportion choosing A in the Known condition of Experiment 1.

The presence of the third player does not reduce the ability of a subject to ensure the equitable outcome. However, the results of Experiment 1 indicate that subjects may not exhibit a stable preference for equitable outcomes and that reducing the transparency of the link between their actions and fair or unfair outcomes may significantly alter their behavior. Therefore, the introduction of the third player may allow subjects (similarly to Experiment 1) to perceive the lack of a direct link between their actions and harm to the receiver.

HYPOTHESIS 4b. The proportion of "A" choosers will be significantly higher than the proportion choosing "A" in the known condition of experiment 1.

Experiment 2 Results

The results of Experiment 2 and the Known condition of Experiment 1 are presented in Table 2. As expected, a significant proportion of subjects (35%) chose action B, indicating some preference for fairness.[6]

As pointed out earlier, a stable preference for fairness implies roughly the same proportion of subjects choosing B in the three player game as in the Known condition of Experiment 1 (Hypothesis 4a). Comparing the results in Table 2, however, it is clear that

[6] This game is similar to one used by Guth, et al., (1998). In their three-person ultimatum game, proposers offered three-way divisions of a pie between themselves, responders, and strategically irrelevant dummy players. Their findings were conceptually similar: low offers to third parties were rarely rejected, and thus proposers and responders acted in coalition to maximize their own payoffs at the expense of the third party.

the proportions are quite different. While 74% of subjects choose B in the Known condition, only 35% chose B in the three-player game of Experiment 2. This difference is statistically significant, confirming hypothesis 4b (Pearson $\chi^2 = 5.87$, df = 1, $p < 0.05$). Further, it seems that those in the role of receiver (Player Z) shared our intuition. All ten of them correctly predicted that A would be the most common choice among Players X and Y. This is in sharp contrast with the receivers' hypothetical actions in the Known condition of Experiment 1, where no recipients (out of 19) indicated they would choose A.[7]

	proportion choosing "A"
Experiment 2	13/20 (65%)
Experiment 1 (Known)	5/19 (26%)

Table 2. Results of Experiment 2 and Known condition in Experiment 1

The results of Experiment 2 further strengthen our claim that our subjects' behavior is inconsistent with a model that assumes a stable preference for giving. In this experiment, the option of ensuring equity was available to all players. However, neither player could ensure inequity by playing in accord with self-interest, thus decoupling self-interest from directly being unfair or harmful. As a result, the option of ensuring fairness was infrequently exercised. It seems that this apparent lack of responsibility for ensuring

[7] Of course, this comparison is slightly awkward since in experiment 1 the passive players indicate their own hypothetical action while in experiment 2 they indicate their expectation of the behavior of the other players. We make the comparison simply to illustrate that the expectations of the passive participants about appropriate/actual behavior show a similar pattern in both experiments.

equity is enough of a justification for subjects to behave selfishly, just as ignorance of payoffs presented an excuse for selfish behavior in Experiment 1. Moreover, the receivers appeared to share our intuition, as none expected that the equity ensuring option would be most frequently chosen.

Toward an improved theory of fair behavior

Behavioral economic models explain generosity in laboratory experiments by assuming that fair outcomes provide utility to the giver. While we replicate findings of generosity in the Known condition of our first experiment, we show that much of this generosity quickly disappears when changes are made to the situation that ought to be largely inconsequential. Our findings are difficult to reconcile with the stable preference for equity or for others' outcomes that is the basis of most "fairness preferences" models.[8] Instead, we believe a better model of social behavior is one in which people seek to maximize their own self-interest subject to "constraints," but will exploit moral "wriggle room" around these constraints. The constraints result from a desire to behave consistently with beliefs about themselves. People seek to maximize their welfare, but cannot do so (or perhaps experience disutility from doing so) if it involves taking an

[8] Rabin (1995) presents a model of "fairness constraints" that is consistent with the results of experiment 1. In his model, decision makers will only take selfish actions if the probability that they will harm others is sufficiently low. When these probabilities are fixed and known, the model predicts behavior that is identical to the models of "fairness preferences" that we challenge with our results. However, Rabin shows that when decision makers can acquire information about the true probability that their action will cause harm – as in our first experiment – fairness preferences predict that people will acquire such information but fairness constraints allow them to manipulate their information acquisition to remain with the belief that their actions will not cause excessive harm. This is similar to what we show in experiment 1, where subjects choose to acquire no new information about whether their preferred action harms the other subject before making a decision. However, it is not clear how Rabin's model can explain the results of experiment 2.

action that is inconsistent with "self-belief." [9] For many people, this includes causing harm to another purely out of self-interest (especially when this other has done nothing to deserve the harm).

