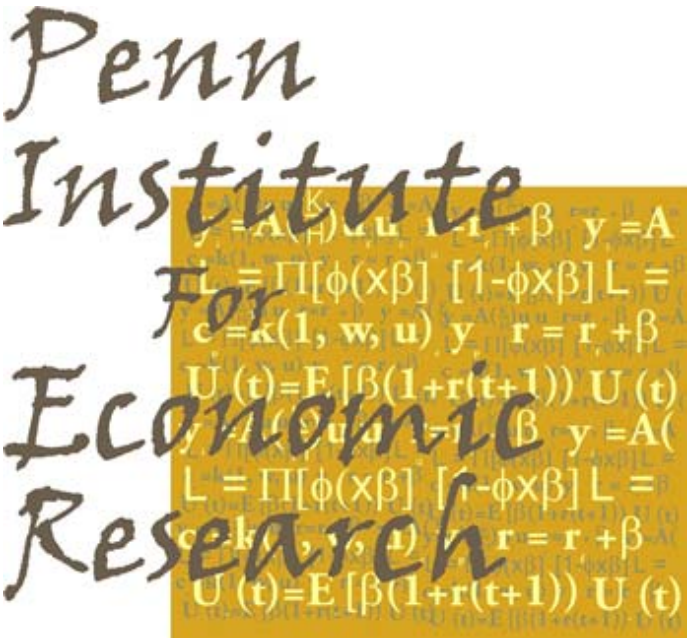

Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://www.econ.upenn.edu/pier

PIER Working Paper 09-014

"Ashamed to be Selfish"
Second Version

by

David Dillenberger and Philipp Sadowski

http://ssrn.com/abstract=1384176

Ashamed to be Selfish*

David Dillenberger† Philipp Sadowski‡

April 14, 2009

Abstract

We study a two-stage choice problem. In the first stage, the decision maker (DM) chooses a set of payoff-allocations between herself and a passive recipient. In the second stage, DM chooses an allocation from the set. The recipient is only aware of the second stage choice. Choosing selfishly in the second stage, in the face of a fairer available alternative, may inflict shame on DM. We axiomatize a representation of DM's preferences over sets that identifies DM's selfish ranking, her norm of fairness and shame. It has been suggested that altruism is a prominent motive for non-selfish choice. We identify a condition under which shame to be selfish can mimic altruism, when the experimenter only records the second stage choice. An additional condition implies that the norm of fairness can be characterized as the Nash solution of a bargaining game induced by the second-stage choice problem. The representation is applied to a simple strategic situation, a game of trust.

JEL Classifications: C78, D63, D64, D80, D81

1. Introduction

1.1. Motivation

The notions of fairness and altruism have attracted the attention of economists in different contexts. The relevance of these motives to decision making is intuitive and has been extensively studied. For example, in a classic "dictator game," where one person gets to anonymously divide, say, \$10 between herself and another person, people tend not to take the whole amount for themselves, but to give a sum between \$0 and \$5 to the other player (for a review, see Camerer (2003)). They act as if they are trading off a concern for fairness

*We thank Roland Benabou and Wolfgang Pesendorfer for their invaluable support. We are also grateful to Eric Maskin, Stephen Morris, Andrew Postlewaite, Charles Roddie and Tymon Tatur for helpful suggestions. This paper was written in part while the authors were graduate prize fellows at the University Center for Human Values, Princeton University. Financial support from the NSF under grant SES-0550540 is gratefully acknowledged.

†Department of Economics, University of Pennsylvania. E-mail: ddill@sas.upenn.edu

‡Department of Economics, Duke University. E-mail: p.sadowski@duke.edu

or for the other person's incremental wealth and a concern for their own. Thus, preferences for fairness as well as preferences for altruism have been considered (for example, Fehr and Schmidt (1999), Anderoni and Miller (2002), and Charness and Rabin (2002)).

Recent experiments, however, have challenged this interpretation. For example, Dana, Cain and Dawes (2006) study a variant of the same dictator game, where the dictator is given the option to exit the game before the recipient learns it is being played. In case she opts out, she is given a prespecified amount of money and the recipient gets nothing. About a third of the participants choose to leave the game when offered \$9 for themselves and \$0 for the recipient. Write this allocation as (\$9, \$0). Such behavior contradicts altruistic concern regarding the recipient's payoff, because then the allocation (\$9, \$1) should be strictly preferred. It also contradicts purely selfish preferences, as (\$10, \$0) would be preferred to (\$9, \$0). Instead, people seem to suffer from behaving selfishly in a choice situation where they could dictate a fairer allocation. Therefore, they try to avoid getting into such a situation, if they can. Two examples of real-life scenarios would be:

• donating to a charity over the phone, but wishing not to have been home when the call came

• crossing the road to avoid meeting a beggar

We contend that whether or not a person's actions are observed by someone who is affected by her choice plays a crucial role in determining her behavior.[1] We term "shame" the motive that distinguishes choice behavior when observed from choice behavior when not observed. In our model, individuals are selfish when not observed. Thus, concern for another person's payoff is motivated not by altruism, but by avoiding the feeling of shame that comes from behaving selfishly when observed.[2,3]The interpretation is that, if people are observed, they feel shame when they do not choose the fairest available alternative. Our explanation is supported by further evidence. In a follow-up to the experiment cited above, Dana *et al.* report that only one out of twenty-four dictators exits the game when second-stage choice is also unknown to the recipient. Similarly, Pillutla and Murningham (1995) find evidence

[1]We disregard any influence on the player's behavior caused by the presence of the experimenter. In our model, observation by the experimenter is not considered a reason for shame, as the experimenter is not affected by DM's choice.

[2]To distinguish shame from guilt, note that guilt is typically understood to involve regret, even in private, while, according to Buss (1980), *"shame is essentially public; if no one else knows, there is no basis for shame. [...] Thus, shame does not lead to self-control in private."* We adopt the interpretation that even observation of a selfish behavior without identification of its purveyor can cause shame.

[3]Of course, various other-regarding preferences that are not impacted by observation could be present as well (for a comprehensive survey, see Levitt and List (2007)). We do not account for those, as our aim is not to describe a range of possible attitudes toward others, but to highlight shame as a motive for giving. Relaxing our assumptions to allow for some other-regarding preferences, even without observation, would not qualitatively change our results.

that people's giving behavior under anonymity depends on the information given to the observing recipient. In experiments related to our leading example, Lazear, Malmendier and Weber (2005) as well as Broberg, Ellingsen and Johannesson (2008) predict and find that the most generous dictators are keenest to avoid an environment where they could share with an observing recipient. Broberg *et al.* further elicit the price subjects are willing to pay in order to exit the dictator game; they find that the mean exit reservation price equals 82% of the dictator game endowment. Tadelis (2008) studies a Trust Game and experimentally verifies a probabilistic version of our prediction: When moving from a game with no observation to a game with observation, both the likelihood of cooperation by the receiver and the likelihood of trust by the sender increase.

To better understand the notion of shame and its interaction with selfish preferences, we need to identify the effects of these two motives. A simple and tractable tool for analysis would be a utility that is additively separable in the moral cost (shame) and the private payoff, and that specifies the properties of the shame component (a similar utility is used, for example, by Levitt and List (2007)). We justify using this convenient form by deriving it from plausible assumptions on both preferences and the underlying norm of fairness. To this end, we consider games like the one conceived by Dana *et al.* as a two-stage choice problem. In the first stage, the decision maker (DM) chooses a "menu," a set of payoff-allocations between herself and the anonymous recipient. This choice is not observed by the recipient. In the second stage, DM chooses an alternative from the menu. This choice is observed, in the sense that the recipient is aware of the menu available to DM.[4,5] DM has well-defined preferences over sets of alternatives (menus). Our interpretation of shame as the motivating emotion allows considerations of fairness to impact preferences only through their effect on second-stage choices, where the presence of a fairer option reduces the attractiveness of an allocation. Our representation results demonstrate how DM's norm of fairness and her choice behavior interact. On the one hand, properties of the norm impact choice; on the other hand, the norm of fairness used by DM can be elicited from her choice behavior.

[4]The observed part of the choice procedure is naturally modelled as stage two: The recipient always learns the ultimate choice, as it determines his payoff. Prior to this, DM might be given the option to constrain the set of allocations available for choice. This preceding decision may or may not be observed. If it is not observed, then there is a meaningful first stage. The passage of physical time is not relevant for the distinction of the two stages. This is in contrast to most other models of choice over menus, where subjective uncertainty might be resolved or temptation may kick in over time.

[5]If the exit option is chosen in the aforementioned experiment by Dana *et al*, as in our setup, the recipient is unaware that there is a dictator who could have chosen another allocation. In their experiment, the recipient is further unaware that another person was involved at all. It would be interesting to see whether informing the recipient that some other person had received \$9 would change the experimental findings. This would correspond to our setup.

1.2. Illustration of Results

Denote a typical menu as $A = \{(a_1, a_2), (b_1, b_2), ...\}$, where the first and second components of each alternative are, respectively, the private payoff for DM and for the recipient. We impose axioms on DM's preferences over menus that allow us to establish a sequence of representation theorems. To illustrate our results, consider a special case of those representations:

$$U(A) = \max_{(a_1,a_2)\in A} [u(a_1) + \beta\varphi(a_1, a_2)] - \beta \max_{(b_1,b_2)\in A} [\varphi(b_1, b_2)], \qquad (*)$$

where u and φ are increasing in all arguments. u is a utility function over private payoffs and $\varphi(a_1, a_2)$ is interpreted as the fairness of the allocation (a_1, a_2).

Alternatively, if we denote by a^* and b^* the two maximizers above, it can be written as:

$$U(A) = \underbrace{u(a_1^*)}_{\text{value of private payoff}} - \underbrace{\beta(\varphi(b_1^*, b_2^*) - \varphi(a_1^*, a_2^*))}_{\text{shame}}.$$

This representation captures the tension between the impulse to maximize private payoff and the desire to minimize shame from not choosing the fairest alternative within a set. It evaluates a menu by the highest utility an allocation on the menu gets, where this utility depends on the menu itself. The utility function that is used to evaluate allocations is additive and has two distinct components. The first component, $u(a_1)$, gives the value of a degenerate menu (a singleton set) that contains the allocation under consideration. When evaluating degenerate menus, which leave DM with a trivial choice under observation, we assume her to be *Selfish*: she prefers one allocation to another if and only if the former gives her a greater private payoff, independent of the recipient's payoff. The second component is "shame." It represents the cost DM incurs when selecting (a_1, a_2) in the face of the fairest available alternative, (b_1^*, b_2^*).

As shame is evoked whenever this fairest available alternative is not chosen, we can relate choice to a second, induced binary relation "fairer than", which represents DM's private norm of fairness. Based on the definition that *"fair implies an elimination of one's own feelings, prejudices, and desires so as to achieve a proper balance of conflicting interests."* (Merriam-Webster Collegiate Dictionary (Tenth Edition, 2001)), DM's private norm of fairness is assumed to satisfy at least the following three properties: *Fairness Ranking*, which implies that the fairness comparison of any two alternatives is independent of the other available options; the *Pareto* criterion on payoffs; and *Compensation*, that allows any variation in the level of one person's payoff to be compensated by appropriate variation in the level of the other person's payoff.

In the special case considered here, the shame from choosing (a_1, a_2) in stage two is

$\beta(\varphi(b_1^*, b_2^*) - \varphi(a_1, a_2))$. This implies that even alternatives that are not chosen may matter for the value of a set, and larger sets are not necessarily better. To see this, consider the representation $(*)$ with $u(a_1) = a_1$, $\beta = \frac{1}{2}$ and $\varphi(a_1, a_2) = a_1 a_2$ and compare the sets $\{(10,1),(4,3)\}$, $\{(10,1)\}$ and $\{(4,3)\}$. Evaluating these sets we find $U\{(10,1),(4,3)\} = 9$, $U\{(10,1)\} = 10$ and $U\{(4,3)\} = 4$. To permit such a ranking, we assume a version of *Left Betweenness*, which allows smaller sets to be preferred over larger sets. Theorem 1 establishes that our weakest representation, which captures the intuition discussed thus far, is equivalent to the collection of all the above assumptions.

Representations similar to $(*)$ have been extensively studied in the literature on temptation, starting with the work of Gul and Pesendorfer (2001, henceforth GP). GP consider preferences over menus of lotteries and, furthermore, impose a version of the independence axiom. Whereas we feel that introducing (for technical reasons) uncertainty to an otherwise riskfree environment is debatable, imposing the independence axiom would be simply inappropriate in our context. For example, suppose that the fairness ranking of alternatives is symmetric. Then $(9,1)$ is as fair as $(1,9)$. Independence implies that any randomization over these two outcomes is as fair as either of them, while common sense suggests that awarding \$9 to either player with probability $\frac{1}{2}$ would be fairer.[6] In addition, the independence axiom implies a menu-independent second-stage choice criterion. Contrary to this, we argue that a higher degree of shame may well lead to a fairer choice. Our most general representation, Theorem 1, accommodates such a context-dependent choice criterion.[7]

The special case considered in $(*)$, on the other hand, does feature a context-independent choice criterion. To see this, regroup the terms as follows:

$$U(A) = \underbrace{\max_{(a_1,a_2)\in A} [u(a_1) + \beta\varphi(a_1, a_2)]}_{\text{second stage choice criterion}} - \underbrace{\beta \max_{(b_1,b_2)\in A} [\varphi(b_1, b_2)]}_{\text{effect of fairest alterative}}.$$

Theorem 2 establishes that, given the assumptions made so far, an additional separability assumption on preferences over sets, *Consistency*, is equivalent to the existence of such a choice criterion. Suppose that only the second stage of the choice procedure is observed (for example, because DM, as in the classic dictator game, never gets to choose between menus). If second-stage choice is context independent, shame might be mistaken for altruism: DM

[6] See section 5 for a context that naturally involves uncertainty and our suggestion for incorporating it into our model.

