8-31-2005

Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness

Jason Dana
University of Illinois at Urbana-Champaign

Roberto A. Weber
Carnegie Mellon University, rweber@andrew.cmu.edu

Jason Xi Kuang
Georgia Institute of Technology - Main Campus

Published In
.

Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness

Jason Dana
Department of Psychology, University of Illinois Urbana-Champaign

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

Jason Xi Kuang
College of Management, Georgia Institute of Technology

August 31, 2005[*]

[*] We thank Cristina Bicchieri, Iris Bohnet, Colin Camerer, Robyn Dawes, Ernst Fehr, George Loewenstein, John Patty, Charlie Plott, Matthew Rabin, and seminar participants at Carnegie Mellon, Berkeley, Cornell, Emory, and Princeton, and participants at the 2003 Public Choice / Economic Science meetings in Nashville, the 2003 Economic Science meetings in Pittsburgh, the 2003 Society for Judgment and Decision Making meetings in Vancouver, and the 2004 Behavioral Decision Research in Management meetings in Durham 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 are often generous. This behavior is often interpreted as reflecting a preference for equitable, efficient, or otherwise desirable social outcomes. We show that a considerable proportion of such fair behavior may be driven by a desire to appear fair without actually wanting a fair outcome. To do so, we first demonstrate a high frequency of fair behavior in a modification of the standard dictator game, but then show that fairness decreases substantially when the connection between choices and outcomes is obfuscated.

Specifically, we show that in a binary version of the dictator game, a majority of subjects choose the fair and efficient outcome. We then show that subjects playing the same game instead choose to maximize their own payoffs, at the expense of fairness and efficiency, when the recipients' payoffs are uncertain, even if this uncertainty can be costlessly resolved. We also find that when either of two subjects can sacrifice to implement a fair and efficient outcome, but neither can ensure fairness or inefficiency, selfishness prevails. Finally, we also find less fair behavior when unfair outcomes can plausibly result from an external factor rather than just the dictator's choice.

Our results indicate that much fair behavior may reflect motivations other than simply a preference for desirable social outcomes. Instead, much of the behavior in our experiments is consistent with a desire to pursue self-interest, while maintaining the illusion of behaving fairly.

Do people share with others because they want to, or because they feel they have to in certain situations? This paper explores the motivation for fair behavior. In particular, we explore whether experimental demonstrations of people sacrificing monetary payoffs to benefit others truly represent evidence of concern for others' welfare or for desirable social outcomes. To do so, we use experimental manipulations that allow subjects to leave the relationship between their actions and others' outcomes uncertain. These manipulations thus allow subjects to choose selfishly without knowing whether they negatively impact others' payoffs. However, in our manipulations subjects may always implement the fair outcome with certainty if they so desire. We find significantly less generous behavior in these manipulations relative to a baseline in which the relationship between actions and others' outcomes is certain, as in the standard dictator game. This finding occurs in spite of the fact that the recipients are anonymous and cannot retaliate. We conclude that people are often fair because they intrinsically dislike appearing unfair, either to themselves or others.

I. Background

Subjects across a variety of experiments show apparent concern for others' welfare, beyond any concerns for reputation or punishment. This phenomenon is perhaps clearest in dictator games, where a subject in the role of "dictator" divides an endowment between herself and an anonymous "recipient" who must accept the division. A purely self-interested dictator should keep the entire endowment. However, a review of several such experiments reveals that a majority of dictators share a positive amount and that the average amount shared is over 20% of the endowment (see Camerer, 2003). Even when elaborate steps are taken to

ensure double-blind anonymity, the amount given to an unknown recipient is still often greater than zero (e.g., Hoffman, et al., 1994).

Many ”social preference” theories have attempted to reconcile such sharing with the traditional economic assumption of self-interest by assuming that people share because they prefer ”good” social outcomes. For instance, people may share with others because they derive utility (e.g., a “warm glow”) from the basic act (Andreoni 1990; see also Andreoni and Miller, 2000). Alternatively, agents may be averse to large payoff differences (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) or may desire to maximize total social payoffs or the lowest payoff to any one party (Charness and Rabin, 2002; Engelmann and Strobel, 2004). Thus, if a dictator gives, for instance, it is assumed that she has revealed her preference for a more equitable distribution of payoffs than pure selfishness ensures.

Yet, there may be other important motives for giving that we cannot adequately capture with payoffs alone. For example, people in situations such as the dictator game might be averse to appearing unfair, either to themselves or to others. Thus, the underlying motivation driving much fair behavior observed in experiments might be self-interest, coupled with a desire to maintain the illusion of not being selfish.

In the next section, we report an experiment demonstrating how the desire to maintain the illusion of fairness can be satisfied even while the outcomes resulting from one’s actions are unfair. That is, we show that we can “turn off” a substantial proportion of fair behavior by introducing minor or irrelevant modifications to the decision context that allow unfair outcomes to be produced while maintaining the illusion of behaving fairly. Specifically, we show that in situations like the standard dictator game, where the link between the dictator’s actions and social outcomes is transparently clear, many subjects in the role of dictator are

fair. However, when we obscure this link, so that a meager payoff to the recipient might have been due, e.g., to nature rather than the dictator, we observe significantly more selfishness.

