

Ashamed to be Selfish*

David Dillenberger[†] Philipp Sadowski[‡]

September 28, 2009

Abstract

We study a decision maker (DM) who has preferences over sets of payoff-allocations between herself and a passive recipient, which represent second-stage choice-problems. The recipient is only aware of second-stage choice of an allocation. Not choosing the fairest available allocation in the second stage inflicts shame on DM. We derive a representation that identifies DM's private ranking of allocations, her norm of fairness and shame. The norm of fairness can be further characterized as the Nash solution of a bargaining game induced by the second-stage choice problem.

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 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)).

*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

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 pre-specified 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 purely 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 crossing the road to avoid meeting a beggar and donating to charity over the phone, but wishing not to have been home when the call came.

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. We term "shame" the motive that distinguishes choice behavior when observed from choice behavior when not observed.¹ Individuals in our model are endowed with two rankings; the first is their private ranking, which determines the value of allocations when not observed. The second is their fairness ranking. Although the private ranking might incorporate other-regarding preferences (such as altruism),² individuals are selfish, in the sense that their private ranking always gives more weight to their own payoffs than the fairness ranking. In the extreme case, the private ranking of an individual is based solely on her own payoffs. When observed, individuals feel shame when they do not choose the fairest available alternative; they are ashamed to be selfish.³

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

¹Our notion of shame is closely related to that of embarrassment. We use shame to highlight that the emotional cost experienced by DM is caused by her own actions and is not triggered by actions of others, which place DM in a socially awkward situation. One way to distinguish between shame and guilt is to view guilt as involving 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.

² Levitt and List (2007) is a comprehensive survey of models that involve various other-regarding preferences that are not necessarily impacted by observation.

³In order to be meaningful for revealed preference analysis, at least part of the choice behavior must be observed by someone, for example the experimenter. Therefore, we can only study the effect of the differential shame from being observed by the recipient in addition to the experimenter. While we do not explicitly model the influence on the player's behavior caused by the presence of the experimenter, any such effect would be contained in the private ranking.

second-stage choice is also unknown to the recipient. Similarly, Pillutla and Murningham (1995) find evidence 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 value of allocations, and that specifies the properties of the shame component. We justify using this convenient form by deriving it from plausible assumptions on preferences. 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.

⁴The 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 resolve or temptation may kick in over time.

⁵If 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 $A = \{\mathbf{a}, \mathbf{b}, \dots\} = \{(a_1, a_2), (b_1, b_2), \dots\}$, 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 these representations:

$$U(A) = \max_{\mathbf{a} \in A} [u(\mathbf{a}) + \beta\varphi(\mathbf{a})] - \beta \max_{\mathbf{b} \in A} [\varphi(\mathbf{b})] \quad (1)$$

where u and φ are increasing in all arguments. u is a utility function over allocations and $\varphi(\mathbf{a})$ is the fairness of the allocation \mathbf{a} .

Alternatively, if we denote by \mathbf{a}^* and \mathbf{b}^* the two maximizers above, (1) can be written as:

$$U(A) = \underbrace{u(\mathbf{a}^*)}_{\text{private value of an allocation}} - \underbrace{\beta(\varphi(\mathbf{b}^*) - \varphi(\mathbf{a}^*))}_{\text{shame}}.$$

This representation captures the tension between DM's impulse to choose the allocation she prefers in private and her desire to minimize shame. Each menu is evaluated 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, $u(\mathbf{a})$, gives the value of a degenerate menu that contains only the allocation under consideration. Since degenerate menus leave DM with a trivial choice under observation, DM's ranking over them can be thought of as her private ranking of allocations. The second component is "shame." It represents the cost DM incurs when selecting \mathbf{a} in the face of the fairest available alternative, \mathbf{b}^* . We assume DM to be more *selfish* in private: among equally fair alternatives she strictly prefers the one that gives her the highest private payoff.

As shame is evoked whenever the 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 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*, which 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 case of (1), the shame from choosing \mathbf{a} in stage two is $\beta(\varphi(\mathbf{b}^*) - \varphi(\mathbf{a}))$. 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, let $u(\mathbf{a}) = a_1$, $\beta = \frac{1}{2}$ and $\varphi(\mathbf{a}) = a_1 a_2$ 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 (1) 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 impose the independence axiom. Instead, our representation is based on choice over less complicated, risk-free objects. This domain is in line with that of our motivating examples. More importantly, imposing the independence axiom would be 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 each 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 (1), on the other hand, does feature a context-independent choice criterion. To see this, regroup the terms as follows:

$$U(A) = \underbrace{\max_{\mathbf{a} \in A} [u(\mathbf{a}) + \beta\varphi(\mathbf{a})]}_{\text{second stage choice criterion}} - \underbrace{\beta \max_{\mathbf{b} \in A} [\varphi(\mathbf{b})]}_{\text{effect of fairest alternative}} .$$

Theorem 2 identifies two additional assumptions on preferences over sets that are equivalent to the existence of such a choice criterion. Suppose that only second-stage choice can be observed (for example, because DM, as in the classic dictator game, never gets to choose between menus), but the experimenter is willing to assume that first-stage choice satisfies all our assumptions. If the choice criterion is context-independent, then the experimenter

⁶See section 5 for a context that naturally involves uncertainty and our suggestion for how to incorporate it into our model.

⁷In 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.

would not be able learn anything about DM's fairness ranking. Conversely, any violation of the weak axiom of revealed preferences in second-stage choice reveals at least part of the fairness ranking.

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, $\varphi(\mathbf{a}) = v_1(a_1)v_2(a_2)$. This particular form means that 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.

Example: Let $u(\mathbf{a}) = a_1$, $\varphi(\mathbf{a}) = v_1(a_1)v_2(a_2) = a_1a_2$ and $\beta = 0.5$. This implies that DM's private value of an allocation, u , is purely selfish in the sense that it depends only on her private payoffs, a_1 . The utilities v_1 and v_2 , 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 $A = \{(10, 0) (9, 1) (8, 2), \dots, (0, 10)\}$ then describes the dictator game. It induces the imaginary bargaining game with possible utility-allocations $\{(10, 0), (9, 1), (8, 2), \dots, (0, 10), (0, 0)\}$, where the imaginary disagreement point is $\lim_{(x,y) \rightarrow 0} (v_1^{-1}(x), v_2^{-1}(y)) = (0, 0)$. 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 $a_1 + a_1a_2$, which is $(6, 4)$. Its fairness is $6 \cdot 4 = 24$ and the shame from choosing it equals 0.5. 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 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, $A, B \in K$, as

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

where $d : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$ 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 captures the idea that shame is a mental cost, which is invoked by unchosen alternatives.

