Title
Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings

Authors
Segal, Uzi
Sobel, Joel

Publication Date
1999-06-01

99-10

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

TIT FOR TAT: FOUNDATIONS OF PREFERENCES FOR RECIPROCITY IN STRATEGIC SETTINGS

BY

UZI SEGAL

AND

JOEL SOBEL

DISCUSSION PAPER 99-10
JUNE 1999

Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings*

Uzi Segal† and Joel Sobel‡

June 3, 1999

Abstract

This paper assumes that in addition to the conventional (selfish) preferences over outcomes, players in a strategic environment have preferences over strategies. In the context of two-player games, it provides conditions under which a player's preferences over strategies can be represented as a weighted average of the individual's selfish payoffs and the selfish payoffs of the opponent. The weight one player places on the opponent's selfish utility depends on the opponent's behavior. In this way, the framework is rich enough to describe the behavior of individuals who repay kindness with kindness and meanness with meanness. The paper assumes that each player has an ordering over his opponent's strategies that describes the niceness of these strategies. It introduces a condition that insures that the weight on opponent's utility increases if and only if the opponent chooses a nicer strategy.

*We are grateful to Miguel Costa-Gomes, David Levine, Joe Ostroy, Ariel Rubinstein, Dan Vincent, and Bill Zame for their comments. We thank SSHRCC and NSF for financial support.

†Department of Economics, University of Western Ontario, London N6A 5C2, Canada. E-mail: usegal@julian.uwo.ca

‡Department of Economics, University of California, San Diego, La Jolla, CA 92093, U.S.A. E-mail: jsobel@weber.ucsd.edu

1 Introduction

The notion that economic agents act rationally is a premise that unites most work in economic theory. The rationality assumption is often stated broadly and implemented narrowly. The broad version of the assumption is that agents are goal oriented and seek to maximize preferences subject to constraints. The narrow version of the assumption is that an individual's preferences are exogenously given and depend only on those aspects of an allocation that directly influence his or her material well being.

This paper lays the foundations for an extension of the narrow view of rationality in strategic settings. No modification of game theory is needed to permit individuals to be motivated by something other than material well being. The utility in standard game theory may be derived from arbitrary preferences over outcome distributions. Our theory goes beyond this. We present a representation theorem in games that incorporates the possibility that preferences will be influenced by the behavior of others.

Game theory always assumes that players have preference relationships defined on lotteries over outcomes. Our starting point is to also assume that players have preferences over strategies. Since the space of (mixed) strategies is a mixture space, it lends itself to the expected utility setup. In other words, we assume that for any three strategies σ^1, σ^2, and σ^3, and for all $\alpha \in (0,1]$, $\sigma^1 \succeq \sigma^2$ iff $\alpha\sigma^1 + (1-\alpha)\sigma^3 \succeq \alpha\sigma^2 + (1-\alpha)\sigma^3$. This, together with continuity and transitivity, implies that preferences over strategies can be represented by an expected utility functional, *where the utility is a utility from strategies.* This utility does not have to agree with the expected utility from payoffs obtained when the player uses this strategy.

We limit attention to two-player games. Section 2 presents the basic representation theorem. We show that in a fixed game G, and given that his opponent is playing σ_j, player i's preferences over his own strategies σ_i will be represented by a utility function of the form

$$u_1(\sigma_i, \sigma_j) + a^G_{i,\sigma_j} u_2(\sigma_i, \sigma_j)$$

the representation is a weighted sum of the two players' utilities, where the weight player i gives to player j's utility depends on j's action. This result is a consequence of a theorem due to Harsanyi [27]. The critical assumption is that if, given a fixed strategy of player j, two of player i's strategies lead to

the same distribution of selfish expected utility for both players, then player i is indifferent between these two strategies. The coefficients a^G_{i,σ_j} represent the degree to which player i is willing to take person j's interests into consideration. In standard theory, $a^G_{i,\sigma_j} \equiv 0$. Positive values of the coefficient suggest that player i is willing to sacrifice his selfish payoff in order to increase the payoff of his opponent. Negative values suggest a willingness to sacrifice selfish payoff in order to lower the opponent's payoff. Since player i's coefficient depends on player j's strategy, the players may exhibit preferences for reciprocity. A player may be willing to make selfish sacrifices to increase or decrease his opponent's payoff in the same strategic setting.

We allow the possibility that one player's preferences over outcomes can reflect concern for the well being of the other player. That is, we allow players to be intrinsically altruistic or spiteful. More important, however, is that we permit a player's preferences over strategies to place a higher weight on opponent's selfish payoffs in response to nice behavior. This is done in Section 3, where we connect the coefficient a^G_{i,σ_j} to the way player i perceives j's behavior. The goal is to formalize the intuition that a player would respond to nice behavior by reducing his selfish utility to benefit his opponent and respond to nasty behavior by reducing his selfish utility to harm his opponent. We assume that player i has preferences over his opponent's strategies, which describe his view of their 'niceness'. Section 3 we identifies conditions under which $a^G_{i,\sigma_j} > a^G_{i,\sigma'_j}$ if and only if σ_i is 'nicer' than σ'_j. This theorem captures the idea that a player is more likely to be kind to an opponent who treats him nicely. In order to prove the result, we introduce a Reciprocal Altruism assumption that makes precise the intuition behind the theorem.

Section 4 discusses possible objections to our model and describes several examples that illustrate its features. Section 5 briefly reviews the most relevant experiments in the vast literature detailing shortcomings of the rational actor model. It provides a more detailed discussion of closely related theoretic responses to the evidence.