Part of the purpose of our paper is to shed light on limitations of current models that posit that the motivation for fair behavior is a desire to implement socially desirable – e.g., equitable or efficient – outcomes. We show that such outcomes are implemented significantly less frequently when minor modifications to the decision context mean that the dictator is not transparently responsible – to either herself or the recipient – for any inequitable or inefficient outcomes. However, it is important to note that in all the modifications in our experiment, a dictator who wants to implement the "desirable" social outcome can do so with certainty.

Of course, this paper is not the first to point out problems with models that account for fair behavior by a distributional preference. For instance, intention-based models, in which the utility or disutility of social outcomes is influenced by individuals' intentions in producing those outcomes are consistent with much experimental data (e.g. Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; see also, Blount, 1995; Falk, Fehr, and Fischbacher, 2003). However, the features that drive fair behavior in such models are held constant in our experiment – for instance, the recipient never takes any action – so these intentions-based models do not explain our results.

Our work is more closely related to that of Konow (2000) and Rabin (1995), whose models posit a fairness standard against which people compare their behavior, but that such a standard serves as a constraint that individuals seek to circumvent rather than a goal that they seek to implement. Thus, as in their models, many subjects in our experiment appear to be motivated by self interest, while attempting to maintain the belief that such behavior is not inconsistent with fairness.

An important property of dictator experiments is that there is a commonly known one-to-one mapping from actions into outcomes. Thus, the dictator knows the precise monetary payoffs associated with any action choice, and the recipient can infer precisely what action the dictator chose upon receiving a payoff. This is a property we term "transparency." Our modifications all relax transparency, allowing either the dictator to not know the precise consequences of a self-interested action or the recipient not to know how an unfair payoff came about. As our results indicate, these kinds of modifications move the modal behavior from fairness to self-interest. Specifically, we show that in situations like the standard dictator game, where the link between the dictator's actions and social outcomes is transparent, many subjects in the role of dictator are fair. However, when we obscure this link, so that bad social outcomes do not certainly and directly result from the dictator's choice, we observe significantly more selfishness.

In our framework, we argue that in transparent situations, much sharing occurs because people feel compelled to do so, without really wanting to. For these people, however, choosing fairness over self-interest is conditional on the situation being transparent. These people do not have a preference for socially desirable outcomes – in fact, they might wish they did not feel compelled to sacrifice in the pursuit of such outcomes. Thus, we demonstrate, using three manipulations that relax transparency, that a significant proportion of people exploit such "moral wiggle room" in order to behave self-interestedly. When we relax transparency, behavior always moves in the direction of self-interest.

In the first manipulation, the payoffs to the recipient are initially hidden, and are either conflicting or mutual with the dictator's interests. We find that many dictators choose not to reveal the true payoffs, even though doing so is costless, which allows them to maximize their

own payoffs while maintaining doubt as to whether this decremented the recipient's payoff. Thus, dictators are able to avoid giving by remaining strategically ignorant of the consequences of their actions.[1]

In a second manipulation, there are two strategic players ("dictators") and a passive recipient. Either dictator can implement a fair outcome, but neither dictator acting alone can implement the unfair outcome. That is, we introduce "diffused responsibility" by breaking the one-to-one mapping between actions and outcomes: a dictator choosing fairly knows precisely what outcome will obtain, but a dictator choosing unfairly essentially leaves it up to the other dictator to ensure fairness. We again find that a majority of dictators choose selfishly.

In the final manipulation, we introduce the possibility that dictators might be randomly "cut off" by the computer before making a choice, but receivers do not know whether the cut-off happened and, therefore, do not know precisely how their payoffs were determined. We again find more selfish choices and also find that many dictators appear to make intentionally slower choices, thus allowing themselves to be cutoff.

We now describe the design and results of our experiment.

II. Experiment

The basic decision faced by subjects in the role of dictator is a simple binary choice between a "fair" and an "unfair" outcome. The fair outcome yields $5 for the dictator and $5 for a passive recipient. The unfair outcome yields $6 for the dictator and $1 for the recipient.[2]

[1] Economists have noted other instances in which people might employ strategic ignorance. For instance, Carillo and Mariotti (2000) model agents who choose to be ignorant in the present so as to decrease the amount of an externality on their future selves. Their model differs from the sort of ignorance we describe in that it relies on time discounting that reflects impatience.

[2] Thus, the fair outcome also maximizes social welfare.

Therefore, our baseline case involves a transparent decision resembling a standard dictator game.[3]

A. Baseline treatment

In our Baseline treatment, dictators chose between one of the two allocations. These allocation choices were labeled with a context-neutral "A" and "B" respectively (see Figure 1). Subjects received the payoffs that resulted from this choice – in addition to a $5 participation bonus – privately at the conclusion of the experiment.