P_1 (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 The second part of the axiom requires that 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 .

Shame refers 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 1: 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

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

⁹The 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 $n-1$ other recipients, is given in the appendix.

¹⁰This 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.

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

$A \succ A \cup \{\mathbf{b}\}$ implies that \mathbf{b} adds to the shame incurred by the original choice in A . In this case \mathbf{b} is 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 . Because \succ_f is an induced binary relation, F axioms are implicit axioms on \succ ; \succ_f is only an expositional device.¹² Since the underlying notion of fairness is at the heart of F -axioms, making assumptions directly on \succ_f and motivating them in that context is natural.¹³

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. 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 many contexts, people would disagree with the statement that the allocation (10, 6) is fairer than (5, 5). In accordance with the notion of fairness that we adopt in this paper, we feel 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. Therefore, in cases where there is no potential for redistribution, we believe that people find the Pareto criterion a reasonable requirement for one allocation to be fairer than another.

¹²Our 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 15 for further details.

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

F_3 (**Compensation**) For all \mathbf{a} and \mathbf{b} , there exist 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 implies 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. 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 across individuals.

The last axiom in this section concerns DM's preferences over singleton sets. A singleton set is a degenerate menu that contains only one feasible allocation. Therefore, choosing between two singleton sets reveals DM's private preferences over allocations. Although we allow private preferences to regard the recipient's payoff, DM's norm of fairness always cares about it more. This difference is captured in the following single-crossing assumption:

F_2 (**Selfishness**) If $a_1 > b_1$ and $\mathbf{b} \not\succeq_f \mathbf{a}$, then $\{\mathbf{a}\} \succ \{\mathbf{b}\}$.

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 \mathbb{R}_+^2 .¹⁵

Definition 2: Consider two real functions g and f on \mathbb{R}_{++}^2 . We say that f is *less selfish than* g if for all $\mathbf{a} \in \mathbb{R}_{++}^2$ and for all $\Delta_1, \Delta_2 > 0$ such that $(a_1 - \Delta_1, a_2 + \Delta_2) \in \mathbb{R}_{++}^2$,

(i) $g(\mathbf{a}) = g(a_1 - \Delta_1, a_2 + \Delta_2)$ implies $f(\mathbf{a}) < f(a_1 - \Delta_1, a_2 + \Delta_2)$ and,

(ii) $g(\mathbf{a}) = g(a_1 + \Delta_1, a_2 - \Delta_2)$ implies $f(\mathbf{a}) > f(a_1 + \Delta_1, a_2 - \Delta_2)$.¹⁶

Definition 3: A function $\varphi : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ 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.

¹⁵To 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.

¹⁶If g and f were differentiable, then the condition would be $\frac{f_1}{f_2}(\mathbf{a}) < \frac{g_1}{g_2}(\mathbf{a})$ for all \mathbf{a} , where f_i denotes the partial derivative of f with respect to its i^{th} component. In this case, the definition of *less selfish than* coincides with the definition of more altruistic than in Cox *et al.* (2008).

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 $u : \mathbb{R}_+^2 \rightarrow \mathbb{R}$, a less selfish continuous fairness function φ , and a continuous function $g : \mathbb{R}_+^2 \times \varphi(\mathbb{R}_+^2) \rightarrow \mathbb{R}$, weakly increasing in its second argument and satisfying $g(\mathbf{a}, x) \geq 0$ whenever $\varphi(\mathbf{a}) \leq x$, such that the function $U : K \rightarrow \mathbb{R}$ defined as

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

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, $U : K \rightarrow \mathbb{R}$ and $\varphi : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ respectively. F_2 and F_3 imply that φ is a fairness function. The combination of *Strong Left Betweenness* (P_1) 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. 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 second component, the function g , in the representation.

Theorem 1 provides a representation of DM's fairness ranking 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 value 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 shame, $g(\mathbf{a}, \varphi(\mathbf{b}))$, 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 second stage choice may be recorded. For example, the standard dictator game corresponds only to second-stage choice in our setup. Suppose that the correspondence that governs second-stage choice is

$$C(A) := \left\{ \arg \max_{\mathbf{a} \in A} \left[u(\mathbf{a}) - g \left(\mathbf{a}, \max_{\mathbf{b} \in A} \varphi(\mathbf{b}) \right) \right] \right\} \quad (2)$$

A natural question is to what extent one can elicit DM's fairness ranking based solely on choice from menus. Consider the induced binary relation " \mathbf{b} alters choice in the face of \mathbf{a} " defined as $\mathbf{b} \succ_a \mathbf{a}$ if and only if there is A such that $\mathbf{a} \in A$, $\mathbf{b} \notin C(A \cup \{\mathbf{b}\})$ and $C(A) \neq C(A \cup \{\mathbf{b}\})$. It follows from (2) that if the addition of \mathbf{b} changes the choice from A , then $\mathbf{b} \in \arg \max_{\mathbf{b}' \in A \cup \{\mathbf{b}\}} \varphi(\mathbf{b}')$ and $\mathbf{b} \succ_a \mathbf{a}$ implies $\mathbf{b} \succ_f \mathbf{a}$. Therefore, a second-stage violation of the weak axiom of revealed preference (warp) partially identifies DM's fairness ranking. Furthermore, enough violations of warp would enable us to fully recover \succ_f .¹⁸ It follows that if there are no violations of warp, observing DM's choice of menus is necessary to elicit the fairness ranking of any two alternatives. In this section we show when this is the case, that is, we identify conditions under which DM's second-stage choice is independent of the choice set. Context independent choice implies the existence of a function $\psi : \mathbb{R}_+^2 \rightarrow \mathbb{R}$, such that \mathbf{a} is a choice from A only if $\psi(\mathbf{a}) \geq \psi(\mathbf{b})$ for all $\mathbf{b} \in A$.