[7] In the context of temptation, Noor and Takeoka (2008) suggest relaxations of the independence axiom that allow menu-dependent choice. In Epstein and Kopylov (2007), the choice objects are menus of acts. They relax independence and characterize a functional form with a convex temptation utility. Independently of our work, Olszewski (2008) studies preferences over subsets of a finite set of deterministic outcomes and finds a representation where both choice and temptation are context dependent.

seems to trade off a selfish concern for his private payoff with a concern for the recipient's welfare. We argue, however, that it is hard to reconcile such an interpretation with any choice reversal in stage two. Thus, when observing stage two in isolation, shame can mimic altruism only if the induced choice ranking is context independent, or equivalently if the ranking of menus satisfies Consistency.

We further specify the norm of fairness by assuming that the private payoffs to the two players have *Independent Fairness Contributions*: Fairness should be concerned with utilities, not monetary payoffs, but interpersonal comparisons of utilities are infeasible. Thus, the fairness contribution of raising one player's monetary payoff can not depend on the level of the other player's payoff. With this additional assumption, Theorem 3 establishes that there are two utility functions, v_1 and v_2, evaluated in the payoff to DM and the recipient respectively, such that the value of their product represents the fairness ranking, $\varphi(a_1, a_2) = v_1(a_1) v_2(a_2)$. Thus, the fairest alternative within a set of alternatives can be characterized as the Nash Bargaining Solution (NBS) of an associated game. Because the utility functions used to generate this game are private, so is the norm. We argue that when based on true selfish utilities, the NBS is a convincing fairness criterion in our context. Those utilities may not be known to DM (especially in anonymous choice situations) but one can assess the descriptive appeal of the representation by asking whether the utilities comprising DM's norm at least resemble selfish utilities.

Example: Let $u(a_1) = a_1$, $\varphi(a_1, a_2) = v_1(a_1) v_2(a_2) = a_1 a_2$ and $\beta = \frac{1}{2}$. This implies that the utilities v, which are used to generate the fairness ranking, coincide with u. Shame is half the difference between the Nash-product of the fairest and the chosen alternatives. In the experiment by Dana *et al.* mentioned above, only whole dollar amounts are possible allocations. The set $A = \{(10,0)(9,1)(8,2), ..., (0,10)\}$ then describes the dictator game. It induces the imaginary bargaining game with possible utility-allocations $\{(10,0), (9,1), (8,2), ..., (0,10), (0,0)\}$, where the imaginary disagreement point is $\lim_{(x,y)\to 0} \left(v_1^{-1}(x), v_2^{-1}(y)\right) = (0,0)$. According to the NBS, $(5,5)$ would be the outcome of the bargaining game. Its fairness is $5 \cdot 5 = 25$. To trade off shame with selfishness, DM chooses the alternative that maximizes the sum of private utility and fairness, $a_1 + a_1 a_2$, which is $(6,4)$. Its fairness is $6 \cdot 4 = 24$ and the shame incurred by choosing it is $\frac{1}{2}$. Hence $U(A) = 5.5$. From the singleton set $B = \{(9,0)\}$, which corresponds to the exit option in the experiment, the choice is trivial and $U(B) = 9$. This example illustrates both the trade-off DM faces when choosing from a non-degenerate menu and the reason why she might prefer a smaller menu.

The organization of the paper is as follows: Section 2 presents the basic model and a representation that captures the concepts of fairness and shame. Section 3 isolates a choice criterion from the choice situation. Section 4 further specifies the fairness ranking. Section 5 suggests an application to a simple strategic situation, a game of trust. Section 6 concludes by pointing out connections to existing literature. An extension of our results to incorporate multiple recipients and all proofs are relegated to the appendix.

2. The Model

Let K be the set of all finite subsets of $\mathbb{R}^2_+$.[8] Any element $A \in K$ is a finite set of alternatives. A typical alternative $\mathbf{a} = (a_1, a_2)$ is interpreted as a payoff pair, where a_1 is the private payoff for DM, and a_2 is the private payoff allocated to the (potentially anonymous) other player, the recipient.[9] Endow K with the topology generated by the Hausdorff metric, which is defined for any pair of non-empty sets, $A, B \in K$, by:

$$d_h(A, B) := \max\left[\max_{\mathbf{a}\in A}\min_{\mathbf{b}\in B} d(\mathbf{a}, \mathbf{b}), \max_{\mathbf{b}\in B}\min_{\mathbf{a}\in A} d(\mathbf{a}, \mathbf{b})\right],$$

where $d : \mathbb{R}^2_+ \to \mathbb{R}_+$ is the standard Euclidian distance.

Let $\succ$ be a continuous, strict preference relation over K. The associated weak preference, $\succeq$ and the indifference relation, $\sim$ are defined in the usual way.

The choice of a menu $A \in K$ is not observed by the recipient, while the choice from any menu is. We call the impact this observation has on choice "shame." The first axiom specifies DM's preferences over singleton sets.

P_1 **(Selfishness)** $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ *if and only if* $a_1 > b_1$.

A singleton set $\{\mathbf{a}\}$ is a degenerate menu that contains only one feasible allocation, (a_1, a_2). It leaves DM with a trivial choice to be made when being observed in the second stage. Therefore, the ranking over singleton sets can be thought of as the ranking over allocations that are imposed on DM. We contend that there is no room for shame in this situation; choosing between two singleton sets reveals DM's "true" preferences over allocation outcomes. The axiom states that DM is not concerned about the payoff to the second player when evaluating such sets; she compares any pair of alternatives based solely on the first component, her private payoff. If, for example, DM had an altruistic concern for fairness

[8] With $\mathbb{R}_+$ we denote the positive reals including 0. $\mathbb{R}_{++}$ denotes the positive reals without 0.

[9] The extension to the case where K is the set of all finite subsets of $\mathbb{R}^n_+$, that is, the case where DM is concerned about the welfare of other n-1 recipients, is given in the appendix.

in the dictator game previously described, she would strictly prefer the menu $\{(9,1)\}$ to $\{(9,0)\}$. P_1 rules out such altruistic concerns. Negative emotions regarding the other player, such as spite or envy, are ruled out as well.

The next axiom captures the idea that shame is a mental cost, which is invoked by unchosen alternatives.

P_2 **(Strong Left Betweenness)** *If $A \succeq B$, then $A \succeq A \cup B$. Further, if $A \succ B$ and $\exists C$ such that $A \cup C \succ A \cup B \cup C$, then $A \succ A \cup B$.*

We assume that adding unchosen alternatives to a set can only increase shame. Therefore, no alternative is more appealing when chosen from $A \cup B$, than when chosen from one of the smaller sets, A or B. Hence, $A \succeq B$ implies $A \succeq A \cup B$.[10] Furthermore, if additional alternatives add to the shame incurred by the original choice from a menu $A \cup C$, then they must also add to the shame incurred by any choice from the smaller menu A. Thus, if there is C such that $A \cup C \succ A \cup B \cup C$ and if $A \succ B$, then $A \succ A \cup B$.

Shame, which is the only motive DM knows beyond selfishness, must refer to some personal norm that determines what the appropriate choice should have been. In our interpretation, this norm is to choose one of the fairest available allocations. Accordingly, we define an induced binary relation "fairer than".

Definition: For $\mathbf{a}, \mathbf{b} > 0$, we say that DM deems $\mathbf{b}$ to be *fairer than* $\mathbf{a}$, written $\mathbf{b} \succ_f \mathbf{a}$, if $\exists A \in K$ with $\mathbf{a} \in A$, such that $A \succ A \cup \{\mathbf{b}\}$.[11]

$A \succ A \cup \{\mathbf{b}\}$ implies that $\mathbf{b}$ adds to the shame incurred by the original choice in A. Our interpretation implies that $\mathbf{b}$ must be fairer than any alternative in A, and in particular $\mathbf{b} \succ_f \mathbf{a}$.

Some of the axioms below are imposed on $\succ_f$ rather than on $\succ$ and are labeled by F instead of P. Since $\succ_f$ is an induced binary relation, F axioms are implicit axioms on $\succ$; $\succ_f$ is only an expositional device.[12] However, the underlying notion of fairness is at the heart of F-axioms,[13] hence making assumptions directly on $\succ_f$, and motivating them in that

[10] This is the "Left Betweenness" axiom. It appears in Dekel, Lipman and Rustichini (2008) and is a weakening of "Set Betweenness" as first posed in GP.

[11] The notion of "fairer than" is analogous to the definition of "more tempting than" in Gul and Pesendorfer (2005).

[12] Our definition of $\succ_f$ implies that F-axioms are imposed only for strictly positive payoffs. Anticipating the implied choice behavior, this is done to avoid requirements on $\succ_f$ that have no testable implications for $\succ$. See footnote 16 for further details.

[13] In everyday language, "fair" is used to capture various notions. As pointed out in the introduction,

context, is natural. The implications of F-axioms on $\succ$ are most easily understood from the representation.

F_1 (**Fairness Ranking**) *$\succ_f$ is an anti-symmetric and negatively transitive binary relation.*

Our discussion rests on the assumption that fairness is a property of an allocation, independent of the menu on which it appears, and that DM can rank alternatives according to their fairness. In $\mathbb{R}^2_+$ and with increasing utility from self-payoffs, this assumption is not unreasonably restrictive. F_1 implies that only one alternative in each menu, the fairest, is responsible for shame.

F_2 (**Pareto**) *If $\mathbf{a} \geq \mathbf{b}$ and $\mathbf{a} \neq \mathbf{b}$, then $\mathbf{a} \succ_f \mathbf{b}$.*[14]

According to this axiom, absolute, as opposed to relative, well-being matters; the Pareto criterion excludes notions such as "strict inequality aversion." The resulting concept of fairness must have some concern for efficiency. In cases where there is truly no potential for redistribution, we believe that people find the Pareto criterion a reasonable requirement for one allocation to be fairer than another.[15]

F_3 (**Compensation**) *If $(a_1, a_2) \not\succ_f (b_1, b_2)$, then there are x and y such that both $(a_1, x) \succ_f (b_1, b_2)$ and $(y, a_2) \succ_f (b_1, b_2)$.*

The axiom states that any variation in the level of one person's payoff can always be compensated by appropriate variation in the level of the other person's payoff. The qualifier takes into account that payoffs are bounded below by 0. F_3 requires $\succ_f$ never to be satiated in either payoff. This assumption captures the idea that any fairness ranking with a concern for efficiency must go beyond the Pareto principle and trade off, in some manner, payoffs

we base our arguments on the definition according to the Merriam-Webster Collegiate Dictionary (Tenth Edition, 2001); "*Fair implies an elimination of one's own feelings, prejudices, and desires so as to achieve a proper balance of conflicting interests.*"

[14] F_2 explicitly rules out purely selfish individuals who evaluate sets according to $A \succ B \Leftrightarrow \max_{\mathbf{a} \in A} a_1 > \max_{\mathbf{b} \in B} b_1$. Even without F_2, for those individuals the norm of fairness would never be reflected in choice.

[15] In many contexts, people would disagree with the statement that the allocation $(10, 6)$ is fairer than $(5, 5)$. On the basis of the definition in footnote 13, however, we claim that the opposition to $(10, 6)$ as a fair allocation is not based on the greater appeal of $(5, 5)$, but rather on the implicit premise that there must be some mechanism to divide the gains more evenly (such a mechanism would imply the availability of a third option, say $(8, 8)$, which would render both of the above allocations unfair.) In an explicit binary choice situation, this premise cannot be sustained.

across individuals.

As $\succ$ is continuous on $\mathbb{R}^2_+$, $\succ_f$ is continuous on its domain, $\mathbb{R}^2_{++}$. Assuming that $\succ_f$ is continuous even in alternatives for which it does not relate to $\succ$ has obviously no implication for choice. For ease of exposition, we consider, in all that follows, the unique continuous extension of $\succ_f$ to all of $\mathbb{R}^2_+$.[16]

Definition: A function $\varphi : \mathbb{R}^2_+ \to \mathbb{R}$ is called a *fairness function* if it is strictly increasing and satisfies $\sup_{x \in \mathbb{R}_+} \varphi(x, a) > \varphi(\mathbf{b})$ and $\sup_{x \in \mathbb{R}_+} \varphi(a, x) > \varphi(\mathbf{b})$ for all $a \in \mathbb{R}_+$ and $\mathbf{b} \in \mathbb{R}^2_+$.

It is clear that $\succ_f$ satisfies $F_1 - F_3$ if it can be represented by a fairness function.

Theorem 1 *$\succ$ and $\succ_f$ satisfy $P_1 - P_2$ and $F_1 - F_3$ respectively, if and only if there exists a continuous, strictly increasing function $u : \mathbb{R}_+ \to \mathbb{R}$, a continuous fairness function φ, and a continuous function $g : \mathbb{R}^2_+ \times \varphi\left(\mathbb{R}^2_+\right) \to \mathbb{R}$, weakly increasing in its second argument and satisfying $g(\mathbf{a}, x) \gtreqless 0$ whenever $\varphi(\mathbf{a}) \lesseqgtr x$, such that the function $U : K \to \mathbb{R}$ defined as*

$$U(A) = \max_{\mathbf{a} \in A} \left[u(a_1) - g\left(\mathbf{a}, \max_{\mathbf{b} \in A} \varphi(\mathbf{b})\right)\right]$$

represents $\succ$ and φ represents $\succ_f$.

