

Ashamed to be Selfish

David Dillenberger Philipp Sadowski

University of Pennsylvania Duke University

September 20, 2010
ERID Working Paper Number 84

This paper can be downloaded without charge from
The Social Science Research Network Electronic Paper Collection:
<http://ssrn.com/abstract=1701946>

**Economic Research Initiatives at Duke
WORKING PAPERS SERIES**



Ashamed to be Selfish*

David Dillenberger[†] Philipp Sadowski[‡]

September 20, 2010

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 normatively best allocation in the second stage inflicts shame on DM. We derive a representation that identifies DM's private ranking of allocations, her subjective norm, and shame. The normatively best allocation 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 notion of other-regarding preferences has attracted the attention of economists in different contexts. The relevance of this motive to decision making is intuitive and has been studied extensively. 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 with a concern for their own. Thus, preferences for fairness as well as altruistic preferences 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. A coeditor and two anonymous referees provided valuable comments that improved the paper significantly. 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 one 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 then (\$10, \$0) would be preferred to (\$9, \$0). Instead, people seem to suffer from behaving selfishly in a choice situation where they could behave pro-socially. 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 a person's actions are observed plays a crucial role in determining her behavior. We term *shame* the motive that distinguishes choice behavior that is observed from choice behavior that is not observed.¹ We identify shame as the moral cost that an individual experiences if, instead of choosing an alternative that she perceives to be in accordance with a social norm (which might include, but is not limited to, considerations of fairness and altruism), she is observed choosing an alternative that favors her own material payoffs. We refer to the criterion that determines whether an alternative causes shame when it is not chosen as the individual's (subjective) norm.

Since at least part of the choice behavior is observed by someone (for example the experimenter) it would be impossible to provide a complete theory of observation-driven shame that is based solely on revealed preferences. We thus study the effect of the differential shame from being observed by the recipient in addition to the experimenter. Correspondingly, we say that a person is observed if her choice behavior is revealed to someone who is directly affected by it. Otherwise, we say that she chooses in private. The significance of our notion of shame is supported by further evidence. In a follow-up to the experiment cited above,

¹Our notion of shame is closely related to that of embarrassment. We use the word shame to highlight that the emotional cost experienced by DM is caused by her own actions and is not triggered by others' actions, 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." The public-private distinction between guilt and shame is also suggested by Gehm and Scherer (1988). We adopt the interpretation that even observation of a selfish behavior without identification of its purveyor can cause shame. It is worth noting that in the psychological literature one can find other criteria for the distinction between guilt and shame. For example, Lewis (1971) suggests that shame focuses on self (what we are) whereas guilt focuses on behavior (what we do). A comprehensive discussion of guilt and shame can be found in Tangney and Dearing (2002).

Dana *et al.* report that only one out of twenty-four dictators exits the game when they are told that the recipient will never learn that a dictator game has been played. Similarly, Pillutla and Murningham (1995) find evidence that even if their identity is not revealed to the recipient, people’s giving behavior depends on the information given to the recipient regarding the source of the payoffs. 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.

To 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 function 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.^{2,3} DM has preferences over sets of alternatives (menus). Shame impacts preferences through its anticipated effect on second-stage choices, where the presence of a normatively better option reduces the attractiveness of an allocation. Our representation results demonstrate how DM’s norm and her choice behavior interact. On the one hand, properties of the norm impact choice; on the other hand, the norm can be elicited from DM’s choice behavior.

²The 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.

³If 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 two people participated in the experiment and that the 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 by $A = \{\mathbf{a}, \mathbf{b}, \dots\} = \{(a_1, a_2), (b_1, b_2), \dots\}$, where the first and second components of each alternative are, respectively, the private payoff for DM and for the recipient. To illustrate our results, consider the following special case of our representations:

$$U(A) = \max_{\mathbf{a} \in A} \left[\underbrace{u(\mathbf{a})}_{\text{private value of an allocation}} - \underbrace{g\left(\varphi(\mathbf{a}), \max_{\mathbf{b} \in A} \varphi(\mathbf{b})\right)}_{\text{shame}} \right] \quad (1)$$

where u and φ are increasing in all arguments. u is a utility function over allocations, and φ represents DM's norm. The function g is shame. It is decreasing in its first argument and increasing in its second argument.

This representation captures the tension between DM's impulse to choose the allocation she prefers in private and her desire to minimize shame. The value of a menu is the sum of two distinct components. The first component, $u(\mathbf{a})$, gives the value of the degenerate menu $\{\mathbf{a}\}$. Since degenerate menus leave DM with no choice to be made 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 one of the normatively best available alternatives, $\mathbf{b}^* \in \arg \max_{\mathbf{b} \in A} \varphi(\mathbf{b})$. Although the private ranking over allocations might incorporate other-regarding preferences (such as altruism), we assume DM to be more *selfish* in private: among normatively equally good alternatives, she strictly prefers the one that gives her the highest private payoff.

According to our interpretation of shame, we can relate choice to a second, induced binary relation, \succ_n , which captures DM's subjective norm. \succ_n is assumed to satisfy the following three properties: *Ranking*, which says that \succ_n is weak order; the *Pareto* criterion on payoffs; and *Compensation*, which requires the norm to be sufficiently responsive to variations in either person's payoff.

In the case of (1), the shame from choosing \mathbf{a} in the second stage is $g\left(\varphi(\mathbf{a}), \max_{\mathbf{b} \in A} \varphi(\mathbf{b})\right)$. This implies that even alternatives that are not chosen may matter for the value of a set, and that larger sets are not necessarily better. To see this, let $u(\mathbf{a}) = 2a_1$, $\varphi(\mathbf{a}) = (a_1 + 1)(a_2 + 1)$, and $g(x, y) = y - x$, so that $U(\{(10, 1), (5, 3)\}) = 18$, $U(\{(10, 1)\}) = 20$ and $U(\{(5, 3)\}) = 10$. 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 studied extensively 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 normative ranking of alternatives is symmetric. Then $(10, 0)$ is as good as $(0, 10)$. Independence implies that any randomization over these two allocations is as good as either of them, while intuition suggests that awarding \$10 to either player with probability 0.5 would be normatively best. In addition, the independence axiom implies that the suggested second-stage choice correspondence satisfies the weak axiom of revealed preferences (WARP); that is, the choice criterion is menu-independent. Contrary to this, we argue that violations of WARP are plausible in our context. Our most general representation (Theorem 1) does accommodate such violations.⁴

In contrast, if we take $g(x, y) = y - x$ in (1), the representation does feature a menu-independent choice criterion:

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

This representation is axiomatized in Theorem 2. The representation puts strong restrictions on the structure of shame-driven preferences, to the extent that second-stage choice alone does not reveal anything about the normative ranking. That is, without knowing DM's preferences over menus, one cannot distinguish between a standard decision maker and one who is susceptible to shame.

We further specify DM's norm by assuming that it satisfies *Independent Normative Contributions*: the contribution of raising one player's monetary payoff to the normative value of an allocation does not depend on the level of the other player's payoff. With this additional assumption, Theorem 3 establishes that there are two positive, increasing, and continuous 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 norm, $\varphi(\mathbf{a}) = v_1(a_1)v_2(a_2)$. Thus, the normatively best alternative within a set of alternatives can be characterized as the Nash Bargaining Solution of an associated game. Since the utility functions used to generate this game are subjective, so is the norm.

⁴In 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.

Example: In representation (1), let, again, $u(\mathbf{a}) = 2a_1$ and $\varphi(\mathbf{a}) = v_1(a_1)v_2(a_2) = (a_1 + 1)(a_2 + 1)$ and $g(x, y) = y - x$. 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. The set A induces an imaginary bargaining game where the disagreement point gives zero utility to each player. According to the Nash Bargaining Solution, $(5, 5)$ would be the outcome of the bargaining game. Its normative value is $6 \cdot 6 = 36$. To trade off shame with selfishness, DM chooses the alternative that maximizes the sum $2a_1 + (a_1 + 1)(a_2 + 1)$, which is $(6, 4)$. Its normative value is $7 \cdot 5 = 35$, and the shame from choosing it equals 1. Hence $U(A) = 11$. From the singleton set $B = \{(9, 0)\}$, which corresponds to the exit option in the experiment, the choice is trivial, and $U(B) = 18$. This example illustrates both the trade-off DM faces when choosing from a nondegenerate 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 interaction of the selfish and normative rankings through shame. Section 3 isolates a choice criterion from the choice situation. Section 4 further specifies the normative ranking. Section 5 concludes by pointing out connections to existing literature. All proofs are relegated to the appendix.

2. The Model

For some $k \in \mathbb{R} \cup \{-\infty\}$, let K be the set of all finite subsets of $(k, \infty)^2$.⁵ 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. 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 $d : (k, \infty)^2 \rightarrow \mathbb{R}_+$ is the standard Euclidean distance.

Let \succeq be a binary relation over K . The symmetric and asymmetric parts of \succeq , \sim and \succ respectively, are defined in the usual way. Our first two axioms on \succeq are standard.

P_1 (**Weak order**) \succeq is complete and transitive.

⁵ $\mathbb{R}_+^2 \subset (k, \infty)^2$, whenever $k < 0$. In particular, $(k, \infty)^2 = \mathbb{R}^2$ for $k = -\infty$.

P_2 (**Continuity**) \succeq is continuous.

