Theoretical Economics 7 (2012), 99–124 1555-7561/20120099

Ashamed to be selfish

David Dillenberger
Department of Economics, University of Pennsylvania

Philipp Sadowski
Department of Economics, Duke University

We study a decision maker (DM) who has preferences over choice problems, which are sets of payoff allocations between herself and a passive recipient. An example of such a set is the collection of possible allocations in the classic dictator game. The choice of an allocation from the set is observed by the recipient, whereas the choice of the set itself is not. Behaving selfishly under observation, in the sense of not choosing the normatively best allocation, inflicts shame on the DM. We derive a representation that identifies the DM's private ranking of allocations, her subjective norm, and her shame. The normatively best allocation can be further characterized as the Nash solution of a bargaining game induced by the second-stage choice problem.

Keywords. Shame, selfishness, subjective norm, dictator game, Nash bargaining solution.

JEL classification. 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

David Dillenberger: ddill@sas.upenn.edu
Philipp Sadowski: p.sadowski@duke.edu

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 co-editor 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.

DOI: 10.3982/TE674

example, Fehr and Schimdt 1999, Andreoni and Miller 2002, and Charnes and Rabin 2002).

Recent experiments, however, have challenged this interpretation. For example, Dana et al. (2006) study a variant of the same dictator game, where the dictator is given the option to exit the game before the recipient learns it is being played. In case she opts out, she is given a prespecified amount of money and the recipient gets nothing. About 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 are 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 a person's behavior may depend on whether it is observed by someone who is directly affected by it. In this case, we say that the choice is observed. Otherwise, we say that the decision maker chooses in private. We identify *shame* as the moral cost 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.[1] We refer to the criterion that determines whether an alternative causes shame when it is not chosen as the individual's (*subjective*) *norm*.

The significance of the effect of observability on behavior is supported by additional evidence. In a follow-up to the experiment cited above, Dana et al. report that only 1 out of 24 potential dictators exits the game, if he is assured that the recipient remains ignorant about the way his payoffs are determined. Similarly, Pillutla and Murnighan (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 et al. (2006) as well as Broberg et al. (2007) 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 to exit the dictator game: they find that the mean exit reservation price equals 82% of the dictator game endowment.

[1] Our notion of shame is closely related to that of embarrassment and guilt. We use the word shame to highlight that the emotional cost experienced by the decision maker (DM) is caused by her own actions. In particular, it is not triggered by others' actions, which place the DM in a socially awkward situation, as is the case with embarrassment. 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).

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 is 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 that 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, the DM chooses an alternative from the menu. This choice is observed, in the sense that the recipient is aware of the menu available to the DM.[2,3] The 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 the 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 the DM's choice behavior.

1.2 *Illustration of results*

Denote a typical menu by $A = \{\mathbf{a}, \mathbf{b}, \ldots\} = \{(a_1, a_2), (b_1, b_2), \ldots\}$, where the first and second components of each alternative are, respectively, the private payoff for the DM and for the recipient. To illustrate our results, consider the special case of our representations

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

where u and φ are increasing in all arguments. The function u is a utility function over allocations and the function φ represents the DM's norm. The function g is interpreted as the shame from choosing $\mathbf{a}$ in face of the alternative that maximizes the DM's norm. It is decreasing in its first argument, increasing in its second argument, and satisfies $g(z, z) = 0$, that is, there is no shame if a normatively best alternative is chosen.

This representation captures the tension between the 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

[2]The key distinction between the two stages is not the passage of physical time, but that the recipient observes the choice set that is available in stage two and not the set of menus that are available in stage one. This is in contrast to most other models of choice over menus, where subjective uncertainty might resolve or temptation may kick in over time.

[3]If the exit option is chosen in the aforementioned experiment by Dana et al., the recipient is unaware that there is a dictator who can chose another allocation. In their experiment, the recipient is further unaware that another person is involved at all. It would be interesting to see whether informing the recipient that two people participate in the experiment and that the other person receives \$9 would change the experimental findings.

degenerate menu $\{\mathbf{a}\}$. A degenerate menu leaves the DM with no choice to be made under observation. We contend that there is no room for shame in this situation; the DM's ranking over degenerate menus can be thought of as her private ranking of allocations. The second component is shame. It represents the cost the 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 the DM to be more *selfish* in private: among normatively equally good alternatives, she 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 the DM's subjective norm. The relation $\succ_n$ 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 $g(\varphi(\mathbf{a}), \max_{\mathbf{b}\in A}\varphi(\mathbf{b}))$. 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 our most general representation, which captures the intuition discussed thus far.

Representations similar to (1) are studied extensively in the literature on temptation, starting with the work of Gul and Pesendorfer (2001, henceforth GP), who 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 is 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 linearity implied by the independence axiom means 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]

[4]In the context of temptation, Noor and Takeoka (2011) explore the connection between the independence axiom and WARP, and suggest relaxations of the independence axiom that allow for a 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 (2011) studies preferences over subsets of a finite set of deterministic outcomes and finds a representation where both choice and temptation are context-dependent.