Our interpretation of the laboratory evidence of generous or fair behavior is simple: Most people believe they are not the type to harm others purely out of self-interest. Therefore, taking an action that harms another (by, for instance, leaving them with a small amount of compensation for participating in an experiment) when there is no justification other than self-interest is inconsistent with the preceding self-belief. If I believe I am not the type of person to harm others out of self-interest, then I find it difficult to reconcile being unfair out of self-interest (the only interpretation for choosing A in the Known condition of experiment 1) with my beliefs. Our argument is that people reject specific actions that result in such inconsistency – or at least experience disutility from doing so.[10]

The decision not to become informed about the consequences of our actions is not the same as harming others out of self-interest. Neither is taking an action that might or might not harm another. Therefore, subjects in our Unrevealed conditions and in

[9] Our argument is consistent with Murnighan, Oesch, and Pillutla's, (1999) "self-impression management," which they use to explain heterogeneity in giving behavior in dictator experiments. Returning to the example given in the introduction, one can avoid feeling selfish – and thereby maintain a positive self-impression – by not being on the donor list, but not giving when one knows they have a match is less excusable.

[10] We note at this point that "not harming another" typically involves social considerations and depends on the expectations of the other party, even when that party is anonymous. Thus, if the receiver were to be unaware of any choices the proposer made, such that any payoffs sort of "fell from the sky," we speculate that proposer would maximize self-interest at the receiver's expense, which differs from many notions of pure morality or fairness preferences.

experiment 2 can behave "selfishly" without violating the constraint of not being unfair out of self-interest.[11]

While our interpretation is similar to Rabin's (1995) notion of "moral constraints" (instead of "moral preferences") it is important to note that under our interpretation the constraints arise only out of a desire to take actions that are consistent with one's self-beliefs and avoid actions that are inconsistent, rather than from any moral beliefs that one holds. In this sense, decision-makers derive utility – or disutility – from actions themselves (specifically, what the actions imply about the decision-maker), independently of the outcomes that result from these actions.

Modeling "self-belief consistency"

Of course, the value for economics of a theory such as the one we outline above lies in the extent to which it can be included in formal models of decision-making. Therefore, we discuss two ways – derived from existing behavioral economic models – in which our argument might be extended to a formal economic decision-making framework.

A key assumption in both of the models below is that people care not only about outcomes (as in the fairness preference models), but also about the choices that they make and what these choices imply about the person who makes them.[12] This idea is not new. An existing body of work in psychology focuses on "self-perception theory," which assumes that people make inferences about themselves by observing the actions that they

[11] The fact that these actions are not really viewed as being "too selfish" or "unfair" is reflected in the expectations of the passive participants.

[12] Weber (2003) finds that people learn how to play games by making choices in games repeatedly, even when they do not observe the outcomes of these choices. We view this result and our interpretation of the results in this paper as related: What we do is often as informative and valuable as the outcomes that result.

take, and as a result, form beliefs about their selves and their attitudes (Bem, 1967).[13] Our main contribution is to apply it to explain "other-regarding" behavior in light of our results.

Personal identification constraints

Akerlof and Kranton (2000) present a model in which people derive utility from their personal identifications (such as group affiliations), I_j, as well as from the combination of own and others' actions ($U_j = U_j(a_j, a_{-j}, I_j)$) and where the strength of their identifications is affected by their own actions ($I_j = I_j(a_j, \_)$). We propose a specific way in which this kind of model might explain apparently altruistic or fair behavior.