The choice of a menu $A \in K$ is not observed by the recipient, whereas the choice from any menu is. We call the impact this observation has on choice *shame*. The next axiom captures the idea that shame is a mental cost, which is invoked by unchosen alternatives.

P_3 (**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 $A \succeq A \cup B$.⁶ 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 . This latter requirement rules out, for example, a situation in which shame only sets in once two allocations are sufficiently different according to, say, the Euclidean distance on $(k, \infty)^2$.

Definition 1. We say that DM is *susceptible to shame* if there exist A and B such that $A \succ A \cup B$.

Shame refers to some personal norm that determines what the appropriate choice should have been. Accordingly, we define an induced binary relation, “normatively better than”.

Definition 2. If DM is susceptible to shame, then we say that DM deems \mathbf{b} to be *normatively better than* \mathbf{a} , written $\mathbf{b} \succ_n \mathbf{a}$, if $\exists A \in K$ with $\mathbf{a} \in A$, such that $A \succ A \cup \{\mathbf{b}\}$.⁷

$A \succ A \cup \{\mathbf{b}\}$ implies that \mathbf{b} adds to the shame incurred by the original choice in A . In this case \mathbf{b} is socially better than any alternative in A , and in particular $\mathbf{b} \succ_n \mathbf{a}$. The

⁶This is the “Left Betweenness” axiom. It appears in Dekel, Lipman and Rustichini (2008).

⁷The notion of “normatively better than” is analogous to “more tempting than” in Gul and Pesendorfer (2005).

induced relations \succeq_n and \sim_n are defined as follows:

$$\begin{aligned} \mathbf{a} \succeq_n \mathbf{b} &\Leftrightarrow \mathbf{b} \not\succeq_n \mathbf{a}, \\ \mathbf{a} \sim_n \mathbf{b} &\Leftrightarrow \text{both } \mathbf{b} \not\succeq_n \mathbf{a} \text{ and } \mathbf{a} \not\succeq_n \mathbf{b} \end{aligned}$$

Some of the axioms below are imposed on \succ_n rather than on \succeq and are labeled by N instead of P . Because \succ_n is an induced binary relation, N axioms are implicit axioms on \succeq ; \succ_n is only an expositional device.⁸ The N -axioms capture basic features of the norm that we assume throughout. Making those assumptions directly on \succ_n and motivating them in that context is natural.

N_1 (**Ranking**) \succ_n is an asymmetric and negatively transitive binary relation.

Axiom N_1 rules out situations in which there are two alternatives, \mathbf{a} and \mathbf{b} , and two menus, A and B , with $\{\mathbf{a}, \mathbf{b}\} \subset A \cap B$, such that \mathbf{a} contributes to shame in A and \mathbf{b} contributes to shame in B . This implies that the normative value of allocations is menu-independent and, furthermore, that multiple alternatives on a menu cannot jointly contribute to shame. Instead, it implies that only one alternative in each menu, one of the normatively best, is responsible for shame.

N_2 (**Pareto**) If $\mathbf{b} \geq \mathbf{a}$ and $\mathbf{b} \neq \mathbf{a}$, then $\mathbf{b} \succ_n \mathbf{a}$.

The axiom states that an alternative with higher payoffs to both individuals is normatively better. In other words, the subjective norm must have some concern for efficiency.

N_3 (**Compensation**) For all \mathbf{a}, \mathbf{b} , there exist x, y such that both $(a_1, x) \succ_n (b_1, b_2)$ and $(y, a_2) \succ_n (b_1, b_2)$.

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. In particular, N_3 requires \succ_n never to be satiated in either payoff.⁹

⁸Since for the standard decision maker it is never the case that $A \succ A \cup B$, she is not susceptible to shame. Therefore, we cannot infer anything about her perceived norm, and the N -axioms put no restrictions on her choice behavior.

⁹Definition 2 (of \succ_n) implies that if truly $\mathbf{a} \sim_n \mathbf{b}$, then the false hypothesis that $\mathbf{a} \succ_n \mathbf{b}$ cannot be refuted in a finite number of observations, as one would have to establish that there exists no menu B with $\mathbf{b} \in B$, such that $B \succ B \cup \{\mathbf{a}\}$. This renders the negative transitivity part of N_1 non-refutable whenever the violation involves indifference. Accepting N_1 , axiom P_4 , which relates \succ and \succ_n , is refutable. Axiom N_2

The next axiom concerns DM's preferences over singleton sets. A singleton set is a degenerate menu that contains only one feasible allocation. Since degenerate menus leave DM with no choice to be made under observation, choosing between two singleton sets reveals DM's private ranking of the allocations. Although we allow private preferences to care about the recipient's payoff, DM's norm always cares about it more. This difference is captured by the following single-crossing assumption:

P_4 (**Selfishness**) If $a_1 > b_1$ and $\mathbf{a} \succ_n \mathbf{b}$, then $\{\mathbf{a}\} \succ \{\mathbf{b}\}$.

Definition 3. Let f and h be two functions on $(k, \infty)^2$. We say that h is *more selfish than* f , if for all \mathbf{a} and for all $\Delta_1 \in (0, a_1 - k)$ and $\Delta_2 \in (0, a_2 - k)$,

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

(ii) $h(\mathbf{a}) = h(a_1 + \Delta_1, a_2 - \Delta_2)$ implies $f(\mathbf{a}) \geq f(a_1 + \Delta_1, a_2 - \Delta_2)$ with strict inequality for at least one pair Δ_1, Δ_2 .

Definition 3 is the discrete version of the requirement that the slope of h in the (a_1, a_2) space is, at any point, greater than that of f .¹⁰

Definition 4. A function $\varphi : (k, \infty)^2 \rightarrow \mathbb{R}$ is called a *subjective norm function* if it is strictly increasing and satisfies $\sup_{x \in (k, \infty)} \varphi(x, y) > \varphi(\mathbf{b})$ and $\sup_{x \in (k, \infty)} \varphi(y, x) > \varphi(\mathbf{b})$ for all $y \in (k, \infty)$ and $\mathbf{b} \in (k, \infty)^2$.

It is clear that if \succ_n satisfies $N_1 - N_3$, then any function $\varphi : (k, \infty)^2 \rightarrow \mathbb{R}$ that represents it should be a subjective norm function.

Theorem 1 \succ and \succ_n satisfy $P_1 - P_4$ and $N_1 - N_3$ respectively, if and only if there exist (i) a continuous subjective norm function φ , (ii) a continuous function $u : (k, \infty)^2 \rightarrow \mathbb{R}$, which is weakly increasing and more selfish than φ , and (iii) a continuous function $g : (k, \infty)^2 \times \varphi((k, \infty)^2) \rightarrow \mathbb{R}$, which is strictly increasing in its second argument and satisfies

can be separated into a weak version (in which the implication is of the of the \succeq_n type), which is refutable, and an axiom that requires that indifference curves are not "thick", which is not refutable. Axiom N_3 is of the "there exists" type and hence is not refutable independently of the induced nature of \succ_n .

¹⁰If h and f were differentiable and the inequalities in both items (i) and (ii) were strict, then the condition would be $\frac{f_1}{f_2}(\mathbf{a}) < \frac{h_1}{h_2}(\mathbf{a})$ 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 *more selfish than* coincides with the definition of less altruistic than in Cox, Friedman and Sadiraj (2008).

$g(\mathbf{a}, x) = 0$ whenever $\varphi(\mathbf{a}) = x$, 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 \succeq and φ represents \succ_n .

Theorem 1 provides a representation of DM’s normative ranking based on revealed preferences. This should help the empirical quest to understand people’s perception of social norms. The representation captures the tension between DM’s impulse to choose the allocation she prefers in private and her desire to minimize shame. There are at most two essential alternatives within a set, to be interpreted as the “chosen” and the “normatively best” alternative, \mathbf{a} and \mathbf{b} respectively. For the latter, only its normative value, $\varphi(\mathbf{b})$, matters for its impact on the set’s value. Since $g(\mathbf{a}, \varphi(\mathbf{a})) = 0$ and g is strictly increasing in its second argument, $g(\mathbf{a}, \varphi(\mathbf{b})) > 0$ whenever $\varphi(\mathbf{a}) < \varphi(\mathbf{b})$, where $\varphi(\mathbf{a})$ is the normative value of the chosen alternative. The representation captures the idea of shame being an emotional cost that emerges whenever the normatively best available allocation is not chosen. 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 utility will never lie below $u(\mathbf{b})$, the utility of the normatively best allocation. By P_4 , any deviations by DM from choosing the normatively best allocation will be in her own favor. In other words, it is being selfish, and not being too generous, that triggers shame. Its magnitude may depend on the chosen allocation.