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 the DM's preferences over menus, one cannot distinguish between a standard DM and a DM who is susceptible to shame.

We further specify the 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 the 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 (Nash 1950) of an associated game. Since the utility functions used to generate this game are subjective, so is the norm.

Example. In representation (1), let $u(\mathbf{a}) = 2a_1$, $\varphi(\mathbf{a}) = v_1(a_1)v_2(a_2) = (a_1 + 1)(a_2 + 1)$, and $g(x, y) = y - x$. In the experiment by Dana et al. mentioned above, only whole dollar amounts are possible allocations. The set $A = \{(10, 0), (9, 1), (8, 2), \ldots, (0, 10)\}$ 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, $(5, 5)$ is the outcome of the bargaining game. Its normative value is $6 \cdot 6 = 36$. To trade off shame with selfishness, the 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 the 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 the 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 $X = (k, \infty)$ be an open interval of monetary prizes.[5] Let $\mathcal{K}$ be the set of all finite subsets of X^2. Any element $A \in \mathcal{K}$ is a finite set of alternatives.

[5] Whenever $k < 0$, $\mathbb{R}_+ \subset X$. In particular, $X = \mathbb{R}$ for $k = -\infty$.

A typical alternative $\mathbf{a} = (a_1, a_2)$ is a payoff pair, where a_1 is the private payoff for the DM and a_2 is the private payoff allocated to the (potentially anonymous) other player, the recipient. Endow $\mathcal{K}$ with the topology generated by the Hausdorff metric, which is defined for any pair of nonempty sets, $A, B \in K$, as

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

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

Let $\succeq$ be a binary relation over $\mathcal{K}$. The symmetric ($\sim$) and asymmetric ($\succ$) parts of $\succeq$ are defined in the usual way. Our first two axioms on $\succeq$ are standard.

Axiom P_1 (Weak order). *The relation $\succeq$ is complete and transitive.*

Axiom P_2 (Continuity). *The relation $\succeq$ 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.

Axiom P_3 (Strong left betweenness). *If $A \succeq B$, then $A \succeq A \cup B$. Further, if $A \succ B$ and there 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 sets in only after two allocations are sufficiently different according to, say, the Euclidean distance on X^2.

Definition 1. We say that the 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 the DM is susceptible to shame, then we say that the 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]

[6] This is the left betweenness axiom. It appears in Dekel et al. (2009).

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

The relationship $A \succ A \cup \{\mathbf{b}\}$ implies that $\mathbf{b}$ adds to the shame incurred by the original choice in A. In this case, $\mathbf{b}$ is normatively better than any alternative in A and, in particular, $\mathbf{b} \succ_n \mathbf{a}$. The induced relations $\succeq_n$ and $\sim_n$ are defined as

$$\mathbf{a} \succeq_n \mathbf{b} \quad \Leftrightarrow \quad \mathbf{b} \not\succ_n \mathbf{a}$$

$$\mathbf{a} \sim_n \mathbf{b} \quad \Leftrightarrow \quad \text{both } \mathbf{b} \not\succ_n \mathbf{a} \text{ and } \mathbf{a} \not\succ_n \mathbf{b}.$$

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.

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

Axiom N_1 rules out situations in which there are two alternatives, a and 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. If this were the case, then we would have $A \setminus \{\mathbf{a}\} \succ A$ and $B \setminus \{\mathbf{b}\} \succ B$, which means, in contradiction to the asymmetry of $\succ_n$, that both $\mathbf{b} \succ_n \mathbf{a}$ and $\mathbf{a} \succ_n \mathbf{b}$. 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, only one alternative in each menu, one of the normatively best, is responsible for shame.

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

This 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.

Axiom 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 $(y, a_2) \succ_n (b_1, b_2)$.*

This 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, Axiom N_3 requires $\succ_n$ never to be satiated in either payoff.

The next axiom concerns the DM's preferences over singleton sets. A singleton set is a degenerate menu that contains only one feasible allocation. Since degenerate menus leave the DM with no choice to be made under observation, choosing between two singleton sets reveals the DM's private ranking of the allocations. Although the private ranking of allocations might incorporate other-regarding preferences (such as altruism), we

[8] Since 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.

assume the DM to be more selfish in private: if alternative **a** gives higher payoffs to the DM than alternative **b** and is also normatively better, then **a** must be preferred to **b** in private.

Axiom P_4 (Selfishness). *If* $a_1 > b_1$ *and* $\mathbf{a} \succ_n \mathbf{b}$, *then* $\{\mathbf{a}\} \succ \{\mathbf{b}\}$.[9,10]

Definition 3. Let f and h be two functions on X^2. We say that h is *more selfish than* f if for all **a** and for all Δ_1 and Δ_2 such that $(a_1 - \Delta_1, a_2 - \Delta_2) \in X^2$,

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

(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 a level curve of h in the (a_1, a_2) plane is, at any point, weakly greater than that of f.[11]

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

It is clear that if $\succ_n$ satisfies Axioms N_1–N_3, then any function $\varphi : X^2 \to \mathbb{R}$ that represents it should be a subjective norm function.