Our modification to the above model is simple. The relationship between I_j and a_j is determined by a set of behavioral proscriptions (or prescriptions) for a particular identity or self-belief. Taking an action in the proscribed set means that the identification I_j is changed, resulting in a decrease in utility if the decision-maker places positive value on that particular identification.[14] [15]

To give an overly simplified example, a devout Christian might hold Christianity as his or her only identification. This person will derive utility from taking actions that

[13] Another area in which self-signaling might drive behavior is work on incentives and the "crowding out" hypothesis (e.g., Gneezy, 2003). Work in this literature finds that people are often less willing to do something for small amounts of money than they are for free. To the extent that engaging in certain activities for free (such as helping a neighbor with yard work) tells us we are nice people, but receiving a small amount of pay ($1) for the same activity means that we cannot draw the same inference about ourselves, the explanation for behavior is the same as under our theory. Also similarly to our experiments, breaking the direct link between behavior and incentives (e.g., replacing the $1 with a cold beer afterwards or with compensation that is not explicitly tied to the behavior), makes it much more likely that we will engage in the activity.

[14] Similarly, it might be the case that taking an action inconsistent with a negative identification (such as a person who ashamedly views herself as a smoker turning down a cigarette) increases utility by decreasing the strength of the negative identification.

[15] Interestingly, Plott (1972) presents a similar theory of social choice in which the goal of social choice is to maintain consistency between policies (actions) and "ethics" (feelings or attitudes about which actions should or should not be taken).

maintain or strengthen this identification, even if this means doing something that goes against self-interest. (As an extreme example, a recent documentary on British television presented a group of Christians – the Jesus Christians – who actively seek out the opportunity to donate one kidney to a stranger as part of strengthening their faith).[16] However, engaging in a behavior that is proscribed by this identification (for instance, violating one of the Ten Commandments) will decrease the strength of her Christian identification (her belief that she is a Christian) and decrease her utility. Note that the reduction in utility comes not from violating the "rule" itself, but instead from the accompanying decrease in the validity of the identification. This example is similar to Akerlof and Kranton's (2000) model, in which people gain utility from behaving in a manner implied by a group with which they identify. Our modification is simply to introduce the possibility that certain behaviors are explicitly proscribed or prescribed (such as guiding one's decisions by the question "what would Jesus do?") and that behaviors not in these sets do not produce a similar effect on I_j – even though they may result in the same outcomes.

This kind of model implies that in cases where there are two (sets of) actions that result in identical outcomes – as in the Known and Unrevealed 1 conditions of our experiment 1 – but where one of the actions is in the proscribed set of behaviors, the agent will avoid implementing a particular outcome in one case but not in the other. This is what we believe occurs in our experiment 1 where choosing A in the Known condition is "being selfish at another's expense," but choosing not to become informed and then choosing A (with unknown consequences) in the Unrevealed conditions has a more ambiguous interpretation. We believe that many people avoid taking the action that can

[16] ."Kidneys for Jesus," Jon Ronson, Channel 4.

only be interpreted as "being selfish at another's expense" because this is inconsistent with being a "fair" person.

Consistent "self-stories"

Another way to model our basic theory is based on Lam's (2002) model of non-Bayesian inference. In this model, people form "beliefs" about reality by evaluating a finite set of propositions, each of which they either agree or disagree with. People do this through "reasoning chains" that specify a relationship between the propositions. For instance, propositions might be of the form x = "the defendant's fingerprints are on the murder weapon" and y = "the defendant is guilty," and a reasoning chain might simply be of the form x → y. Lam applies his model to learning; people in his model respond to new information (a signal) by iterating through the reasoning process until they come up with a plausible "story" that is both consistent with the signal and internally consistent (i.e., the reasoning chain reaches a fixed point).

One approach to modeling self-belief consistency is to modify Lam's model to allow propositions to be about ourselves (e.g., "I am a fair person") and about the relationships between people and actions (e.g., "fair people do not hurt others out of self interest"). Actions can then include either a utility from behaving consistently with our current "self-story" (i.e., a "story" in Lam's sense, but about ourselves rather than about an external state of the world) or a disutility from having to reason to a new "self-story." Such disutility might even be infinite, resulting in a constraint preventing us from taking any action that is inconsistent with our self-story. Under this kind of model, people would seek to take actions that are consistent with the way they view themselves, which

might include some people who consider themselves "self-interested" (or "rational economists") keeping all the money in a dictator game.

In both of the above models the exact kind of behavior that is consistent with an identification or "self-story" can be vague and may or may not apply to a particular situation or behavior depending on how it is defined. However, it is precisely this vagueness that could underlie heterogeneity in other-regarding behavior: some people may find certain behaviors consistent with an identification or self-belief, while others may not. As an example, the belief about one's self: "I am a good spouse" can have very different implications for what constitutes acceptable behavior (with respect to cheating, spousal abuse, etc.) depending on the individual.