1. Experimental design

Subjects were 38 undergraduates (19 pairs) 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. Upon entering a large room with several computer terminals, subjects were randomly assigned identification numbers that they entered into their interfaces. These numbers determined their role assignment ("Player X" (dictator) or "Player Y" (recipient)). Subjects were instructed that they would be playing a simple game with one other person in the room, with whom they would be matched anonymously and randomly. Subjects were told that both players would be paid according to the choice made by Player X. After receiving instructions describing a generic payoff table (see appendix for instructions), a short quiz was given to ensure that the task and the payoff

[3] In the standard dictator game a sender divides a fixed endowment ($10) 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 three reasons. First, by giving the receiver a large marginal benefit from the sender's sacrifice of one dollar, we gave our subjects more of a reason to give than in the standard dictator game, thus also allowing for greater contrast when we remove transparency. Second, the binary choice allows for a simple classification of actions as either selfish or generous. Third, a binary choice game makes it easier to implement manipulations such as payoff uncertainty.

representation were understood. Subjects were then shown the actual payoffs for the experiment. The experimental stimuli were presented via computer interface, and all interaction occurred via the computers. As in all subsequent treatments, immediately prior to making a choice, subjects were given sixty seconds – during which the payoff matrix and choice interface were displayed on the screen – before the software allowed them to make their choice. While dictators made their choices, recipients chose hypothetically between the two options, serving in part to maintain the anonymity of the roles. Upon completion of the game, subjects were paid privately as they exited the room.

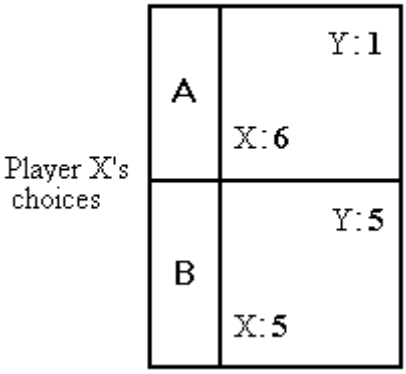


Figure 1. Subject's view of payoffs in baseline treatment

2. Results

As expected, we observed a significant amount of fair behavior. Of the 19 subjects in the role of dictator (Player X), 14 (74%) chose the fair option (*B*). All 19 recipients (Player Y) hypothetically chose the fair option (*B*).

The large amount of sharing observed above is consistent with a preference for implementing equitable or efficient outcomes. However, it is also consistent with the idea that dictators feel compelled to give in transparent situations. To test this alternate

hypothesis, we examine how minor modifications to the decision context influence the amount of sharing. We anticipate that relaxing transparency will provide subjects with the "moral wiggle room" to choose self-interestedly, and thus will significantly decrease giving.

B. Hidden information treatment

In our first manipulation of the baseline game, dictators also faced a choice in which they could receive \$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. That is, dictators did not know if the actual payoffs were (*A*:\$6,\$1; *B*:\$5,\$5), as in the baseline treatment, or (*A*:\$6,\$5; *B*:\$5,\$1). Each case was equally likely and dictators could reveal the true payoffs costlessly.

If the giving we see in the baseline treatment reflects a preference for socially desirable outcomes, then we would expect to see a similar proportion of subjects in this manipulation reveal the true payoffs and choose the most equitable action. That is, dictators who prefer the fair outcome in the baseline (i.e., the 74 percent who chose *B*) should acquire the costless payoff information and then choose the option that yields \$5 for the recipient.[4]

1. Experimental design

Subjects were 64 undergraduates (32 pairs) at the University of Pittsburgh. Basic procedures were identical to the baseline case (see appendix). The only difference was that subjects were told that the true payoffs were given by one of two tables (see Figure 2), and had been determined by a coin flip prior to the session. Subjects were instructed that the true

[4] That is, if the true payoffs are (*A*:\$6,\$1; *B*:\$5,\$5) a dictator who preferred a fair payoff distribution would want to know and choose *B*, while if the true payoffs are (*A*:\$6,\$5; *B*:\$5,\$1) such a dictator would want to know and choose *A*.

payoffs would not be revealed publicly, but that Player X could, if he or she wanted, click a button to privately reveal which game was being played. Subjects were also informed that Player Y would not know whether Player X revealed the payoffs.

As before, subjects in the role of Player Y made a hypothetical choice. These subjects were told – for this choice – to assume that the true payoffs were the same as in the baseline.

Figure 2 about here

2. Results

The results are summarized in Table 1, which shows the proportion of subjects making each choice, depending on the true underlying payoffs, as well as the corresponding proportion in the baseline treatment. Of the four sessions conducted, two had the same payoffs as the baseline treatment and the other two had the alternate payoffs.[5] As a result, 16 dictators faced each payoff table. The bottom of Table 1 reports the hypothetical choices made by recipients in the two treatments.

Of the 16 dictators who faced the same payoffs as in the baseline, ten (63%) chose *A*, resulting in the inequitable ($6,$1) outcome. This behavior resulted in spite of the fact that dictators could costlessly reveal that the payoffs were exactly the same as in the baseline treatment, where a majority of dictators (74%) chose the fair option (*B*). The difference in these proportions is statistically significant (Pearson $\chi^2 = 4.64$, df = 1, $p < 0.05$).

[5] Prior to conducting the four sessions, we flipped a coin to determine which two would have each set of payoffs. In this way, each set of payoffs was equally likely and the actual payoffs were determined by a fair coin flip.