Theorem 1. *Relations* $\succeq$ *and* $\succ_n$ *satisfy Axioms* 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* $u : X^2 \to \mathbb{R}$, *which is weakly increasing and more selfish than* φ, *and (iii) a continuous function* $g : X^2 \times \varphi(X^2) \to \mathbb{R}$, *which is strictly increasing in its second argument and satisfies* $g(\mathbf{a}, x) = 0$ *whenever* $\varphi(\mathbf{a}) = x$, *such that the function* $U : K \to \mathbb{R}$, *defined as*

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

represents $\succeq$ *and* φ *represents* $\succ_n$.

[9] Coupled with continuity (Axiom P_2), Axiom P_4 implies that if $a_1 > b_1$ and $\mathbf{a} \sim_n \mathbf{b}$, then $\{\mathbf{a}\} \succeq \{\mathbf{b}\}$. That is, among equally good alternatives according to the normative ranking, the DM weakly prefers the one that gives her the highest private payoff.

[10] Definition 2 (of $\succ_n$) implies that if, in fact, $\mathbf{a} \sim_n \mathbf{b}$, 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 Axiom N_1 non-refutable whenever the violation involves indifference. Accepting Axiom N_1, Axiom P_4, which relates $\succ$ and $\succ_n$, is refutable. Axiom N_2 can be separated into a weak version (in which the implication is 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 contains an existential quantifier and hence is not refutable independently of the induced nature of $\succ_n$.

[11] If h and f were differentiable and the inequalities in both items (i) and (ii) were strict, then the condition would be $(f_1/f_2)(\mathbf{a}) < (h_1/h_2)(\mathbf{a})$ for all **a**, where f_i denotes the partial derivative of f with respect to its ith component. In this case, the definition of *more selfish than* coincides with the definition of less altruistic than in Cox et al. (2008).

Theorem 1 provides a representation of the 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 the 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 (nonpositive). The other max operator implies that the DM's utility never lies below $u(\mathbf{b})$, the utility of the normatively best allocation. Put differently, any deviations by the DM from choosing the normatively best allocation are in her own favor. Therefore, it is being selfish, and not being too generous, that triggers shame. The magnitude of shame 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, $U : K \to \mathbb{R}$ and $\varphi : X^2 \to \mathbb{R}$, respectively. The combination of Axioms N_2 and N_3 implies that φ is a subjective norm function. Note that by the asymmetry part of Axiom 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 Axioms 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 2. 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 Axiom 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 $\{\mathbf{c}\} \not\succ \{\mathbf{a}, \mathbf{c}\}$. 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 $\mathbf{a} \not\succ_n \mathbf{c}$ for all $\mathbf{c}$ and, in particular, for $\mathbf{c} < \mathbf{a}$, violating Axiom 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. Last, 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 $[\underline{\varphi}, \overline{\varphi}]$. Then we can find $\mathbf{b}$ and $\mathbf{b}'$ such that $\varphi(\mathbf{b}) = \underline{\varphi}$, $\varphi(\mathbf{b}') = \overline{\varphi}$, 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 Axiom 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

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

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 she 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 the DM's second-stage choice rule is set-independent, that is, it satisfies WARP.

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

AXIOM 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 the DM chooses ($\mathbf{a}$) and the alternative she deems to be 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 Axioms P_1–P_5 and N_1–N_3, an additional assumption is equivalent to a set-independent choice ranking.

AXIOM 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 $\mathbf{a}'$ is never the unique choice in the face of $\mathbf{a}$.

THEOREM 2. *Relations* $\succeq$ *and* $\succ_n$ *satisfy Axioms* 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*

$u : X^2 \to \mathbb{R}$, which is weakly increasing and more selfish than φ, such that the function $U : K \to \mathbb{R}$, defined as

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

represents $\succeq$ and φ represents $\succ_n$.

The representation in Theorem 2 suggests a choice criterion that is independent of the choice problem: the 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.[12] A natural question is, To what extent can one elicit the 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 (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 the DM's normative ranking: $\mathbf{b} \succ_a \mathbf{a}$ implies $\mathbf{b} \succ_n \mathbf{a}$. Furthermore, enough violations of WARP fully identify $\succ_n$.[13] If there are no violations of WARP, then observing the 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 the 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.

Axiom 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 $(a_1', b_2) \sim_n (a_1, b_2')$.*

[12] The remark is not specific to our model: it generalizes to all models that study preferences over sets of alternatives.

[13] The relation $\succ_n$ 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 Axiom 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 $\mathbf{b} \succ_n \mathbf{a} \Leftrightarrow \mathbf{b} \succ_a \mathbf{a}$.

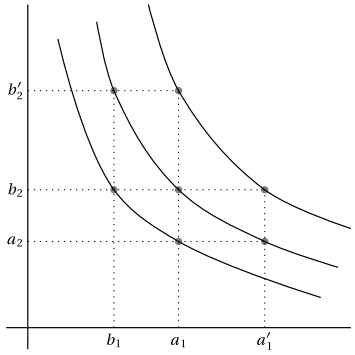

FIGURE 1. Independent normative contributions.