All detailed proofs are in the appendix. We now highlight the important steps. Since both $\succ$ and $\succ_f$ are continuous preference relations, they can be represented by continuous functions, $U : K \to \mathbb{R}$ and $\varphi : \mathbb{R}^2_+ \to \mathbb{R}$ respectively. F_2 and F_3 imply that φ is a fairness function. The combination of *Strong Left Betweenness* (P_2) and *Fairness Ranking* (F_1) implies GP's *Set Betweenness* (SB) property: $A \succeq B$ implies $A \succeq A \cup B \succeq B$. GP demonstrate that imposing SB on preferences over sets makes every set indifferent to a certain subset of it, which includes at most two elements (Lemma 2 in their paper). Hence we confine our attention to a subset of the domain that includes all sets with cardinality no greater than 2. Selfishness (P_1) and the definition of $\succ_f$ imply that a set $\{\mathbf{a}, \mathbf{b}\}$ is strictly inferior to $\{\mathbf{a}\}$ if and only if $a_1 > b_1$ and $\mathbf{b} \succ_f \mathbf{a}$. Based on this observation we employ $F_1 - F_3$ to show that any set is indifferent to some two-element set that includes one of the fairest allocations in the original set. Furthermore, the impact of this alternative on the menu's value depends only on its fairness, φ. Finally, we establish the continuity of the

[16] To see why we had to restrict $\succ_f$ to $\mathbb{R}_+ \times \mathbb{R}_{++}$, note that the fairness of an alternative $\mathbf{b}$ is relevant for choice, if and only if there is another alternative $\mathbf{a}$ with $\mathbf{a} \prec_f \mathbf{b}$ and $a_1 > b_1$, which requires $a_2 < b_2$. Thus $b_2 > 0$ is necessary for the construction of $\mathbf{a}$. Continuity implies that the unique continuous extension of $\succ_f$ from $\mathbb{R}^2_{++}$ to $\mathbb{R}_+ \times \mathbb{R}_{++}$ coincides with the ranking we would have found on that domain.

second component, the function g, in the representation.

Theorem 1 provides a representation of DM's fairness ranking entirely based on revealed preferences. This should help the empirical quest to understand people's norm of fairness. It also highlights the basic trade-off between private payoff and shame as the only concepts DM may care about. There are at most two essential alternatives within a set, to be interpreted as the "chosen" and the "fairest" alternative, $\mathbf{a}$ and $\mathbf{b}$ respectively. For the latter, its fairness, $\varphi(\mathbf{b})$, is a sufficient statistic for its impact on the set's value. DM suffers from shame, measured by $g(\mathbf{a}, \varphi(\mathbf{b}))$, whenever $\varphi(\mathbf{a}) < \varphi(\mathbf{b})$, where $\varphi(\mathbf{a})$ is the fairness of the chosen alternative. The representation captures the idea of shame being an emotional cost that emerges whenever the fairest available allocation is not chosen. Note that the properties of the function g and the max operator inside imply that the second term is always a cost (non-positive). The other max operator implies that DM's payoff will never lie below b_1, which is her payoff as suggested by the fairest allocation. Thus, any deviations by DM from choosing the fairest allocation will be in her own favor. These observations justify labeling said cost as "shame." Its magnitude may depend on the chosen allocation.

From the representation in Theorem 1, it is easy to see that actual choice may be context-dependent, in the sense that a higher degree of shame may affect choice. For example, consider DM that chooses $(8,2)$ from $\{(10,0),(8,2),(5,5)\}$, and chooses $(10,0)$ from $\{(10,0),(8,2)\}$; While he finds his preferred allocation to be $(10,0)$ when the fairest available alternative is $(8,2)$, choosing it becomes too costly in the presence of $(5,5)$, making $(8,2)$ the best compromise. This type of violation of the weak axiom of revealed preferences is plausible when shame is taken into account.[17] In the next section we spell out the implications of enforcing a context-independent criterion for choice.

3. A Second-Stage Choice Ranking

In many situations, only the second stage choice may be recorded. For example, the standard dictator game corresponds only to the second stage choice in our setup. Typical behavior in various versions of this game, where subjects tend to give part of the endowment to the recipient, is often interpreted as evidence of an altruistic motive. Based on the definition of altruism as "*unselfish regard for or devotion to the welfare of others*" (Merriam-Webster Collegiate Dictionary (Tenth Edition, 2001)), we interpret altruism to imply that the recipient's welfare is a good, just as selfishness implies that DM's private payoff is a good. If DM had those two motives, she would have to make a trade-off between them. As in

[17] Consider charities that call for donations by listing suggested amounts such as \$20, \$50 and \$100. They might raise donation levels simply by including a large sum on the menu, which no one is expected to choose, but the purpose of which is to moderate DM's selfish choice.

the case of two generic goods, very basic assumptions would lead to a context-independent choice ranking of alternatives. Relating to the discussion at the end of section 2, we can define a binary relation "better choice than," $\succ_c$, by $\mathbf{a} \succ_c \mathbf{b}$ if $\exists B$ with $\mathbf{b} \in B$, such that $B \cup \{\mathbf{a}\} \succ B$. This binary relation need not be acyclic: Different choice problems, A and B, may lead to different second-stage rankings of $\mathbf{a}$ and $\mathbf{b}$ for $\mathbf{a}, \mathbf{b} \in A \cap B$. If no such cycles occur, second-stage behavior might look as if it were generated by, for instance, a trade-off between selfishness and altruism, even though observation of the first stage choice would rule this out. If, on the other hand, cycles are observed in the second stage choice, simple altruistic motives cannot be solely responsible for behavior that is not purely selfish. In this section, we identify a condition on preferences that makes DM's second-stage choice independent of the choice set. This implies finding a function $\psi : \mathbb{R}^2_+ \to \mathbb{R}$ that assigns a value to each $\mathbf{a} \in A$, such that $\mathbf{a}$ is a choice from A only if $\psi(\mathbf{a}) \geq \psi(\mathbf{b})$ for all $\mathbf{b} \in A$.

For any set of two allocations $\{\mathbf{a}, \mathbf{b}\}$, we interpret the preference ordering $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$ as an indication of a discrepancy between what DM chooses ($\mathbf{a}$) and the alternative she deems to be the fairest ($\mathbf{b}$), which causes her choice to bear shame. This shame, however, is not enough to make her choose $\mathbf{b}$.

Notation: We write $\langle \mathbf{a}, \mathbf{b} \rangle$ to denote a menu $\{\mathbf{a}, \mathbf{b}\}$ with $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$.[18]

Theorem 1 establishes that choice between sets depends on the fairness of the fairest alternative in each set. The next axiom also relates choice to the fairness of the chosen alternative: The fairer DM's choice, the less shame she feels.

P_3 **(Fairer is Better)** *If* $\{\mathbf{a}\} \sim \{\mathbf{a}'\}$ *and* $\mathbf{a} \succ_f \mathbf{a}'$, *then* $\langle \mathbf{a}, \mathbf{b} \rangle \succ \langle \mathbf{a}', \mathbf{b} \rangle$.

Axiom P_3 implies that only the fairness of the chosen alternative matters for its impact on shame.

Given $P_1 - P_3$ and $F_1 - F_3$, an additional separability assumption is equivalent to separable shame, and thus to a set-independent choice ranking.

P_4 **(Consistency)** *If* $\langle \mathbf{a}, \mathbf{b} \rangle \sim \langle \mathbf{a}', \mathbf{b}' \rangle$ *and* $\langle \mathbf{a}, \mathbf{d} \rangle \sim \langle \mathbf{a}', \mathbf{d}' \rangle$ *then*

$$\langle \mathbf{c}, \mathbf{b} \rangle \succ \langle \mathbf{c}', \mathbf{b}' \rangle \Leftrightarrow \langle \mathbf{c}, \mathbf{d} \rangle \succ \langle \mathbf{c}', \mathbf{d}' \rangle .$$

The axiom requires independence between the impact of the chosen and the fairest al-

[18] In the axioms, we write $\langle \mathbf{a}, \mathbf{b} \rangle$ short for "any menu $\{\mathbf{a}, \mathbf{b}\}$ for which $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2_+$ satisfy $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$ as well as all other explicit qualifiers".

ternative on the set ranking. Assume, without loss of generality, that $\{\mathbf{a}\} \succ \{\mathbf{a}'\}$. Suppose there are two pairs of fairer and less attractive alternatives, $\mathbf{b}, \mathbf{b}'$ and $\mathbf{d}, \mathbf{d}'$, such that for each of them, pairing their members with $\mathbf{a}$ and $\mathbf{a}'$, respectively, gives rise to two indifferent menus, from which those fairer alternatives are not chosen. In the context of Theorem 1, this implies that both pairs induce the same shame differential, which exactly cancels the selfish preference of $\{\mathbf{a}\}$ over $\{\mathbf{a}'\}$: $\langle \mathbf{a}, \mathbf{b} \rangle \sim \langle \mathbf{a}', \mathbf{b}' \rangle$ and $\langle \mathbf{a}, \mathbf{d} \rangle \sim \langle \mathbf{a}', \mathbf{d}' \rangle$. The axiom then states that pairing the members of $\mathbf{b}, \mathbf{b}'$ or $\mathbf{d}, \mathbf{d}'$ with any other chosen and less fair alternatives $\mathbf{c}$ and $\mathbf{c}'$, respectively, must also lead to the same differential in shame. In particular, $\langle \mathbf{c}, \mathbf{b} \rangle \succ \langle \mathbf{c}', \mathbf{b}' \rangle$ implies $\langle \mathbf{c}, \mathbf{d} \rangle \succ \langle \mathbf{c}', \mathbf{d}' \rangle$. We make no claim about the normative or descriptive appeal of this assumption. Instead, we view it as an empirical criterion: Theorem 2 below suggests that given the other axioms, observation of the second stage choice does not suffice to distinguish altruism from shame as the motive behind DM's other-regarding behavior, if and only if this condition is met.

Theorem 2 *$\succ$ and $\succ_f$ satisfy $P_1 - P_4$ and $F_1 - F_3$ respectively, if and only if there exist a continuous and strictly increasing function $u : \mathbb{R}_+ \to \mathbb{R}$ and a continuous fairness function φ, such that the function $U : K \to \mathbb{R}$ defined as*

$$U(A) = \max_{\mathbf{a} \in A} [u(a_1) + \varphi(\mathbf{a})] - \max_{\mathbf{b} \in A} [\varphi(\mathbf{b})]$$

represents $\succ$ and φ represents $\succ_f$.

The representation isolates a choice criterion that is independent of the choice problem: DM's behavior is governed by maximizing

$$\psi(\mathbf{a}) = u(a_1) + \varphi(\mathbf{a}).$$

The value of the set is reduced by

$$\max_{\mathbf{b} \in A} \varphi(\mathbf{b}),$$

a term that depends solely on the fairest alternative in the set. Grouping the terms differently reveals the trade-off between self-payoff, $u(a_1)$, and the shame involved with choosing $\mathbf{a}$ from the set A:

$$\max_{\mathbf{b} \in A} [\varphi(\mathbf{b}) - \varphi(\mathbf{a})] \geq 0.$$

Note that now shame takes an additively separable form, depends only on the fairness of both alternatives, and is increasing in the fairness of the fairest and decreasing in that of the chosen alternative.

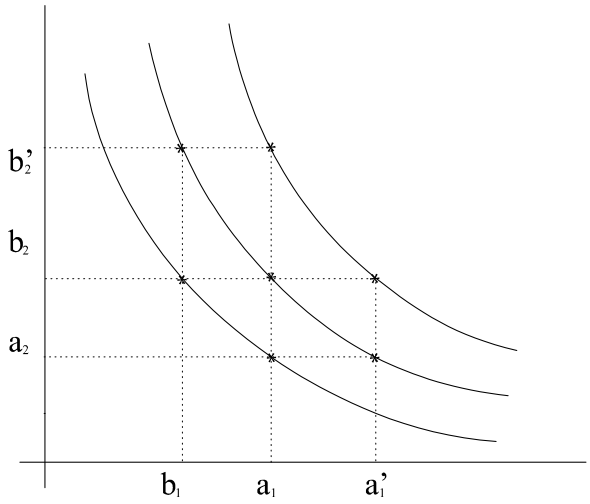

Figure 1: Independent Fairness Contributions.

4. Specifying a Fairness Ranking

In this section, we impose one more axiom on $\succ_f$ to further characterize the fairness ranking. It asserts that the fairness contribution of one person's marginal payoff cannot depend on the initial payoff levels.

F_4 (**Independent Fairness Contributions**) *If* $(a_1, a_2) \sim_f (b_1, b_2)$ *and* $(a_1', a_2) \sim_f (a_1, b_2) \sim_f (b_1, b_2')$, *then* $(a_1', b_2) \sim_f (a_1, b_2')$.

The axiom is illustrated in figure 1. If $a_1 = a_1'$ or $b_2 = b_2'$, this axiom is implied by F_1, F_2 and the continuity of $\succ_f$. For $a_1 \neq a_1'$ and $b_2 \neq b_2'$, the statement is more subtle. Consider first a stronger assumption:

F_4' (**Strong Independent Fairness Contributions**) $(a_1, a_2) \sim_f (b_1, b_2)$ *and* $(a_1', a_2) \sim_f (b_1, b_2')$ *imply* $(a_1', b_2) \sim_f (a_1, b_2')$.

The fairness contribution of one person's marginal payoff cannot depend on the initial payoff level of the other person: It is unclear to DM how much an increase in monetary payoff means to the recipient, because even if the (marginal) utility of the recipient were known to DM, she could not compare it to her own, as interpersonal utility comparisons are infeasible. The qualifier in F_4' establishes that DM considers the fairness contribution of changing her own payoff from a_1 to a_1' given the allocation (a_1, a_2) to be the same as that of changing the recipient's payoff from b_2 to b_2' given (b_1, b_2). F_4' then states that starting from

the allocation (a_1, b_2), changing a_1 to a_1' should again be as favorable in terms of fairness as changing b_2 to b_2'. This is the essence of *Independent Fairness Contributions*. The stronger qualifier $(b_1, b_2') \sim_f (a_1, b_2) \sim_f (a_1', a_2)$ in F_4 weakens the axiom. For example, the fairness ranking $(a_1, a_2) \succ_f (b_1, b_2)$ if and only if $\min(a_1, a_2) > \min(b_1, b_2)$ is permissible under F_4, but not under F_4'.

Kranz *et al* (1971) provide an additive representation based on F_4', which they refer to as the *Thomsen condition*. Karni and Safra (1998) demonstrate that the weaker condition F_4, which they term the *Hexagon condition,* implies F_4' in the context of their axioms. The next Theorem is based on those results:

Theorem 3 *$\succ_f$ satisfies $F_1 - F_4$, if and only if there are continuous and unbounded functions $v_1, v_2 : \mathbb{R}_+ \to \mathbb{R}_{++}$, strictly increasing on $\mathbb{R}_{++}$, such that $\varphi(\mathbf{a}) = v_1(a_1)\, v_2(a_2)$ represents $\succ_f$.*

If $\varphi'(\mathbf{a}) = v_1'(a_1)\, v_2'(a_2)$ also represents $\succ_f$, then there are α, β_1, β_2, all strictly positive, such that $v_1' = \beta_1 v_1^{\alpha}$ and $v_2' = \beta_2 v_2^{\alpha}$.