We conclude this section by mentioning the main steps of the proof. We provide intuition only for the most instructive steps. Continuity of \succeq implies that \succ_n is also a continuous preference relation. Therefore, they can both be represented by continuous functions, $U : K \rightarrow \mathbb{R}$ and $\varphi : (k, \infty)^2 \rightarrow \mathbb{R}$, respectively. The combination of N_2 and N_3 implies that φ is a subjective norm function. Note that by the asymmetry part of N_1 , if $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$, then $\{\mathbf{a}, \mathbf{b}\} \succeq \{\mathbf{b}\}$. We generalize this observation to show that the combination of P_3 and N_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. Hence we confine our attention to a subset of the domain that includes all sets with cardinality no greater than two. A key step for the remainder of the proof is to show that u is more selfish than φ and, in particular, that for any \mathbf{a} there is a region in which part (ii) of Definition 3 is satisfied with strict inequality. To see this, note that if the inequality is never strict, then by P_4 , $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ implies $\mathbf{a} \succ_n \mathbf{b}$. Suppose that $\{\mathbf{c}\} \succ \{\mathbf{a}, \mathbf{c}\}$ for some \mathbf{c} . Then by SB $\{\mathbf{c}\} \succ \{\mathbf{a}\}$, which implies

that $\mathbf{c} \succ_n \mathbf{a}$, or $\{\mathbf{c}\} \not\succeq \{\mathbf{a}, \mathbf{c}\}$. Therefore $\{\mathbf{a}, \mathbf{c}\} \succeq \{\mathbf{c}\}$ for all \mathbf{c} . We generalize this conclusion to show that $C \cup \{\mathbf{a}\} \succeq C$ for all C , which implies that $\mathbf{a} \not\succeq_n \mathbf{c}$ for all \mathbf{c} and in particular for $\mathbf{c} < \mathbf{a}$, violating N_2 . After establishing that u is more selfish than φ , we show that any set A is indifferent to some two-element subset of it that includes one of the normatively best allocations in A . Furthermore, only the normative value, φ , of this alternative affects the value of A . Lastly, we show that the shame function, g , must be strictly increasing in its second component. To see this, assume to the contrary that $g(\mathbf{a}, \varphi)$ is positive and constant on some interval $[\underline{\varphi}, \bar{\varphi}]$. Then we can find \mathbf{b} and \mathbf{b}' such that $\varphi(\mathbf{b}) = \underline{\varphi}$, $\varphi(\mathbf{b}') = \bar{\varphi}$ and $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}, \mathbf{b}'\}$. By SB, $\{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$. Since u is more selfish than φ , there exists \mathbf{c} such that $\{\mathbf{c}\} \succ \{\mathbf{a}\}$, $\{\mathbf{c}\} \succ \{\mathbf{c}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$, $\{\mathbf{c}, \mathbf{b}'\} \succ \{\mathbf{a}, \mathbf{b}'\}$ and $\varphi(\mathbf{c}) \in (\varphi(\mathbf{b}), \varphi(\mathbf{b}'))$. Then $\{\mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}', \mathbf{c}\}$ and by the second part of P_3 we have $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, which is a contradiction.

3. A Second-Stage Choice Ranking

In many situations, second-stage choice may also be observed by the experimenter. Suppose that, as suggested by Theorem 1, the correspondence that governs second-stage choice is

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

If the function g is increasing and convex in its second argument, then our model can accommodate a DM who does not suffer much if he deviates slightly from the normatively right behavior but considers large deviations from the norm to be unacceptable. For example, she might choose $(8, 2)$ from the set $\{(10, 0), (8, 2), (5, 5)\}$ and $(10, 0)$ from the set $\{(10, 0), (8, 2)\}$; while she finds her preferred allocation to be $(10, 0)$ when the normatively best 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 (WARP) is plausible when shame is taken into account. That said, it is also plausible that DM would like her second-stage choice rule to be set-independent, that is, to satisfy WARP.

The next axiom strengthens the role of the normative value of the chosen alternative in determining shame: the greater the normative value of DM's choice, the less shame she feels.

P_5 (**Mitigating Shame**) Suppose $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$. If $\{\mathbf{a}'\} \sim \{\mathbf{a}\}$ and $\mathbf{a}' \succ_n \mathbf{a}$ then $\{\mathbf{a}', \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}\}$.

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 normatively best (\mathbf{b}), which causes her choice to bear shame. This shame, however, is not enough to make her choose \mathbf{b} . Axiom P_5 then implies that only the normative value of the chosen alternative matters for its impact on shame.

Given $P_1 - P_5$ and $N_1 - N_3$, an additional assumption is equivalent to a set-independent choice ranking.

P_6 (**Consistency**) If $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\} \succ \{\mathbf{a}', \mathbf{b}\}$ then for any \mathbf{b}' , either $\{\mathbf{a}, \mathbf{a}', \mathbf{b}'\} \succ \{\mathbf{a}', \mathbf{b}'\}$ or $\{\mathbf{b}'\} \succeq \{\mathbf{a}', \mathbf{b}'\}$.

The qualifier says that the addition of \mathbf{a} improves the value of $\{\mathbf{a}', \mathbf{b}\}$, which implies that \mathbf{a} is the (unique) choice from $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$. The axiom requires that \mathbf{a} improves any other (two-element) set that includes \mathbf{a}' , unless the third alternative, \mathbf{b}' is sufficiently good. In terms of the suggested second-stage choice, the axiom implies that if \mathbf{a} is ever the unique choice when \mathbf{a}' is available, then \mathbf{a}' will never be the unique choice in the face of \mathbf{a} .

Theorem 2 \succ and \succ_n satisfy $P_1 - P_6$ and $N_1 - N_3$ respectively, if and only if there exist a continuous subjective norm function φ and a continuous function $u : (k, \infty)^2 \rightarrow \mathbb{R}$, which is weakly increasing and more selfish than 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_n .

The representation in Theorem 2 suggests 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 normatively best 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 and depends only on the normative value of both alternatives.

We conclude this section by remarking on the identifiability of the normative ranking.¹¹

¹¹The remark is not specific to our model: it generalizes to all models that study preferences over sets of

A natural question is to what extent one can elicit DM’s normative 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 the definition of the choice correspondence (equation (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 $\mathbf{a} \notin \arg \max_{\mathbf{b}' \in A \cup \{\mathbf{b}\}} \varphi(\mathbf{b}')$. Therefore, a second-stage violation of WARP partially identifies DM’s normative ranking: $\mathbf{b} \succ_a \mathbf{a}$ implies $\mathbf{b} \succ_n \mathbf{a}$. Furthermore, enough violations of WARP would fully identify \succ_n .¹² If there are no violations of WARP, then observing DM’s choice of menus is necessary to elicit the normative ranking of any two alternatives. This is the case where Theorem 2 applies.

4. Specifying a Normative Ranking

We now impose another axiom on \succ_n to further specify DM’s subjective norm. The axiom is known in the literature as the Hexagon condition (Karni and Safra (1998)). In our context, it asserts that the contribution of one person’s marginal payoff to the normative value of an allocation cannot depend on the initial payoff levels.

N_4 (**Independent Normative Contributions**) If $(a_1, a_2) \sim_n (b_1, b_2)$ and $(a'_1, a_2) \sim_n (a_1, b_2) \sim_n (b_1, b'_2)$, then $(a'_1, b_2) \sim_n (a_1, b'_2)$.

The axiom is illustrated in figure 1. If $a_1 = a'_1$ or $b_2 = b'_2$, this axiom is implied by N_1 , N_2 , and the continuity of \succ_n . For $a_1 \neq a'_1$ and $b_2 \neq b'_2$, the statement is subtler. Consider first a stronger assumption, which is also known as the *Thomsen condition* (Krantz, Luce, Suppes and Tversky (1971)):

N'_4 (**Strong Independent Normative Contributions**) $(a_1, a_2) \sim_n (b_1, b_2)$ and $(a'_1, a_2) \sim_n (b_1, b'_2)$ imply $(a'_1, b_2) \sim_n (a_1, b'_2)$.

To motivate this axiom, assume that DM constructs her subjective norm based on two perceived utility functions over monetary payoffs (not over allocations): one for herself and one for the recipient. At the same time, she either cannot, or is not willing to, compare their alternatives.

¹² \succ_n 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 $\mathbf{b} \sim_n \mathbf{a}$ then $\mathbf{b} \sim_a \mathbf{a}$. Take any $\mathbf{b} \succ_n \mathbf{a}$. By N_2 there exist $\mathbf{c} > \mathbf{d}$ such that $\mathbf{c} \sim_n \mathbf{b} \succ_n \mathbf{a} \sim_n \mathbf{d}$. By Pareto, $\mathbf{c} \succ_a \mathbf{d}$ and by transitivity $\mathbf{b} \succ_a \mathbf{a}$. In the text we argue that the other direction is also true, hence $\mathbf{b} \succ_n \mathbf{a} \Leftrightarrow \mathbf{b} \succ_a \mathbf{a}$.

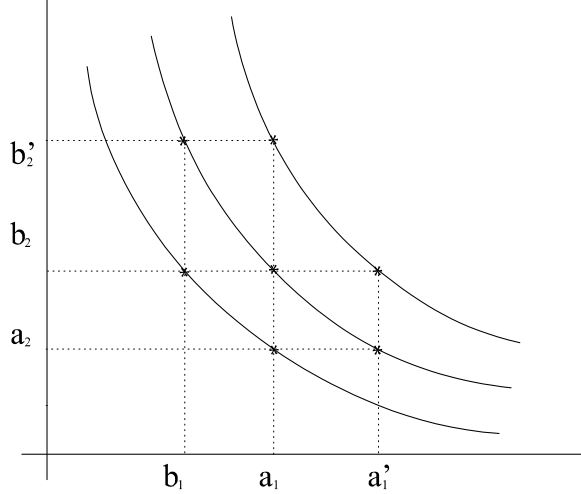


Figure 1: Independent Normative Contributions.

relative intensities. In other words, she does not make interpersonal comparisons of utilities.¹³ The qualifier in N'_4 establishes that DM considers the 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) . N'_4 then states that starting from the allocation (a_1, b_2) , changing a_1 to a'_1 should again be as favorable as changing b_2 to b'_2 . This is the essence of *Independent Normative Contributions*. The stronger qualifier $(b_1, b'_2) \sim_n (a_1, b_2) \sim_n (a'_1, a_2)$ in N_4 weakens the axiom. For example, the normative ranking $(a_1, a_2) \succ_n (b_1, b_2)$ if and only if $\min(a_1, a_2) > \min(b_1, b_2)$ is permissible under N_4 , but not under N'_4 .