This axiom is illustrated in Figure 1. If $a_1 = a_1'$ or $b_2 = b_2'$, this axiom is implied by Axioms N_1 and 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 et al. 1971).

AXIOM N_4' (Strong independent normative contributions). *The relationships* $(a_1, a_2) \sim_n (b_1, b_2)$ *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 the 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 relative intensities. In other words, she does not make interpersonal comparisons of utilities.[14] The qualifier in Axiom N_4' establishes that the 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). Axiom 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 Axiom 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 Axiom N_4, but not under Axiom N_4'.

Krantz et al. (1971) provide an additive representation based on Axiom N_4'. Karni and Safra (1998) demonstrate that the weaker condition, Axiom N_4, implies Axiom N_4' in the context of their axioms. The next theorem is based on those results.

THEOREM 3. *The relation* $\succ_n$ *is continuous and satisfies Axioms N1–N4 if and only if there are continuous, strictly increasing, and unbounded functions* $v_1, v_2 : X \to \mathbb{R}_{++}$, *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 the DM is concerned about: she behaves as if she has in mind two positive, increasing,

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

and unbounded utility functions on X: one for herself and one for the recipient. By mapping the alternatives within each set into the associated utility space, any choice set induces a finite bargaining game, where the imaginary disagreement point corresponds to zero utility payoffs.[15] The 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, the DM might refer to the prediction of a symmetric mechanism for generating allocations. For example, the DM might ask what the allocation would be if both she and the recipient 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; the DM is not ashamed of payoffs, but of using her stronger position to distribute the gains. The intuitive and possibly descriptive appeal of the NBS in many bargaining situations then makes it normatively appealing to the DM.[16]

5. Related literature

Other-regarding preferences are considered extensively in economic literature. In particular, inequality aversion, as studied by Fehr and Schimdt (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 the 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 Bénabou and Tirole (2006).

Neilson (2009) 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: the DM might prefer a smaller menu over a larger menu either because avoiding shame compels her to be generous when choosing from the larger menu or because being selfish when choosing from the larger menu bears the cost of shame.

The structure of our representation resembles the representation of preferences with self-control under temptation, as axiomatized in GP. They study preferences over sets

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

[16] The descriptive value of the NBS has been tested empirically. For a discussion, see Davis and Holt (1993, pp. 247–255). Further, multiple seemingly natural implementations of NBS have been proposed (Nash 1953, Osborne and Rubinstein 1994).

of lotteries and show that their axioms lead to a representation of the form

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

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} represents the temptation ranking. While the two works yield representations with a similar structure, their domains—and therefore the axioms—are different. They 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 is not plausible. However, one of GP's axioms is the set betweenness axiom, $A \succcurlyeq B \Rightarrow A \succcurlyeq A \cup B \succcurlyeq B$. We show that our axioms, strong left betweenness (Axiom P_3) and normative ranking (Axiom 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.[17] 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 the DM privately weakly 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. If this preference is strict, it contradicts the central notion that the DM is ashamed to be selfish, which is captured by the assumption that she is more selfish in private (Axiom P_4). Put differently, the two models qualitatively disagree—unless the normative ranking coincides with the private ranking—once the recipient gets higher payoff than the DM. 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. Regardless whether of such rankings are reasonable, many other private and normative rankings are excluded, which defies the purpose of eliciting these rankings from behavior.

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.

[17] We thank a referee for suggesting this connection between the two models.

Last, 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 et al. (1998), 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 (2008), 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 et al. (2001) 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."[18] 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.

Appendix

A.1 *Proofs*

Proof of Theorem 1. Let $U : K \to \mathbb{R}$ be a continuous function that represents $\succ$. Without loss of generality, 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 $\varphi : X^2 \to \mathbb{R}$ be a continuous function that represents $\succ_n$. By Axiom N_2, φ is strictly increasing. Because $\succ_n$ is continuous, Axiom N_3 immediately implies that if $(a_1, a_2) \not\succ_n (b_1, b_2)$, 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 Axiom N_3 without further discussion.

Claim 1 (Right betweenness). *The implication $A \succeq B \Rightarrow A \cup B \succeq B$ holds.*

Proof. There are two cases to consider:

Case 1. For all $\mathbf{a} \in A$, $\exists \mathbf{b} \in B$ such that $\mathbf{b} \succ_n \mathbf{a}$. Let $A = \{\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^N\}$ and $C_0 = B$. Define $C_n = C_{n-1} \cup \{\mathbf{a}^n\}$ for $n = 1, 2, \ldots, N$. According to Axiom N_1, for all $\mathbf{a}^n$ there exists $\mathbf{b} \in B$ such that $\mathbf{a}^n \not\succ_n \mathbf{b}$. By the definition of $\succ_n$, $C_{n-1} \not\succ C_n$. By negative transitivity of $\succ$, $C_0 \not\succ C_N$ or $A \cup B \succeq B$.