¹⁷Consider 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.

¹⁸ \succ_f is uniquely identified from second stage choice, if and only if \succ_a is a continuous weak order that satisfies the Pareto property. To see this, note that (2) and completeness of \succ_a imply that if $\mathbf{b} \sim_f \mathbf{a}$ then $\mathbf{b} \sim_a \mathbf{a}$. Take any $\mathbf{b} \succ_f \mathbf{a}$. By F_2 there exist $\mathbf{c} > \mathbf{d}$ such that $\mathbf{c} \sim_f \mathbf{b} \succ_f \mathbf{a} \sim_f \mathbf{d}$. By Pareto, $\mathbf{c} \succ_a \mathbf{d}$ and by transitivity $\mathbf{b} \succ_a \mathbf{a}$. In the text we show that the other direction is also true, hence $\mathbf{b} \succ_f \mathbf{a} \Leftrightarrow \mathbf{b} \succ_a \mathbf{a}$.

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 $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$.¹⁹

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

¹⁹In the axioms, we use $\langle \mathbf{a}, \mathbf{b} \rangle$ to denote 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.

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 $u : \mathbb{R}_+ \rightarrow \mathbb{R}$ and a less selfish continuous fairness function φ , such that the function $U : K \rightarrow \mathbb{R}$ defined as

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

represents \succ and φ represents \succ_f .

The representation in Theorem 2 isolates a choice criterion that is independent of the choice problem: DM's behavior is governed by maximizing $\psi(\mathbf{a}) = u(\mathbf{a}) + \varphi(\mathbf{a})$. 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, $u(\mathbf{a})$ 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.

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

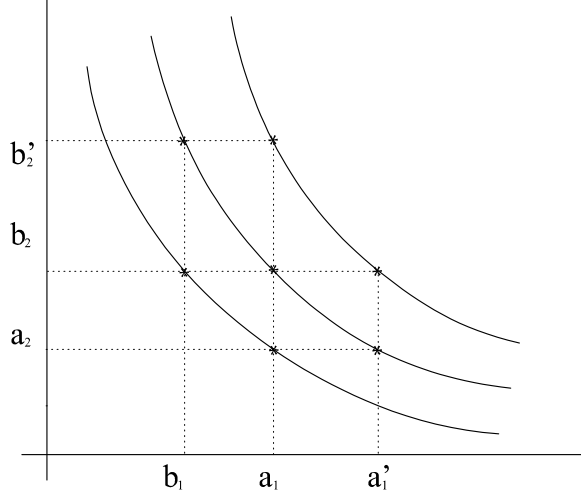


Figure 1: Independent Fairness Contributions.

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 $v_1, v_2 : \mathbb{R}_+ \rightarrow \mathbb{R}_{++}$, 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.²⁰ 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 $(a_1, \tilde{a}_2) \sim_f (a'_1, \tilde{a}'_2)$, then $(\tilde{a}_1, a_2) \sim_f (\tilde{a}'_1, a'_2) \Leftrightarrow (\tilde{a}_1, \tilde{a}_2) \sim_f (\tilde{a}'_1, \tilde{a}'_2)$. With this knowledge, we can create a monotone and increasing mapping $a_2 \rightarrow \gamma(a_2)$ 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.²¹ Quasi-linearity implies that there is an increasing continuous function $\xi : \mathbb{R}_+ \rightarrow \mathbb{R}$, 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 $v_1, v_2 : \mathbb{R}_+ \rightarrow \mathbb{R}_{++}$ 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 and one for the recipient. By mapping the alternatives within each 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, 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. Theorem 3 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. Further, any fairness

²⁰See Theorem 2 in Chapter 6 of Kranz et al (1971).

²¹For brevity, we do not reproduce their argument in more detail in this paper. A direct proof of Theorem 3 is available upon request.

²²See Nash (1950). The imaginary disagreement point is determined by $\lim_{(x,y) \rightarrow 0} (v_1^{-1}(x), v_2^{-1}(y))$. 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.

ranking must acknowledge 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 This criterion 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.

A more appealing interpretation of the NBS as the fairest allocation 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. For example, DM might ask what would be the allocation if both her 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. The intuitive and possibly descriptive appeal of the NBS in many bargaining situations then makes it normatively appealing to DM.²⁴ Theorem 3 establishes the behavioral equivalence of this interpretation and our axioms.

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.

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

²⁴The 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)).

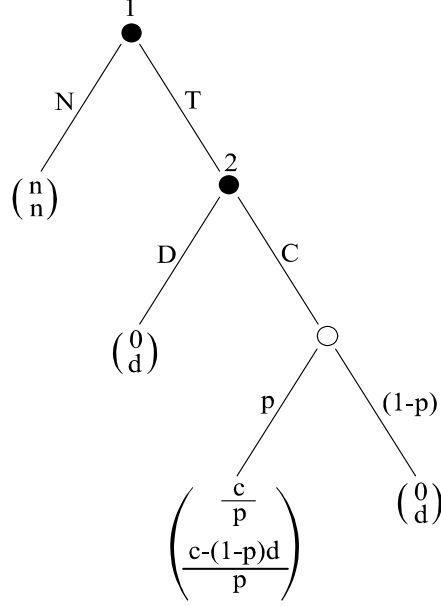


Figure 2: A game of trust with uncertain outcomes.

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.

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 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,

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

where $\tilde{\mathbf{a}} := E(a)$, the expected allocation generated by player 2's choice rule. 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 $\tilde{a}_2 + \beta \tilde{a}_1 \tilde{a}_2$. Player 1, on the other hand, cares only about her expected payoff, \tilde{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 assume that out of equilibrium beliefs attribute unexpected outcomes or actions to pure strategy deviations. The next proposition characterizes the equilibria of this game.

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 she 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 2.²⁵ 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 (2008) is motivated by the same experimental evidence. He also considers menus of allocations as objects of choice. Neilson does not axiomatize a representation result, but qualitatively 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:

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

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, $A \succ B \Rightarrow A \succ A \cup B \succ B$. We show that our axioms, *Strong Left Betweenness* (P_1) 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.

²⁵Neilson (2006) axiomatizes a reference-dependent preference, that can be interpreted in terms of Fehr and Schmidt's objective function.