Krantz, Luce, Suppes and Tversky (1971) provide an additive representation based on N'_4 . Karni and Safra (1998) demonstrate that the weaker condition, N_4 , implies N'_4 in the context of their axioms. The next theorem is based on those results:

Theorem 3 \succ_n is continuous and satisfies $N_1 - N_4$, if and only if there are continuous, strictly increasing, and unbounded functions $v_1, v_2 : (k, \infty) \rightarrow \mathbb{R}_{++}$, such that $\varphi(\mathbf{a}) = v_1(a_1)v_2(a_2)$ represents \succ_n .

This representation suggests an appealing interpretation of the normative ranking DM is concerned about: she behaves as if she had in mind two positive, increasing and unbounded utility functions on (k, ∞) : 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

¹³This argument resembles the idea, common in social choice theory, that interpersonal utility comparisons are infeasible.

payoffs.¹⁴ DM then identifies the normatively best alternative within a set as the Nash Bargaining Solution (NBS) of the game. Moreover, all alternatives can be normatively ranked according to the same functional, the Nash product.

One justification of the NBS as the normative best allocation is related to Gauthier’s (1986) principle of “moral by agreement”: trying to assess what is normative, but finding herself unable, or unwilling, 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.¹⁵

5. 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. Both works attach a cost to any deviation from choosing the normatively best alternative. In Fehr and Schmidt’s work, the normatively best allocation is any equal-split and need not be feasible. In our work, the normatively best *available* allocation is responsible for shame. The dependency of the normatively best 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

¹⁴Since v_1 and v_2 are positive, the imaginary disagreement point does not correspond to any allocation in our domain.

¹⁵The 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).)

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 expected utility functionals. 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_3) and Normative Ranking (N_1) imply Set Betweenness. Hence, GP’s Lemma 2 can be employed, allowing us to confine attention to sets with only two elements.

Epstein and Kopylov (2007) study preferences over menus of Anscombe and Aumann acts. Their representation captures the intuition that people become pessimistic as the time of consumption approaches. By treating state i as the payoff to player i , and by redefining the mixture operator as convex combinations of allocations instead of the usual convex combinations of acts, their axioms can be applied to our domain.¹⁶ While a version of their main representation can accommodate the specific dictator game experiment of Dana *et al.*, the two works differ significantly. Applying Epstein and Kopylov’s Axiom 7 (constant acts cannot be tempted) to our context implies that DM privately prefers an allocation that favors the recipient over an allocation that gives the same amount to both players, if the two are equally good according to the normative ranking. This contradicts the central notion that DM is ashamed to be selfish, which is captured by the assumption that she is more selfish in private (our Axiom P_4). In addition, due to the definition of the mixture operation on \mathbb{R}^2 , their axioms restrict the private ranking to be represented by a linear function, u , and the social trade-off function, φ , to be piecewise linear with a kink on the main diagonal. Whether or not such rankings are reasonable, many other private and normative rankings are excluded, which defies the purpose of eliciting these rankings from behavior.

¹⁶We thank a referee for suggesting this connection between the two models.

Empirically, the assumption that only two elements of a choice set matter for the magnitude of shame (the normatively best 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.”¹⁷ 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.

6. Appendix

6.1. Proofs

Proof of Theorem 1

Let $U : K \rightarrow \mathbb{R}$ be a continuous function that represents \succ . Without loss of generality,

¹⁷Surely 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.

we assume that U is bounded from below. Define $u(\mathbf{a}) \equiv U(\{\mathbf{a}\})$. Axiom P_2 implies that \succ_n is continuous and hence admits a continuous representation. Let $\varphi : (k, \infty)^2 \rightarrow \mathbb{R}$ be a continuous function that represents \succ_n . By N_2 , φ is strictly increasing. Because \succ_n is continuous, N_3 immediately implies that if $(a_1, a_2) \not\succeq_n (b_1, b_2)$, then there are x and y such that $(a_1, x) \sim_n (b_1, b_2) \sim_n (y, a_2)$. In all that follows, we use this stronger version of N_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 $\mathbf{b} \succ_n \mathbf{a}$. 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 N_1 , for all \mathbf{a}^n there exists $\mathbf{b} \in B$ such that $\mathbf{a}^n \not\succeq_n \mathbf{b}$. By the definition of \succ_n , $C_{n-1} \not\succeq C_n$. By negative transitivity of \succ , $C_0 \not\succeq C_N$ or $A \cup B \succeq B$.

Case 2) $\exists \mathbf{a} \in A$ such that $\mathbf{a} \succ_n \mathbf{b}, \forall \mathbf{b} \in B$. 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 the definition of \succ_n and N_1 , $\forall C$ such that $\mathbf{a} \in C, C \not\succeq C \cup \{\mathbf{b}^m\}$. Hence, $C_{m-1} \not\succeq C_m$. By negative transitivity of \succ , $C_0 \not\succeq C_M$ or $A \cup B \succeq A \succeq B$, hence $A \cup B \succeq B$.||

Combining Claim 1.1 with P_3 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 \succeq 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 $\min_{\mathbf{b}' \in A} \max_{\mathbf{a}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$.*

We now use SB, Lemma 1.1, P_4 and the monotonicity of φ to show that u is more selfish than φ according to Definition 3.

Claim 1.2: u is weakly increasing and more selfish than φ .

Proof: First, suppose that u is not weakly increasing. Then there is $\mathbf{b} > \mathbf{a}$, such that $u(\mathbf{a}) > u(\mathbf{b})$, and therefore $\{\mathbf{a}\} \succ \{\mathbf{b}\}$. But $\mathbf{b} > \mathbf{a}$ implies that both $\mathbf{b} \succ_n \mathbf{a}$ by P_2 and $b_1 > a_1$, a contradiction to P_4 .

To show that item (i) in Definition 3 holds, assume that $u(\mathbf{a}) = u(a_1 - \Delta_1, a_2 + \Delta_2)$ which implies that $\{\mathbf{a}\} \sim \{a_1 - \Delta_1, a_2 + \Delta_2\}$. Since $a_1 - \Delta_1 < a_1$, it must be that

$(a_1 - \Delta_1, a_2 + \Delta_2) \succeq_n \mathbf{a}$ or $\varphi(a_1 - \Delta_1, a_2 + \Delta_2) \geq \varphi(\mathbf{a})$. The weak inequality in item (ii) is shown similarly. Let $U^\succ(\mathbf{a}) := \{\mathbf{a}' : u(\mathbf{a}') > u(\mathbf{a})\}$ and $U^{\succ_n}(\mathbf{a}) := \{\mathbf{a}' : \varphi(\mathbf{a}') > \varphi(\mathbf{a})\}$. To establish that one inequality in item (ii) must be strict, we show that we cannot have \mathbf{a} for which $u(\mathbf{a}) = u(a_1 + \Delta_1, a_2 - \Delta_2)$ implies $\varphi(\mathbf{a}) = \varphi(a_1 + \Delta_1, a_2 - \Delta_2)$ for all Δ_1 and Δ_2 . Suppose it was the case. Then by P_4 , $U^\succ(\mathbf{a}) \subseteq U^{\succ_n}(\mathbf{a})$.

Step 1: There is no \mathbf{c} such that $\{\mathbf{c}\} \succ \{\mathbf{c}, \mathbf{a}\}$.

Proof of Step 1: Suppose instead that there exists \mathbf{c} such that $\{\mathbf{c}\} \succ \{\mathbf{c}, \mathbf{a}\}$. Then by SB $\{\mathbf{c}\} \succ \{\mathbf{a}\}$, and, since $U^\succ(\mathbf{a}) \subseteq U^{\succ_n}(\mathbf{a})$, $\mathbf{c} \succ_n \mathbf{a}$. Therefore, by definition $\{\mathbf{c}\} \not\succeq \{\mathbf{c}, \mathbf{a}\}$, which is a contradiction.

Step 2: If there is no \mathbf{c} such that $\{\mathbf{c}\} \succ \{\mathbf{c}, \mathbf{a}\}$, then there is no C such that $C \succ C \cup \{\mathbf{a}\}$.