The self-beliefs that people hold are likely more complicated and varied than those we use in the examples above. For instance, some people might not believe that harming others is inconsistent with their self-beliefs. Murnighan, Oesch, and Pillutla (1999) show that certain people's selfish behavior in dictator experiments is consistent with an "economist's" self-impression that it is acceptable to be self-interested. Similarly, our beliefs about ourselves may include group identifications that make it acceptable – or even desirable – to hurt members of other groups.

Situations themselves might also have implications for what behavior is consistent with our self-beliefs. For instance, the influence of social norms has been demonstrated to be a significant influence on positive or negative social behavior (e.g., Cialdini, Kallgren and Reno, 1991). Social norms might provide justification for harmful action in certain circumstances or situations (times of war, when others have harmed us in the

past). These social norms may replace statements such as "good people do not harm others" with "good people do not harm others, unless these others have harmed us in the past." While the impact of individual heterogeneity and social norms undoubtedly complicates our model, we believe that their effect is important, and that properly incorporating them into a model of social behavior is desirable.

In spite of the drawbacks of the above modeling approaches, we can make a couple of predictions based on our theory. First, when people's identifications are primed, their behavior and beliefs about qualities they possess should shift in the direction of being more consistent with the identifications (e.g., Hogg and Turner, 1987; Simon and Hamilton, 1994). Specifically, when we force people to think of what they believe about themselves, their behavior should become even more extreme in adhering to the prescriptions or proscriptions related to this identity.

In addition, when there is a behavior that is inconsistent with a belief that a majority of people hold about themselves, few people should engage in this behavior. This implies that if we can manipulate the extent to which the behavior is viewed as inconsistent with this self-belief, we should also observe changes in the frequency of the behavior. This is what we believe we do in our experiments, where we allow people to behave more self-interestedly by relaxing the applicability of the description "harmful to others" to their behavior.

While the above models represent only the first stages in how to model the phenomenon demonstrated in our experiments, our broader point about the validity of "fairness preference" models is clearly supported by our experiments. Subjects in our experiments simply do not exhibit a preference for implementing fair or equitable

outcomes. In fact, changing the situation only slightly – to break the direct link between actions and harmful consequences – results in a significant change in behavior in the direction of self-interest.

Conclusion

We demonstrate situations in which preferences for others' welfare vary in a manner inconsistent with theories of fairness preferences. While the situations captured by our experiments make up a small part of the total array of real-world giving environments, we feel they are an important part and that the challenges we raise for current theories of fairness and altruism are significant. While the precise mechanism driving our subjects' behavior is as yet inconclusive, we suggest that a better description of why most people give is that they seek to avoid committing acts that are unambiguously interpretable as harming others out of self-interest, because this is inconsistent with the beliefs they hold about themselves.

A final point deals with the applicability of experimental results outside of the laboratory. While a large body of economic research uses laboratory tests of situations like the dictator game to develop new theories of behavior, an obvious drawback to these theories has to do with how general situations like the dictator game really are. We argue that situations in which someone is forced into an unambiguous choice of how selfish to be are rare outside the laboratory. In the real world, people often make earlier decisions that might or might not lead them to such a situation – such as whether to acquire information – and the link between actions and outcomes in those situations might be much less clear than it is in the dictator game. Paying attention to these kinds of issues –

and incorporating them into laboratory experiments – is both important and useful for developing more general theories of behavior.

References

Akerlof, G. and R. Kranton. (2000). Economics and identity. The Quarterly Journal of Economics, 115(3): 715-753.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm glow giving. *The Economic Journal*, 100, 464-477.

Andreoni, J. (1995). Warm glow versus cold prickle: The effect of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, 110, 1-21.

Andreoni, J., Miller, J. (2002). Giving according to GARP: an experimental test of the consistency of preferences for altruism. *Econometrica*, 70, 737-753.

Babcock and Loewenstein. (1997). "Explaining bargaining impasse: The role of self-serving biases." *Journal of Economic Perspectives*, 11(1).

Bolton,G. E., Ockenfels, A. (2000). A theory of equity reciprocity and competition. *American Economic Review*, 100, 166-193.

Byrne, M.M., Thompson, P. (2001). Screening and preventable illness. *Journal of Health Economics*, 20, 1077-1088.