Dictators (Player X) choices

Treatment	Proportion Choosing "A" (unfair choice)	Proportion Revealing True Payoffs
Baseline	5/19 (26%)	
Hidden information (Matrix 1 – baseline payoffs)	10/16 (63%)	8/16 (50%)
Hidden information (Matrix 2 – alternate payoffs)	13/16 (81%)	10/16 (63%)

Recipients (Player Y) hypothetical choices

Baseline	0/19 (0%)
Hidden information	13/32 (41%)

Table 1. Comparison of Baseline and Hidden Information treatments

To see why behavior shifted in the direction of self-interest, consider a dictator's decision of whether or not to acquire the payoff information. Table 2 presents choices broken down by true payoffs and payoff acquisition. As we discuss above, dictators motivated by a preference for socially desirable outcomes should reveal the true payoffs. However, we find that of 32 dictators, only 18 (56%) opted to reveal the true state.[6]

If all dictators who share are motivated by the goal of implementing socially desirable outcomes, then the proportion of dictators choosing to "act fairly" should be the same except for sampling error across the two treatments. That is, dictators in the hidden information treatment who both reveal the true state and choose the fair option (*B* in matrix 1, *A* in matrix

[6] Interestingly, two dictators revealed matrix 1, but still chose *A*. Because these dictators would presumably have also chosen *A* ($6, $5) with the alternate payoffs, these dictators acquired the information even though it did not impact their choices. Our interpretation of this behavior is curiosity (and self-interest). Additionally, one dictator chose *B* after revealing the payoffs to be the alternate payoffs. This dictator might have done so out of a desire to end with higher relative payoffs

2) should be at least as high as the proportion choosing *B* in the baseline treatment (74%).[7] However, as Table 2 shows, only 15 of 32 dictators (47%) revealed the true state *and* chose the other-regarding option. This proportion is significantly lower than the proportion who chose fairly (14/19) in the baseline treatment (Pearson $\chi^2 = 3.49$, df = 1, $p < 0.07$).

Actual payoffs	Information acquisition choice	Proportion choosing "A"
Matrix 1 (baseline payoffs)	Chose to reveal (8/16, 50%)	2/8 (25%)
	Chose not to reveal (8/16, 50%)	8/8 (100%)
Matrix 2 (alternate payoffs)	Chose to reveal (10/16, 63%)	9/10 (90%)
	Chose not to reveal (6/16, 38%)	4/6 (67%)

Table 2. Allocation Choices by Revelation Choices in Hidden Information Treatment

Interestingly, the behavior of dictators is mirrored by the hypothetical responses of recipients. While all recipients stated they would choose the fair option in the baseline treatment, only 59 percent said they would do so when the information was initially hidden. These proportions also differ significantly (Pearson $\chi^2 = 10.36$, df = 1, $p < 0.01$).

The results of our first manipulation are inconsistent with the idea that all fair choices in the baseline represent people who value implementing socially desirable outcomes. An apparent taste for equity is greatly reduced when allocations can be made – if the dictator desires – in ignorance of a recipient's payoffs. If the dictator remains ignorant, then he or she can believe that the responsibility for the recipient's outcome lies, at least partly, with the coin flip, eliminating the dictator's direct responsibility for outcomes. Moreover, the dictator will

[7] In fact, we would expect this proportion to be higher because choosing *A* with matrix 2 is both selfish (it yields the highest payoff for the sender) and other-regarding (it yields the highest payoff for the receiver).

never learn the recipient's outcome and the recipient will never know what the dictator knew at the time of the choice.

The hidden information treatment demonstrated that many subjects will choose to be selfish if this choice does not transparently determine adverse consequences for others.[8] What if, however, subjects always knew they could determine equity? That is, if dictators could ensure equity, but not doing so did not ensure inequity, would most behave fairly? We explore this possibility through the addition of a second strategic player.

C. Diffusion of Responsibility Treatment

In this manipulation, we introduce a second strategic player, making the dictator no longer solely responsible for implementing outcomes. In this situation, either "dictator" can ensure the fair outcome, which means that neither alone can ensure the unfair outcome. That is, both dictators must act selfishly in order for the unfair outcome to result.

The payoff table for this treatment (as presented to subjects) is depicted in Figure 3. In this game, there are two strategic players (X and Y), each of whom chooses between an unfair and a fair action – *A* and *B*, respectively. A third, passive, recipient (Player Z) receives a payoff based on the choices of the two dictators. If either dictator chooses *B*, then the equitable outcome results ($5, $5, $5). However, only if both dictators choose *A* does the unfair outcome result ($6, $6, $1). Thus, if an individual dictator values fair outcomes sufficiently, then he or she can choose *B*, but choosing the self-interested *A* option does not map one-to-one with the recipient getting $1. If sharing in the baseline treatment reflects a

[8] The difference between the baseline and hidden information treatments has been replicated using double-blind anonymity (Larson, 2005) and using a within-subjects design with varying probabilities of the two states (Munyan, 2005).