Proof of Step 2: Suppose instead that $C \succ C \cup \{\mathbf{a}\}$ for some C . By SB, there exist $\mathbf{c}, \mathbf{c}', \mathbf{c}'' \in C$ such that $C \sim \{\mathbf{c}', \mathbf{c}''\}$ and $C \cup \{\mathbf{a}\} \sim \{\mathbf{c}, \mathbf{a}\}$. (\mathbf{c}, \mathbf{c}' , and \mathbf{c}'' need not be distinct). Without loss of generality assume that \mathbf{c}'' is the minimizer in C (clearly \mathbf{a} is the minimizer in $C \cup \{\mathbf{a}\}$). Then $\{\mathbf{c}'\} \succeq \{\mathbf{c}', \mathbf{c}''\} \sim C \succ C \cup \{\mathbf{a}\} \sim \{\mathbf{c}, \mathbf{a}\} \succeq \{\mathbf{c}', \mathbf{a}\}$, where the last \succeq is because \mathbf{c} is one of the maximizers in $C \cup \{\mathbf{a}\}$. By transitivity, $\{\mathbf{c}'\} \succ \{\mathbf{c}', \mathbf{a}\}$, which is a contradiction.

Hence, by Definition 1, $\mathbf{a} \not\succeq_n \mathbf{c}$ for all \mathbf{c} , and in particular for $\mathbf{c} = (a_1 - \varepsilon, a_2 - \varepsilon)$, contradicting the strict Pareto criterion (N_2).||

Claim 1.3: (i) $[\varphi(\mathbf{a}) < \varphi(\mathbf{b}) \text{ and } \{\mathbf{a}\} \succ \{\mathbf{b}\}] \Leftrightarrow \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$

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

(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_3 $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$. Conversely if $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$, then by SB $\{\mathbf{a}\} \succ \{\mathbf{b}\}$, and by definition $\mathbf{b} \succ_n \mathbf{a}$, or $\varphi(\mathbf{a}) < \varphi(\mathbf{b})$.

(ii) If $\{\mathbf{b}\} \succeq \{\mathbf{a}\}$, then by SB $\{\mathbf{b}\} \succeq \{\mathbf{a}, \mathbf{b}\}$. Since $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$, N_1 implies that there is no B such that $\mathbf{b} \in B$ and $B \succ B \cup \{\mathbf{a}\}$. Thus, $\{\mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}\}$.

(iii) $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ implies by SB $\{\mathbf{a}\} \succeq \{\mathbf{a}, \mathbf{b}\}$. As $\varphi(\mathbf{a}) = \varphi(\mathbf{b})$, part (i) implies that not $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$, and therefore $\{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\}$.||

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

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

(By Lemma 1.1, $(\mathbf{b}^*(A), \mathbf{a}^*(A))$ also solves $\min_{\mathbf{b}' \in A} \max_{\mathbf{a}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$.)

Claim 1.4: 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} and \mathbf{c} such that $\{\mathbf{a}, \mathbf{c}\} \sim A$, and $(\mathbf{a}, \mathbf{c}) = (\mathbf{a}^*(A), \mathbf{b}^*(A))$. Therefore,

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

where the first strict inequality is because \mathbf{b} is not one of the minimizers. But $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ implies that $\mathbf{c} \succ_n \mathbf{b}$, which is a contradiction. ||

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

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

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 $\mathbf{b} \succ_n \mathbf{c}$ 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 $\mathbf{b} \sim_n \mathbf{c}$ imply $\{\mathbf{a}, \mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}\}$.
- (3) $\mathbf{b} \in Y(\mathbf{a}, \varphi)$, $\{\mathbf{b}\} \succ \{\mathbf{b}'\}$ and $\mathbf{b}' \sim_n \mathbf{b}$ 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 $\mathbf{b}' \sim_n \mathbf{b}$, 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 $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$. Then by SB $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, and, therefore, $\mathbf{c} \succ_n \mathbf{b}$, which is a contradiction. If (2) did not hold, then either $\{\mathbf{a}, \mathbf{c}\} \succ \{\mathbf{a}, \mathbf{b}\}$ and $\{\mathbf{a}, \mathbf{c}\} \succ \{\mathbf{c}\}$, which imply, using SB, that $\{\mathbf{a}, \mathbf{c}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, or $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{c}\}$ and $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, which imply, again using SB, that $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. In both cases, we get a contradiction to $\mathbf{b} \sim_n \mathbf{c}$. 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 $\mathbf{b}' \sim_n \mathbf{b}$. 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, $\{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, which is again a contradiction to $\mathbf{b}' \sim_n \mathbf{b}$. To verify (4), assume $\{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$. Then, by Claim 1.3 (i) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$, and 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 $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}'\}$, the combination of $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$ and SB imply that $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, which is a contradiction to $\mathbf{b}' \in I_n(\varphi(\mathbf{b}))$. Note that we cannot have $\{\mathbf{b}'\} \succ \{\mathbf{a}, \mathbf{b}'\}$. Otherwise, we would have $\mathbf{a} \succ_n \mathbf{b}' \sim_n \mathbf{b} \succ_n \mathbf{a}$, which is a contradiction. ||

Define $f : (k, \infty)^2 \times (k, \infty)^2 \rightarrow \mathbb{R}$ such that $f(\mathbf{a}, \mathbf{b}) = u(\mathbf{a}) - \tilde{U}(\mathbf{a}, \mathbf{b})$, where $\tilde{U} : (k, \infty)^2 \times (k, \infty)^2 \rightarrow \mathbb{R}$ is a function satisfying:¹⁸

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

By definition we have $f(\mathbf{a}, \mathbf{a}) = 0$ for every $\mathbf{a} \in (k, \infty)$. Note as well that

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

as otherwise we would have:

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

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

Claim 1.5: There exists a function \tilde{U} satisfying the condition specified above such that $f(\mathbf{a}, \mathbf{b}) = g(\mathbf{a}, \varphi(\mathbf{b}))$ for some $g : (k, \infty)^2 \times \mathbb{R} \rightarrow \mathbb{R}$, which is strictly increasing in its second argument.

Proof: Given \mathbf{b} and \mathbf{c} such that $\varphi(\mathbf{b}) > \varphi(\mathbf{c})$, we show that for all \mathbf{a} , $f(\mathbf{a}, \mathbf{b}) > f(\mathbf{a}, \mathbf{c})$ is consistent with \succ .

We first show that if $\varphi(\mathbf{b}) \geq \varphi(\mathbf{a}) \geq \varphi(\mathbf{c})$, then $f(\mathbf{a}, \mathbf{b}) \geq 0 > f(\mathbf{a}, \mathbf{c})$ is consistent with \succ , and if $\varphi(\mathbf{a}) \geq \varphi(\mathbf{b}) > \varphi(\mathbf{c})$, then $0 \geq f(\mathbf{a}, \mathbf{b}) > f(\mathbf{a}, \mathbf{c})$ is consistent with \succ . To see this, we consider pairs (\mathbf{a}, \mathbf{b}) , (\mathbf{a}, \mathbf{c}) and identify the restrictions imposed by all combinations of corresponding values of u and ϕ . Consider first the pair (\mathbf{a}, \mathbf{b}) :

(i) $u(\mathbf{a}) > u(\mathbf{b})$ and $\phi(\mathbf{b}) > \phi(\mathbf{a})$. By Claim 1.3 (i), $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$, so that $f(\mathbf{a}, \mathbf{b}) > 0$.

(ii) $u(\mathbf{a}) \leq u(\mathbf{b})$ and $\phi(\mathbf{b}) > \phi(\mathbf{a})$. By Claim 1.3 (ii), $\{\mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}\}$. By the formula for $U(\{\mathbf{a}, \mathbf{b}\})$, a sufficient condition is that $f(\mathbf{a}, \mathbf{b}) > 0 > f(\mathbf{b}, \mathbf{a})$.

(iii) $u(\mathbf{a}) > u(\mathbf{b})$ and $\phi(\mathbf{b}) = \phi(\mathbf{a})$. By Claim 1.3 (iii), $\{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$. By the formula for $U(\{\mathbf{a}, \mathbf{b}\})$, a sufficient condition is that $f(\mathbf{a}, \mathbf{b}) = 0 = f(\mathbf{b}, \mathbf{a})$.

¹⁸Note that $\max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} U(\{\mathbf{a}, \mathbf{b}\}) = \max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} \left[\max_{\mathbf{a}' \in \{\mathbf{a}, \mathbf{b}\}} \min_{\mathbf{b}' \in \{\mathbf{a}, \mathbf{b}\}} \tilde{U}(\mathbf{a}', \mathbf{b}') \right] = \max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} \tilde{U}(\mathbf{a}, \mathbf{b})$.

(iv) $u(\mathbf{a}) \leq u(\mathbf{b})$ and $\phi(\mathbf{b}) = \phi(\mathbf{a})$. By Claim 1.3 (ii), $\{\mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}\}$. By the formula for $U(\{\mathbf{a}, \mathbf{b}\})$, a sufficient condition is that $f(\mathbf{a}, \mathbf{b}) = 0 = f(\mathbf{b}, \mathbf{a})$.

Similar calculations can be done for (\mathbf{a}, \mathbf{c}) , by replacing \mathbf{b} with \mathbf{c} .

Suppose $\phi(\mathbf{b}) \geq \phi(\mathbf{a}) \geq \phi(\mathbf{c})$. Then from cases (i) – (iv), $f(\mathbf{a}, \mathbf{b}) \geq 0 \geq f(\mathbf{a}, \mathbf{c})$. But since one of the weak inequalities is strict, either $f(\mathbf{a}, \mathbf{b}) > 0$ or $f(\mathbf{a}, \mathbf{c}) < 0$ (or both), hence indeed $f(\mathbf{a}, \mathbf{b}) > f(\mathbf{a}, \mathbf{c})$ is consistent with \succ . Suppose $\phi(\mathbf{a}) \geq \phi(\mathbf{b}) > \phi(\mathbf{c})$. Then $0 \geq f(\mathbf{a}, \mathbf{b}) > f(\mathbf{a}, \mathbf{c})$ is consistent with \succ (that is, we can, arbitrarily, choose the values of $f(\mathbf{a}, \mathbf{b})$ and $f(\mathbf{a}, \mathbf{c})$ as such).