We establish that our axioms imply the axioms of Karni and Safra and, according to their Lemma, those imply the axioms of Kranz et al. Hence, an additively separable representation exists, where the utilities are unique up to translation and a common linear transformation.[19] A direct proof of this result, based on our axioms, repeatedly uses axioms F_3 and F_4 to establish that if $(a_1, a_2) \sim_f (a_1', a_2')$ and $(a_1, \widetilde{a}_2) \sim_f (a_1', \widetilde{a}_2')$, then $(\widetilde{a}_1, a_2) \sim_f (\widetilde{a}_1', a_2') \Leftrightarrow (\widetilde{a}_1, \widetilde{a}_2) \sim_f (\widetilde{a}_1', \widetilde{a}_2')$. With this knowledge, we can create a monotone and increasing mapping $a_2 \to \gamma(a_2)$ that transforms the original indifference map to be quasi-linear with respect to the first coordinate in the $(a_1, \gamma(a_2))$ plane. Keeney and Raiffa (1976) refer to the construction of this transformation as the lock-step procedure.[20] Quasi-linearity implies that there is an increasing continuous function $\xi : \mathbb{R}_+ \to \mathbb{R}$, such that $\varphi(\mathbf{a}) := \xi(a_1) + \gamma(a_2)$ represents $\succ_f$. Given the additively separable representation, define $v_1(a_1) := \exp(\xi(a_1))$ and $v_2(a_2) := \exp(\gamma(a_2))$. Then $v_1, v_2 : \mathbb{R}_+ \to \mathbb{R}_{++}$ are increasing and continuous and if we redefine $\varphi(\mathbf{a}) := v_1(a_1)\, v_2(a_2)$, it represents $\succ_f$ and is unique in the sense of the theorem. That the representation satisfies our axioms is easy to verify.

This representation suggests an appealing interpretation of the fairness ranking DM is concerned about: She behaves *as if* she had in mind two increasing and unbounded utility functions, one for herself[21] and one for the recipient. By mapping the alternatives within each

[19] See Theorem 2 in Chapter 6 of Kranz et al (1971).

[20] For brevity, we do not reproduce their argument in more detail in this paper. A direct proof of Theorem 3 is available upon request.

[21] This utility function, v_1, need not agree with her true utility for personal payoffs, u. The interpretation

set into the associated utility space, any choice set induces a finite bargaining game, where the imaginary disagreement point corresponds to zero utility payoffs. DM then identifies the fairest alternative within a set as the Nash Bargaining Solution of the game.[22] Moreover, the fairness of all alternatives can be ranked according to the same functional, namely the Nash product.

The tension of having to trade off marginal payoffs (F_1) without being able to compare their welfare contribution (F_4) is common in a range of social-choice problems (for a review, see Hammond (1990). Our axioms are weak in the sense that they do not constrain DM in this trade-off, as long as she takes into account that the fairness contribution of increasing one person's payoff should not depend on the other's payoff. The power of Theorem 3 is that it bases a representation on these weak assumptions.

To underline the appeal of the Nash product as a descriptive representation of fairness, we now point out how DM might reason within the constraints of the axioms. We justified the Pareto criterion, F_2, as a plausible axiom for the fairness ranking. As argued above, concern for fairness requires the acknowledgment of some form of interpersonal comparability of the intensity of preferences. If utilities were known cardinally, symmetry in terms of utility payoffs is the other criterion we would expect the ranking to satisfy.[23] In our context, this implies independence of the role a person plays, dictator or recipient. However, utilities are inherently ordinal, rendering such a comparison infeasible. At best, if we assume people to have cardinal utilities that reflect their attitudes toward risk, we can determine marginal utilities up to scaling. Mariotti (1999), for example, considers a context in which *"interpersonal comparisons of utility are meaningful; that is, there exists an (unknown) rescaling of each person's utility which makes utilities interpersonally comparable."* But at the same time, *"interpersonal comparisons of utility are not feasible."* Assume there is a correct interpersonal utility scaling, but DM cannot determine it. Can she guarantee that for this unknown scaling both symmetry and Pareto are satisfied? They would have to be satisfied for all potential scalings. Mariotti establishes that the NBS is the only criterion with this property.

---

is that DM is concerned about the recipient's perception of her choice. The recipient, however, may not know DM's true utility, especially under anonymity. However, our interpretation might be more convincing when they resemble each other empirically. In particular, it is more appealing if DM's actual utility from self-payoff u is unbounded.

[22] See Nash (1950). The imaginary disagreement point is determined by $\lim_{(x,y)\to 0}\left(v_1^{-1}(x), v_2^{-1}(y)\right)$. It could be some finite and weakly positive pair of monetary payoffs. In particular it could be $(0,0)$, which corresponds to DM imagining that players walk away in the case that no agreement is reached. It could also be negative. This corresponds to DM imagining that players have an extra incentive to find an agreement: there is a cost to disagreement.

[23] This reasoning leads Rawls (1971) to suggest Pareto and Symmetry as the two criteria a decision maker (under a veil of ignorance) should respect.

Even more appealing is an interpretation of the NBS as the fairest allocation that is related to Gauthier's (1986) principle of "moral by agreement": Trying to assess what is fair, but finding herself unable to compare utilities across individuals, DM might refer to the prediction of a symmetric mechanism for generating allocations. In particular, DM might ask what would be the allocation if both she and the recipient were to bargain over the division of the surplus. To answer this question, she does not need to assume the intensities of the two preferences. This is a procedural interpretation that is not built on the axioms; DM is not ashamed of payoffs, but of using her stronger position in distributing the gains. It is, then, the intuitive and possibly descriptive appeal of the NBS in many bargaining situations that makes it normatively appealing to DM in our context.[24] Theorem 3 establishes the behavioral equivalence of this interpretation and our axioms.

Remark: Any concern DM has about fairness originates from being observed. Consequently, DM should expect a potentially anonymous observer to share her notion of what is fair. Her private norm of fairness, which we observe indirectly, should reflect her concern about not violating a social norm. If the observed choice situation is anonymous, DM does not know the recipient's identity and is aware that the recipient does not know hers. Therefore, the ranking cannot depend on either identity. Combining this with the idea that fairness of an allocation should not depend on the role a person plays, whether dictator or recipient, one might want to impose symmetry of the fairness ranking in terms of direct payoffs.

F_5 **(Symmetry)** $(a_1, a_2) \sim_f (a_2, a_1)$.

Adding this assumption constrains $v_1(a) = v_2(a)$ in the representation of Theorem 3. The numerical example given in the introduction features the combination of Theorem 2 and Theorem 3, where all functions involved are the identity. For brevity, we will not repeat it here.

5. Application, A Game of Trust

In this section, we demonstrate that shame, as a motive for other-regarding behavior, also has implications for strategic environments. As an example, consider the game of trust, which is depicted in Figure 2 and is a variant of a game suggested by Tadelis (2008): In the first stage, player 1 can either trust (T) or not trust (N). Action N ends the game and leads to payoff n for both player 1 and 2. Write this outcome as (n, n). If trusted, player 2 can

[24] The descriptive value of the NBS has been tested empirically. For a discussion see Davis and Holt (1993) pages 247-55. Further, multiple seemingly natural implementations of NBS have been proposed (Nash (1953), Osborne and Rubinstein (1994)).

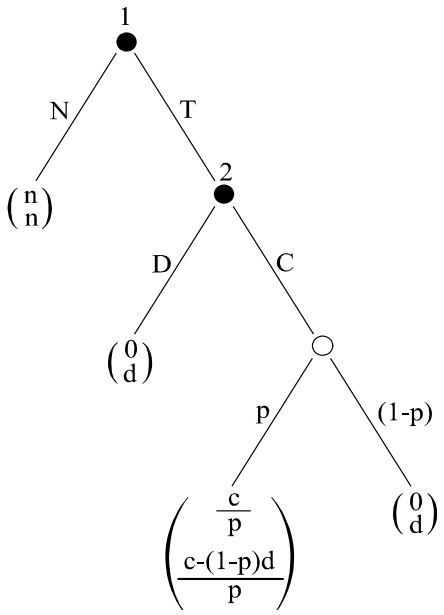

Figure 2: A game of trust with uncertain outcomes.

either cooperate (C) or defect (D). Action D generates the outcome $(0, d)$ with certainty. Action C leads to the cooperative outcome $\left(\frac{c}{p}, \frac{c-(1-p)d}{p}\right)$ only with probability p, and to the uncooperative outcome $(0, d)$ otherwise, where $d > c > n > 0$.

Player 1 is not susceptible to shame, but player 2 is. Since this setting involves uncertainty, we need to specify how the players evaluate risky prospects. For simplicity, assume that both players are risk neutral and satisfy the expected utility axioms, so that lotteries are evaluated by their expected value. In this case, player 1's options can be written as the two menus $N = \{(n, n)\}$ and $T = \{(c, c), (0, d)\}$.

Unlike the setting considered so far, in which DM chooses an allocation in two stages, the game of trust is a strategic situation: Player 1 chooses a menu from which player 2 will choose in the second stage. If instead player 2 chooses over menus in an unobserved first stage, we assume that she evaluates menus according to a variant of our representation,

$$U(A) = \max_{\mathbf{a} \in A} [a_2 + \beta \widetilde{a}_1 \widetilde{a}_2] - \beta \max_{\mathbf{b} \in A} [b_1 b_2]$$

where $\widetilde{\mathbf{a}}$ is player 1's expectation of the allocation $\mathbf{a}$ generated by player 2's choice. As we point out in the discussion of Theorem 2, this suggests that player 2's choice from menu A is governed by maximizing the term $a_2 + \beta \widetilde{a}_1 \widetilde{a}_2$. Player 1, on the other hand, evaluates allocation $\mathbf{a}$ according to her payoff, a_1.

We consider two cases: In the observed case, player 1 observes player 2's action, whereas

in the unobserved case, player 1 only observes the outcome of the game. In what follows, we restrict the strategy space of each player to pure strategies. The next proposition characterizes the (pure strategies) equilibria of this game.[25]

Proposition:
i) In the observed case, the unique equilibrium is T, C, *if* $\frac{d-c}{\beta c^2} < 1$. *If* $\frac{d-c}{\beta c^2} > 1$ *it is* N, D, *and if* $\frac{d-c}{\beta c^2} = 1$ *both equilibria exist.*
ii) In the unobserved case, there are no equilibria if $\frac{d-c}{\beta c^2} < p$. *If* $\frac{d-c}{\beta c^2} \geq p$ *the unique equilibrium is* N, D.

The interesting case is where $\frac{d-c}{\beta c^2} \in [p, 1)$, in which case player 1's equilibrium behavior is to trust (T), if and only if he can observe player 2's action. This behavior can be explained by player 1's correct anticipation of shame as a motivating force, leading player 2 to cooperate only under observation. It cannot be explained by emotions like altruism or guilt, which do not depend on observation.

6. Related Literature

Other-regarding preferences have been considered extensively in economic literature. In particular, inequality aversion, as studied by Fehr and Schmidt (1999), is based on an objective function with a similar structure to the representation of second-stage choice in Theorem 2.[26] Both works attach a cost to any deviation from choosing the fairest alternative. In Fehr and Schmidt's work, the fairest allocation need not be feasible and is independent of the choice situation. In our work, the fairest allocation is always a feasible choice and it is identified through the axioms. This dependence of the fairest allocation on the choice situation allows us to distinguish observed from unobserved choice.

The idea that there may be a discrepancy between DM's preference to behave "pro-socially" and her desire to be viewed as behaving pro-socially is not new to economic literature. For a model thereof, see Benabou and Tirole (2006).

Neilson's (2008) work is motivated by the same experimental evidence as ours. He also considers menus of allocations as objects of choice. Neilson does not axiomatize a representation result, but points out how choices among menus should relate to choices from menus,

[25] It follows immediately from the arguments given in the proof of the proposition (see appendix) that there can be no mixed-strategy equilibria of this game. Restricting the strategy spaces to pure strategies only serves the purpose of determining out-of-equibrium beliefs, as is needed for part (ii) of the proposition.

[26] Neilson (2006) axiomatizes a reference-dependent preference, that can be interpreted in terms of Fehr and Schmidt's objective function.

if shame were the relevant motive. He relates the two aspects of shame that also underlie the *Set Betweenness* property in our work; DM might prefer a smaller menu over a larger menu either because avoiding shame compels her to be generous when choosing from the larger menu, or because being selfish when choosing from the larger menu bears the cost of shame.

The structure of our representation resembles the representation of preferences with self-control under temptation, as axiomatized in GP. GP study preferences over sets of lotteries and show that their axioms lead to a representation of the following form:

$$U^{GP}(A) = \max_{a \in A} \left\{ u^{GP}(a) + v^{GP}(a) \right\} - \max_{b \in A} \left\{ v^{GP}(b) \right\}$$

with u^{GP} and v^{GP} both linear in the probabilities and where A is now a set of lotteries. In their context, u^{GP} represents the "commitment"- and v^{GP} the "temptation"-ranking. While the two works yield representations with a similar structure, their domains - and therefore the axioms - are different. GP impose the independence axiom and indifference to the timing of the resolution of uncertainty. This allows them to identify the representation above that consists of two functions that are linear in the probabilities. Each of these functions is an expected utility functional. The objects in our work, in contrast, are sets of monetary allocations and there is no uncertainty. Even if we did consider risky prospects, we argue in the introduction that imposing the independence axiom would not be plausible. However, one of GP's axioms is the Set Betweenness axiom, $A \succcurlyeq B \Rightarrow A \succcurlyeq A \cup B \succcurlyeq B$. We show that our axioms, Strong Left Betweenness (P_2) and Fairness Ranking (F_1) imply Set Betweenness. Hence, GP's Lemma 2 can be employed, allowing us to confine attention to sets with only two elements.