preference for the fair outcome, then we should expect a similar proportion of *B* choices here.[9] But if sharing in the baseline treatment does not reflect such a preference, then the presence of another dictator might partially relieve them from the duty of having to give.[10]

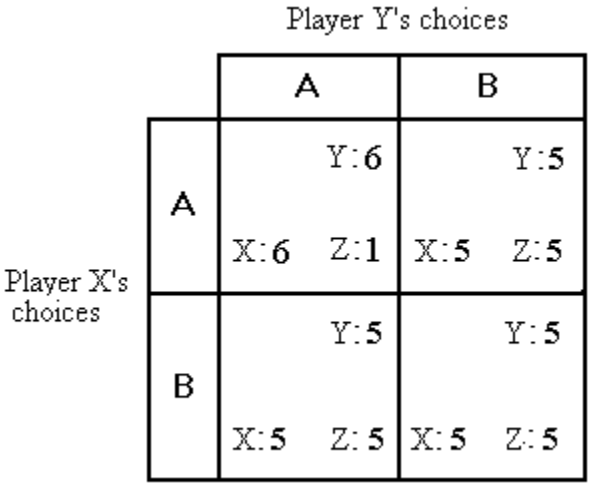


Figure 3. Payoff table for diffusion of responsibility treatment

1. Experimental design

Subjects were 30 undergraduates at the University of Pittsburgh. Two experimental sessions were each run with 15 subjects present. The procedures were the same as those in previous treatments, except that subjects were presented with the game depicted in Figure 3.

[9] Of course, there is a difference between the two treatments 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 baseline condition treatment, which we saw were a minority.

[10] A large body of literature in social psychology (e.g., Darley & Latane, 1968; Latane & Nida, 1981) demonstrates the negative social effects of "diffusion of responsibility." In situations involving bystander intervention (when individuals can intervene to help others or create a socially desirable outcome), pro-social behavior decreases as more people are present and able to help. This often produces the perverse result that *less* help is rendered when *more* people are available to help. In bystander intervention problems, there are also incentives to free-ride; the present manipulation involves, in a sense, a more pure test of diffusion of responsibility.

Two thirds of the subjects (20) were assigned to strategic player roles, the rest were passive recipients. The roles were introduced as Player X, Player Y, and Player Z, with Player Z the strategically irrelevant player. Subjects were told that all three players would be paid according to the combined choices of Players X and Y. All roles, groupings, and choices made in the experimental sessions were anonymous. While subjects assigned to the role of X or Y made their choices, those assigned to the role of Z indicated which option they thought the majority of players would choose. Subjects were paid privately as they exited the room.

2. Results

The results of the diffusion of responsibility and baseline treatments are presented in Table 3. While 74% of subjects chose *B* in the baseline, only 35% chose *B* in the three-player game. This difference is statistically significant (Pearson $\chi^2 = 5.87$, df = 1, $p < 0.05$). Further, it seems that recipients (Player Z) shared our intuition. All ten correctly predicted that *A* would be the most common choice by strategic players. This response pattern is in sharp contrast with the receivers' hypothetical actions in the baseline treatment, where no recipients (out of 19) indicated they would choose *A*.[11]

	Proportion choosing "A"
Diffusion of responsibility	13/20 (65%)
Baseline	5/19 (26%)

Table 3. Choices by dictators in baseline and diffusion of responsibility treatments

[11] Of course, this comparison is slightly awkward since in the baseline the recipients indicate their own hypothetical action while in this treatment they indicate their expectation of the behavior of the other players (we elicited the expectation because we realized after running the baseline that hypothetical responses might be influenced by social desirability). We make the comparison simply to illustrate that the expectations of the passive participants (in all three treatments thus far) about appropriate/actual behavior show a similar pattern to the actual behavior of dictators.

The results of the diffusion of responsibility treatment further strengthen our hypothesis that a great deal of giving does not result from a desire to implement socially desirable outcomes. In this treatment, the option of ensuring equity and efficiency was available to all strategic players. However, it was not transparent that choosing self-interestedly (*A*) adversely affected the recipient; the other strategic player could still ensure equity. As a result, the option of ensuring fairness was infrequently exercised. Moreover, the recipients appeared to share our intuition, as none expected that the equity ensuring option would be most frequently chosen.

In both manipulations thus far, allowing dictators uncertainty about the consequences of their choices, and not allowing recipients to map outcomes uniquely to dictators' choices, resulted in decreased giving. Of course, this raises an interesting question: Would dictators choose fairly if they knew the consequences of their choice with certainty, but recipients were unable to map outcomes uniquely to choices? To explore this question, we turn to a third manipulation in which dictators are faced with the possibility of losing their ability to make a choice, and recipients do not know whether this ability has been lost. This final manipulation allows us to measure the extent to which dictators share less in the first two manipulations because they are capitalizing on the recipients' incomplete information (i.e., engaging in "other–deception"), or because they are engaging in "self-deceptive" thinking at the time of their choice (e.g., "perhaps my choice will not harm the recipient").