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

Case 2. There exists $\mathbf{a} \in A$ such that $\mathbf{a} \succ_n \mathbf{b}$ $\forall \mathbf{b} \in B$. Let $B = \{\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^M\}$. Define $C_0 = A$ and $C_m = C_{m-1} \cup \{\mathbf{b}^m\}$ for $m = 1, 2, \ldots, M$. By the definition of $\succ_n$ and Axiom N_1, $\forall C$ such that $\mathbf{a} \in C$, $C \not\succ C \cup \{\mathbf{b}^m\}$. Hence, $C_{m-1} \not\succ C_m$. By negative transitivity of $\succ$, $C_0 \not\succ C_M$ or $A \cup B \succeq A \succeq B$, hence $A \cup B \succeq B$. □

Combining Claim 1 with Axiom 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's Lemma 2, which states that any set is indifferent to a specific two-element subset of it.

Lemma 1 (GP's 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, Axiom P_4, and the monotonicity of φ to show that u is more selfish than φ according to Definition 3.

Claim 2. *The function 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 Axiom P_2 and $b_1 > a_1$, a contradiction to Axiom 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 Axiom P_4, $U^{\succ}(\mathbf{a}) \subseteq U^{\succ_n}(\mathbf{a})$.

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

To prove 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, $\{\mathbf{c}\} \not\succ \{\mathbf{c}, \mathbf{a}\}$, 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 $C \succ C \cup \{\mathbf{a}\}$.

To prove 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, $\mathbf{a} \not\succ_n \mathbf{c}$ for all $\mathbf{c}$ and, in particular, for $\mathbf{c} = (a_1 - \varepsilon, a_2 - \varepsilon)$, contradicting the strict Pareto criterion (Axiom N_2). □

Claim 3. *(i) The implication $[\varphi(\mathbf{a}) < \varphi(\mathbf{b})$ and $\{\mathbf{a}\} \succ \{\mathbf{b}\}] \Leftrightarrow \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$ holds.*

(ii) The implication $[\varphi(\mathbf{a}) < \varphi(\mathbf{b})$ and $\{\mathbf{b}\} \succeq \{\mathbf{a}\}] \Rightarrow \{\mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}\}$ holds.

(iii) The implication $[\varphi(\mathbf{a}) = \varphi(\mathbf{b})$ *and* $\{\mathbf{a}\} \succ \{\mathbf{b}\}] \Rightarrow \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$ *holds.*

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 Axiom 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 $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$, Axiom 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) Relationship $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ implies by SB that $\{\mathbf{a}\} \succeq \{\mathbf{a}, \mathbf{b}\}$. As $\varphi(\mathbf{a}) = \varphi(\mathbf{b})$, part (i) implies that $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$ 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, $(\mathbf{b}^*(A), \mathbf{a}^*(A))$ also solves $\min_{\mathbf{b}' \in A} \max_{\mathbf{a}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$.)

CLAIM 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 $\mathbf{c} \succ_n \mathbf{b}$, 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.

O1. The relationships $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}\}$, and $\mathbf{b} \succ_n \mathbf{c}$ imply $\{\mathbf{a}, \mathbf{c}\} \succeq \{\mathbf{a}, \mathbf{b}\}$.

O2. The relationships $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}\} \succ \{\mathbf{c}\}$, and $\mathbf{b} \sim_n \mathbf{c}$ imply $\{\mathbf{a}, \mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}\}$.

O3. The relationships $\mathbf{b} \in Y(\mathbf{a}, \varphi)$, $\{\mathbf{b}\} \succ \{\mathbf{b}'\}$, and $\mathbf{b}' \sim_n \mathbf{b}$ imply $\mathbf{b}' \in Y(\mathbf{a}, \varphi)$.

O4. 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 O1 does 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 O2 does 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 O3 does 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 O4, assume $\{\mathbf{a},\mathbf{b}'\} \succ \{\mathbf{b}'\}$. Then, by Claim 3(i), $\{\mathbf{a}\} \succ \{\mathbf{a},\mathbf{b}'\} \succ \{\mathbf{b}'\}$, and by observation O2, $\{\mathbf{a},\mathbf{b}'\} \sim \{\mathbf{a},\mathbf{b}\}$. If, alternatively, $\{\mathbf{a},\mathbf{b}'\} \sim \{\mathbf{b}'\}$, 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 $f : X^2 \times X^2 \to \mathbb{R}$ such that $f(\mathbf{a},\mathbf{b}) = u(\mathbf{a}) - \widetilde{U}(\mathbf{a},\mathbf{b})$, where $\widetilde{U} : X^2 \times X^2 \to \mathbb{R}$ is a function satisfying[19]

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

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

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

as otherwise we have

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

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

CLAIM 5. *There exists a function $\widetilde{U}$ satisfying the condition specified above such that $f(\mathbf{a},\mathbf{b}) = g(\mathbf{a},\varphi(\mathbf{b}))$ for some $g : X^2 \times \mathbb{R} \to \mathbb{R}$, 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 $(\mathbf{a},\mathbf{b})$ and $(\mathbf{a},\mathbf{c})$, and identify the restrictions imposed by all combinations of corresponding values of u and ϕ. Consider first the pair $(\mathbf{a},\mathbf{b})$.