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 growing 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."²⁶ 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 $N - 1 \geq 2$ 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.

²⁶Surely the opposite is also conceivable: Subjects might be particularly keen to be selfless when the experimenter observes their behavior. This example is just ment to draw attention to the difficulties faced by experimenters in the context of norms.

A typical alternative $\mathbf{a} = (a_1, a_2, \dots, a_N)$ 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**) For all \mathbf{b} , \mathbf{a} and m , there exists a_m such that $(a_m, \mathbf{a}_{-m}) \succ_f \mathbf{b}$.

This property requires the fairness ranking never to be satiated in any individual payoff.

Definition 4: 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 $m, n \in \{1, \dots, N\}$, 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 $u : \mathbb{R}_+^N \rightarrow \mathbb{R}$ and $\varphi : \mathbb{R}_+^N \rightarrow \mathbb{R}$ such that the function $U : K \rightarrow \mathbb{R}$ defined as $U(A) = \max_{\mathbf{a} \in A} [u(\mathbf{a}) + \varphi(\mathbf{a})] - \max_{\mathbf{b} \in A} [\varphi(\mathbf{b})]$ 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 $v_1, \dots, v_N : \mathbb{R}_+ \rightarrow \mathbb{R}_{++}$, where v_1, \dots, 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 $U : K \rightarrow \mathbb{R}$ be a continuous function that represents \succ . Define $u(\mathbf{a}) \equiv U(\{\mathbf{a}\})$.

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 $\varphi : \mathbb{R}_{++}^2 \rightarrow \mathbb{R}$ be a continuous function that represents \succ_f . By F_2 , φ is strictly increasing and by P_2 , u is strictly increasing as well.

Because \succ_f is continuous, F_3 immediately implies that if $(a_1, a_2) \not\sim_f (b_1, b_2)$, 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 $A = \{\mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^N\}$ and $C_0 = B$. Define $C_n = C_{n-1} \cup \{\mathbf{a}^n\}$ for $n = 1, 2, \dots, N$. According to F_1 , for all \mathbf{a}^n there exists $\mathbf{b} \in B$ such that $\mathbf{a}^n \not\sim_f \mathbf{b}$. By P_1 , $C_{n-1} \not\sim_f C_n$. By negative transitivity of \succ , $C_0 \not\sim_f C_N$ 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 $B = \{\mathbf{b}^1, \mathbf{b}^2, \dots, \mathbf{b}^M\}$. Define $C_0 = A$ and $C_m = C_{m-1} \cup \{\mathbf{b}^m\}$ for $m = 1, 2, \dots, M$. By P_1 , $\forall C$ such that $\mathbf{a} \in C, C \not\sim_f C \cup \{\mathbf{b}^m\}$. Hence, $C_{m-1} \not\sim_f C_m$. By negative transitivity of \succ , $C_0 \not\sim_f C_M$ or $A \cup B \succeq A \succeq B$, hence $A \cup B \succeq B$.||

Combining Claim 1.1 with P_1 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 $\max_{\mathbf{a}' \in A} \min_{\mathbf{b}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$ and (\mathbf{b}, \mathbf{a}) solves $\min_{\mathbf{b}' \in A} \max_{\mathbf{a}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$.*

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

$$U(\{\mathbf{a}, \mathbf{b}\}) = \max_{\mathbf{a}' \in \{\mathbf{a}, \mathbf{b}\}} \min_{\mathbf{b}' \in \{\mathbf{a}, \mathbf{b}\}} \tilde{U}(\mathbf{a}', \mathbf{b}') = \min_{\mathbf{b}' \in \{\mathbf{a}, \mathbf{b}\}} \max_{\mathbf{a}' \in \{\mathbf{a}, \mathbf{b}\}} \tilde{U}(\mathbf{a}', \mathbf{b}').^{27}$$

²⁷Note that $\max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} U(\{\mathbf{a}, \mathbf{b}\}) = \max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} \left[\max_{\mathbf{a}' \in \{\mathbf{a}, \mathbf{b}\}} \min_{\mathbf{b}' \in \{\mathbf{a}, \mathbf{b}\}} \tilde{U}(\mathbf{a}', \mathbf{b}') \right] = \max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} \tilde{U}(\mathbf{a}, \mathbf{b})$.

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:

$$U(\{\mathbf{a}, \mathbf{b}\}) = \max \left\{ \begin{array}{l} u(\mathbf{a}) - \max_{f(\mathbf{a}, \mathbf{a})=0} \{f(\mathbf{a}, \mathbf{b})\} \\ u(\mathbf{b}) - \max_{f(\mathbf{b}, \mathbf{b})=0} \{f(\mathbf{b}, \mathbf{a})\} \end{array} \right\} \geq u(\mathbf{a}) - \max \left\{ \begin{array}{l} f(\mathbf{a}, \mathbf{a}) = 0 \\ f(\mathbf{a}, \mathbf{b}) \end{array} \right\} = U(\{\mathbf{a}\}).$$

- Claim 1.2:** (i) $[\varphi(\mathbf{a}) < \varphi(\mathbf{b}) \text{ and } \{\mathbf{a}\} \succ \{\mathbf{b}\}] \Leftrightarrow \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$
(ii) $[\varphi(\mathbf{a}) < \varphi(\mathbf{b}) \text{ and } \{\mathbf{b}\} \succeq \{\mathbf{a}\}] \Rightarrow \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\}$
(iii) $[\varphi(\mathbf{a}) = \varphi(\mathbf{b}) \text{ and } \{\mathbf{a}\} \succ \{\mathbf{b}\}] \Rightarrow \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$.

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 $\{\mathbf{a}\} \succ \{\mathbf{b}\}$, by P_1 $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$. Conversely if $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$, then $\mathbf{b} \succ_f \mathbf{a}$ and hence $\varphi(\mathbf{a}) < \varphi(\mathbf{b})$.

(ii) If $\{\mathbf{b}\} \succeq \{\mathbf{a}\}$ 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) $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ implies 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,

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

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) $\mathbf{b} \in Y(\mathbf{a}, \varphi)$, $\{\mathbf{b}\} \succ \{\mathbf{b}'\}$ and $\mathbf{b}' \sim_f \mathbf{b}$ 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 $\mathbf{b}' \sim_f \mathbf{b}$, 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 contradiction. 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 exists a function \tilde{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 $g : \mathbb{R}_+^2 \times \mathbb{R} \rightarrow \mathbb{R}$ which is weakly increasing in its second argument.