Camerer, C. (2003). Behavioral Game Theory: Experiments on Strategic Interaction. Princeton, NJ: Princeton University Press.

Carrillo, J., Mariotti, T. (2000). Strategic ignorance as a self-disciplining device. *Review of Economic Studies*, 67, 529-544.

Charness, G., Rabin, M. (2000). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117, 817-869.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58,* 1015-1026.

Dawes, R. M., J. McTavish and H. Shaklee. 1977. "Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation." Journal of Personality and Social Psychology, 35: 1-11.

Fehr, E., U. Fishbacher and S. Gächter. (2002). Strong reciprocity, human cooperation and the enforcement of social norms. Human Nature 13: 1-25.

Fehr, E., G. Kirchsteiger and A. Riedl. (1996). Involuntary unemployment and non-compensating wage differentials in an experimental labour market. The Economic Journal, 106(434): 106-121.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition and co-operation. *Quarterly Journal of Economics*, 114, 817-868.

Forsythe, R., Horowitz, J., Savin, N., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6, 347-369.

Guth, W., van Damme, E. (1998). Information, strategic behavior, and fairness in ultimatum bargaining: An experimental study. *Journal of Mathematical Psychology*, 42, 227-247.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior*, 7, 346-380.

Hogg, M. A. and J. C. Turner. (1987). Intergroup behaviour, self-stereotyping and the salience of social categories. British Journal of Social Psychology, 26: 325-340.

Jovanovic, B., Stolyarov, D. (2000). Ignorance is bliss.

Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness and the assumptions of economics. *Journal of Business*, 59, 285-300.

Levine, D. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1, 593-622.

Loewenstein, G., Adler, D. (1995). A bias in the prediction of tastes. *Economic Journal*, 105, 929-937.

Murnighan, J.K., Oesch, J.M., Pillutla, M. (1999). Player types and self impression management in dictator games: Two experiments. *Games and Economic Behavior*, 37, 388-414.

Plott, C. R. (1972). Ethics, social choice theory and the theory of economic policy. *Journal of Mathematical Sociology*, 2: 181-208.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases. Unpublished manuscript.

Rabin, M. (1993). Fairness equilibirum…

Simon, B. and D. Hamilton. (1994). Social identity and self-stereotyping: The effects of relative group size and group status. Journal of Personality and Social Psychology, 66: 699-711.

Thaler, R.H. (1992). The winner's curse: Paradoxes and anomalies of economic life. Princeton, NJ: Princeton Univ. Press.

Weber, R. A. (2003). Learning and transfer of learning with no feedback: An experimental test across games. Working paper.

Appendix A – Instructions.

All Conditions

This is an experiment in the economics of decision-making. Several research institutions have provided funds for this research. You will be paid for your participation in the experiment. The exact amount you will be paid will depend on your and/or others' decisions. Your payment will consist of the amount you accumulate plus a $5 participation bonus. You will be paid privately in cash at the conclusion of the experiment.

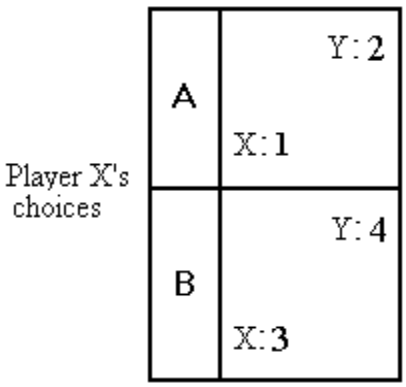
If you have a question during the experiment, raise your hand and an experimenter will assist you. Please do not talk, exclaim, or try to communicate with other participants during the experiment. Please put away all outside materials (such as book bags, notebooks) before starting the experiment. Throughout the experiment, please do not click "Continue" until the experimenter tells you to do so. Participants violating the rules will be asked to leave the experiment and will not be paid.

Known and Unrevealed

In this experiment, each of you will play a game with one other person in the room. Before playing, we will randomly match people into pairs. The grouping will be anonymous, meaning that no one will ever know which person in the room they played with. Each of you will be randomly assigned a role in this game. Your role will be player X or player Y. This role will also be kept anonymous. The difference between these roles will be described below. Thus, exactly one half of you will be a Player X and one half a Player Y. Also, each of you will be in a pair that includes exactly one of each of these types.