(i) Take $u(\mathbf{a}) > u(\mathbf{b})$ and $\phi(\mathbf{b}) > \phi(\mathbf{a})$. By Claim 3(i), $\{\mathbf{a}\} \succ \{\mathbf{a},\mathbf{b}\}$, so that $f(\mathbf{a},\mathbf{b}) > 0$.

[19] Note that

$$\max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} U(\{\mathbf{a},\mathbf{b}\}) = \max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} \Big[\max_{\mathbf{a}' \in \{\mathbf{a},\mathbf{b}\}} \min_{\mathbf{b}' \in \{\mathbf{a},\mathbf{b}\}} \widetilde{U}(\mathbf{a}',\mathbf{b}') \Big] = \max_{\mathbf{a} \in A} \min_{\mathbf{b} \in A} \widetilde{U}(\mathbf{a},\mathbf{b}).$$

(ii) Take $u(\mathbf{a}) \leq u(\mathbf{b})$ and $\phi(\mathbf{b}) > \phi(\mathbf{a})$. By Claim 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) Take $u(\mathbf{a}) > u(\mathbf{b})$ and $\phi(\mathbf{b}) = \phi(\mathbf{a})$. By Claim 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 $f(\mathbf{a}, \mathbf{b}) = 0 = f(\mathbf{b}, \mathbf{a})$.

(iv) Take $u(\mathbf{a}) \leq u(\mathbf{b})$ and $\phi(\mathbf{b}) = \phi(\mathbf{a})$. By Claim 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 $Y(\mathbf{a}, \varphi(\mathbf{b})) = \varnothing$. Let $\underline{u} := \inf_{\mathbf{a}' \in X} u(\mathbf{a}')$ (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 $Y(\mathbf{a}, \varphi(\mathbf{c})) \neq \varnothing$, then $f(\mathbf{a}, \mathbf{c}) < u(\mathbf{a}) - u(\mathbf{c}) < u(\mathbf{a}) - \underline{u} < f(\mathbf{a}, \mathbf{b})$. If $Y(\mathbf{a}, \varphi(\mathbf{c})) = \varnothing$, 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 $Y(\mathbf{a}, \varphi(\mathbf{b})) \neq \varnothing$. 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 O2, $f(\mathbf{a}, \mathbf{b}') = f(\mathbf{a}, \mathbf{b}'')$ $\forall \mathbf{b}', \mathbf{b}'' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$. If $\mathbf{b}'' \in I_n(\varphi(\mathbf{b})$ but $\mathbf{b}'' \notin Y(\mathbf{a}, \varphi(\mathbf{b}))$, then, by observation O4, 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 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 Axiom 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 3(i) and observation O1 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 O4 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 $\varphi(\mathbf{b}) \in [\underline{\varphi}, \overline{\varphi}]$, $Y(\mathbf{a}, \varphi') \neq \varnothing$ 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 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 Axiom 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 $S := \{(\mathbf{a}, \varphi) : Y(\mathbf{a}, \varphi) \neq \varnothing\}$. Note that S is an open set.

CLAIM 6. *There is $g(\mathbf{a},\varphi)$, which is continuous.*

PROOF. If $Y(\mathbf{a},\varphi)\neq\varnothing$, 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 $Y(\mathbf{a},\varphi)=\varnothing$, then $\varphi\leq\varphi(\mathbf{a})$ implies $g(\mathbf{a},\varphi)\leq 0$, while $\varphi>\varphi(\mathbf{a})$ implies $g(\mathbf{a},\varphi)\geq u(\mathbf{a})-u(\mathbf{0})$. Define a switch point $(\widehat{\mathbf{a}},\widehat{\varphi})$ to be a boundary point of S such that there exists $\mathbf{b}^*\in X^2$ with $\varphi(\mathbf{b}^*)=\widehat{\varphi}$. For $\widehat{\varphi}=\varphi(\widehat{\mathbf{a}})$, define $g(\widehat{\mathbf{a}},\widehat{\varphi}):=0$ and for $\widehat{\varphi}>\varphi(\widehat{\mathbf{a}})$, define $g(\widehat{\mathbf{a}},\widehat{\varphi}):=u(\widehat{\mathbf{a}})-\inf_{\mathbf{b}\in I_n(\widehat{\varphi})}u(\mathbf{b})$.