Proof: Such \tilde{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 $\mathbf{a} = 0$, 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 $\mathbf{a} > 0$ and $\varphi(\mathbf{b}) > \varphi(\mathbf{c}) > \varphi(\mathbf{a})$.

If there is no $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$ with $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\}$, then $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c}) \geq 0$ is consistent

with \succ . If there exists $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$ such that $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\}$, then 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 .) 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 $\{\mathbf{b}'\} \succ \{\mathbf{c}'\}$ by Axiom P_2 . Claim 1.2 (i) and observation (1) imply $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}'\} \succeq \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\} \succ \{\mathbf{c}'\}$. If $\{\mathbf{a}, \mathbf{c}\} \succeq \{\mathbf{a}, \mathbf{b}\}$ then $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$ is consistent with \succ . If $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{c}\}$, then observation (4) implies that $\{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{b}\}$ and hence $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$ is consistent with \succ .

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

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 $g(\mathbf{a}, \varphi) = u(\mathbf{a}) - U(\{\mathbf{a}, \mathbf{b}\})$ 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 $g(\mathbf{a}, \varphi) \geq u(\mathbf{a}) - u(\mathbf{0})$. 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 $\varphi(\widehat{\mathbf{b}}) = \widehat{\varphi}$. For $\widehat{\varphi} = \varphi(\widehat{\mathbf{a}})$ define $g(\widehat{\mathbf{a}}, \widehat{\varphi}) := 0$ and for $\widehat{\varphi} > \varphi(\widehat{\mathbf{a}})$ define $g(\widehat{\mathbf{a}}, \widehat{\varphi}) := u(\widehat{\mathbf{a}}) - \inf_{\mathbf{b} \in I_f(\widehat{\varphi})} u(\mathbf{b})$.

Consider a sequence $\{(\mathbf{a}^n, \varphi^n)\} \rightarrow (\widehat{\mathbf{a}}, \widehat{\varphi})$ in S . Pick a sequence $\{\mathbf{b}^{n'}\}$ with $\mathbf{b}^{n'} \in Y(\mathbf{a}^n, \varphi^n) \forall n$. Define $\{b_1^n\} = \left\{ \min \left[\frac{1}{n}, b_1^{n'}, \widehat{b}_1 \right] \right\}$. 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) and P_2 , $\mathbf{b}^n = (b_1^n, b_2^n) \in Y(\mathbf{a}^n, \varphi^n)$. Lastly, let $\widehat{\mathbf{b}}^n$ solve $\widehat{b}_1^n = b_1^n$ and $\varphi(\widehat{\mathbf{b}}^n) = \widehat{\varphi}$. We have $U(\{\mathbf{a}^n, \mathbf{b}^n\}) = u(\mathbf{a}^n) - g(\mathbf{a}^n, \varphi^n)$. If in the switch point $\widehat{\varphi} = \varphi(\widehat{\mathbf{a}})$, then $U(\{\widehat{\mathbf{a}}, \widehat{\mathbf{b}}^n\}) = u(\widehat{\mathbf{a}})$. By continuity, $U(\{\mathbf{a}^n, \mathbf{b}^n\}) - U(\{\widehat{\mathbf{a}}, \widehat{\mathbf{b}}^n\}) \xrightarrow{n \rightarrow \infty} 0$, hence

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

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

continuity argument

$$\lim_{n \rightarrow \infty} g(\mathbf{a}^n, \varphi^n) = \lim_{n \rightarrow \infty} \left[u(\mathbf{a}^n) - u(\widehat{\mathbf{b}}^n) \right] = u(\widehat{a}_1) - \inf_{\mathbf{b} \in I_f(\widehat{\varphi})} u(\mathbf{b}) = g(\widehat{\mathbf{a}}, \widehat{\varphi}).$$

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 $\{(\mathbf{a}^n, \varphi^n)\} \rightarrow (\widehat{\mathbf{a}}, \widehat{\varphi})$, we have $\lim_{n \rightarrow \infty} g(\mathbf{a}^n, \varphi^n) = 0$.||

This establishes the existence of a continuous representation

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

of \succ on K_{++} with the properties as 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 Krantz *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 , $(a_1, a) \succeq_f (b_1, a) \Leftrightarrow a_1 \geq b_1 \Leftrightarrow (a_1, b) \succeq_f (b_1, b)$ 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 $a \neq b$, $(a_i, a) \sim_f (a_{i+1}, b)$ for all i . A

standard sequence is bounded if there exist \underline{a} and \bar{a} such that for all i , $a_i \in (\underline{a}, \bar{a})$. 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 \bar{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 $(\bar{a}, a) \sim_f (x, b)$. By F_2 , $x > \bar{a}$. Since $\{a_i\}$ is an increasing and bounded sequence, it must converge to its least upper bound, \bar{a} . By continuity, there exists a sub-sequence $\{a_{ik}\}$ that converges to x . In particular, there exists K such that for $k > K$, $x - a_{ik} < \varepsilon := \frac{x - \bar{a}}{2}$, 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_2 implies that $(\underline{a}, b) \prec_f (a_{i+1}, b)$. Therefore, $(a_i, a) \prec_f (a_{i+1}, b)$, contradiction.

Essentiality: Not $(a', b) \sim_f (a, b)$ for all b , or $(a, b) \sim_f (a, b')$ 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 $\varphi : \mathbb{R}_+^N \rightarrow \mathbb{R}$ 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 $\bar{\varphi} := \sup_{\mathbf{a} \in \mathbb{R}_{++}^N} \varphi(\mathbf{a})$ and $\underline{\varphi} := \inf_{\mathbf{a} \in \mathbb{R}_{++}^N} \varphi(\mathbf{a})$, if they are well defined. Otherwise, take $\bar{\varphi} = \infty$ and $\underline{\varphi} = -\infty$. Define $U_S : \mathbb{R}_{++}^N \times (\underline{\varphi}, \bar{\varphi}) \rightarrow \mathbb{R}$.