2 Representation Theorems

Assume two players. Let X_i be the space of outcomes to player i, $i = 1, 2$. Each player has "selfish" preferences $\succeq_i^{sel}$ over $\Delta(X_i)$, the space of lotteries over X_i. A game G is a collection $\mathbf{s}_i^G = \{s_i^{G,1}, \ldots, s_i^{G,n_i^G}\}$ of strategies for

player i, $i = 1, 2$, together with the payoff function $O^G : \mathbf{s}_1^G \times \mathbf{s}_2^G \to X_1 \times X_2$. Let Σ_i^G be the space of mixed strategies of player i for game G, and extend O^G to be from $\Sigma_1^G \times \Sigma_2^G$ to $\Delta(X_1) \times \Delta(X_2)$. Throughout the paper, $i, j \in \{1, 2\}$, and $i \neq j$.

Given a game G and his opponent's (mixed) strategy $\sigma_j \in \Sigma_j$, player i has a complete and transitive preference relation $\succeq_{i,\sigma_j}^G$ over Σ_i^G. As long as G is fixed, we omit the superscript G, and use the notations $\mathbf{s}_i$, Σ_i, and $\succeq_{i,\sigma_j}$. The preferences $\succeq_{i,\sigma_j}$ over strategies need not be linked to the selfish preferences $\succeq_i^{sel}$ over outcomes. We assume that these preferences satisfy the following axioms.[1]

(C) Continuity (a) For every $\sigma_i^* \in \Sigma_i$, the sets $\{(\sigma_i, \sigma_j) \in \Sigma_i \times \Sigma_j : \sigma_i \succeq_{i,\sigma_j} \sigma_i^*\}$ and $\{(\sigma_i, \sigma_j) \in \Sigma_i \times \Sigma_j : \sigma_i^* \succeq_{i,\sigma_j} \sigma_i\}$ are closed subsets of $\Sigma_i \times \Sigma_j$.

(IND) Independence $\forall \sigma_i^1, \sigma_i^2, \sigma_i^3 \in \Sigma_i$, $\forall \sigma_j \in \Sigma_j$, and $\forall \alpha \in (0, 1]$, $\sigma_i^1 \succeq_{i,\sigma_j} \sigma_i^2$ iff $\alpha\sigma_i^1 + (1-\alpha)\sigma_i^3 \succeq_{i,\sigma_j} \alpha\sigma_i^2 + (1-\alpha)\sigma_i^3$.

As selfish preferences $\succeq_i^{sel}$ are defined over $\Delta(X_i)$, these preferences exist independently of the strategic environment. Preferences over strategies $\succeq_{i,\sigma_j}^G$, on the other hand, depend on the game being played. In this framework, a Nash Equilibrium is a strategy profile in which each agent's strategy is maximal according to $\succeq_{i,\sigma_j}^G$. Lemma 1 asserts that a Nash Equilibrium exists in our framework. We omit the proof, which follows from standard arguments.

Lemma 1 *If, for a given game G, both players' preferences satisfy the Continuity and the Independence axioms, then Nash Equilibrium exists for this game.*

We make two more assumptions.

(EU) Expected Utility The preferences $\succeq_i^{sel}$ satisfy the assumptions of expected utility theory.

It follows by this axiom that there are vN–M utility functions $u_i : X_i \to \mathbb{R}$ such that the preferences $\succeq_i^{sel}$ over lotteries over X_i are represented by the expected value of the utility u_i from their payoffs. To simplify notation, denote by $u_i(\sigma_i, \sigma_j)$ the expectation of the utility u_i player i receives from the lottery $O_i(\sigma_i, \sigma_j)$ (O_i is the lottery person i receives from O). Let $u(\sigma_i, \sigma_j) = (u_i(\sigma_i, \sigma_j), u_j(\sigma_i, \sigma_j))$.

[1] For part (b) of the Continuity assumption, see Section 3 below.

(SI) Self Interest Suppose that $u_j(\sigma_i', \sigma_j) = u_j(\sigma_i, \sigma_j)$. Then $\sigma_i' \succeq_{i,\sigma_j} \sigma_i$ if, and only if, $u_i(\sigma_i', \sigma_j) \geqslant u_i(\sigma_i, \sigma_j)$.

This axiom is weaker than the one usually used. In standard game theory, it is assumed that $\sigma_i' \succeq_{i,\sigma_j} \sigma_i$ iff $u_i(\sigma_i', \sigma_j) \geqslant u_i(\sigma_i, \sigma_j)$. That is, the preferences of person i over his own set of strategies, given that player j is playing σ_j, are fully determined by i's payoff. Here we only require that player i's preferences over strategies agree with his selfish preferences when player j is (selfishly) indifferent between σ_i and σ_i'. Axiom **SI** implies in particular

$(\star)$ If $u(\sigma_i', \sigma_j) = u(\sigma_i, \sigma_j)$, then $\sigma_i' \sim_{i,\sigma_j} \sigma_i$.

The structure of the model so far resembles that of Harsanyi's social choice theory [27]. In his model, members of society have preferences over (lotteries) over social states, and these preferences are expected utility. There are social preferences over the same domain, and these preferences too are expected utility. Finally, a Pareto assumption connects these preferences, where it is assumed that if all members of society are indifferent between two social policies, then so is society. From these assumptions Harsanyi got the "utilitarian" social welfare function $\sum \alpha_i u_i$. Similarly, we get here

Fact 1 *Given the Expected Utility and Independence axioms and the* $(\star)$ *condition, the preferences* $\succeq_{i,\sigma_j}$ *over* Σ_i *can be represented by*

$$a^i_{i,\sigma_j} u_i(\sigma_i, \sigma_j) + a^j_{i,\sigma_j} u_j(\sigma_i, \sigma_j) \tag{1}$$

Proof See Border [6] or Fishburn [20]. ■

Note that the weights depend on j's strategy σ_j. In standard game theory, $a^i_{i,\sigma_j} \equiv 1$ and $a^j_{i,\sigma_j} \equiv 0$. Here we can only retain the first of these two identities.