Empirically, the assumption that only two elements of a choice set matter for the magnitude of shame (the fairest available alternative and the chosen alternative) is clearly simplifying: Oberholzer-Gee and Eichenberger (2008) observe that dictators choose to make much smaller transfers when their choice set includes an unattractive lottery. In other words, the availability of an unattractive allocation seems to lessen the incentive to share.

Lastly, it is necessary to qualify our leading example: The experimental evidence on the effect of (anonymous) observation on the level of giving in dictator games is by no means conclusive. Behavior tends to depend crucially on surroundings, like the social proximity of the group of subjects and the phrasing of the instructions, as, for example, Bolton, Katok and Zwick (1994); Burnham (2003); and Haley and Fessler (2005) record. While supported by the body of evidence mentioned in the introduction, our interpretation is in contrast to evidence collected by Koch and Normann (2005), who claim that altruistic behavior persists

at an almost unchanged level when observability is credibly reduced. Similarly, Johannesson and Persson (2000) find that incomplete anonymity - not observability - is what keeps people from being selfish. Ultimately, experiments aimed at eliciting a norm share the same problem: Since people use different (and potentially contradictory) norms in different contexts, it is unclear whether the laboratory environment triggers a different set of norms than would other situations: Frohlich, Oppenheimer and Moore (2000) point out that money might become a measure of success rather than a direct asset in the competition-like laboratory environment, such that the norm might be "do well" rather than "do not be selfish."[27] More theoretically, Miller (1999) suggests that the phrasing of instructions might determine which norm is invoked. For example, the reason that Koch and Normann do not find an effect of observability might be that their thorough explanation of anonymity induces a change in the regime of norms, in effect telling people "be rational," which might be interpreted as "be selfish." Then being observed might have no effect on people who, under different circumstances, might have been ashamed to be selfish.

7. Appendix

7.1. Extension to Multiple Recipients

The underlying idea is that DM (without loss of generality individual 1) is concerned about $N-1 \geqslant 2$ other individuals, whose payoffs depend on her choice. In analogy to section 2, let K be the set of all finite subsets of $\mathbb{R}^N_+$. Any element $A \in K$ is a finite set of alternatives. A typical alternative $\mathbf{a} = (a_1, a_2, ..., a_N)$ is interpreted as a payoff vector, where a_n is the payoff allocated to individual n. We write, for example, $(a_m, a_n, \mathbf{a}_{-m,n})$ as the alternative with payoff a_m to individual m, payoff a_n to individual n and $\mathbf{a}_{-m,n} \in \mathbb{R}^{N-2}_+$ lists all other individuals' payoffs in order. We endow K with the topology generated by the Hausdorff metric.

Let $\succ$ be a continuous preference relation over K. All axioms we impose on $\succ$ in section 2 can be readily applied to $\succ$ on this new domain. We define $\succ_f$ in analogy to the previous definition. Instead of F_3 we write

F_3^N **(Weak Solvability)** *If* $(a_n, \mathbf{0}) \not\succ_f \mathbf{b}$ *then for all* $m \neq n$, there exists a_m *such that* $(a_m, a_n, \mathbf{0}) \succ_f \mathbf{b}$.

[27] Surely the opposite is also conceivable: Subjects might be particularly keen to be selfless when the experimentor observes their behavior. This example is just ment to draw attention to the difficulties faced by experimenters in the context of norms.

The axiom states that it is always possible to increase the fairness of an allocation with payoff to only one individual beyond that of an initially fairer allocation by giving appropriate payoffs to any second individual. This property requires the fairness ranking never to be satiated in any individual payoff.

Definition: The pair of possible payoffs to individuals m and n is *Preferentially Independent with respect to its Complement* (P.I.C.), if the fairness ranking in the (a_m, a_n)-space is independent of $\mathbf{a}_{-m,n}$.

F_4^N **(Pairwise Preferential Independence)** *For all $m, n \in \{1, .., N\}$, the pair of possible payoffs to individuals m and n is P.I.C.*

Similarly to F_4**,** this axiom must hold if the contribution of one person's marginal private payoff to the fairness of an allocation cannot depend on another person's private payoff level.

Theorem 4 *Assume $N \geq 3$.*
(i) $\succ$ *and* $\succ_f$ *satisfy* $P_1 - P_4$ *and* F_1, F_2 and F_3^N *respectively, if and only if there exist continuous and strictly increasing functions* $u : \mathbb{R}_+ \to \mathbb{R}$ *and* $\varphi : \mathbb{R}_+^N \to \mathbb{R}$ *such that the function* $U : K \to \mathbb{R}$ *defined as* $U(A) = \max_{\mathbf{a} \in A} [u(a_1) + \varphi(a_1, a_2, ..., a_n)] - \max_{\mathbf{b} \in A} [\varphi(b_1, b_2, ..., b_n)]$ *represents* $\succ$ *and* φ *represents* $\succ_f$.
(ii) $\succ_f$ also satisfies F_4^N *if and only if there exist continuous and strictly increasing functions* $v_1, .., v_N : \mathbb{R}_+ \to \mathbb{R}_{++}$*, where* $v_1, .., v_N$ *are unbounded such that* $\varphi(\mathbf{a}) = \prod_{i=1}^{N} v_i(a_i)$.

The proof is in the next section

7.2. Proofs

Proof of Theorem 1
Let $U : K \to \mathbb{R}$ be a continuous function that represents $\succ$. Define $u(a_1) \equiv U(\{(a_1, 0)\})$. By P_1, $u(a_1) = U(\{(a_1, a_2)\})$ independent of a_2, with $u(a_1)$ continuous and strictly increasing.

Because $\succ_f$ is defined only on $\mathbb{R}_{++}^2$ we first construct a representation for $\succ$ on the set of all finite subsets of $\mathbb{R}_{++}^2$ (denoted by K_{++}). Continuity then allows us to extend it to K. Let $\varphi : \mathbb{R}_{++}^2 \to \mathbb{R}$ be a continuous function that represents $\succ_f$. By F_2, φ is also strictly increasing.

Because $\succ_f$ is continuous, F_3 immediately implies that if $(a_1, a_2) \not\succ_f (b_1, b_2)$, then there are x and y such that $(a_1, x) \sim_f (b_1, b_2) \sim_f (y, a_2)$. In all that follows, we use this stronger

version of F_3 without further discussion.

Claim 1.1 (Right Betweenness): $A \succeq B \Rightarrow A \cup B \succeq B$.

Proof: There are two cases to consider:

Case 1) $\forall \mathbf{a} \in A,\ \exists \mathbf{b} \in B$ such that $\mathbf{b} \succ_f \mathbf{a}$. Let $A = \left\{\mathbf{a}^1, \mathbf{a}^2 ..., \mathbf{a}^N\right\}$ and $C_0 = B$. Define $C_n = C_{n-1} \cup \{\mathbf{a}^n\}$ for $n = 1, 2, .., N$. According to F_1, for all $\mathbf{a}^n$ there exists $\mathbf{b} \in B$ such that $\mathbf{a}^n \not\succ_f \mathbf{b}$. By P_2, $C_{n-1} \not\succ C_n$. By negative transitivity of $\succ$, $C_0 \not\succ C_N$ or $A \cup B \succeq B$.

Case 2) $\exists \mathbf{a} \in A$ such that $\mathbf{a} \succ_f \mathbf{b},\ \forall \mathbf{b} \in B$. Let $B = \left\{\mathbf{b}^1, \mathbf{b}^2 ..., \mathbf{b}^M\right\}$. Define $C_0 = A$ and $C_m = C_{m-1} \cup \{\mathbf{b}^m\}$ for $m = 1, 2, .., M$. By P_2, $\forall C$ such that $\mathbf{a} \in C$, $C \not\succ C \cup \{\mathbf{b}^m\}$. Hence, $C_{m-1} \not\succ C_m$. By negative transitivity of $\succ$, $C_0 \not\succ C_M$ or $A \cup B \succeq A \succeq B$, hence $A \cup B \succeq B$.‖

Combining Claim 1.1 with P_2 guarantees Set Betweenness (SB): $A \succeq B \Rightarrow A \succeq A \cup B \succeq B$. Having established Set Betweenness, we can apply GP Lemma 2, which states that any set is indifferent to a specific two-element subset of it.

Lemma 1.1 (GP Lemma 2): *If $\succ$ satisfies SB, then for any finite set A, there exist $\mathbf{a}, \mathbf{b} \in A$ such that $A \sim \{\mathbf{a}, \mathbf{b}\}$, $(\mathbf{a}, \mathbf{b})$ solves $\underset{\mathbf{a}' \in A}{\max}\underset{\mathbf{b}' \in A}{\min} U\left(\{\mathbf{a}', \mathbf{b}'\}\right)$ and $(\mathbf{b}, \mathbf{a})$ solves $\underset{\mathbf{b}' \in A}{\min}\underset{\mathbf{a}' \in A}{\max} U\left(\{\mathbf{a}', \mathbf{b}'\}\right)$.*

Define $f : \mathbb{R}^2_+ \times \mathbb{R}^2_+ \to \mathbb{R}$ such that $f(\mathbf{a}, \mathbf{b}) = u(a_1) - \widetilde{U}(\mathbf{a}, \mathbf{b})$, where $\widetilde{U} : \mathbb{R}^2_+ \times \mathbb{R}^2_+ \to \mathbb{R}$ is a function satisfying:

$$U\left(\{\mathbf{a}, \mathbf{b}\}\right) = \max_{\mathbf{a}' \in \{\mathbf{a},\mathbf{b}\}} \min_{\mathbf{b}' \in \{\mathbf{a},\mathbf{b}\}} \widetilde{U}\left(\mathbf{a}', \mathbf{b}'\right) = \min_{\mathbf{b}' \in \{\mathbf{a},\mathbf{b}\}} \max_{\mathbf{a}' \in \{\mathbf{a},\mathbf{b}\}} \widetilde{U}\left(\mathbf{a}', \mathbf{b}'\right).^{28}$$

By definition we have $f(\mathbf{a}, \mathbf{a}) = 0$ for every $\mathbf{a} \in \mathbb{R}^2_+$. Note as well that

$$\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \Rightarrow f(\mathbf{a}, \mathbf{b}) > 0,$$

as otherwise we would have:

$$U\left(\{\mathbf{a}, \mathbf{b}\}\right) = \max \begin{Bmatrix} u(a_1) - \max\left\{{}^{f(\mathbf{a},\mathbf{a})=0}_{f(\mathbf{a},\mathbf{b})}\right\} \\ u(b_1) - \max\left\{{}^{f(\mathbf{b},\mathbf{a})}_{f(\mathbf{b},\mathbf{b})=0}\right\} \end{Bmatrix} \geq u(a_1) - \max \begin{Bmatrix} f(\mathbf{a}, \mathbf{a}) = 0 \\ f(\mathbf{a}, \mathbf{b}) \end{Bmatrix} = U\left(\{\mathbf{a}\}\right).$$

Claim 1.2: (i) $[\varphi(\mathbf{a}) < \varphi(\mathbf{b}) \text{ and } a_1 > b_1] \Leftrightarrow \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$

(ii) $[\varphi(\mathbf{a}) < \varphi(\mathbf{b}) \text{ and } a_1 \leq b_1] \Rightarrow \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\}$

[28] Note that $\underset{\mathbf{a} \in A}{\max}\underset{\mathbf{b} \in A}{\min} U\left(\{\mathbf{a}, \mathbf{b}\}\right) = \underset{\mathbf{a} \in A}{\max}\underset{\mathbf{b} \in A}{\min}\left[\underset{\mathbf{a}' \in \{\mathbf{a},\mathbf{b}\}}{\max}\underset{\mathbf{b}' \in \{\mathbf{a},\mathbf{b}\}}{\min} \widetilde{U}\left(\mathbf{a}', \mathbf{b}'\right)\right] = \underset{\mathbf{a} \in A}{\max}\underset{\mathbf{b} \in A}{\min} \widetilde{U}(\mathbf{a}, \mathbf{b})$.

(iii) $[\varphi(\mathbf{a}) = \varphi(\mathbf{b}) \text{ and } a_1 > b_1] \Rightarrow \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$.

Proof: (i) If $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$ then there exists A such that $\mathbf{a} \in A$ and $A \succ A \cup \{\mathbf{b}\}$. As $a_1 > b_1 \Leftrightarrow \{\mathbf{a}\} \succ \{\mathbf{b}\}$, by P_2 $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$. Conversely if $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$, then $\mathbf{b} \succ_f \mathbf{a}$ and hence $\varphi(\mathbf{a}) < \varphi(\mathbf{b})$. Further from SB and P_1, $a_1 > b_1$.

(ii) If $a_1 \leq b_1$ then by SB $\{\mathbf{b}\} \succeq \{\mathbf{a}, \mathbf{b}\}$. Since $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$, there is no B such that $\mathbf{b} \in B$ and $B \succ B \cup \{\mathbf{a}\}$, hence $\{\mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}\}$.

(iii) By P_1 $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ and then by SB $\{\mathbf{a}\} \succeq \{\mathbf{a}, \mathbf{b}\}$. As $\varphi(\mathbf{a}) = \varphi(\mathbf{b})$, using (i) we have $\{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\}$.‖

Let $(\mathbf{a}^*(A), \mathbf{b}^*(A))$ be the solution of

$$\max_{\mathbf{a}' \in A} \min_{\mathbf{b}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$$

so $(\mathbf{b}^*(A), \mathbf{a}^*(A))$ solves $\min_{\mathbf{b}' \in A} \max_{\mathbf{a}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$.

Claim 1.3: There exists $\mathbf{b} \in \arg\max_{\mathbf{a}' \in A} \varphi(\mathbf{a}')$ such that $A \sim \{\mathbf{a}, \mathbf{b}\}$ for some $\mathbf{a} \in A$ and $\mathbf{b}^*(A) = \mathbf{b}$.