D. Plausible deniability treatment

This manipulation employed a game identical to the baseline treatment, but with an additional "cutoff" feature. Dictators were provided with a 10 second interval during which

to make their choices. At a random point during the interval, the software would intervene and randomly choose between the options (*A* and *B*) if the dictator had not already chosen. Dictators did not know when the cutoff would occur, but knew that it could happen at any point during the 10 second interval. While recipients were aware of the possibility of a cutoff, only the dictator would know whether it actually occurred. Thus, a recipient could not be sure if her payoff was due to the dictator's choice or the software cutoff.

This design allows us, to some degree, to test between two alternate explanations for the "moral wiggling" observed previously. If the decrease in sharing in the previous two treatments was due to the fact that recipients were unaware of how outcomes resulted, then we would expect a significant proportion of dictators to capitalize on recipients' uncertainty and choose selfishly (*A*) immediately. However, if the decreased sharing in previous treatments is really due to dictators fooling themselves into believing that they did not cause bad outcomes, then we would expect many dictators to "dither" and allow themselves to be cutoff. The cutoff would produce the selfish outcome with probability 0.5, but the dictators would not be responsible for such an outcome, and with probability 0.5, the cutoff would leave them the fair outcome they would have felt compelled to choose anyway. Of course, if dictators prefer the fair outcome, the manipulation should prove largely irrelevant – meaning that most dictators would choose *B* prior to the cutoff (as we discuss below, the cutoff times were unlikely to be binding for dictators who did not want to be cut off).

1. Experimental design

Subjects were 58 undergraduates (29 pairs) at the University of Pittsburgh. Three experimental sessions were run with at least 18 subjects present at each. The procedures were

similar to those in the baseline, except that subjects were now also instructed and quizzed about the cutoff feature (see appendix). Specifically, subjects were told that dictators would have some amount of time of up to 10 seconds to make a choice, but that at a randomly selected time in this interval the computer would cut them off. They were told that if they were cut off, the software would choose randomly among options *A* and *B* with equal probability. While dictators made their choices, recipients made hypothetical choices, also with the possibility of a cutoff. As in earlier treatments, subjects were given one minute to think about the game (while viewing the choice interface) before playing. Further, the choice interval did not begin until subjects clicked a "begin game" button.

The cutoff times were determined by a discretized normal distribution.[12] Our goal was to allow most dictators the opportunity to choose before being cutoff. To test how long dictators normally took to make a choice, we conducted pre-testing in which we asked people to quickly make a choice (A or B) after clicking the "begin game" button. In this pre-testing, all choices were made in less than 2 seconds. The probability of being cutoff in less than 2 seconds in our design was about 3×10^{-5}, and in fact no one was actually cut off in less than 4 seconds, meaning that a dictator truly interested in implementing a particular outcome would almost certainly have enough time to do so. Thus, it was unlikely that time constraints should play a role in being cut off. Subjects who were cut off were asked to indicate how they would have responded, A or B, if they had not been cut off.

2. Results

[12] To be precise, cutoffs occurred exactly on one of the 10 seconds, with the greatest mass at 5 and 6 seconds. The mass placed on seconds 5 and 6 was equivalent to the area in the first standard unit (.34), seconds 4 and 7 the second standard unit (about .13), etc.

The results are presented in Table 4. Among those dictators who were not cut off (22/29), a majority (55%) chose the selfish action *A*. This differs significantly from the 26% who did so in the baseline treatment (Pearson $\chi^2 = 3.35$, df = 1, $p < 0.07$). Thus, recipients' uncertainty about how payoffs are determined appears sufficient to promote increased self-interest, even when dictators know the consequences of their actions.

	Dictators (n=29)	Recipients hypothetical choices (n=29)
Proportion cutoff	7/29 (24%)	11/29 (38%)
Average cutoff time for those cutoff	4.30	4.64
Proportion of A choices by those not cutoff	12/22 (55%)	5/18 (28%)
Total number of A outcomes	17/29 (59%)	
Proportion of those cutoff stating they would have chosen A	1/7 (14%)	3/11 (27%)

Table 4. Results in the plausible deniability treatment

However, a significant proportion of dictators (24%) allowed themselves to be cut off and therefore did not have to make a choice. The average cutoff time for these dictators was approximately 4.3 seconds, with none of the cutoffs occurring before 4 seconds, indicating that these dictators intended to wait at least this long before making a choice.[13]

In this manipulation, we again find significantly more selfish behavior than in the baseline. Compared to the 74 percent of fair choices in the baseline, we find here that only 10 out of 29 choices (34 percent) are truly consistent with a desire to implement the fair outcome.

[13] Recall that subjects were given a one-minute period prior to making their choice in which they looked at the locked interface and also had to click a "begin game" button to start the interval, meaning that the delay is unlikely to be due to confusion.

There also appears to be heterogeneity in how the unfair outcomes result – while some dictators directly choose *A,* others allow themselves to be cutoff by the computer, which then means a 0.5 probability of obtaining the self-interested but inequitable outcome.