Consider a sequence $\{(\mathbf{a}^n,\varphi^n)\}\to(\widehat{\mathbf{a}},\widehat{\varphi})$ in S. Pick a sequence $\{\mathbf{b}^{n\prime}\}$ with $\mathbf{b}^{n\prime}\in Y(\mathbf{a}^n,\varphi^n)\ \forall n$. Define $\{b_1^n\}=\{\min[k+1/n,b_1^{n\prime},b_1^*]\}$. Define b_2^n to be a solution to $\varphi(b_1^n,b_2^n)=\varphi^n$. By Axioms N_2 and N_3, b_2^n is well defined. Note that by observation O3 and Axiom P_4, $\mathbf{b}^n=(b_1^n,b_2^n)\in Y(\mathbf{a}^n,\varphi^n)$. Last, let $\widehat{\mathbf{b}}^n$ solve $\widehat{b}_1^n=b_1^n$ and $\varphi(\widehat{\mathbf{b}}^n)=\widehat{\varphi}$. We have $U(\{\mathbf{a}^n,\mathbf{b}^n\})=u(\mathbf{a}^n)-g(\mathbf{a}^n,\varphi^n)$. If in the switch point $\widehat{\varphi}=\varphi(\widehat{\mathbf{a}})$, then $U(\{\widehat{\mathbf{a}},\widehat{\mathbf{b}}^n\})=u(\widehat{\mathbf{a}})$. By Axiom P_2, $U(\{\mathbf{a}^n,\mathbf{b}^n\})-U(\{\widehat{\mathbf{a}},\widehat{\mathbf{b}}^n\})\underset{n\to\infty}{\to}0$, hence

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

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

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

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

This establishes the existence of a continuous representation

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

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 Axioms P_1 and P_2 is obvious. We have already shown that if g is strictly increasing, then Axiom P_3 is satisfied. For the necessity of Axiom P_4, let $H(\mathbf{b}):=\{\mathbf{a}:\varphi(\mathbf{a})\geq\varphi(\mathbf{a})\}\cap\{\mathbf{a}:a_1\geq b_1\}$. 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 Axioms N_1–N_3, it is 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), $L^{\succ_n}(\mathbf{a})\cap U^{\succ}(\mathbf{a})\neq\phi$. 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 Axiom 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, Axiom 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 7. *There are $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{a}}'$ such that $\varphi(\widehat{\mathbf{a}}) = \varphi(\mathbf{a})$, $\varphi(\widehat{\mathbf{a}}') = \varphi(\mathbf{a}')$, conditions (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 $\widetilde{\mathbf{a}}'$ such that $\varphi(\widetilde{\mathbf{a}}') = \varphi(\mathbf{a}')$ and $u(\widetilde{\mathbf{a}}') = \max\{u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})), u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))\}$. By continuity, there is $\widehat{\mathbf{a}}'$ such that $\varphi(\widehat{\mathbf{a}}') = \varphi(\mathbf{a}')$ and for sufficiently small $\varepsilon > 0$, $u(\widehat{\mathbf{a}}') = \max\{u(\mathbf{b}) + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b})), u(\mathbf{b}') + g(\varphi(\mathbf{a}'), \varphi(\mathbf{b}'))\} + \varepsilon$.

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

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

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

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

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

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

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

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

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

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

Claim 7 implies that

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

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 Axiom P_6.

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 Axiom P_5. To verify that Axiom 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 Axiom P_6 and concludes the proof of Theorem 2. $\square$

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).

In addition to Axioms N_1 (Weak order) and N_4 (their hexagon condition), Karni and Safra require the following axioms.

- *Independence.* The relationship $(a_1, a) \succeq_n (b_1, a)$ for some a implies $(a_1, b) \succeq_n (b_1, b)$ for all b.

Independence is implied since by Axiom 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 Axiom 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 $\overline{a}$ such that for all i, $a_i \in (\underline{a}, \overline{a})$. Define similarly standard (and bounded) sequences by varying the second component.

- *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 $b > a$, then $\{a_i\}$ is a decreasing sequence. Suppose that $\{a_i\}$ is bounded away from k and ∞. Let $\overline{a}$ and $\underline{a}$ be the least upper bound and greatest lower bound, respectively.

Case 1, $a > b$. By Axiom N_3, there exists x such that $(\overline{a}, a) \sim_n (x, b)$. By Axiom N_2, $x > \overline{a}$. Since $\{a_i\}$ is an increasing and bounded sequence, it must converge to its least upper bound, $\overline{a}$. By continuity, there exists a subsequence $\{a_{ik}\}$ that converges to x. In particular, there exists K such that for $k > K$, $x - a_{ik} < \varepsilon := (x - \overline{a})/2$, which is a contradiction.

Case 2, $a < b$. By Axiom 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, Axiom 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 Axiom N_3(ii).

Karni and Safra (1998) show (see the lemma in their paper) that their axioms imply the axioms of Krantz et al. (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 $a_2 \to \gamma(a_2)$ 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.[20] Quasilinearity implies that there is an increasing continuous function $\xi : X \to \mathbb{R}$, 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 $v_1, v_2 : X \to \mathbb{R}_{++}$ are increasing and continuous, and if we redefine $\varphi(\mathbf{a}) := v_1(a_1)v_2(a_2)$, it represents $\succ_n$. Note that if $\varphi'(\mathbf{a}) = v_1'(a_1)v_2'(a_2)$ also represents $\succ_n$, then there are α, β_1, and 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 $\alpha = 1$. □

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

References

Andreoni, James and John H. Miller (2002), "Giving according to GARP: An experimental test of the consistency of preferences for altruism." *Econometrica*, 70, 737–753. [100]