Proof: Assume not, then there exist $\mathbf{a}, \mathbf{c}$ such that $\{\mathbf{a}, \mathbf{c}\} \sim A$, $(\mathbf{a}, \mathbf{c}) = (\mathbf{a}^*(A), \mathbf{b}^*(A))$. Therefore,

$$\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \sim A \ \forall \mathbf{b} \in \arg\max_{\mathbf{a}' \in A} \varphi(\mathbf{a}')$$

and and hence $\mathbf{c} \succ_f \mathbf{b}$, a contradiction.‖

For the remainder of the proof, let $I_f(\varphi) := \{\mathbf{b}' : \varphi(\mathbf{b}') = \varphi\}$. Define

$$Y(\mathbf{a}, \varphi) = \{\mathbf{b}' \in I_f(\varphi) : \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}\}$$

We make the following four observations:

(1) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}\}$ and $\mathbf{b} \succ_f \mathbf{c}$ imply $\{\mathbf{a}, \mathbf{c}\} \succeq \{\mathbf{a}, \mathbf{b}\}$.

(2) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}\} \succ \{\mathbf{c}\}$ and $\mathbf{b} \sim_f \mathbf{c}$ imply $\{\mathbf{a}, \mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}\}$.

(3) $\mathbf{b} \in Y(\mathbf{a}, \varphi)$, $\mathbf{b}' \sim_f \mathbf{b}$ and $\{\mathbf{b}\} \succ \{\mathbf{b}'\}$ imply $\mathbf{b}' \in Y(\mathbf{a}, \varphi)$.

(4) If $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{b}'\} \succ \{\mathbf{b}\}$ and $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$, then either $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}'\}$ or $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{b}'\} \succeq \{\mathbf{a}, \mathbf{b}\}$.

To verify these observations, suppose first that (1) did not hold. Then $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{c}\}$ and $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, hence by SB $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ and therefore $\mathbf{c} \succ_f \mathbf{b}$, which is a con-

tradiction. If (2) did not hold, we would get a contradiction to $\mathbf{b} \sim_f \mathbf{c}$ immediately. Next suppose that (3) did not hold. Then either $\{\mathbf{b}'\} \sim \{\mathbf{a}, \mathbf{b}'\}$ or $\{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}'\}$. In the first case $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\} \succ \{\mathbf{b}'\} \sim \{\mathbf{a}, \mathbf{b}'\}$and by SB $\{\mathbf{b}\} \succeq \{\mathbf{b}, \mathbf{b}'\}$ and, applying SB again , $\{\mathbf{b}\} \succeq \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$. But then $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, contradicting $\mathbf{b}' \sim_f \mathbf{b}$. In the second case $\{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\} \succ \{\mathbf{b}'\}$ and, using SB twice, $\{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, again contradicting $\mathbf{b}' \sim_f \mathbf{b}$. To verify (4), assume $\{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$. Then by Claim 1.2 (i) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$ and then by observation (2) $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{a}, \mathbf{b}\}$. If on the other hand $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{b}'\}$, then if $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}'\}$, $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$ and SB imply $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, a contradiction to $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$. Note that by Claim 1.3 we cannot have $\{\mathbf{b}'\} \succ \{\mathbf{a}, \mathbf{b}'\}$.‖

Next we claim that $\varphi(\mathbf{b}^*)$ is a sufficient statistic for the impact of $\mathbf{b}^*$ on a two element set.

Claim 1.4: There exits a function $\widetilde{U}$ satisfying the condition specified above such that $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$ implies $f(\mathbf{a}, \mathbf{b}) = g(\mathbf{a}, \varphi(\mathbf{b}))$ for some $g : \mathbb{R}^2_+ \times \mathbb{R} \to \mathbb{R}$ which is weakly increasing in its second argument.

Proof: Such $\widetilde{U}$ exists, if and only if $f(\mathbf{a}, \mathbf{b}) = g(\mathbf{a}, \varphi(\mathbf{b}))$ is consistent with $\succ$. Therefore it is enough to consider the constraints $\succ$ puts on f. Given $\mathbf{a}$ and $\mathbf{b}$, look at all $\mathbf{c}$ such that $\varphi(\mathbf{b}) > \varphi(\mathbf{c})$. We should show that $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$.

First note that if $\varphi(\mathbf{b}) \geq \varphi(\mathbf{a}) \geq \varphi(\mathbf{c})$, then $f(\mathbf{a}, \mathbf{b}) \geq 0 \geq f(\mathbf{a}, \mathbf{c})$ is consistent with $\succ$. If $\varphi(\mathbf{a}) \geq \varphi(\mathbf{b}) > \varphi(\mathbf{c})$, then $0 \geq f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$ is consistent with $\succ$. If $a_1 = 0$, then $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c}) \geq 0$ is consistent with $\succ$. Therefore, confine attention to the case where $a_1 > 0$ and $\varphi(\mathbf{b}) > \varphi(\mathbf{c}) > \varphi(\mathbf{a})$.

By Claim 1.2 (i), F_2 and F_3, there exists $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$ such that $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\}$. Thus, there are two cases to consider:

Case 1) Suppose $Y(\mathbf{a}, \varphi(\mathbf{b})) \neq \emptyset$. Define $f(\mathbf{a}, \mathbf{b}) := f(\mathbf{a}, \mathbf{b}')$ for some $\mathbf{b}' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$ (note that by observation (2) $f(\mathbf{a}, \mathbf{b}') = f(\mathbf{a}, \mathbf{b}'')$ $\forall \mathbf{b}', \mathbf{b}'' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$ and using observations (3) and (4), this definition is consistent with $\succ$.) If $Y(\mathbf{a}, \varphi(\mathbf{c})) \neq \emptyset$ then by observation (1) $\{\mathbf{a}, \mathbf{c}\} \succeq \{\mathbf{a}, \mathbf{b}\}$ and hence $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$. If $Y(\mathbf{a}, \varphi(\mathbf{c})) = \emptyset$ then by F_2 and continuity of $\succ_f$, there exists $\mathbf{c}' \in I_f(\mathbf{c})$ with $c'_1 < b'_1$ for some $\mathbf{b}' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$. Then by Claim 1.2 (i), P_1 and observation (1) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}'\} \succeq \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\} \succ \{\mathbf{c}'\}$, so $\mathbf{c}' \in Y(\mathbf{a}, \varphi(\mathbf{c}))$. Contradiction.

Case 2) Suppose $Y(\mathbf{a}, \varphi(\mathbf{b})) = \emptyset$. Define $f(\mathbf{a}, \mathbf{b}) := u(\mathbf{a}_1) - u(0)$. If $Y(\mathbf{a}, \varphi(\mathbf{c})) \neq \emptyset$, then $f(\mathbf{a}, \mathbf{c}) < u(a_1) - u(c_1) < u(a_1) - u(0) = f(\mathbf{a}, \mathbf{b})$. If $Y(\mathbf{a}, \varphi(\mathbf{c})) = \emptyset$ then set $f(\mathbf{a}, \mathbf{c}) = u(a_1) - u(0) = f(\mathbf{a}, \mathbf{b})$.‖

Let $S := \{(\mathbf{a}, \varphi) : Y(\mathbf{a}, \varphi) \neq \emptyset\}$. Note that S is an open set.

Claim 1.5: There is $g(\mathbf{a}, \varphi)$, which is continuous.

Proof: If $Y(\mathbf{a}, \varphi) \neq \emptyset$, then $g(\mathbf{a}, \varphi) = u(a_1) - U(\{\mathbf{a}, \mathbf{b}\})$ for some $\mathbf{b} \in Y(\mathbf{a}, \varphi)$ is clearly continuous. If $Y(\mathbf{a}, \varphi) = \emptyset$, then $\varphi \leq \varphi(\mathbf{a})$ implies $g(\mathbf{a}, \varphi) \leq 0$, while $\varphi > \varphi(\mathbf{a})$ implies $g(\mathbf{a}, \varphi) \geq u(a_1) - u(0)$. Define a switch point $(\widehat{\mathbf{a}}, \widehat{\varphi})$ to be a boundary point of S such that there exists $\widehat{\mathbf{b}} \in \mathbb{R}^2_{++}$ with $\varphi\left(\widehat{\mathbf{b}}\right) = \widehat{\varphi}$. For $\widehat{\varphi} = \varphi(\widehat{\mathbf{a}})$ define $g(\widehat{\mathbf{a}}, \widehat{\varphi}) := 0$ and for $\widehat{\varphi} > \varphi(\widehat{\mathbf{a}})$ define $g(\widehat{\mathbf{a}}, \widehat{\varphi}) := u(\widehat{a}_1) - u(0)$.

Consider a sequence $\{(\mathbf{a}^n, \varphi^n)\} \to (\widehat{\mathbf{a}}, \widehat{\varphi})$ in S. Pick a sequence $\{\mathbf{b}^{n\prime}\}$ with $\mathbf{b}^{n\prime} \in Y(\mathbf{a}^n, \varphi^n) \, \forall n$. Define $\{b_1^n\} = \left\{\min\left[\frac{1}{n}, b_1^{n\prime}, \widehat{b}_1\right]\right\}$. Define b_2^n to be a solution to $\varphi(b_1^n, b_2^n) = \varphi^n$. By F_2 and F_3, b_2^n is well defined. Note that by observation (3) $\mathbf{b}^n = (b_1^n, b_2^n) \in Y(\mathbf{a}^n, \varphi^n)$. Lastly, let $\widehat{b}_1^n \equiv b_1^n$ and $\widehat{b}_2^n$ be the solution to $\varphi\left(\widehat{b}_1^n, \widehat{b}_2^n\right) = \widehat{\varphi}$. We have $U(\{\mathbf{a}^n, \mathbf{b}^n\}) = u(a_1^n) - g(\mathbf{a}^n, \varphi^n)$. If in the switch point $\widehat{\varphi} = \varphi(\widehat{\mathbf{a}})$, then $U\left(\left\{\widehat{\mathbf{a}}, \widehat{\mathbf{b}}^n\right\}\right) = u(\widehat{a}_1)$. By continuity, $U(\{\mathbf{a}^n, \mathbf{b}^n\}) - U\left(\left\{\widehat{\mathbf{a}}, \widehat{\mathbf{b}}^n\right\}\right) \underset{n\to\infty}{\to} 0$, hence

$$\lim_{n\to\infty} g(\mathbf{a}^n, \varphi^n) = \lim_{n\to\infty} [u(a_1^n) - u(\widehat{a}_1)] = u(\widehat{a}_1) - u(\widehat{a}_1) = 0 = g(\widehat{\mathbf{a}}, \widehat{\varphi}).$$

If in the switch point $\widehat{\varphi} > \varphi(\widehat{\mathbf{a}})$, then $U\left(\left\{\widehat{\mathbf{a}}, \widehat{\mathbf{b}}^n\right\}\right) = u\left(\widehat{b}_1^n\right) = u(b_1^n)$. By the same continuity argument

$$\lim_{n\to\infty} g(\mathbf{a}^n, \varphi^n) = \lim_{n\to\infty} [u(a_1^n) - u(b_1^n)] = u(\widehat{a}_1) - u(0) = g(\widehat{\mathbf{a}}, \widehat{\varphi}).$$

For $\varphi < \varphi(\mathbf{a})$ let $g(\mathbf{a}, \varphi) < 0$. This satisfies the constraint on f. So g can be continuous in both arguments and increasing in φ and such that for any sequence $\{(\mathbf{a}^n, \varphi^n)\}$ in S, with $\{(\mathbf{a}^n, \varphi^n)\} \to (\widehat{\mathbf{a}}, \widehat{\varphi})$, we have $\lim_{n\to\infty} g(\mathbf{a}^n, \varphi^n) = 0.\|$

This establishes the existence of a continuous representation

$$U(A) = \max_{\mathbf{a}\in A}\left[u(a_1) - g\left(\mathbf{a}, \max_{\mathbf{b}\in A}\varphi(\mathbf{b})\right)\right]$$

of $\succ$ on K_{++} with the properties a specified in the theorem. Continuity of $\succ$ implies that it's unique continuous extension to K represents $\succ$ on K. This extension can be found by extending φ to $\mathbb{R}^2_+$ and g to $\mathbb{R}^2_+ \times \mathbb{R}$. That the representation satisfies the axioms is easy to verify. This completes the proof of Theorem 1.■

Proof of Theorem 2

Theorem 2 and Theorem 4 (i) are analogous, where Theorem 2 covers the case $N = 2$, while Theorem 4 (i) covers the case $N \geq 3$. We prove Theorem 4 (i) below by first establishing that the analogous version of Theorem 1 holds. From there on the proof of Theorem 2 is identical to the proof of Theorem 4 (i), with a_2 substituted for $\mathbf{a}_{-1}$.

Proof of Theorem 3

Here we show that $F_1 - F_4$ imply those in the Lemma of Karni and Safra (1998). The existence and uniqueness of the representation follows immediately from their Lemma, and Theorem 2 in Chapter 6 of Kranz *et al.* (1971).

Beside F_1 (weak order) and F_4 (their Hexagon Condition.), Karni and Safra require the following axioms:

Independence: $(a_1, a) \succeq_f (b_1, a)$ for some a implies $(a_1, b) \succeq_f (b_1, b)$ for all b.

Independence is implied since by F_2, $(a_1, a) \succeq_f (b_1, a) \Leftrightarrow a_1 \geqslant b_1 \Leftrightarrow (a_1, b) \succeq_f (b_1, b)$ for all b.

Restricted Solvability: If $(a_1, a_2) \succeq_f (b_1, b_2) \succeq_f (a_1', a_2)$ *then there is* x *such that* $(b_1, b_2) \sim_f (x, a_2)$. And if $(a_1, a_2) \succeq_f (b_1, b_2) \succeq_f (a_1, a_2')$ *then there is* y *such that* $(b_1, b_2) \sim_f (a_1, y)$

Restricted Solvability is immediately implied by F_3.

A sequence $\{a_i\}$ is a standard sequence, if for some $a \neq b$, $(a_i, a) \tilde{}_f (a_{i+1}, b)$ for all i. A standard sequence is bounded if there exist $\underline{a}$ and $\overline{a}$ such that for all i, $a_i \in (\underline{a}, \overline{a})$. Define similarly standard (And bounded) sequences by varying the second component.