Therefore, confine attention to the case where $\varphi(\mathbf{b}) > \varphi(\mathbf{c}) > \varphi(\mathbf{a})$.

If there is no $\mathbf{b}' \in I_n(\varphi(\mathbf{b}))$ with $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\}$, then $f(\mathbf{a}, \mathbf{b}) > f(\mathbf{a}, \mathbf{c}) \geq 0$ is consistent with \succ . Suppose there exists $\mathbf{b}' \in I_n(\varphi(\mathbf{b}))$ such that $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\}$. There are two cases to consider:

Case 1) Suppose $Y(\mathbf{a}, \varphi(\mathbf{b})) = \emptyset$. Let $\underline{u} := \inf_{\mathbf{a}' \in (k, \infty)} u(\mathbf{a}')$ (recall that U is chosen to be bounded from below). Define $f(\mathbf{a}, \mathbf{b})$ such that $f(\mathbf{a}, \mathbf{b}) > u(\mathbf{a}) - \underline{u}$. If $Y(\mathbf{a}, \varphi(\mathbf{c})) \neq \emptyset$, then $f(\mathbf{a}, \mathbf{c}) < u(\mathbf{a}) - u(\mathbf{c}) < u(\mathbf{a}) - \underline{u} < f(\mathbf{a}, \mathbf{b})$. If $Y(\mathbf{a}, \varphi(\mathbf{c})) = \emptyset$ then we can choose $f(\mathbf{a}, \mathbf{b}) > \max\{u(\mathbf{a}) - \underline{u}, f(\mathbf{a}, \mathbf{c})\}$, so that $f(\mathbf{a}, \mathbf{c}) < f(\mathbf{a}, \mathbf{b})$ is consistent with \succ .

Case 2) 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) $f(\mathbf{a}, \mathbf{b}') = f(\mathbf{a}, \mathbf{b}'') \forall \mathbf{b}', \mathbf{b}'' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$. If $\mathbf{b}'' \in I_n(\varphi(\mathbf{b}))$ but $\mathbf{b}'' \notin Y(\mathbf{a}, \varphi(\mathbf{b}))$, then, by observation (4), it must be that $\{\mathbf{b}''\} \sim \{\mathbf{a}, \mathbf{b}''\}$. The constraint is thus $f(\mathbf{a}, \mathbf{b}'') > u(\mathbf{a}) - u(\mathbf{b}'')$. Using the the definition above we have $f(\mathbf{a}, \mathbf{b}'') = f(\mathbf{a}, \mathbf{b}) > u(\mathbf{a}) - u(\mathbf{b}) > u(\mathbf{a}) - u(\mathbf{b}'')$, hence the definition is consistent with \succ . By N_2 and continuity of \succ_n , there exists $\mathbf{c}' \in I_n(\mathbf{c})$ 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_4 . Claim 1.3 (i) and observation (1) imply that $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}'\} \succeq \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\} \succ \{\mathbf{c}'\}$. Therefore, $f(\mathbf{a}, \mathbf{b}') \geq f(\mathbf{a}, \mathbf{c}')$. Note that 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 , even if $\mathbf{b} \notin Y(\mathbf{a}, \varphi(\mathbf{b}))$ or $\mathbf{c} \notin Y(\mathbf{a}, \varphi(\mathbf{c}))$. If $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{c}\}$, then observation (4) implies that $\{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{b}\}$, and, therefore, $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$ is consistent with \succ . We still need to argue that $f(\mathbf{a}, \mathbf{b}) \neq f(\mathbf{a}, \mathbf{c})$. Consider a nontrivial interval $[\underline{\varphi}, \overline{\varphi}]$ with $\varphi(\mathbf{b}) \in [\underline{\varphi}, \overline{\varphi}]$, $Y(\mathbf{a}, \varphi') \neq \emptyset$ for all $\varphi' \in [\underline{\varphi}, \overline{\varphi}]$, and a function $g(\mathbf{a}, \varphi)$ that is constant on $[\underline{\varphi}, \overline{\varphi}]$. Choose $\mathbf{b} \in Y(\mathbf{a}, \underline{\varphi})$ and $\mathbf{b}' \in Y(\mathbf{a}, \overline{\varphi})$. By assumption $g(\mathbf{a}, \varphi(\mathbf{b})) = g(\mathbf{a}, \varphi(\mathbf{b}'))$. Suppose that $f(\mathbf{a}, \cdot) \equiv g(\mathbf{a}, \varphi(\cdot))$. Then $\{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}'\} \succ \mathbf{b}'$ and by SB $\{\mathbf{a}, \mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$. By claim 1.2, there exists \mathbf{c} such that $\{\mathbf{c}\} \succ \{\mathbf{a}\}$, $\{\mathbf{c}\} \succ \{\mathbf{c}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$, and $\{\mathbf{c}, \mathbf{b}'\} \succ \{\mathbf{a}, \mathbf{b}'\}$ (that is, $u(\mathbf{c}) - g(\mathbf{c}, \varphi(\mathbf{b}')) > u(\mathbf{a}) - g(\mathbf{a}, \varphi(\mathbf{b}'))$ and $\varphi(\mathbf{c}) \in (\varphi(\mathbf{b}), \varphi(\mathbf{b}'))$). But then $\{\mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}', \mathbf{c}\}$, and by the second part of P_3 $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, which is a contradiction.

Hence, g must be strictly increasing in its second argument whenever strict SB holds, and can be chosen to be strictly increasing anywhere else.||

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

Claim 1.6: 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 $(\hat{\mathbf{a}}, \hat{\varphi})$ to be a boundary point of S such that there exists $\mathbf{b}^* \in (k, \infty)^2$ with $\varphi(\mathbf{b}^*) = \hat{\varphi}$. For $\hat{\varphi} = \varphi(\hat{\mathbf{a}})$ define $g(\hat{\mathbf{a}}, \hat{\varphi}) := 0$ and for $\hat{\varphi} > \varphi(\hat{\mathbf{a}})$ define $g(\hat{\mathbf{a}}, \hat{\varphi}) := u(\hat{\mathbf{a}}) - \inf_{\mathbf{b} \in I_n(\hat{\varphi})} u(\mathbf{b})$.

Consider a sequence $\{(\mathbf{a}^n, \varphi^n)\} \rightarrow (\hat{\mathbf{a}}, \hat{\varphi})$ in S . Pick a sequence $\{\mathbf{b}^{n'}\}$ with $\mathbf{b}^{n'} \in Y(\mathbf{a}^n, \varphi^n) \forall n$. Define $\{b_1^n\} = \{\min[k + \frac{1}{n}, b_1^{n'}, b_1^*]\}$. Define b_2^n to be a solution to $\varphi(b_1^n, b_2^n) = \varphi^n$. By N_2 and N_3 , b_2^n is well-defined. Note that by observation (3) and P_4 , $\mathbf{b}^n = (b_1^n, b_2^n) \in Y(\mathbf{a}^n, \varphi^n)$. Lastly, let $\hat{\mathbf{b}}^n$ solve $\hat{b}_1^n = b_1^n$ and $\varphi(\hat{\mathbf{b}}^n) = \hat{\varphi}$. We have $U(\{\mathbf{a}^n, \mathbf{b}^n\}) = u(\mathbf{a}^n) - g(\mathbf{a}^n, \varphi^n)$. If in the switch point $\hat{\varphi} = \varphi(\hat{\mathbf{a}})$, then $U(\{\hat{\mathbf{a}}, \hat{\mathbf{b}}^n\}) = u(\hat{\mathbf{a}})$. By P_2 , $U(\{\mathbf{a}^n, \mathbf{b}^n\}) - U(\{\hat{\mathbf{a}}, \hat{\mathbf{b}}^n\}) \xrightarrow{n \rightarrow \infty} 0$, hence

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

If at the switch point $\hat{\varphi} > \varphi(\hat{\mathbf{a}})$, then $U(\{\hat{\mathbf{a}}, \hat{\mathbf{b}}^n\}) = u(\hat{\mathbf{b}}^n)$. Note that for any $\tilde{\mathbf{b}} \in I_n(\hat{\varphi})$ there exists N such that $\hat{b}_1^n < \tilde{b}_1$ for all $n > N$. Since u is more selfish than φ , $u(\hat{\mathbf{b}}^n) < u(\tilde{\mathbf{b}})$ for all $n > N$. This implies that $\lim_{n \rightarrow \infty} u(\hat{\mathbf{b}}^n) = \inf_{\mathbf{b} \in I_n(\hat{\varphi})} u(\mathbf{b})$. By the same continuity argument

$$\lim_{n \rightarrow \infty} g(\mathbf{a}^n, \varphi^n) = \lim_{n \rightarrow \infty} [u(\mathbf{a}^n) - u(\hat{\mathbf{b}}^n)] = u(\hat{\mathbf{a}}) - \inf_{\mathbf{b} \in I_n(\hat{\varphi})} u(\mathbf{b}) = g(\hat{\mathbf{a}}, \hat{\varphi}).$$

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, increasing in φ , and such that for any sequence $\{(\mathbf{a}^n, \varphi^n)\}$ in S , with $\{(\mathbf{a}^n, \varphi^n)\} \rightarrow (\hat{\mathbf{a}}, \hat{\varphi})$, 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.