Lemma 2 *Given the Self Interest assumption,* a^i_{i,σ_j} *can be chosen to be positive,* $i = 1, 2$, $j \neq i$.

Proof For a given σ_j, the set $S_{\sigma_j} = \{u(\sigma_i, \sigma_j) : \sigma_i \in \Sigma_i\}$ is either a chord in $\mathbb{R}^2$, or it is convex with a non empty interior. If the latter happens, let $\sigma_i, \sigma_i' \in \Sigma_i$ such that $u_i(\sigma_i', \sigma_j) > u_i(\sigma_i, \sigma_j)$ and $u_j(\sigma_i', \sigma_j) = u_j(\sigma_i, \sigma_j)$. Hence,

by Axiom **SI**, $\sigma_i' \succ_{i,\sigma_j} \sigma_i$. By eq. (1), $a^i_{i,\sigma_j} u_i(\sigma_i', \sigma_j) > a^i_{i,\sigma_j} u_i(\sigma_i, \sigma_j)$ (recall that $u_j(\sigma_i', \sigma_j) = u_j(\sigma_i, \sigma_j)$). Since $u_i(\sigma_i', \sigma_j) > u_i(\sigma_i, \sigma_j)$, it follows that $a^i_{i,\sigma_j} > 0$.

Suppose that S_{σ_j} is a chord in the line $k_i u_i + k_j u_j = C$. If $k_i = 0$, then similarly to above, Axiom **SI** implies that a^i_{i,σ_j} is positive. If $k_j = 0$, then a^i_{i,σ_j} can be any number, in particular, it can be positive. Finally, if $k_i, k_j \neq 0$, Axiom **IND** implies that the order of $\succeq_{i,\sigma_j}$ on S_{σ_j} is either always increasing with u_i or always decreasing with u_i. In the first case, let $a^i_{i,\sigma_j} = 1$ and $a^j_{i,\sigma_j} = 0$. In the second case, let $a^i_{i,\sigma_j} = 1$ and $a^j_{i,\sigma_j} > k_j/k_i$ if k_i and k_j have the same sign, and $a^j_{i,\sigma_j} < k_j/k_i$ if k_i and k_j have different signs. ■

Conclusion 1 We may assume, without loss of generality, that $a^i_{i,\sigma_j} \equiv 1$. That is, the preferences $\succeq_{i,\sigma_j}$ can be represented by

$$u_i(\sigma_i, \sigma_j) + a^j_{i,\sigma_j} u_j(\sigma_i, \sigma_j)$$

Note, however, that a^j_{i,σ_j} may be negative. For simplicity, we omit the superscript j, and let a_{i,σ_j} be the weight player i gives to the utility of player j.

3 Reciprocal Altruism

We now assume that player i has continuous preferences $\succeq_i^{opp}$ over Σ_j, the set of player j's strategies, $i = 1, 2$, $j \neq i$. (The superscript *opp* stands for "opponent"). The interpretation of the statement "$\sigma_j^1 \succeq_i^{opp} \sigma_j^2$" is that player i considers j to be nicer to him when j is using σ_j^1 than when she is using σ_j^2. In this section we analyze the connection between these preferences and the weight a_{i,σ_j} player i gives to j's utility. The main results of the section are theorems that provide conditions that formalize the statement: The weight player i puts on player j's utility is an increasing function of the niceness of player j's strategy.

3.1 Nice Behavior

In this subsection we offer some examples for what we mean by nice behavior, but our results apply to a much larger set of preferences. Given that player

j is using strategy σ_j, person i is offered the following vector of utilities

$$u_i(\boldsymbol{s}_i, \sigma_j) = (u_i(s_i^1, \sigma_j), \ldots, u_i(s_i^{n_i}, \sigma_j)) \in \mathbb{R}^{n_i}$$

In [45] we offer axioms implying that the preferences $\succeq_i^{opp}$ can be represented by one of the following two functionals.

$$U_i(u_i(s_i^1, \sigma_j), \ldots, u_i(s_i^{n_i}, \sigma_j)) = \max_k \{u_i(s_i^k, \sigma_j)\} \quad (2)$$

$$U_i(u_i(s_i^1, \sigma_j), \ldots, u_i(s_i^{n_i}, \sigma_j)) = \min_k \{u_i(s_i^k, \sigma_j)\} \quad (3)$$

These functional forms offer two different notions of nice behavior. According to the first, person j's behavior is nicer if she offers player i higher possible utility. This representation is consistent with the concept of fairness adopted by Rabin [39]. The second notion of niceness is for player j to let player i have high minimal utility. A possible justification for this is that if player j believes that i does not understand the game, she can protect him by increasing his security level.[2]

A more general evaluation of the niceness of the opponent's behavior may involve the opponent's outcome. Consider the following game.

$3,3$	$0,4$	$3,5$
$4,0$	$1,1$	$4,2$

Choosing left may indicate a nicer behavior by person j (the column player) than choosing right, because the latter benefits her as well as i. Such preferences may be represented by

$$U_i(u(s_i^1, \sigma_j), \ldots, u(s_i^{n_i}, \sigma_j)) = \max_k \{u_i(s_i^k, \sigma_j)\} - \max_k \{u_j(s_i^k, \sigma_j)\}$$

We do not consider such preferences in our analysis, and will assume throughout that the preferences $\succeq_i^{opp}$ depend only on the vector $u_i(\boldsymbol{s}_i, \sigma_j)$.