Archimedian property:every bounded standard sequence is finite.

To show that the Archimedian property is implied, fix $a \neq b$ and let $\{a_i\}$ be a standard sequence. If $a > b$ then $\{a_i\}$ is an increasing sequence and if $b > a$ $\{a_i\}$ is a decreasing sequence. Suppose that $\{a_i\}$ is bounded away from 0 and ∞. Let $\overline{a}$ and $\underline{a}$ be the least upper bound and greatest lower bound respectively.

Case 1, $a > b$. By F_3, there exists x such that $(\overline{a}, a) \tilde{}_f (x, b)$. By F_2, $x > \overline{a}$. Since $\{a_i\}$ is an increasing and bounded sequence, it must converge to its least upper bound, $\overline{a}$. By continuity, there exists a sub-sequence $\{a_{ik}\}$ that converges to x. In particular, there exists K such that for $k > K$, $x - a_{ik} < \varepsilon := \frac{x - \overline{a}}{2}$, contradiction.

Case 2, $a < b$. By F_2, $(\underline{a}, a) \prec_f (\underline{a}, b)$. Since $\{a_i\}$ is a decreasing and bounded sequence, it must converge to its greatest lower bound, $\underline{a}$. By continuity, there exists I such that $i > I$ implies $(a_i, a) \prec_f (\underline{a}, b)$. Since $\underline{a}$ is the greatest lower bound, F_2implies that $(\underline{a}, b) \prec_f (a_{i+1}, b)$. Therefore, $(a_i, a) \prec_f (a_{i+1}, b)$, contradiction.

Essentiality: Not $(a', b) \tilde{}_f (a, b)$ for all b, or $(a, b) \tilde{}_f (a, b')$ for all a.

Essentiality is immediately implied by F_2.

Proof of Theorem 4

(i) The analogue of Theorem 1 can be established by substituting $\mathbf{a}_{-1}$ for a_2 in the theorem and in the proof, where now $\varphi : \mathbb{R}^N_+ \to \mathbb{R}$ represents the unique continuous extension of $\succ_f$ to $\mathbb{R}^N_+$.

As in the proof of Theorem 1 and its analogue, we first show that there is a representation as in Theorem 4 on the domain K_{++}. Continuity will then allow us to extend it to all of K. Unless mentioned otherwise, let $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^N_{++}$ for the remainder of this proof.

Given φ, let $\overline{\varphi} := \sup_{\mathbf{a} \in \mathbb{R}^N_{++}} \varphi(\mathbf{a})$ and $\underline{\varphi} := \inf_{\mathbf{a} \in \mathbb{R}^N_{++}} \varphi(\mathbf{a})$, if they are well defined. Otherwise, take $\overline{\varphi} = \infty$ and $\underline{\varphi} = -\infty$. As before, let $S := \{(\mathbf{a}', \varphi') : Y(\mathbf{a}', \varphi') \neq \emptyset\}$. By F_3^N and the representation analogous to Theorem 1, $u(a_1) - u(0) > g(\mathbf{a}, \varphi)$ for $(\mathbf{a}, \varphi) \in S$.

Let $\succ_S$ be a binary relation on S defined by $(\mathbf{a}, \varphi) \succ_S (\widetilde{\mathbf{a}}, \widetilde{\varphi}) \Leftrightarrow \{\mathbf{a}, \mathbf{b}\} \succ \left\{\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}\right\}$ $\forall \mathbf{b} \in Y(\mathbf{a}, \varphi)$ and $\forall \widetilde{\mathbf{b}} \in Y(\widetilde{\mathbf{a}}, \widetilde{\varphi})$.

Define $U_S : \mathbb{R}^N_{++} \times \left(\underline{\varphi}, \overline{\varphi}\right) \to \mathbb{R}$ such that $U_S(\mathbf{a}, \varphi) := U(\{\mathbf{a}, \mathbf{b}\})$ for some $\mathbf{b} \in Y(\mathbf{a}, \varphi)$. By Theorem 1, $\succ_S$ is a weak order that can be represented by U_S. Note that the Consistency axiom (P_4) is relevant precisely on this domain. For $(\mathbf{a}, \varphi) \notin S$ define

$$U_S(\mathbf{a}, \varphi) := \begin{cases} u(0) & \text{for } \varphi(\mathbf{a}) < \varphi \\ u(a_1) & \text{for } \varphi(\mathbf{a}) \geq \varphi \end{cases}.$$

Claim 4.1: U_S is continuous in all arguments.

Proof: Since the utility function is continuous on S, and because outside of S the function was chosen to be either a constant (hence continuous) or a continuous function, the only candidates for discontinuity are points on the boundary of S. There are two cases:

Case 1) $\varphi(\mathbf{a}) \geqslant \varphi$: Take $(\mathbf{a}, \varphi) \in bdr(S)$. Since $(\mathbf{a}, \varphi)$ is a boundary point, it must be that $\varphi(\mathbf{a}) = \varphi$. Now let $\{\mathbf{a}^n, \varphi^n\}$ be a sequence in S which converges to $(\mathbf{a}, \varphi)$. By the definition of S, $U_s\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right)$. Because preferences are continuous and using the properties of g from Theorem 1, we have $\lim_{n \to \infty} u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u(a_1)$ as required.

Case 2) $\varphi(\mathbf{a}) < \varphi$: Take $(\mathbf{a}, \varphi) \in bdr(S)$. Again, let $\{\mathbf{a}^n, \varphi^n\}$ be an arbitrary sequence in S which converges to $(\mathbf{a}, \varphi)$. By the definition of S,

$$U_s\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) > \inf_{\mathbf{b} \in \mathbb{R}^n_{++}} \{u(b_1) : \varphi(\mathbf{b}) = \varphi^n \text{ and } b_1 < a_1^n\}.$$

Since $\succ$ is continuous, we have

$$\begin{aligned} \lim_{n\to\infty} u\left(a_1^n\right) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) &= u\left(a_1\right) - g\left(\left(a_1, \mathbf{a}_{-1}\right), \varphi\right) \geq \\ \inf_{\mathbf{b}\in\mathbb{R}^n_{++}} \left\{u\left(b_1\right) : \varphi\left(\mathbf{b}\right) = \varphi \text{ and } b_1 < a_1\right\} &= u\left(0\right). \end{aligned}$$

where the last equality is implied by F_3^N. As $(\mathbf{a},\varphi) \notin S$, we claim that $u\left(a_1\right) - g\left(\left(a_1, \mathbf{a}_{-1}\right), \varphi\right) \leq \inf_{\mathbf{b}\in\mathbb{R}^n_{++}} \left\{u\left(b_1\right) : \varphi\left(\mathbf{b}\right) = \varphi \text{ and } \{\mathbf{b}\} \sim \{\mathbf{a},\mathbf{b}\}\right\} = u\left(0\right)$. If not then there exists $\mathbf{c}$ with $c_1 > 0$ and $\varphi\left(\mathbf{c}\right) = \varphi$, such that $u\left(a_1\right) - g\left(\left(a_1, \mathbf{a}_{-1}\right), \varphi\right) = u\left(c_1\right) > u\left(0\right)$. But using F_3^N, we could find $\mathbf{c}'$ with $c_1' < c_1$ and $\varphi\left(\mathbf{c}'\right) = \varphi\left(\mathbf{c}\right) = \varphi$. Using Theorem 1, this would imply that $(\mathbf{a},\varphi) \in S$, which is a contradiction. Combining we have $\lim_{n\to\infty} u\left(a_1^n\right) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u\left(0\right)$, as required.‖

Definition: For $(\mathbf{a},\varphi) \in S$, define $I_S\left(\mathbf{a},\varphi\right) := \{(\mathbf{a}',\varphi') \in S : (\mathbf{a}',\varphi') \sim_S (\mathbf{a},\varphi)\}$. That is, $I_S\left(\mathbf{a},\varphi\right)$ is the $\succ_S$ equivalence class of (**a**,φ).

Let $a_1^* : \mathbb{R}^2_+ \times \left(\underline{\varphi}, \overline{\varphi}\right) \to \mathbb{R}_+$ be the solution to

$$u\left(a_1^*\left(\mathbf{a},\varphi\right)\right) = u\left(a_1\right) - g\left(\mathbf{a},\varphi\right) = U_S\left(\mathbf{a},\varphi\right).$$

a_1^* is the "first component equivalent" functional on S.[29] Since $u\left(a_1\right) > u\left(a_1\right) - g\left(\mathbf{a},\varphi\right) > u\left(0\right)$ and $\succ_S$ is continuous, a_1^* is well defined and we have $(\mathbf{a},\varphi) \succ_S (\widetilde{\mathbf{a}},\widetilde{\varphi}) \Leftrightarrow a_1^*\left(\mathbf{a},\varphi\right) > a_1^*\left(\widetilde{\mathbf{a}},\widetilde{\varphi}\right)$.

Claim 4.2: Shame $g\left(\mathbf{a},\varphi\right)$ is strictly increasing in φ.

Proof: Assume to the contrary that there is $\varphi' > \varphi$ and $(\mathbf{a},\varphi') \sim_S (\mathbf{a},\varphi)$ for some $\mathbf{a}$. Then for $\varphi' > \varphi'' > \varphi''' > \varphi$ we must have $(\mathbf{a},\varphi'') \sim_S (\mathbf{a},\varphi''')$ as shame is weakly increasing in φ. Now pick $\mathbf{a}'$ such that $(\mathbf{a}',\varphi) \succ_S (\mathbf{a}',\varphi')$ and $(\mathbf{a}',\varphi), (\mathbf{a}',\varphi') \in S$. This is possible by continuity of U_S, since for $\mathbf{a}''$ such that $\varphi\left(\mathbf{a}''\right) = \varphi$ the definition of U_S yields $U_S\left(\mathbf{a}'',\varphi\right) > U_S\left(\mathbf{a}'',\varphi'\right)$. Then by P_4, $(\mathbf{a}',\varphi''') \succ_S (\mathbf{a}',\varphi'')$, a contradiction to shame being weakly increasing in φ.‖

Definition: Given $\mathbf{a} = (a_1, a_2, ..., a_k)$, let $\boldsymbol{\varepsilon}\left(\mathbf{a}\right) := \frac{\min\{a_1,a_2,...,a_k\}}{2}\left(1, ..., 1\right)$.

Claim 4.3: For all $(\mathbf{a},\varphi)$ and $\widetilde{\varphi} \in \left(\varphi\left(a_1, \boldsymbol{\varepsilon}\left(\mathbf{a}_{-1}\right)\right), \overline{\varphi}\right)$ there exists $\widetilde{\mathbf{a}}$ such that $(\widetilde{\mathbf{a}},\widetilde{\varphi}) \in$

[29] Formally, $\forall \mathbf{x} \in \mathbb{R}^{N-1}_+$, $\{(a_1^*\left(\mathbf{a},\varphi\right), \mathbf{x})\} \sim \{\mathbf{a},\mathbf{b}\}, \forall \mathbf{b} \in Y\left(\mathbf{a},\varphi\right)$

$I_S(\mathbf{a}, \varphi)$.

Proof: Define φ^* implicitly by $U_s((a_1, \boldsymbol{\varepsilon}(\mathbf{a}_{-1})), \varphi^*) = U_s(\mathbf{a}, \varphi)$. This is possible by the Intermediate Value Theorem, as $U_s((a_1, \boldsymbol{\varepsilon}(\mathbf{a}_{-1})), \varphi(a_1, \boldsymbol{\varepsilon}(\mathbf{a}_{-1}))) = u(a_1) > U_s(\mathbf{a}, \varphi) > U_s((a_1, \boldsymbol{\varepsilon}(\mathbf{a}_{-1})), \varphi)$, where the last inequality is due to P_4 and Claim 4.2. There are two cases to consider:

Case 1) $\widetilde{\varphi} \geq \varphi^*$: Then $U_s((a_1, \boldsymbol{\varepsilon}(\mathbf{a}_{-1})), \widetilde{\varphi}) \leq U_s(\mathbf{a}, \varphi)$ according to the monotonicity of shame. By F_3^N there is $\overline{a}_2(\widetilde{\varphi})$ that solves $\varphi(a_1, \overline{a}_2(\widetilde{\varphi}), \boldsymbol{\varepsilon}(\mathbf{a}_{-1,2})) = \widetilde{\varphi}$. Then $U_s((a_1, \overline{a}_2(\widetilde{\varphi}), \boldsymbol{\varepsilon}(\mathbf{a}_{-1,2})), \widetilde{\varphi}) \geq U_s(\mathbf{a}, \varphi)$ and by the Intermediate Value Theorem there is $\widetilde{a}_2(\widetilde{\varphi}) \in \left[\frac{\min\{a_1, a_2, \ldots, a_n\}}{2}, \overline{a}_2(\widetilde{\varphi})\right)$ such that

$$U_s((a_1, \widetilde{a}_2(\widetilde{\varphi}), \boldsymbol{\varepsilon}(\mathbf{a}_{-1,2})), \widetilde{\varphi}) = U_s(\mathbf{a}, \varphi).$$

Case 2) $\widetilde{\varphi} < \varphi^*$: Then

$$U_s((a_1^*(\mathbf{a}, \varphi), \boldsymbol{\varepsilon}(\mathbf{a}_{-1})), \widetilde{\varphi}) \leq U_s(\mathbf{a}, \varphi) < U_s((a_1, \boldsymbol{\varepsilon}(\mathbf{a}_{-1})), \widetilde{\varphi}).$$

By the Intermediate Value Theorem there is $\widetilde{a}_1(\widetilde{\varphi}) \in [a_1^*(\mathbf{a}, \varphi), a_1]$ such that

$$U_s((\widetilde{a}_1(\widetilde{\varphi}), \boldsymbol{\varepsilon}(\mathbf{a}_{-1})), \widetilde{\varphi}) = U_s(\mathbf{a}, \varphi). \|$$

Combining the two cases we see that $\widetilde{\varphi}$ parametrizes a path

$$\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi}) := \begin{cases} (\widetilde{a}_1(\widetilde{\varphi}), \boldsymbol{\varepsilon}(\mathbf{a}_{-1})) & \text{for } \widetilde{\varphi} < \varphi^* \\ (a_1, \widetilde{a}_2(\widetilde{\varphi}), \boldsymbol{\varepsilon}(\mathbf{a}_{-1,2})) & \text{for } \widetilde{\varphi} \geq \varphi^* \end{cases}$$

of allocations. According to Claim 4.2 $\varphi(\mathbf{a})$ must be strictly increasing along this path. This implies $\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ is strictly increasing in its first component for $\widetilde{\varphi} < \varphi^*$ and in its second component for $\widetilde{\varphi} \geq \varphi^*$.