We now show the necessity of the axioms to the representation. The necessity of P_1 and P_2 is obvious. We have already shown that if g is strictly increasing, then P_3 is satisfied. For the necessity of P_4 , let $H(\mathbf{b}) := \{\mathbf{a} : \varphi(\mathbf{a}) \geq \varphi(\mathbf{b})\} \cap \{\mathbf{a} : a_1 \geq b_1\}$. Since u is more selfish than φ and both are increasing, $H(\mathbf{b}) \subseteq U^\succ(\mathbf{b})$ hence $\mathbf{a} \in H(\mathbf{b})$ implies $u(\mathbf{a}) > u(\mathbf{b})$ or $\{\mathbf{a}\} \succ \{\mathbf{b}\}$. For $N_1 - N_3$, it would be sufficient to show that if $\mathbf{a} \succ_n \mathbf{c}$, then the representation provides a set C such that $\mathbf{c} \in C$ and $C \succ C \cup \{\mathbf{a}\}$. Let $L^{\succ n}(\mathbf{a}) := \{\mathbf{a}' : \varphi(\mathbf{a}') < \varphi(\mathbf{a})\}$. By Definition 3 (ii), $L^{\succ n}(\mathbf{a}) \cap U^\succ(\mathbf{a}) \neq \emptyset$. There are two cases to consider:

Case 1: If $\mathbf{c} \in L^{\succ n}(\mathbf{a}) \cap U^\succ(\mathbf{a})$, then $\max\{u(\mathbf{c}) - g(\mathbf{c}, \varphi(\mathbf{a})), u(\mathbf{a})\} < u(\mathbf{c})$, which implies that $\{\mathbf{c}\} \succ \{\mathbf{c}, \mathbf{a}\}$, so take $C = \{\mathbf{c}\}$.

Case 2: Suppose that $\mathbf{d} \in L^{\succ n}(\mathbf{a}) \cap L^\succ(\mathbf{a})$. The representation implies that there is $\mathbf{c} \in L^{\succ n}(\mathbf{a}) \cap U^\succ(\mathbf{a})$ such that $u(\mathbf{c}) - g(\mathbf{c}, \varphi(\mathbf{a})) > u(\mathbf{a})$. Then $u(\mathbf{c}) - g(\mathbf{c}, \varphi(\mathbf{a})) < \min\{u(\mathbf{c}) - g(\mathbf{c}, \varphi(\mathbf{d})), u(\mathbf{c})\}$, hence $C = \{\mathbf{c}, \mathbf{d}\}$.

To show P_4 , let $H(\mathbf{b}) := \{\mathbf{a} : \varphi(\mathbf{a}) > \varphi(\mathbf{b})\} \cap \{\mathbf{a} : a_1 > b_1\}$. Since u is more selfish than φ and both are weakly increasing, $H(\mathbf{b}) \subseteq U^\succ(\mathbf{b})$. Therefore, $\mathbf{a} \in H(\mathbf{b})$ implies $u(\mathbf{a}) > u(\mathbf{b})$ or $\{\mathbf{a}\} \succ \{\mathbf{b}\}$. This completes the proof of Theorem 1. ■

Proof of Theorem 2

Given Theorem 1, P_5 implies that $g(\mathbf{a}, \varphi(\mathbf{b})) < g(\mathbf{a}', \varphi(\mathbf{b})) \Leftrightarrow \varphi(\mathbf{a}) > \varphi(\mathbf{a}')$. Therefore, $g(\mathbf{a}, \varphi(\mathbf{b})) = g(\varphi(\mathbf{a}), \varphi(\mathbf{b}))$, where $g(\varphi(\mathbf{a}), \varphi(\mathbf{b}))$ is strictly decreasing in $\varphi(\mathbf{a})$. We now show that the function g is linear whenever it is relevant for choice (outside of that region it can always be chosen to be linear). If not, then there are $\mathbf{a}, \mathbf{a}', \mathbf{b}$ and \mathbf{b}' such that:

- (i) $u(\mathbf{a}) > u(\mathbf{a}) - g(\varphi(\mathbf{a}), \varphi(\mathbf{b})) > u(\mathbf{b})$
- (ii) $u(\mathbf{a}') > u(\mathbf{a}') - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})) > u(\mathbf{b})$
- (iii) $u(\mathbf{a}) > u(\mathbf{a}) - g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) > u(\mathbf{b}')$
- (iv) $u(\mathbf{a}') > u(\mathbf{a}') - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) > u(\mathbf{b}')$
- (v) $g(\varphi(\mathbf{a}), \varphi(\mathbf{b})) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})) > g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) > 0$.

Conditions (i) – (iv) imply that the value of g is relevant for the four possible pairings. Condition (v) captures the nonlinearity of g . The following claim implies that the nonlinearity of g must lead to context-dependent choice.

Claim 2.1: There are $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{a}}'$ such that $\varphi(\widehat{\mathbf{a}}) = \varphi(\mathbf{a})$, $\varphi(\widehat{\mathbf{a}}') = \varphi(\mathbf{a}')$, (ii) – (iv) hold where $\widehat{\mathbf{a}}$ replaces \mathbf{a} and $\widehat{\mathbf{a}}'$ replaces \mathbf{a}' , and

$$(g(\varphi(\mathbf{a}), \varphi(\mathbf{b})) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}))) > u(\widehat{\mathbf{a}}) - u(\widehat{\mathbf{a}}') > (g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))).$$

Proof: By conditions (i) – (iv), $\varphi(\mathbf{b}) > \varphi(\mathbf{a}')$ and

$u(\mathbf{a}') > \max\{u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})), u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))\}$. This observation, in addition to the property that u is more selfish than φ , implies that there is $\tilde{\mathbf{a}}'$ such that $\varphi(\tilde{\mathbf{a}}') = \varphi(\mathbf{a}')$ and $u(\tilde{\mathbf{a}}') = \max\{u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})), u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))\}$. By continuity, there is $\hat{\mathbf{a}}'$ such that $\varphi(\hat{\mathbf{a}}') = \varphi(\mathbf{a}')$ and for sufficiently small $\varepsilon > 0$, $u(\hat{\mathbf{a}}') = \max\{u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})), u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))\} + \varepsilon$.

We now show that for sufficiently small $\delta > 0$, there is $\hat{\mathbf{a}}$ such that $\varphi(\hat{\mathbf{a}}) = \varphi(\mathbf{a})$ and $u(\hat{\mathbf{a}}) = u(\hat{\mathbf{a}}') + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) + \delta$. A similar continuity argument as in the case of $\hat{\mathbf{a}}'$ applies if $u(\hat{\mathbf{a}}') + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) < u(\mathbf{a})$. To establish this, note that

$$\begin{aligned} u(\mathbf{a}) &> \max\{u(\mathbf{b}) + g(\varphi(\mathbf{a}), \varphi(\mathbf{b})), u(\mathbf{b}') + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}'))\} \\ &> \max\{u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})), u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))\}. \end{aligned}$$

If $u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) > u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}))$, then

$$\begin{aligned} u(\mathbf{a}) &> u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) = u(\hat{\mathbf{a}}') - \varepsilon \\ &> u(\hat{\mathbf{a}}') - \varepsilon + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')). \end{aligned}$$

Thus, for $\varepsilon > 0$ small enough, $u(\mathbf{a}) > u(\hat{\mathbf{a}}') + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))$.

If $u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) \leq u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}))$, then

$$\begin{aligned} u(\mathbf{a}) &> u(\mathbf{b}) + g(\varphi(\mathbf{a}), \varphi(\mathbf{b})) \\ &> u(\hat{\mathbf{a}}') - \varepsilon - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})) + g(\varphi(\mathbf{a}), \varphi(\mathbf{b})) \\ &> u(\hat{\mathbf{a}}') - \varepsilon - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})) + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')) \\ &= u(\hat{\mathbf{a}}') - \varepsilon + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')). \end{aligned}$$

Again, for $\varepsilon > 0$ small enough, $u(\mathbf{a}) > u(\hat{\mathbf{a}}') + g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))$.

Therefore, for $\delta < g(\varphi(\mathbf{a}), \varphi(\mathbf{b})) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})) - (g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')))$,

$$g(\varphi(\mathbf{a}), \varphi(\mathbf{b})) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})) > u(\hat{\mathbf{a}}) - u(\hat{\mathbf{a}}') > g(\varphi(\mathbf{a}), \varphi(\mathbf{b}')) - g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}')). \quad \parallel$$

Claim 2.1 implies that

$$u(\hat{\mathbf{a}}') - g(\varphi(\hat{\mathbf{a}}'), \varphi(\mathbf{b})) > \max\{u(\mathbf{b}), u(\hat{\mathbf{a}}) - g(\varphi(\hat{\mathbf{a}}), \varphi(\mathbf{b}))\}$$

and that

$$u(\widehat{\mathbf{a}}) - g(\varphi(\widehat{\mathbf{a}}), \varphi(\mathbf{b}')) > \max\{u(\mathbf{b}'), u(\widehat{\mathbf{a}}') - g(\varphi(\widehat{\mathbf{a}}'), \varphi(\mathbf{b}'))\}.$$

Consequently, $\{\widehat{\mathbf{a}}'\} \succ \{\widehat{\mathbf{a}}, \widehat{\mathbf{a}}', \mathbf{b}\} \succ \{\widehat{\mathbf{a}}, \mathbf{b}\}$, $\{\widehat{\mathbf{a}}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$ and $\{\widehat{\mathbf{a}}, \mathbf{b}'\} \succ \{\widehat{\mathbf{a}}, \widehat{\mathbf{a}}', \mathbf{b}'\}$, which violates P_6 . \parallel

Linearity of g together with the properties of g implied by Theorem 1, that is, g is increasing in its second argument and $g(\varphi, \varphi) = 0$, imply that $g(\varphi, \varphi') = \beta(\varphi - \varphi')$. Renormalization of u and g yields the representation in Theorem 2.