Below we introduce two axioms. The first, called Reciprocal Altruism, connects the preferences $\succeq_{i,\sigma_j}$ player i has over his set of strategies, to the

[2] Another possibility is $U_i(u_i(s_i^1, \sigma_j), \ldots, u_i(s_i^{n_i}, \sigma_j)) = \sum_k u_i(s_i^k, \sigma_j)$. This notion suggests that player j is interested in maximizing player i's average utility. The reason may be that since she believes that i does not know what to do, it is best to offer him the highest possible utility assuming that he will randomize. It is inconsistent with the Irrelevance axioms we make below.

preferences $\succeq_i^{opp}$ he has over j's behavior. The second, called Irrelevance, connects the preferences $\succeq_{i,\sigma_j}$ in different games. The combination of the two axioms rules out some possible preferences $\succeq_i^{opp}$, but it is lenient enough to permit the functional forms of eq. (2) and eq. (3).

3.2 The Reciprocal Altruism Axiom

In this subsection we describe the Reciprocal Altruism axiom. This axiom formalizes the idea that players are willing to reward nice behavior, and to punish mean behavior.

(RA) Reciprocal Altruism Suppose

(a) $u(\sigma_i^1, \sigma_j^1) = u(\sigma_i^2, \sigma_j^2)$ and $u(\bar{\sigma}_i^1, \sigma_j^1) = u(\bar{\sigma}_i^2, \sigma_j^2)$;
(b) $\sigma_j^2 \succ_i^{opp} \sigma_j^1$ [resp. $\sigma_j^2 \sim_i^{opp} \sigma_j^1$];
(c) $\sigma_i^1 \sim_{i,\sigma_j^1} \bar{\sigma}_i^1$.

Then $\sigma_i^2 \succeq_{i,\sigma_j^2} \bar{\sigma}_i^2$ iff $u_j(\sigma_i^2, \sigma_j^2) \geqslant u_j(\bar{\sigma}_i^2, \sigma_j^2)$ [resp. $\sigma_i^2 \sim_{i,\sigma_j^2} \bar{\sigma}_i^2$].

The axiom requires that if "all things are equal," then when player j plays a nicer strategy, player i will prefer strategies that lead to larger selfish payoffs to player j. In this way, the axiom formalizes the notion that player i repays kindness with kindness. The conditions in the statement of the axiom formalize the notion of what it means for all things to be equal. The axiom places a restriction on player i's ranking in one situation, indicated by the superscript 2 on strategies, using information about his ranking in another situation, indicated by the superscript 1 on strategies (condition (c)). Condition (a) requires that the information about σ_i^1 and $\bar{\sigma}_i^1$ is comparable to the information about σ_i^2 and $\bar{\sigma}_i^2$. It states that when player j switches from σ_j^1 to σ_j^2 neither players' selfish utility changes when player i changes from σ_i^1 to σ_i^2 or from $\bar{\sigma}_i^1$ to $\bar{\sigma}_i^2$. If conditions (a) and (c) hold, then Reciprocal Altruism requires that if player i likes player j's new behavior better than her old behavior, then i should follow j's preferences, in the sense that σ_i^2 should be preferred to $\bar{\sigma}_i^2$ if, and only if, j's selfish utility under σ_i^2 is higher than her utility under $\bar{\sigma}_i^2$.

One can check that if player i has conventional selfish preferences, so that $\sigma_i \succeq_{i,\sigma_j} \bar{\sigma}_i$ iff $u_i(\sigma_i, \sigma_j) \geqslant u_i(\bar{\sigma}_i, \sigma_j)$, then **RA** is satisfied provided that

player i views all of j's strategies as equally kind ($\sigma_j^2 \sim_i^{opp} \sigma_j^1$ for all σ_j^k). To see this, simply note that condition (c) implies that $u_i(\sigma_i^1, \sigma_j^1) = u_i(\bar{\sigma}_i^1, \sigma_j^1)$. Condition (a) now implies $u_i(\sigma_i^2, \sigma_j^2) = u_i(\bar{\sigma}_i^2, \sigma_j^2)$, hence $\sigma_i^2 \sim_{i,\sigma_j^2} \bar{\sigma}_i^2$. In the appendix we describe a nontrivial example of preferences that satisfy all of our assumptions.

Condition (a) in the Reciprocal Altruism axiom is restrictive. For example, when $n_i = 2$ it is possible to satisfy condition (a) for $\sigma_j^1 \neq \sigma_j^2$ only for nongeneric payoffs. Put differently, if for a given strategy σ_j of player j, the utility opportunity set that can be generated by i strategies is a line in the two-dimensional selfish utility space, then it is impossible to determine the value of a_{i,σ_j} used by person i. The restrictive nature of condition (a) suggests that the Reciprocal Altruism axiom is relatively weak. We must combine it with other assumptions to link player i's preferences over his opponent's choice of strategy to a_{i,σ_j}, the weight person i is gives to j's utility.

Since all the relevant information of a game is summarized by the selfish utility payoffs the two players receive, we can view games as elements of $\mathbb{R}^{2 \times n_1 \times n_2}$. On the set of games $\mathcal{G}^{n_1 \times n_2}$ with n_i pure strategies for player i, $i = 1, 2$, we use the Euclidean topology. The set of mixed strategies Σ_i for a game $\mathcal{G}^{n_i \times n_j}$ can be viewed as the simplex $\Delta^{n_i} = \{(p_1, \ldots, p_{n_i}) \in \mathbb{R}_+^{n_i} : \sum p_k = 1\}$. With a little abuse of notation, we will write $\sigma_i \in \Delta^{n_i}$.