Consider first $(\mathbf{a}, \varphi) \in S$, where as before $S := \{(\mathbf{a}, \varphi) : Y(\mathbf{a}, \varphi) \neq \emptyset\}$, and let $U_S(\mathbf{a}, \varphi) := U(\{\mathbf{a}, \mathbf{b}\})$ for some $\mathbf{b} \in Y(\mathbf{a}, \varphi)$. Let \succ_S be a binary relation on S defined by $(\mathbf{a}, \varphi) \succ_S (\tilde{\mathbf{a}}, \tilde{\varphi}) \Leftrightarrow \{\mathbf{a}, \mathbf{b}\} \succ \{\tilde{\mathbf{a}}, \tilde{\mathbf{b}}\} \forall \mathbf{b} \in Y(\mathbf{a}, \varphi)$ and $\forall \tilde{\mathbf{b}} \in Y(\tilde{\mathbf{a}}, \tilde{\varphi})$. Let $U_S(\mathbf{a}, \varphi) := U(\{\mathbf{a}, \mathbf{b}\})$ for some $\mathbf{b} \in Y(\mathbf{a}, \varphi)$, where, according to the representation analogous to Theorem 1, $U(\{\mathbf{a}, \mathbf{b}\}) = u(\mathbf{a}) - g(\mathbf{a}, \varphi) > \inf_{\mathbf{b} \in I_f(\varphi)} u(\mathbf{b})$. By Theorem 1, \succ_S is a weak order that can be represented by U_S . Note that *Consistency* (P_4) is relevant precisely on this domain.

For $(\mathbf{a}, \varphi) \notin S$ define

$$U_S(\mathbf{a}, \varphi) := \begin{cases} \inf_{\mathbf{b} \in I_f(\varphi)} u(\mathbf{b}) & \text{for } \varphi(\mathbf{a}) < \varphi \\ u(\mathbf{a}) & \text{for } \varphi(\mathbf{a}) \geq \varphi \end{cases}$$

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) $\varphi(\mathbf{a}) \geq \varphi$: Take $(\mathbf{a}, \varphi) \in bdr(S)$. Since (\mathbf{a}, φ) 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}, φ) . By the definition of S , $U_s(\mathbf{a}^n, \varphi^n) = u(\mathbf{a}^n) - g(\mathbf{a}^n, \varphi^n)$. Because preferences are continuous and using the properties of g from Theorem 1, we have $\lim_{n \rightarrow \infty} u(\mathbf{a}^n) - g(\mathbf{a}^n, \varphi^n) = u(\mathbf{a})$ 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}, φ) . By the definition of S ,

$$U_s(\mathbf{a}^n, \varphi^n) = u(\mathbf{a}^n) - g(\mathbf{a}^n, \varphi^n) > \inf_{\mathbf{b} \in I_f(\varphi^n)} u(\mathbf{b}).$$

Since \succ is continuous, we have

$$\lim_{n \rightarrow \infty} u(\mathbf{a}^n) - g(\mathbf{a}^n, \varphi^n) = u(\mathbf{a}) - g(\mathbf{a}, \varphi) \geq \inf_{\mathbf{b} \in I_f(\varphi)} u(\mathbf{b}).$$

As $(\mathbf{a}, \varphi) \notin S$, we claim that

$$u(\mathbf{a}) - g(\mathbf{a}, \varphi) \leq \inf_{\mathbf{b} \in I_f(\varphi)} u(\mathbf{b}).$$

If not, then there exists $\mathbf{c} \in I_f(\varphi)$ with $u(\mathbf{a}) - g(\mathbf{a}, \varphi) = u(\mathbf{c}) > \inf_{\mathbf{b} \in I_f(\varphi)} u(\mathbf{b})$. But using F_3^N , we could find $\mathbf{c}' \in I_f(\varphi)$ with $u(\mathbf{c}') < u(\mathbf{c})$. Using Theorem 1, this would imply that $(\mathbf{a}, \varphi) \in S$, which is a contradiction. Combining we have $\lim_{n \rightarrow \infty} u(\mathbf{a}^n) - g(\mathbf{a}^n, \varphi^n) = \inf_{\mathbf{b} \in I_f(\varphi)} u(\mathbf{b})$, as required. ||

Definition 5: For $(\mathbf{a}, \varphi) \in S$, define $I_S(\mathbf{a}, \varphi) := \{(\mathbf{a}', \varphi') \in S : (\mathbf{a}', \varphi') \sim_S (\mathbf{a}, \varphi)\}$. That is, $I_S(\mathbf{a}, \varphi)$ is the \succ_S equivalence class of (\mathbf{a}, φ) .

Claim 4.2: Shame $g(\mathbf{a}, \varphi)$ 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 $\varphi(\mathbf{a}'') = \varphi$ the definition of U_S yields $U_S(\mathbf{a}'', \varphi) > U_S(\mathbf{a}'', \varphi')$. Then by P_4 , $(\mathbf{a}', \varphi''') \succ_S (\mathbf{a}', \varphi'')$, a contradiction. \parallel

Definition 6: Let

$$\varphi_{\min}(\mathbf{a}) := \inf \{ \varphi \mid \exists \hat{\mathbf{a}} \text{ with } u(\hat{\mathbf{a}}) = u(\mathbf{a}) \text{ and } \varphi(\hat{\mathbf{a}}) = \varphi \}$$

and

$$\varphi_{\max}(\mathbf{a}) := \sup \{ \varphi \mid \exists \hat{\mathbf{a}} \text{ with } u(\hat{\mathbf{a}}) = u(\mathbf{a}) \text{ and } \varphi(\hat{\mathbf{a}}) = \varphi \}$$

Claim 4.3: For all $(\mathbf{a}, \varphi) \in S$ and $\tilde{\varphi} \in (\varphi_{\min}(\mathbf{a}), \varphi_{\max}(\mathbf{a}))$ there exists $\tilde{\mathbf{a}}$ such that $(\tilde{\mathbf{a}}, \tilde{\varphi}) \in I_S(\mathbf{a}, \varphi)$.