Bénabou, Roland and Jean Tirole (2006), "Incentives and prosocial behavior." *American Economic Review*, 96, 1652–1678. [111]

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. [113]

Broberg, Thomas, Tore Ellingsen, and Magnus Johannesson (2007), "Is generosity involuntary?" *Economics Letters*, 94, 32–37. [100]

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. [113]

Buss, Arnold H. (1980), *Self-Consciousness and Social Anxiety*. W. H. Freeman, San Francisco. [100]

Camerer, Colin (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, Princeton. [99]

Charnes, Gary and Mathew Rabin (2002), "Understanding social preferences with simple tests." *Quarterly Journal of Economics*, 117, 817–869. [100]

Cox, James C., Daniel Friedman, and Vjollca Sadiraj (2008), "Revealed altruism." *Econometrica*, 76, 31–69. [106]

Dana, Jason, Daylian M. Cain, and Robyn M. Dawes (2006), "What you don't know won't hurt me: Costly (but quiet) exit in dictator games." *Organizational Behavior and Human Decision Processes*, 100, 193–201. [100]

Davis, Douglas D. and Charles A. Holt (1993), *Experimental Economics*. Princeton University Press, Princeton. [111]

Dekel, Eddie, Barton L. Lipman, and Aldo Rustichini (2009), "Temptation-driven preferences." *Review of Economic Studies*, 76, 937–971. [104]

Epstein, Larry G. and Igor Kopylov (2007), "Cold feet." *Theoretical Economics*, 2, 231–259. [102, 112]

Fehr, Ernst and Klaus M. Schimdt (1999), "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114, 817–868. [100, 111]

Frohlich, Norman, Joe Oppenheimer, and J. Bernard Moore (2001), "Some doubts about measuring self-interest using dictator experiments: The costs of anonymity." *Journal of Economic Behavior and Organization*, 46, 271–290. [113]

Gauthier, David (1986), *Morals by Agreement*. Oxford University Press, Oxford. [111]

Gehm, Theodor L. and Klaus R. Scherer (1988), "Relating situation evaluation to emotion differentiation: Nonmetric analysis of cross-cultural questionnaire data." In *Facets of Emotion* (Klaus R. Scherer, ed.), 61–77, Lawrence Erlbaum, Hillsdale, New Jersey. [100]

Gul, Faruk and Wolfgang Pesendorfer (2001), "Temptation and self-control." *Econometrica*, 69, 1403–1435. [102]

Gul, Faruk and Wolfgang Pesendorfer (2005), "The simple theory of temptation and self-control." Unpublished paper, Department of Economics, Princeton University. [104]

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. [113]

Johannesson, Magnus and Bjorn Persson (2000), "Non-reciprocal altruism in dictator games." *Economics Letters*, 69, 137–142. [113]

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. [109, 110, 120, 121]

Keeney, Ralph L. and Howard Raiffa (1976), *Decisions With Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York. [121]

Koch, K. Alexander and Hans Theo Normann (2008), "Giving in dictator games: Regard for others or regard by others?" *Southern Economic Journal*, 75, 223–231. [113]

Krantz, David H., R. Duncan Luce, Patrick C. Suppes, and Amos Tversky (1971), *Foundations of Measurement*, Vol. I. Academic Press, New York. [110, 121]

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber (2006), "Sorting in experiments with application to social preferences." Unpublished paper, University of California, Berkeley. [100]

Lewis, Helen B. (1971), *Shame and Guilt in Neurosis*. International Universities Press, New York. [100]

Miller, Dale T. (1999), "The norm of self-interest." *American Psychologist*, 54, 1053–1060. [113]

Nash, John F. (1950), "The bargaining problem." *Econometrica*, 18, 155–162. [103]

Nash, John F. (1953), "Two-person cooperative games." *Econometrica*, 21, 128–140. [111]

Neilson, William S. (2009), "A theory of kindness, reluctance, and shame for social preferences." *Games and Economic Behavior*, 66, 394–403. [111]

Noor, Jawwad and Norio Takeoka (2011), "Menu-dependent self-control." Unpublished paper, Boston University. [102]

Oberholzer-Gee, Felix and Reiner Eichenberger (2008), "Fairness in extended dictator game experiments." *The B.E. Journal of Economic Analysis and Policy (Contributions)*, 8 (1), Article 16. [112]

Olszewski, Wojciech (2011), "A model of consumption-dependent temptation." *Theory and Decision*, 70, 83–93. [102]

Osborne, Martin J. and Ariel Rubinstein (1994), *A Course in Game Theory*. MIT Press, Cambridge, Massachusetts. [111]

Pillutla, Madan M. and J. Keith Murnighan (1995), "Being fair or appearing fair: Strategic behavior in ultimatum bargaining." *Academy of Management Journal*, 38, 1408–1426. [100]

Tangney, June P. and Ronda L. Dearing (2002), *Shame and Guilt*. Guilford Press, New York. [100]

Submitted 2009-11-3. Final version accepted 2010-12-8. Available online 2010-12-8.