Now we construct a $\succ_S$ indifference curve close to the original one:

Claim 4.4: For $\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ as defined above, $\widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ that solves

$$\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi}), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})\right) \in I_S(\mathbf{a}, \varphi + d\varphi)$$

is increasing in $\widetilde{\varphi}$.

Proof: Assume $\widetilde{\varphi}' > \widetilde{\varphi}$. There are two cases to consider:

Case 1) $\widetilde{\varphi}' > \varphi^*$: Then $\widetilde{a}_{1(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) = a_1$, $\widetilde{a}_{1(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right) \leq a_1$ and $\widetilde{a}_{2(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) > \widetilde{a}_{2(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right)$. P_4 implies

$$\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right)\right) \prec_S \left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right)\right).$$

Case 2) $\widetilde{\varphi}' \leq \varphi^*$: Then $\widetilde{a}_{2(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) = \widetilde{a}_{2(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right) = \frac{\min\{a_1,a_2,\ldots,a_n\}}{2}$ and $\widetilde{a}_{1(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) > \widetilde{a}_{1(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right)$.
As $\succ_S$ is increasing in a_1,

$$\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right)\right) \prec_S \left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right)\right).$$

As shame increases in φ, we must have $\widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) > \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}\right)$ in both cases.‖

The $\succ_S$-indifference curve through $(\mathbf{a}, \varphi)$ constructed above is increasing in φ. Claim 4.4 then implies that we can construct another indifference curve, parametrized by the same path in the $\mathbf{a}$-hyperplane, which is arbitrarily close in terms of φ and is also increasing in φ. For those two indifference curves we can define a re-scaling $\varphi \mapsto \gamma(\varphi)$ that makes them parallel, where $\gamma : \mathbb{R} \to \mathbb{R}$ is strictly increasing and continuous. Due to P_4 the definition of $\gamma : \mathbb{R} \to \mathbb{R}$ is independent of the choice of $(\mathbf{a}, \varphi)$. Thus, $\varphi \mapsto \gamma(\varphi)$ transform the original indifference map of $U_S(\mathbf{a}, \varphi)$ to be quasi-linear in $\gamma(\varphi)$.[30]

Remember that $U_S(\mathbf{a}, \varphi)$ is strictly decreasing in φ. Therefore, there exists $H : \mathbb{R}^N_{++} \to \mathbb{R}$, such that $H(\mathbf{a}) - \gamma(\varphi)$ represents $\succ_S$ on S.

Define $u(a_1) := H(\mathbf{a}) - \lim_{\varphi \to \varphi(\mathbf{a})} \gamma(\varphi)$. Because of P_1,

$$U(\{\mathbf{a}, \mathbf{b}\}) := \begin{cases} u(a_1) \text{ if } \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\} \\ H(\mathbf{a}) - \gamma(\varphi(\mathbf{b})) \text{ if } \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\} \\ u(b_1) \text{ if } \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{b}\} \end{cases}$$

represents $\succ$ confined to the collection of all two element sets in K_{++}. Therefore, $H(\mathbf{a}) \equiv u(a_1) + \gamma(\varphi(\mathbf{a}))$ must hold. Hence

$$U(A) = \max_{\mathbf{a} \in A} [u(a_1) + \gamma(\varphi(\mathbf{a}))] - \max_{\mathbf{b} \in A} [\gamma(\varphi(\mathbf{b}))]$$

represents $\succ$ on K_{++}, where φ represents $\succ_f$, and u and γ are strictly increasing. Since φ represents $\succ_f$, so does $\gamma(\varphi)$. Hence, there is a representation φ of $\succ_f$, such that γ is the

[30] The proof that this rescaling exists is again an application of the lock-step procedure exlained in Keeney and Raiffa (1976). For brevity we do not reproduce it here.

identity and

$$U(A) = \max_{\mathbf{a}\in A}[u(a_1) + \varphi(\mathbf{a})] - \max_{\mathbf{b}\in A}[\varphi(\mathbf{b})]$$

represents $\succ$ on K_{++}. As $\succ$ is continuous, the unique continuous extension of U to K represents $\succ$. This extension can be found by extending u and φ to $\mathbb{R}^n_+$.

(ii) To establish the analogue of Theorem 3, namely that there are N increasing unbounded functions $v_1, .., v_N$, such that the fairness ranking $\succ_f$ can be represented by $\varphi(\mathbf{a}) = v_1(a_1) \cdot ... \cdot v_N(a_n)$, if and only if it satisfies F_1, F_2, F_3^N and F_4^N we apply the Theorem of Luce and Tukey, just as in the proof of Theorem 3. It establishes the existence of an additive representation $\xi_1(a_1) + ... + \xi_N(a_N)$ of $\succ_f$. Define $v_n(a_n) := \exp(\xi_n(a_n))$ for all $n \in \{1, .., N\}$. Then $v_1, .., v_N : \mathbb{R}_+ \to \mathbb{R}_{++}$ are increasing and continuous and if we re-define $\varphi(\mathbf{a}) := v_1(a_1) \cdot ... \cdot v_N(a_N)$, it represents $\succ_f$. By F_3^N, the functions $v_1, .., v_N$ must be unbounded.

That the representations satisfy the axioms is easy to verify.■

Proof of Proposition

i) In the observed case, trusted player 2 can either choose C, which carries no shame and generates direct utility c, or he can choose D, which carries shame βc^2 and generates direct utility d. Hence player 2 cooperates if $c > d - \beta c^2$, is indifferent between cooperating and defecting if $c = d - \beta c^2$ and defects if $c < d - \beta c^2$. Anticipating this, player 1 chooses T if $c > d - \beta c^2$ and N if $c < d - \beta c^2$. If $c = d - \beta c^2$, there are two equilibria, T, C and N, D.

ii) Consider the two possible equilibria of the unobserved case. Suppose player 2 was required to play C in equilibrium. Then, player 1 expects to see either the outcome (c, c) or $(0, d)$, so neither outcome makes her think that player 2 deviated. Since the expected outcome $\widetilde{\mathbf{a}} = (c, c)$ is not affected by player 2's action, it is profitable for player 2 to deviate and play D, generating a higher direct utility (d instead of c) without increasing shame. Therefore, there is no equilibrium where player 2 chooses C. Suppose then that player 2 is required to play D in equilibrium. In that case, player 1 expects to see the outcome $(0, d)$ for sure, and $\widetilde{\mathbf{a}} = (0, d)$. Playing D, therefore, generates direct utility d and carries shame βc^2. If, however, player 1 observes (c, c), then this can only be explained by player 2 having deviated from D to C,[31] and accordingly $\widetilde{\mathbf{a}} = (c, c)$.[32] If player 2 plays C he receives direct utility c and with probability p there is no shame, while with probability $(1 - p)$ shame is still βc^2. Hence, player 2 is willing to play D in equilibrium, if and only if $d - \beta c^2 \geq c - (1 - p)\beta c^2$

[31] This is true only because we exclude mixed strategies from the players' strategy spaces. Otherwise, (c, c) could be explained by a continuum of mixed strategies and we would have to specify out-of-equilibrium beliefs.

[32] More precisely, $\widetilde{\mathbf{a}}$ is player 2's belief about player 1's perception of his action. Whether or not player 1 actually updates her beliefs in this manner is irrelevant.

or $\frac{d-c}{\beta c^2} \geq p$. In that case player 1 anticipates player 2 to play D and chooses N.■

References

[1] Anderoni, James and John H. Miller (2002) "Giving according to GARP: An experimental test of the consistency of preference for altruism." *Econometrica*, 70, 737-753.

[2] Benabou, Roland and Jean Tirole (2006) "Incentives and Prosocial Behavior." *American Economic Review*, 96(5), 1652-1678.

[3] Bolton, Gary E., Elena Katok and Rami Zwick (1998) "Dictator game giving: Rules of fairness versus acts of kindness." *International Journal of Game Theory,* 27: 269-299.

[4] Broberg, Thomas, Tore Ellingsen and Magnus Johannesson (2007) "Is Generosity Involuntary?" *Economics Letters,* 94, 32-37.

[5] Burnham, Terence C. (2003) "Engineering altruism: A theoretical and experimental investigation of anonymity and gift giving." *Journal of Economic Behavior and Organization,* 50, 133-144.

[6] Buss, Arnold H. (1980) "Self-Consciousness and Social Anxiety." San Francisco, W. H. Freeman.

[7] Camerer, Colin (2003) "Behavioral Game Theory: Experiments in Strategic Interaction". Princeton University Press.

[8] Charnes, Gary and Mathew Rabin (2002) "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 117 (3), 817-870.

[9] Dana, Jason D., Dalian M. Cain and Robin M. Dawes (2006) "What you don't Know Won't Hurt me: Costly (but quiet) Exit in a Dictator Game." *Organizational Behavior and Human Decision Processes*, 100(2), 193-201.

[10] Davis, Douglas D. and Charles A. Holt (1993) " Experimental economics." Princeton University Press.

[11] Dekel, Eddie, Barton L. Lipman and Aldo Rustichini (2008) "Temptation Driven Preferences." *Review of Economic Studies*, forthcoming

[12] Epstein, Larry G. and Igor Kopylov (2007) "Cold Feet" *Theoretical Economics*, Volume 2, 231–259

[13] Fehr Ernst, Klaus M. Schimdt (1999) "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics*, 114, 817-868.

[14] Frohlich, Norman, Joe Oppenheimer and J. Bernard Moore (2001) "Some doubts measuring self-interst using dictator experiments: The cost of anonymity." *Journal of Economic Behavior and Organization,* 46, 271-290.

[15] Gauthier, David and Robert Sugden (editors) (June 1993) "Rationality, Justice and Social Contract: Themes from Morals by Agreement." University of Michigan Press.

[16] Gul, Faruk and Wolfgang Pesendorfer (2005) "The Simple Theory of Temptation and Self-Control." mimeo

[17] —— (2001) "Temptation and Self Control." *Econometrica*, Vol. 69, No. 6, 1403-1435

[18] Haley, Kevin J. and Daniel M.T. Fessler (2005) "Nobody's watching? Subtle cues affect generosity in an anonymous economic game." *Evolution and Human Behavior,* 26, 245-256.

[19] Hammond, Peter J.(1991), "Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made", in Elster and Roemer, "Interpersonal Comparisons of Well Being". Cambridge: Cambridge University Press, pp 200-254.

[20] Johannesson, Magnus and Bjoran Persson (2000) "Non-reciprocal altruism in dictator games." *Economics Letters,* 69, 137-142.

[21] Karni, Edi and Zvi Safra (1998) "The Hexagon Condition and Additive Representation for Two Dimensions: An Algebraic Approach." *Journal of Mathematical Psychology,* 42, 393-399.

[22] Keeney, Ralph L. and Howard Raiffa, with a contribution by Richard F. Meyer (1976) "Decisions with multiple objectives : preferences and value trade-offs". New York : Wiley.

[23] Koch, K. Alexander and Hans-Theo Norman (2005) "Giving in Dictator Games: Regard *for* Others or Regard *by* others?" mimeo.

[24] Krantz, David H., R. Duncan Luce, Patrick C. Suppes and Amos Tversky (1971) " Foundations of Measurments, Vol 1." Academic Press, New York.

[25] Lazear, Edward P., Ulrike Malmendier and Roberto A. Weber (2005) "Sorting in Experiments with Application to Social Preferences." mimeo.

[26] Levitt, Steven D and John A. List (2007) "What do laboratory experiments measuring social prefernces reveal about the real world?" *Journal of Economic Perspective,* 21(2): 153–174

[27] Luce, R. Duncan and John W. Tukey (1964) "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology,* 1,1-27.

[28] Mariotti, Marco (1999) "Fair Bargains: Distributive Justice and Nash Bargaining Theory." *Review of Economic Studies*, Vol.66, 733-41.

[29] Miller, Dale T. (1999) "The Norm of Self-Interest." *American Psychologist*, Vol. 54, No. 12, 1053-1060.

[30] Nash, John F. (1953) "Two-Person Cooperative Games." *Econometrica,* Vol. 21, No. 1, 128-140.

[31] —— (1950) "The Bargaining Problem." *Econometrica*, Vol. 18, No. 2, 155-162.

[32] Neilson, William S. (2006) "Axiomatic Reference Dependence in Behavior towards Others and Toward Risk." *Economic Theory* 28, 681-692.

[33] —— (2008) "A Theory of Kindness, Reluctance, and Shame in Dictator Games." *Games and Economic Behavior*, forthcoming .

[34] Noor, Jawwad and Norio Takeoka (2008) "Menu-Dependent Self-Control" mimeo.

[35] Oberholzer-Gee, Felix and Reiner Eichenberger (2008) "Fairness in Extended Dictator Game Experiments," *The B.E. Journal of Economic Analysis & Policy*: Vol. 8: Iss. 1 (Contributions), Article 16. Available at: http://www.bepress.com/bejeap/vol8/iss1/art16

[36] Osborne, Martin J. and Ariel Rubinstein (1994) "A course in Game Theory." MIT press, ISBN 0-262-65040-1.

[37] Olszewski Wojciech (2008) "A Model of Consumption-Dependent Temptation" mimeo

[38] Pillutla, Madan M. and J. Keith Murningham (1995) "Being fair or appearing fair: Strategic behavior in ultimatum bargaining." *Academy of Managment Journal,* 38,1408-1426.

[39] Rawls, John (1971) "A Theory of Justice." *The Belknap Press of Harvard University Press.*

[40] Tadelis, Steve (2008) "The Power of Shame and the Rationality of Trust." mimeo