(C) Continuity (b) Fix n_i and n_j. For every $\sigma_i^* \in \Delta^{n_i}$ and $\sigma_j \in \Delta^{n_j}$, the sets $\{(\sigma_i, G) \in \Delta^{n_i} \times \mathcal{G}^{n_i \times n_j} : \sigma_i \succeq_{i,\sigma_j}^G \sigma_i^*\}$ and $\{(\sigma_i, G) \in \Delta^{n_i} \times \mathcal{G}^{n_i \times n_j} : \sigma_i^* \succeq_{i,\sigma_j}^G \sigma_i\}$ are closed subsets of $\Delta^{n_i} \times \mathcal{G}^{n_i \times n_j}$.

Axiom **C**(a) requires that preferences are continuous within a fixed game. The present axiom requires that $\succeq_{i,\sigma_j}^G$ be continuous as G changes (but with σ_j held fixed).

3.3 Theorems

As we have mentioned above, the set of selfish utility payoffs may be too thin to apply the **RI** axiom. To solve this problem, we will replace the game G with a game that will duplicate one of player i's strategies. Then we will use Axiom **C**(b), and create a thicker set of possible utility payoffs by changing this new strategy in a small neighborhood.

Given a game G, let G_i^ℓ be the enlarged game obtained from G by adding the strategy $s_i^{G_i^\ell, n_i^G+1}$, where

$$u(s_i^{G_i^\ell, n_i^G+1}, s_j^{G_i^\ell, k}) = u(s_i^{G,\ell}, s_j^{G,k})$$

for each k (recall the notation $s_h^{G,k}$, which is the k-th pure strategy of person h in game G). The game G_i^ℓ is obtained from G by duplicating strategy ℓ of person i. For $h \in \{i, j\}$, there are natural isomorphisms between Σ_h^G and $\Sigma_h^{G_h^\ell}$, so we will let $\widetilde{\sigma}_h$ denote the strategies of player h in $\Sigma_h^{G_h^\ell}$ that correspond to σ_h in Σ_h^G. The following axiom states that a player's preferences between any two strategies will not change if a strategy of any one of the players is duplicated.[3]

(IR) Irrelevance (a) $\widetilde{\sigma}_i^1 \succeq_{i,\widetilde{\sigma}_j}^{G_i^\ell} \widetilde{\sigma}_i^2$ iff $\sigma_i^1 \succeq_{i,\sigma_j}^G \sigma_i^2$.

This axiom states that a player's preferences between any two strategies will not change if a strategy of any one of the players is duplicated. This axiom trivially rules out the functional form for nice behavior of footnote 2, where person i is interested in the average value of the weighted sum of his and his opponent utility levels. The reason is that duplicating a strategy will change the average payoff, hence, by Axiom **RI**, it will also change player i's ranking of his own strategies.

Our aim is to prove that $a_{i,\sigma_j^1} \geqslant a_{i,\sigma_j^2}$ iff $\sigma_j^1 \succeq_i^{opp} \sigma_j^2$. The Irrelevance (a) axiom is sufficiently strong for this, provided we restrict attention to the case where the preferences $\succeq_i^{opp}$ are monotonic on the segment connecting σ_j^1 and σ_j^2. Formally, we define:

Definition 1 The two strategies σ_j^1 and σ_j^2 are linearly ordered by $\succeq_i^{opp}$ if either for all $0 \leqslant \alpha \leqslant \beta \leqslant 1$, $\alpha\sigma_j^1 + (1-\alpha)\sigma_j^2 \succeq_i^{opp} \beta\sigma_j^1 + (1-\beta)\sigma_j^2$, or for all such α and β, $\beta\sigma_j^1 + (1-\beta)\sigma_j^2 \succeq_i^{opp} \alpha\sigma_j^1 + (1-\alpha)\sigma_j^2$.

Theorem 1 *Given the Continuity, Independence, Expected Utility, Self Interest, Reciprocal Altruism, and Irrelevance (a) axioms, if the two strategies σ_j^1 and σ_j^2 are linearly ordered, then $a_{i,\sigma_j^1} \geqslant a_{i,\sigma_j^2}$ iff $\sigma_j^1 \succeq_i^{opp} \sigma_j^2$.*

[3]The second part of the Irrelevance axiom appears after Theorem 2 below.

For a given $\sigma_j \in \Sigma_j$, let $S_i(\sigma_j) = \{u(\sigma_i, \sigma_j) : \sigma_i \in \Sigma_i\}$ be the utility opportunity set player i can generate given σ_j. Let $\sigma_j^1, \sigma_j^2 \in \Sigma_j$, and define $S^* = S_i(\sigma_j^1) \cap S_i(\sigma_j^2)$.

There are three parts to the proof of this theorem. In the first part, we assume that the utility opportunity sets are sufficiently rich that we can apply Axiom **RA** directly. In this case, it is possible to find strategies that satisfy the condition of **RA**. The desired conclusion is a straightforward computation. To carry out this argument, we need S^* to have a nonempty interior. The second part of the part uses the linear ordering property to obtain the conclusion of the theorem when S^* has an empty interior, but S_i has a nonempty interior for all strategies on a segment connecting σ_j^1 to σ_j^2. Finally, in the third part of the proof, we use the Irrelevance and Continuity assumptions to approximate an arbitrary game by a game in which the richness property used in the second part of the proof holds.