The game your pair will play will be like the one pictured below. Player X will choose one of two options: "A" or "B". Player Y will not make any choice. Both players will receive payments based on the choice of Player X. The numbers in the table are the payments players receive. The payments in this table were chosen only to demonstrate how the game works. In the actual game, the payments will be different.

For example, if player X chooses "B", then we should look in the right square for the earnings. Here, Player X receives 3 dollars and Player Y receives 4 dollars. Notice that player X's payment is in the lower left corner of the square, player Y's payment is in the upper right corner.



At this point, to make sure that everyone understands the game, please answer the following questions:

In this example, if Player X chooses "B" then:
Player X receives __
Player Y receives __

In this example, if Player X chooses "A" then:
Player X receives __
Player Y receives __

<answers read aloud>

The actual game you will play is pictured below. Note that in this game, Players X gets the highest payment of $6 by choosing A, but this gives Player Y the lowest payment of $1. However, if player X chooses B, player X gets a lower payment of $5, while Player Y also gets a payment of $5. Since we will only play this game once and then end the experiment, please take a minute to think about the game.

<subjects see figure 1, matrix 1>

Unrevealed Only

The actual game you will play will be one of the two pictured below. Notice that both games are the same except that Player Y's payments are flipped between the two. Note that in both games, Player X gets his or her highest payment of $6 by choosing A. In the game on the left, this gives Player Y his or her lowest payment of $1. In the game on the right this gives Player Y his or her highest payment of $5. In both games, if Player X chooses B, he or she gets a lower payment of $5. In the game on the left, this gives Player Y the highest payment of $5. In the game on the right, this gives Player Y the lowest payment of $1.

You do not know which of the games you will be playing. However, note that for Player X, the payments will be identical. The only thing that differs is the payments for Player Y.

The actual game you will play was determined by a coin flip before the experiment. However, we will not reveal publicly which game you are actually playing. Before playing, Player X can choose to find out which game is being played, if they want to do so, by clicking a button. This choice will be anonymous, thus Player Y will not know if X knows which game is being played. Player X is not required to find out and may choose not to do so. When the game ends, we will pay each player privately.

<subjects see figure 1>

At this point, to make sure that everyone understands the game, please answer the following questions:

In both games, which action gives player X his or her highest payment of $6? __

If Player X chooses B, then Player Y receives __

a) $5
b) $1
c) either $5 or $1

Three Player Only

In this experiment, each of you will play a game with two other people in the room. Before playing, we will randomly match groups of three people. The grouping will be anonymous, meaning that no one will ever know which two people in the room they played with. Each of you will be randomly assigned a role in this game. Your role will be player X, player Y, or player Z. This role will also be kept anonymous. The difference between these roles will be described below.

The game your group will play will be like the one pictured below. Player X and Player Y will separately and independently choose one of two options: "A" or "B". Both will make their choices at the same time without knowing the other's choice. Player Z will not make any choice. All 3 players will receive payments based on the combined choices of Player X and Player Y. The numbers in the table are the payments players receive. The payments in this table were chosen only to demonstrate how the game works. In the actual game, the payments will be different.

For example, if player X chooses "B" and player Y chooses "A", then we should look in the bottom left square for the earnings. Here, Player X receives 7 dollars, Player Y receives 8 dollars, and Player Z receives 9 dollars. Notice that player X's payment is in the lower left corner of the square, player Y's payment is in the upper right corner and player Z's payment is in the lower right corner.

		Player Y's choices: A	Player Y's choices: B
Player X's choices	A	Y:2 X:1 Z:3	Y:5 X:4 Z:6
	B	Y:8 X:7 Z:9	Y:11 X:10 Z:12

In this example, if Player X chooses "A" and Player Y chooses "B" then:

Player X receives __
Player Y receives __
Player Z receives __

In this example, if Player X chooses "B" and Player Y chooses “A” then:

Player X receives __
Player Y receives __
Player Z receives __

The actual game you will play is pictured below. Note that in this game, Players X and Y get their highest payment of $6 by choosing A, but this gives Player Z the lowest payment of $1. However, if either player chooses B, they both get a lower payment of $5, while Player Z also gets a payment of $5. Since we will only play this game once and then end the experiment, please take a minute to think about the game.

<subjects see figure 2>