Obviously the representation implies P_5 . To verify that P_6 must hold, let $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\} \succ \{\mathbf{a}', \mathbf{b}\}$. According to the representation in Theorem 2, this implies that $u(\mathbf{a}) + \varphi(\mathbf{a}) > \max\{u(\mathbf{a}') + \varphi(\mathbf{a}'), u(\mathbf{b}) + \varphi(\mathbf{b})\}$. Then for any \mathbf{b}' , either (i) $u(\mathbf{b}') + \varphi(\mathbf{b}') < u(\mathbf{a}) + \varphi(\mathbf{a})$ and hence $\{\mathbf{a}, \mathbf{a}', \mathbf{b}'\} \succ \{\mathbf{a}', \mathbf{b}'\}$, or (ii) $u(\mathbf{b}') + \varphi(\mathbf{b}') \geq u(\mathbf{a}) + \varphi(\mathbf{a})$. Case (ii) implies that $u(\mathbf{b}') + \varphi(\mathbf{b}') > u(\mathbf{a}') + \varphi(\mathbf{a}')$. Then $U(\{\mathbf{a}', \mathbf{b}'\}) = u(\mathbf{b}') + \varphi(\mathbf{b}') - \max\{\varphi(\mathbf{b}'), \varphi(\mathbf{a}')\} \leq U(\{\mathbf{a}', \mathbf{b}'\})$ and hence $\{\mathbf{b}'\} \succeq \{\mathbf{a}', \mathbf{b}'\}$. This establishes P_6 . \blacksquare

Proof of Theorem 3

That the representation satisfies the axioms is easy to verify. For the other direction, we first show that our axioms $N_1 - N_4$ imply the axioms posed by Karni and Safra (1998).

Beside N_1 (weak order) and N_4 (their Hexagon Condition), Karni and Safra require the following axioms:

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

Independence is implied since by N_2 , $(a_1, a) \succeq_n (b_1, a) \Leftrightarrow a_1 \geq b_1 \Leftrightarrow (a_1, b) \succeq_n (b_1, b)$ for all b .

Restricted Solvability: If $(a_1, a_2) \succeq_n (b_1, b_2) \succeq_n (a'_1, a_2)$ then there is x such that $(b_1, b_2) \sim_n (x, a_2)$. And if $(a_1, a_2) \succeq_n (b_1, b_2) \succeq_n (a_1, a'_2)$ then there is y such that $(b_1, b_2) \sim_n (a_1, y)$.

Restricted Solvability is immediately implied by N_3 .

A sequence $\{a_i\}$ is a standard sequence, if for some $a \neq b$, $(a_i, a) \sim_n (a_{i+1}, b)$ 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.

Archimedean property: every bounded standard sequence is finite.

To show that the Archimedean 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 k and ∞ . Let \bar{a} and \underline{a} be the least upper

bound and greatest lower bound respectively.

Case 1, $a > b$. By N_3 , there exists x such that $(\bar{a}, a) \sim_n (x, b)$. By N_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 subsequence $\{a_{ik}\}$ that converges to x . In particular, there exists K such that for $k > K$, $x - a_{ik} < \varepsilon := \frac{x - \bar{a}}{2}$, which is a contradiction.

Case 2, $a < b$. By N_2 , $(\underline{a}, a) \prec_n (\underline{a}, b)$. 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 $(a_i, a) \prec_n (\underline{a}, b)$. Since \underline{a} is the greatest lower bound, N_2 implies that $(\underline{a}, b) \prec_n (a_{i+1}, b)$. Therefore, $(a_i, a) \prec_n (a_{i+1}, b)$, which is a contradiction.

Essentiality: Not $(a', b) \sim_n (a, b)$ for all b , or $(a, b) \sim_n (a, b')$ for all a .

Essentiality is immediately implied by N_3 (ii).||

Karni and Safra (1998) show (see the lemma in their paper) that their axioms imply the axioms of Krantz, Luce, Suppes and Tversky (1971). Hence, an additively separable representation exists, where the utilities are unique up to translation and a common linear transformation (See Theorem 2 in Chapter 6 of Krantz *et al.* (1971). 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 quasilinear 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.¹⁹ Quasilinearity implies that there is an increasing continuous function $\xi : (k, \infty) \rightarrow \mathbb{R}$, such that $\varphi(\mathbf{a}) := \xi(a_1) + \gamma(a_2)$ represents \succ_n . Given the additively separable representation, define $v_1(a_1) := \exp(\xi(a_1))$ and $v_2(a_2) := \exp(\gamma(a_2))$. Then $v_1, v_2 : (k, \infty) \rightarrow \mathbb{R}_{++}$ are increasing and continuous, and, if we redefine $\varphi(\mathbf{a}) := v_1(a_1)v_2(a_2)$, it represents \succ_n . Note that if $\varphi'(\mathbf{a}) = v_1'(a_1)v_2'(a_2)$ also represents \succ_n , then there are α, β_1, β_2 , all strictly positive, such that $v_1' = \beta_1 v_1^\alpha$ and $v_2' = \beta_2 v_2^\alpha$. The linear structure of Theorem 2 further requires that $\alpha = 1$. ■

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, 1652-1678.

¹⁹For brevity, we do not reproduce their argument in more detail in this paper. A direct proof of Theorem 3 is available upon request.

- [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, 817-870.
- [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, 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 (2009) "Temptation Driven Preferences." *Review of Economic Studies*, 76, 937-971.
- [13] Epstein, Larry G. and Igor Kopylov (2007) "Cold Feet" *Theoretical Economics*, 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] Gehm, T. L., and Scherer, K. R. (1988) "Relating Situation Evaluation to Emotion Differentiation: Nonmetric Analysis of Cross-Cultural Data." In K. R. Scherer (Ed.) *Facets of Emotion* (61-77). Hillsdale, NJ: Lawrence Erlbaum Associates.
- [18] Gul, Faruk and Wolfgang Pesendorfer (2005) "The Simple Theory of Temptation and Self-Control." mimeo
- [19] ——— (2001) "Temptation and Self Control." *Econometrica*, 69, 1403-1435
- [20] 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.
- [21] 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, 200-254.
- [22] Johannesson, Magnus and Bjorn Persson (2000) "Non-reciprocal altruism in dictator games." *Economics Letters*, 69, 137-142.
- [23] 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.
- [24] 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.
- [25] Koch, K. Alexander and Hans-Theo Norman (2008) "Giving in Dictator Games: Regard for Others or Regard by others?" *Southern Economic Journal*, 75, 223-231.
- [26] Krantz, David H., R. Duncan Luce, Patrick C. Suppes and Amos Tversky (1971) "Foundations of Measurements, Vol 1." Academic Press, New York.
- [27] Lazear, Edward P., Ulrike Malmendier and Roberto A. Weber (2006) "Sorting in Experiments with Application to Social Preferences." mimeo.
- [28] Lewis, Helen B. (1971), "Shame and Guilt in Neurosis." New York: International Universities Press.

- [29] Luce, R. Duncan and John W. Tukey (1964) "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology*, 1,1-27.
- [30] Miller, Dale T. (1999) "The Norm of Self-Interest." *American Psychologist*, 54, 1053-1060.
- [31] Nash, John F. (1953) "Two-Person Cooperative Games." *Econometrica*, 21, 128-140.
- [32] ——— (1950) "The Bargaining Problem." *Econometrica*, 18, 155-162.
- [33] Neilson, William S. (2009) "A Theory of Kindness, Reluctance, and Shame in Dictator Games." *Games and Economic Behavior*, 66, 394-403.
- [34] Noor, Jawwad and Norio Takeoka (2008) "Menu-Dependent Self-Control" mimeo.
- [35] Oberholzer-Gee, Felix and Reiner Eichenberger (2008) "Fairness in Extended Dictator Game Experiments," *The B.E. Journal of Economic Analysis & Policy*: Vol. 8: Iss. 1 (Contributions), Article 16. Available at: <http://www.bepress.com/bejeap/vol8/iss1/art16>
- [36] Osborne, Martin J. and Ariel Rubinstein (1994) "A course in Game Theory." MIT press, Cambridge, Massachusetts.
- [37] Olszewski, Wojciech (2009) "A Model of Consumption-Dependent Temptation." *Theory and Decision*.
- [38] 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.
- [39] Tangney, June P., and Ronda L. Dearing. (2002) "Shame and Guilt." The Guilford Press, New York.