III. Conclusion

Many behavioral economic models explain a dictator's generosity by assuming that she prefers socially desirable – e.g., fair or efficient – monetary outcomes. However, another possible explanation for giving in experiments is that dictators feel compelled to give in that particular context (e.g., because of guilt from behaving unfairly), but do not really value the fair outcome. To test this possibility, we relaxed the transparency – the common knowledge of a one-to-one mapping between actions and outcomes – of standard dictator experiments. Our hypothesis was that relaxing such transparency might allow those who feel compelled to give the "moral wiggle room" to behave self-interestedly while maintaining the illusion of fairness.

In our experiment, we first demonstrated that a significant proportion of subjects chose fairly when put in a transparent situation similar to the standard dictator game, thus replicating previous evidence. We then explored how this behavior changed with minor modifications to the decision context. We found that these modifications all produced significantly less sharing.

In all of our treatments, a dictator had the power to implement a fair outcome with certainty. Table 5 presents the proportion of dictators in each treatment that did so. As the comparison suggests, the manipulations resulted in significantly smaller proportions of subjects choosing fairly. Such behavior is clearly inconsistent with the idea that all sharing in

the baseline case is due to a preference for fair outcomes, but is consistent with the idea that many dictators – roughly a third – share without really wanting to.

Treatment	*Proportion implementing fair outcome*
Baseline	14/19 (74%)
Hidden information (baseline payoffs)	6/16 (38%)
Diffusion of responsibility	7/20 (35%)
Plausible deniability	10/29 (34%)

Table 5. Proportion of dictators implementing fair outcome across treatments

Two important caveats must accompany our results. First, a significant amount of sharing *is* consistent with the idea that people value implementing socially desirable outcomes. Across treatments, a robust percentage of dictators (roughly one third – see Table 5) opt to implement the fair outcome, even when the context leads others not to do so. Second, the fact that the proportion of dictators choosing to actively give decreases substantially does not mean that most subjects are choosing the unfair outcome. A large part of the change in behavior in our treatments occurs because dictators can choose in accord with self-interest while not knowingly decrementing others' payoffs. The proportion of dictators who choose – with certainty – to implement the unfair outcome is never greater than half.[14]

We draw three general conclusions from our experiment:

First, there is a difference between behaving consistently with fairness and acting to implement fair outcomes. Most current models of fairness, in which individuals are

[14] It is 24 percent in the baseline, 13 percent in the hidden information treatment (baseline payoffs), and 41 percent in the plausible deniability treatment. In the diffusion of responsibility treatment, it is impossible for a dictator acting alone to implement the unfair outcome.

motivated by a desire to implement socially "good" outcomes, make no such differentiation. However, as is evident in Table 5, the number of our subjects who seek fair outcomes in the three modifications is less than half than the number who do so in the baseline. Clearly, then, people care about fairness but will capitalize on uncertainty to be more selfish (cf. Schweitzer and Hsee, 2002). This suggests a need for improving on current theoretical models of fair behavior.[15]

Second, our results suggest that there is more than one motive for sharing in contexts like the dictator game. Some dictators behave consistently with a preference for fair outcomes. Others appear more concerned with the appearance of fairness than the realization of fair monetary outcomes (cf. Batson, Thompson, and Chen, 2002; Kagel, Kim, and Moser, 1996). This is consistent with other recent work in which players in a standard dictator game are given the option of not playing the game and receiving a fixed sum (in which case the recipient never finds out about the game). In such situations, a significant proportion of subjects choose to opt out – even if the sum is less than the dictator endowment (Dana, Cain & Dawes, 2004; Lazear, Malmendier & Weber, 2004).[16] However, a significant proportion of subjects (again, roughly a third) choose to continue sharing.

Finally, fairness often depends on normative considerations. In a given context, people are likely to be influenced by what others consider appropriate and what they expect others do. (e.g., Shang and Croson, 2003; Krupka & Weber, 2004). These normative considerations can be powerful determinants of pro-social behavior (Cialdini, Reno, and

[15] One theory that is consistent with some of our results is Rabin's (1995) model of "moral constraints." Agents will choose to consume only if the probability that it causes social harm is sufficiently low. When these probabilities are fixed and known, the model predicts behavior that is identical to the fairness preferences models. However, Rabin shows that when agents update their beliefs about the probability that their action will cause social harm – as in the hidden information treatment – they may refuse costless information or selectively choose signals so as to avoid believing the probability of harm is too high.

[16] See also, Oberholzer-Gee and Eichenberger (2004).

Kallgren, 1990). They may also cause decision makers to consider the implications of their actions for their own identifications as "good" people (Murnighan, Oesch, and Pillutla, 1999; Akerlof and Kranton, 2000). Indeed, various self-signaling models (Prelec and Bodner, 2003; Benabou and Tirole, 2005) consider the impact of choice on individuals' future perceptions of themselves, e.g. through the impact on one's self-respect. Thus, the expectations of even an anonymous other subject who cannot punish may play heavily in one's decision to give.

Our results could have several implications for understanding and improving ethical conduct. We briefly consider two examples. Education and testing are used as methods to prevent the spread of sexually transmitted diseases such as HIV infection. We imagine that many people would feel compelled to constrain their sexual behavior if they knew they were infected, which raises the possibility that they may not want information and testing. Another example is curbing financial fraud. In the spate of recent scandals, often high-level figures accused of transgressions must be shown to have known about harms in order to be held liable. We note that this ignores the efforts that executives may take to remain ignorant. In general, we feel that many people would feel constrained to be ethical, even without threat of punishment, if they were unable to remain ignorant of the consequences of their actions, or else if ignorance were not an excuse from moral responsibility.