Proof Let $\sigma_j^1, \sigma_j^2 \in \Sigma_j$, and suppose first that $S^* = S_i(\sigma_j^1) \cap S_i(\sigma_j^2)$ has a non-empty interior. It follows that there exist σ_i^1, σ_i^2, $\bar{\sigma}_i^1$, $\bar{\sigma}_i^2$, and two distinct points t and t' in S^* such that $t = (t_i, t_j) = u(\sigma_i^1, \sigma_j^1) = u(\sigma_i^2, \sigma_j^2)$, $t' = (t_i', t_j') = u(\bar{\sigma}_i^1, \sigma_j^1) = u(\bar{\sigma}_i^2, \sigma_j^2)$, $t_j \neq t_j'$,[4] and

$$t_i + a_{i,\sigma_j^1} t_j = t_i' + a_{i,\sigma_j^1} t_j' \tag{4}$$

Hence, by Fact 1, $\sigma_i^1 \sim_{i,\sigma_j^1} \bar{\sigma}_i^1$. By the same Fact, $\sigma_i^2 \succeq_{i,\sigma_j^2} \bar{\sigma}_i^2$ if and only if

$$t_i + a_{i,\sigma_j^2} t_j \geqslant t_i' + a_{i,\sigma_j^2} t_j' \tag{5}$$

Subtract inequality (5) from eq. (4) to obtain

$$(a_{i,\sigma_j^1} - a_{i,\sigma_j^2})(t_j' - t_j) \geqslant 0 \tag{6}$$

Suppose now that $\sigma_j^1 \succ_i^{opp} \sigma_j^2$. Then, by **RA**, inequality (5) holds if and only if

$$t_j = u_j(\sigma_i^2, \sigma_j^2) \leqslant u_j(\bar{\sigma}_i^2, \sigma_j^2) = t_j' \tag{7}$$

[4] If the condition $t_j \neq t_j'$ cannot be satisfied, then by Fact 1, indifference curves of $\succeq_{i,\sigma_j^1}$ are constant in person j's utility. In other words, fixing j's strategy and utility while changing i's (strategy and) utility will not change the desirability of i's strategy in the $\succeq_{i,\sigma_j^1}$ order, a violation of Axiom **SI**.

Since $t_j \neq t'_j$, it follows from inequalities (6) and (7) that $a_{i,\sigma_j^1} \geqslant a_{i,\sigma_j^2}$. Similarly, if $\sigma_j^1 \sim_i^{opp} \sigma_j^2$, then **RA** implies that both (5) and (6) hold as equations. Since $t_j \neq t'_j$, $a_{i,\sigma_j^1} = a_{i,\sigma_j^2}$, which establishes the result when S^* has a nonempty interior.

Assume now that for every $\alpha \in [0,1]$, $S_i(\alpha\sigma_j^1 + (1-\alpha)\sigma_j^2)$ has a non-empty interior. By the compactness of $[0,1]$ and Axiom **C**, there are $0 = \alpha_1 < \cdots < \alpha_{n_i} = 1$ such that for $k = 1, \ldots, n_i - 1$,

$$S_i(\alpha_k\sigma_j^1 + (1-\alpha_k)\sigma_j^2) \cap S_i(\alpha_{k+1}\sigma_j^1 + (1-\alpha_{k+1})\sigma_j^2) \neq \varnothing$$

The claim now follows from the first part of this proof. For example, if $\sigma_j^1 \succeq_i^{opp} \sigma_j^2$, then by linear ordering, for $k = 1, \ldots, n_i - 1$,

$$\alpha_k\sigma_j^1 + (1-\alpha_k)\sigma_j^2 \succeq_i^{opp} \alpha_{k+1}\sigma_j^1 + (1-\alpha_{k+1})\sigma_j^2$$

For $k = 1, \ldots, n_i$, let

$$a_k = a_{i,\alpha_k\sigma_j^1 + (1-\alpha_k)\sigma_j^2}$$

and obtain that $a_{i,\sigma_j^1} = a_1 \geqslant \cdots \geqslant a_{n_i} = a_{i,\sigma_j^2}$.

Suppose now that for some α, the interior of $S_i(\alpha\sigma_j^1 + (1-\alpha)\sigma_j^2)$ is empty. Then consider the game G_i^ℓ for some $1 \leqslant \ell \leqslant n_i$ and a sequence $G^m \to G_i^\ell$ in $\mathcal{G}^{(n_i+1)\times n_j}$ such that for every $\alpha \in [0,1]$ and for every m, $S_i(\alpha\sigma_j^1 + (1-\alpha)\sigma_j^2)$ for the game G^m has a non-empty interior. The Theorem now follows by **C**(b), **IR**(a), and the first part of this proof. ■

Theorem 1 shows that when combined with other assumptions, Reciprocal Altruism implies that there is a link between the coefficient a and player i's preferences over his opponent's strategies. The next result proves that **RA** is exactly the assumption we must make in order to establish this link.

Theorem 2 *Suppose that the Continuity (a), Independence, Expected Utility, and Self Interest axioms are satisfied. If $a_{i,\sigma_j^1} \geqslant a_{i,\sigma_j^2}$ iff $\sigma_j^1 \succeq_i^{opp} \sigma_j^2$, then the Reciprocal Altruism axiom holds.*

Proof Let $\sigma^1 = (\sigma_i^1, \sigma_j^1)$, $\sigma^2 = (\sigma_i^2, \sigma_j^2)$, $\bar{\sigma}^1 = (\bar{\sigma}_i^1, \sigma_j^1)$, and $\bar{\sigma}^2 = (\bar{\sigma}_i^2, \sigma_j^2)$ satisfy conditions (a) and (c) of **RA**. By (c) and Fact 1,

$$u_i(\sigma_i^1, \sigma_j^1) + a_{i,\sigma_j^1} u_j(\sigma_i^1, \sigma_j^1) = u_i(\bar{\sigma}_i^1, \sigma_j^1) + a_{i,\sigma_j^1} u_j(\bar{\sigma}_i^1, \sigma_j^1)$$

and therefore by (a)

$$u_i(\sigma_i^2, \sigma_j^2) + a_{i,\sigma_j^1} u_j(\sigma_i^2, \sigma_j^2) = u_i(\bar{\sigma}_i^2, \sigma_j^2) + a_{i,\sigma_j^1} u_j(\bar{\sigma}_i^2, \sigma_j^2) \tag{8}$$