Proof: Given \mathbf{a} , let $\underline{\mathbf{a}}$ be such that $\varphi(\underline{\mathbf{a}}) < \varphi(\mathbf{a})$ and $u(\underline{\mathbf{a}}) = u(\mathbf{a})$. Define φ^* implicitly by $U_S(\underline{\mathbf{a}}, \varphi^*) = U_S(\mathbf{a}, \varphi)$. This is possible by the intermediate value theorem, as $U_S(\underline{\mathbf{a}}, \varphi(\underline{\mathbf{a}})) = u(\underline{\mathbf{a}}) > U_s(\mathbf{a}, \varphi) > U_s(\underline{\mathbf{a}}, \varphi)$, where the last inequality is due to P_3 . There are two cases to consider:

Case 1) $\tilde{\varphi} \geq \varphi^*$: Then $U_S(\underline{\mathbf{a}}, \tilde{\varphi}) \leq U_S(\mathbf{a}, \varphi)$ according to Claim 4.2. Because $\tilde{\varphi} \in (\varphi_{\min}(\mathbf{a}), \varphi_{\max}(\mathbf{a}))$ and by F_3^N there is $\hat{\mathbf{a}}$ that solves $\varphi(\hat{\mathbf{a}}) = \tilde{\varphi}$ and $u(\hat{\mathbf{a}}) = u(\mathbf{a})$. Then $U_s(\hat{\mathbf{a}}, \tilde{\varphi}) \geq U_s(\mathbf{a}, \varphi)$ and by the intermediate value theorem there is $\tilde{\mathbf{a}}$ such that $U_S(\tilde{\mathbf{a}}, \tilde{\varphi}) = U_S(\mathbf{a}, \varphi)$.

Case 2) $\tilde{\varphi} < \varphi^*$: Then by Claim 4.2 $U_S(\mathbf{a}, \varphi) < U_S(\underline{\mathbf{a}}, \tilde{\varphi})$. Let \mathbf{a}^* be a solution to $u(\mathbf{a}^*) = u(\mathbf{a}) - g(\mathbf{a}, \varphi) = U_S(\mathbf{a}, \varphi)$ and $\mathbf{a}^* = \kappa \mathbf{a}$ for some $\kappa < 1$. Then $U_S(\mathbf{a}^*, \tilde{\varphi}) \leq U_S(\mathbf{a}, \varphi)$. By the intermediate value theorem there is $\tilde{\mathbf{a}}$ such that $U_s(\tilde{\mathbf{a}}, \tilde{\varphi}) = U_s(\mathbf{a}, \varphi)$. \parallel

The two cases imply that, given \mathbf{a} and φ , we can apply the intermediate value theorem to define a continuous path of allocations $\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi})$, such that $\tilde{\mathbf{a}} = \tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi})$ satisfies $(\tilde{\mathbf{a}}, \tilde{\varphi}) \in I_S(\mathbf{a}, \varphi)$, and

- (1) For $\tilde{\varphi} \geq \varphi^*$, $u(\tilde{\mathbf{a}}) = u(\mathbf{a})$ and $\tilde{a}_1 \in (\underline{a}_1, \hat{a}_1)$.
- (2) For $\tilde{\varphi} < \varphi^*$, $\tilde{\mathbf{a}} = \lambda \mathbf{a}$ with $\lambda \in (\kappa, 1)$.

According to Claim 4.2 and F_2 , $\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi})$ is strictly increasing for $\tilde{\varphi} < \varphi^*$, while $u(\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi}))$ is constant and $\varphi(\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi}))$ is strictly increasing for $\tilde{\varphi} \geq \varphi^*$.

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

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

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

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

Proof: Assume $\tilde{\varphi}' > \tilde{\varphi}$. Whether or not $\tilde{\varphi}' > \varphi^*$ it must be that $u(\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi}')) \geq u(\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi}))$ and $\varphi(\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi}')) > \varphi(\tilde{\mathbf{a}}_{(\mathbf{a}, \varphi)}(\tilde{\varphi}))$. P_3 then implies

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

Because of Claim 4.2, we must have $\widetilde{\varphi + d\varphi_{(\mathbf{a}, \varphi)}}(\tilde{\varphi}') > \widetilde{\varphi + d\varphi_{(\mathbf{a}, \varphi)}}(\tilde{\varphi})$.

The \succ_S -indifference curve through (\mathbf{a}, φ) 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 $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing and continuous. Due to P_4 the definition of $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ is independent of the choice of (\mathbf{a}, φ) . Thus, $\varphi \mapsto \gamma(\varphi)$ transform the original indifference map of $U_S(\mathbf{a}, \varphi)$ to be quasi-linear in $\gamma(\varphi)$.²⁸

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

Define $u(\mathbf{a}) := H(\mathbf{a}) - \lim_{\varphi \rightarrow \varphi(\mathbf{a})} \gamma(\varphi) = H(\mathbf{a}) - \gamma(\varphi(\mathbf{a}))$. Then

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

represents \succ confined to the collection of all two element sets in K_{++} . Hence

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

represents \succ on K_{++} , where φ represents \succ_f , and u and γ are strictly increasing. Since φ

²⁸The proof that this rescaling exists is again an application of the lock-step procedure explained in Keeney and Raiffa (1976). For brevity we do not reproduce it here.

represents \succ_f , so does $\gamma(\varphi)$. Hence, there is a representation φ of \succ_f , such that γ is the identity and

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

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, \dots, v_N , such that the fairness ranking \succ_f can be represented by $\varphi(\mathbf{a}) = v_1(a_1) \cdot \dots \cdot v_N(a_n)$, 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 $\xi_1(a_1) + \dots + \xi_N(a_N)$ of \succ_f . Define $v_n(a_n) := \exp(\xi_n(a_n))$ for all $n \in \{1, \dots, N\}$. Then $v_1, \dots, v_N : \mathbb{R}_+ \rightarrow \mathbb{R}_{++}$ are increasing and continuous and if we re-define $\varphi(\mathbf{a}) := v_1(a_1) \cdot \dots \cdot v_N(a_N)$, it represents \succ_f . By F_3^N , the functions v_1, \dots, v_N must be unbounded.

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

Proof of Proposition

Suppose that in equilibrium, player 2 is required to mix between D and C with respective weights q and $(1 - q)$. The perceived fairness of her strategy is then independent of her action (in the observed case) or the realized outcome (in the unobserved case). Therefore, it is profitable for player 2 to deviate and to choose $q = 0$, guaranteeing herself higher private payoffs without additional shame. Consequently, $q \in \{0, 1\}$ must hold in any equilibrium.