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

Batson, D., Thompson, E., and H. Chen. (2002). Moral Hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, 83,330-339.

Benabou, R., and J. Tirole. (2005). Incentives and prosocial behavior. Unpublished manuscript.

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

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

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

Charness, G., and M. Rabin. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117, 817-869.

Cialdini, R. B., Reno, R. R., and C.A. Kallgren. (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.

Croson, R. and J. Shang. (2003). Social comparison and public goods provision in the field. Unpublished manuscript.

Dana, J., Cain, D. M., and R.M. Dawes. What you don't Know Won't Hurt me: Costly (but quiet) Exit in a Dictator Game. (October 17, 2004). http://ssrn.com/abstract=494422

Darley, J. and B. Latane. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8, 377-383.

Dufwenberg, M., and G. Kirchsteiger. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47, 268-298.

Engelmann, D., and M. Strobel. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94, 857-869.

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

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

Kagel, J., C. Kim and D. Moser. (1996). Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs. *Games and Economic Behavior*, 13, 100-110.

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

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90, 1072-1091.

Krupka, E. and R. A. Weber. (2004). Norm Salience and Observational History in Dictator Allocation Decisions. Unpublished manuscript.

Latane, B. and S. Nida. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89(2), 308-324.

Lazear, E., U. Malmendier, and R. A. Weber (2004). Sorting Opportunities in Decisions Inolving Fairness. Unpublished manuscript.

Lind, E. A. and T. R. Tyler. (1988). *The Social Psychology of Procedural Justice*. New York: Plenum.

Munyan, Lauren. (2005). Patterns of information avoidance in binary choice dictator games. Working paper.

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.

Oberholzer-Gee, F. and Eichenberger, R., "Fairness in Extended Dictator Game Experiments," *Working Paper* 2004.

Prelec, D. and R. Bodner. (2003). Self-signaling and self-control. in *Time and Decision*, G. Loewenstein, D. Read, & R.F. Baumeister (eds.), Russell Sage Press, New York.

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

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83, 1281-1302.

Schweitzer, M., and C. Hsee. (2002). Stretching the truth: Elastic justification and motivated communication of uncertain information. *The Journal of Risk and Uncertainty*, 25, 185-201.

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.

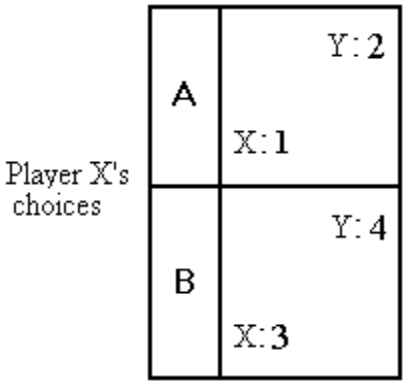
If you have a question during the experiment, raise your hand and an experimenter will assist you. Please do not talk, exclaim, or try to communicate with other participants during the experiment. Please put away all outside materials (such as book bags, notebooks) before starting the experiment. Throughout 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.

Baseline & Hidden Information

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>

Hidden Information 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

Diffusion of Responsibility 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

Player X's choices	A	B
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>

Plausible Deniability only

There is one additional feature to this game. Once the game starts, Player X will be given some amount of time up to 10 seconds to choose A or B. At a random "cutoff" point during that 10 second period, the computer will "step in" and make a choice for Player X if Player X has not already chosen. That is, the computer will randomly select a time between 0 and 10 seconds, and if Player X has not made a choice at this time, then the computer will choose for Player X.

If the cutoff occurs, the computer will randomly choose A or B with equal probability. Player X will then not be able to make a choice. However, if Player X chooses before the cutoff, Player X's choice will count and the computer will not be able to choose. Whether the computer or Player X makes the choice, both players will be paid according to whether A or B is chosen. Player Y will simply know how much money he or she gets from the choice of A or B, which could have been made either by the computer or Player X.

For example, suppose the cutoff happens at 3 seconds. If Player X hasn't chosen before 3 seconds have passed, then the computer randomly chooses A or B with equal probability. However, if Player X has chosen before 3 seconds have passed, his or her choice will be executed and the computer will not be able to affect it. Similarly, suppose the cutoff is 7 seconds. The computer will choose randomly for Player X if Player X has not chosen before 7 seconds, but will do nothing to the choice if Player X has chosen before 7 seconds.

Note that Player X has no way of knowing when the cutoff will occur. Also note that once the cutoff occurs, Player X will not be able to affect the computer's choice. Finally, note that Player Y will not know whether or not Player X was cut off.

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

The most time player X could have to make a choice is __ seconds.
The cutoff will happen at the end of the 10 seconds. (true/false).
The cutoff will be randomly selected by the computer and could come anywhere during the 10 seconds. (true/false).
If the computer cuts off player X, player X can affect the computer's choice. (true/false).

Figure 2. Interface for hidden information treatment