If $\sigma_j^1 \sim_i^{opp} \sigma_j^2$, then, by assumption, $a_{i,\sigma_j^1} = a_{i,\sigma_j^2}$ and therefore eq. (8) implies that

$$u_i(\sigma_i^2, \sigma_j^2) + a_{i,\sigma_j^2} u_j(\sigma_i^2, \sigma_j^2) = u_i(\bar{\sigma}_i^2, \sigma_j^2) + a_{i,\sigma_j^2} u_j(\bar{\sigma}_i^2, \sigma_j^2).$$

Hence, $\sigma_i^2 \sim_{i,\sigma_j^2} \bar{\sigma}_i^2$ by Fact 1.

If $\sigma_j^1 \succ_i^{opp} \sigma_j^2$, then, by assumption, $a_{i,\sigma_j^2} > a_{i,\sigma_j^1}$. By Fact 1, $\sigma_i^2 \succeq_{i,\sigma_j^2} \bar{\sigma}_i^2$ if and only if

$$u_i(\sigma_i^2, \sigma_j^2) + a_{i,\sigma_j^2} u_j(\sigma_i^2, \sigma_j^2) \geqslant u_i(\bar{\sigma}_i^2, \sigma_j^2) + a_{i,\sigma_j^2} u_j(\bar{\sigma}_i^2, \sigma_j^2) \tag{9}$$

Therefore, by eq. (8) and eq. (9), $\sigma_i^2 \succeq_{i,\sigma_j^2} \bar{\sigma}_i^2$ if and only if

$$(a_{i,\sigma_j^2} - a_{i,\sigma_j^1})(u_j(\sigma_i^2, \sigma_j^2) - u_j(\bar{\sigma}_i^2, \sigma_j^2)) \geqslant 0. \tag{10}$$

Since $a_{i,\sigma_j^2} > a_{i,\sigma_j^1}$, **RA** follows from eq. (10). ■

The analysis is more complicated when the strategies are not linearly ordered. The reason is that if $S_i(\sigma_j^1) \cap S_i(\sigma_j^2) = \varnothing$,[5] then it is not possible to use the **RA** axiom to compare a_{i,σ_j^1} with a_{i,σ_j^2}. However, if the preferences $\succeq_i^{opp}$ can be represented by either eq. (2) or by eq. (3), then they satisfy in particular the following more general conditions.

($\star\star$)
1. If $\sigma_j^1 \sim_i^{opp} \sigma_j^2$, then $S_i(\sigma_j^1)$ and $S_i(\sigma_j^2)$ share (at least one) utility level for person i.
2. For every σ_j^1 and σ_j^2 there is a finite sequence $0 = \alpha_1 < \cdots < \alpha_n = 1$ such that for $i = 1, \ldots, n-1$, $\alpha_i \sigma_j^1 + (1 - \alpha_i)\sigma_j^2$ and $\alpha_{i+1}\sigma_j^1 + (1 - \alpha_{i+1})\sigma_j^2$ are linearly ordered.

When preferences are represented by either eq. (2) or eq. (3), the first condition of ($\star\star$) follows immediately. Since the utility function of eq. (2) is quasi convex and that of eq. (3) is quasi concave, the second condition of ($\star\star$) can be satisfied with $n = 3$.

[5] As before, $S_i(\sigma_j)$ is the utility opportunity set given σ_j.

The assumption that follows permits us to conclude that $S_i(\sigma_j^1)$ and $S_i(\sigma_j^2)$ share at least one utility level for player j as well. To do so, we introduce our second way to compare games with different strategy sets.

According to our analysis, the preferences $\succeq_i^{opp}$ are concerned with what is available to person i. Therefore, changing the utility level available to j will not affect these preferences. We create the game $G_i^\ell(\omega)$ by adding a new strategy for player i that gives player i the same selfish utilities as the existing strategy σ_i^ℓ, but leads to a constant utility ω for player j. That is, for every $k = 1, \ldots, n_j^G$,

$$u(s_i^{G_i^\ell(\omega), n_i^G+1}, s_j^{G_i^\ell(\omega), k}) = (u_i(s_i^{G,\ell}, s_j^{G,k}), \omega)$$

As before, we denote by $\tilde{\sigma}_h$ the strategy of player h in $G_i^\ell(\omega)$ that is corresponding to σ_h in G. Define

$$\omega_j(G) = \min\left\{u_j(\sigma_i^{k_i}, \sigma_j^{k_j})\right\} \quad k_i = 1, \ldots, n_i^G; \; k_j = 1, \ldots, n_j^G$$

to be the minimal utility level person j can reach in game G. We add another axiom.

(IR) Irrelevance (b) For $\omega \leqslant \omega_j(G)$, $\tilde{\sigma}_i^1 \succeq_{i,\tilde{\sigma}_j}^{G_i^\ell(\omega)} \tilde{\sigma}_i^2$ iff $\sigma_i^1 \succeq_{i,\sigma_j}^{G} \sigma_i^2$.

This axiom requires that if we add a strategy to player i that from his selfish perspective is the same as a strategy he already has, and such that this strategy yields player j always the same 'bad' outcome, then it will not change the way player i chooses from *the old set of strategies*. This does not mean that he is not going to choose the new strategy, only that its existence does not affect his ranking of the other strategies.