(i) The (expected) fairness generated by the mixed strategy $\langle q, D; (1 - q), C \rangle$ is $(qd + (1 - q)c)(1 - q)c$ and this expression is maximized at $q^* = \max\left\{\frac{d-2c}{2(d-c)}, 0\right\}$. Let $\varphi(\tilde{\mathbf{b}}) := (q^*d + (1 - q^*)c)(1 - q^*)c$. Trusted player 2 can either choose C , which generates direct utility c and shame $\beta\left(\varphi(\tilde{\mathbf{b}}) - c^2\right)$, or D , which generates direct utility d and shame $\beta\varphi(\tilde{\mathbf{b}})$. Hence, under the assumption on out-of-equilibrium beliefs, player 2 cooperates if $c + \beta c^2 > d$, 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 $\tilde{\mathbf{a}} = (c, c)$ 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 $\tilde{\mathbf{a}} = (0, d)$. Playing D , therefore, generates direct utility d and carries shame $\beta\varphi(\tilde{\mathbf{b}})$. If, however, player 1 observes $\left(\frac{c}{p}, \frac{c-(1-p)d}{p}\right)$, then, by the assumption on out of equilibrium beliefs, this can only be explained by player 2 having deviated from D to C ,²⁹ and accordingly $\tilde{\mathbf{a}} = (c, c)$.³⁰ If player 2 plays C she receives direct expected utility c and with probability p the shame is $\beta\varphi(\tilde{\mathbf{b}}) - \beta c^2$, while with probability $(1-p)$ shame is $\beta\varphi(\tilde{\mathbf{b}})$. Hence, player 2 is willing to play D in equilibrium, if and only if $d \geq c + p\beta c^2$ 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] Charness, Gary and Mathew Rabin (2002) "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 117 (3), 817-870.

²⁹This 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.

³⁰More precisely, $\tilde{\mathbf{a}}$ 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.

- [9] Cox, James C., Daniel Friedman, and Vjollca Sadiraj (2008) "Revealed Altruism." *Econometrica*, 76, 31 - 69.
- [10] 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.
- [11] Davis, Douglas D. and Charles A. Holt (1993) " Experimental economics." Princeton University Press.
- [12] Dekel, Eddie, Barton L. Lipman and Aldo Rustichini (2008) "Temptation Driven Preferences." *Review of Economic Studies*, forthcoming
- [13] Epstein, Larry G. and Igor Kopylov (2007) "Cold Feet" *Theoretical Economics*, Volume 2, 231–259
- [14] Fehr, Ernst and Klaus M. Schimdt (1999) "A Theory of Fairness, Competition and Co-operation." *Quarterly Journal of Economics*, 114, 817-868.
- [15] Frohlich, Norman, Joe Oppenheimer and J. Bernard Moore (2001) "Some doubts measuring self-interest using dictator experiments: The cost of anonymity." *Journal of Economic Behavior and Organization*, 46, 271-290.
- [16] Gauthier, David and Robert Sugden (editors) (June 1993) "Rationality, Justice and Social Contract: Themes from *Morals by Agreement*." University of Michigan Press.
- [17] Gul, Faruk and Wolfgang Pesendorfer (2005) "The Simple Theory of Temptation and Self-Control." mimeo
- [18] ——— (2001) "Temptation and Self Control." *Econometrica*, Vol. 69, No. 6, 1403-1435
- [19] 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.
- [20] 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.
- [21] Johannesson, Magnus and Bjorn Persson (2000) "Non-reciprocal altruism in dictator games." *Economics Letters*, 69, 137-142.

- [22] 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.
- [23] 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.
- [24] Koch, K. Alexander and Hans-Theo Norman (2005) "Giving in Dictator Games: Regard *for* Others or Regard *by* others?" mimeo.
- [25] Krantz, David H., R. Duncan Luce, Patrick C. Suppes and Amos Tversky (1971) "Foundations of Measurements, Vol 1." Academic Press, New York.
- [26] Lazear, Edward P., Ulrike Malmendier and Roberto A. Weber (2005) "Sorting in Experiments with Application to Social Preferences." mimeo.
- [27] Levitt, Steven D and John A. List (2007) "What do laboratory experiments measuring social preferences reveal about the real world?" *Journal of Economic Perspectives*, 21(2): 153–174
- [28] Luce, R. Duncan and John W. Tukey (1964) "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology*, 1,1-27.
- [29] Mariotti, Marco (1999) "Fair Bargains: Distributive Justice and Nash Bargaining Theory." *Review of Economic Studies*, Vol.66, 733-41.
- [30] Miller, Dale T. (1999) "The Norm of Self-Interest." *American Psychologist*, Vol. 54, No. 12, 1053-1060.
- [31] Nash, John F. (1953) "Two-Person Cooperative Games." *Econometrica*, Vol. 21, No. 1, 128-140.
- [32] ——— (1950) "The Bargaining Problem." *Econometrica*, Vol. 18, No. 2, 155-162.
- [33] Neilson, William S. (2006) "Axiomatic Reference Dependence in Behavior towards Others and Toward Risk." *Economic Theory* 28, 681-692.
- [34] ——— (2008) "A Theory of Kindness, Reluctance, and Shame in Dictator Games." *Games and Economic Behavior*, forthcoming .
- [35] Noor, Jawwad and Norio Takeoka (2008) "Menu-Dependent Self-Control" mimeo.

- [36] 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>
- [37] Osborne, Martin J. and Ariel Rubinstein (1994) "A course in Game Theory." MIT press, ISBN 0-262-65040-1.
- [38] Olszewski Wojciech (2008) "A Model of Consumption-Dependent Temptation" mimeo
- [39] Pillutla, Madan M. and J. Keith Murningham (1995) "Being fair or appearing fair: Strategic behavior in ultimatum bargaining." *Academy of Management Journal*, 38,1408-1426.
- [40] Rawls, John (1971) "A Theory of Justice." *The Belknap Press of Harvard University Press*.
- [41] Tadelis, Steve (2008) "The Power of Shame and the Rationality of Trust." mimeo