Theorem 3 *Assume that conditions* $(\star\star)$ *1–2, and the Continuity, Independence, Expected Utility, Self Interest, Reciprocal Altruism, and Irrelevance* (*both parts*) *axioms are satisfied. Then* $a_{i,\sigma_j^1} \geqslant a_{i,\sigma_j^2}$ *iff* $\sigma_j^1 \succeq_i^{opp} \sigma_j^2$.

Proof As before, for a given $\sigma_j \in \Sigma_j$, let $S_i(\sigma_j) = \{u(\sigma_i, \sigma_j) : \sigma_i \in \Sigma_i\}$. If the interior of $S^* = S_i(\sigma_j^1) \cap S_i(\sigma_j^2)$ is not empty, then the proof is the same as the first part of the proof of Theorem 1. So suppose that the interior of S^* is empty. Invoking **IR**(a) and **C**(b), we may assume that the interiors of $S_i(\sigma_j^1)$ and $S_i(\sigma_j^2)$ are not empty.

We prove first that if $\sigma_j^1 \sim_i^{opp} \sigma_j^2$, then $a_{i,\sigma_j^1} = a_{i,\sigma_j^2}$. Note that by the first part of condition ($\star\star$), $\sigma_j^1 \sim_i^{opp} \sigma_j^2$ implies that there is a utility level u_i^* and u_j^1, u_j^2 such that $(u_i^*, u_j^k) \in S_i(\sigma_j^k)$, $k = 1, 2$. Denote the strategies of player i that may make this utility level possible ℓ_1 and ℓ_2.

Define now a new game as follows. Let $\omega < \omega_j(G)$, and add to G two strategies $s_i^{n_i+k}$, $k = 1, 2$, where $s_i^{n_i+k}$ is the same for person i as ℓ_k, but always yields player j the selfish utility level ω. That is, consider the game $G^* = (G_i^{\ell_1}(\omega))_i^{\ell_2}(\omega)$. This game has the required non-empty intersection. Therefore,[6] $\tilde{\sigma}_j^1 \sim_i^{opp} \tilde{\sigma}_j^2$ iff $a_{i,\tilde{\sigma}_j^1} = a_{i,\tilde{\sigma}_j^2}$. By **IR**(b) it follows that the weights $a_{i,\tilde{\sigma}_j^2}$ from G^* apply to G and that $\tilde{\sigma}_j^1 \succeq_i^{opp} \tilde{\sigma}_j^2$ if and only if $\sigma_j^1 \succeq_i^{opp} \sigma_j^2$. Therefore $\sigma_j^1 \sim_i^{opp} \sigma_j^2$ implies $a_{i,\sigma_j^1} = a_{i,\sigma_j^2}$.

Suppose next that $\sigma_j^1 \succ_i^{opp} \sigma_j^2$. For $\alpha \in [0, 1]$, let $\sigma_j(\alpha) = \alpha\sigma_j^1 + (1-\alpha)\sigma_j^2$. We construct inductively a sequence $0 = \beta_0, \ldots, \beta_m$ as follows. Suppose we already defined $\beta_0, \ldots, \beta_k$ such that $\sigma_j^1 \succ_i^{opp} \sigma_j(\beta_k)$, and there is no $\beta > \beta_k$ for which $\sigma_j(\beta) \sim_i^{opp} \sigma_j(\beta_k)$ (except maybe for $k = 0$). Define $\beta'_{k+1} > \beta_k$ to be the maximal number such that

1. $\sigma_j(\beta_k)$ and $\sigma_j(\beta'_{k+1})$ are linearly ordered, and
2. $\sigma_j^1 \succeq_i^{opp} \sigma_j(\beta'_{k+1})$.

Define β_{k+1} to be the maximal number for which $\sigma_j(\beta_{k+1}) \sim_i^{opp} \sigma_j(\beta'_{k+1})$. By Theorem 1 and the first part of this proof it follows that $a_{i,\sigma_j(\beta_k)} \geqslant a_{i,\sigma_j^2}$ for all k, and if $\sigma_j(\beta_k) \succ_i^{opp} \sigma_j^1$, then $a_{i,\sigma_j(\beta_k)} > a_{i,\sigma_j^2}$. By the second part of condition ($\star\star$), there is a step k where $\beta_k = 1$, that is, $\sigma_j(\beta_k) = \sigma_j^1$, hence the theorem. ■

Remark Theorem 3 holds under weaker conditions than ($\star\star$)–2. All we need is that for each α there is a segment $[\alpha_1, \alpha_2]$ with nonempty interior containing α such that $\sigma_j(\alpha_1)$ and $\sigma_j(\alpha_2)$ are linearly ordered by $\succeq_i^{opp}$.

4 Discussion

This section discusses some possible questions and objections that relate to our model.

[6] Here $\tilde{\sigma}_\cdot$ denote strategies in the two-step extension game G^*.

In what sense is this model different from standard game theory? One possible response to this question is that the present model permits players to use strategies that are dominated with respect to their selfish payoffs in equilibrium. It is straightforward to construct preferences over strategies, consistent with our assumptions, that permit the joint cooperation outcome to be an equilibrium in the prisoner's dilemma.[7] Intuitively, cooperation is a way of responding nicely to nice behavior.

Precisely the same prediction would follow, however, if we redefined the payoffs associated with outcomes, and use standard game theory. For example, if we treat the players in the prisoner's dilemma as risk-neutral agents who maximize monetary payoffs, the game matrix may look like

$3,3$	$0,4$
$4,0$	$1,1$

However, if we permit more general preferences over outcomes, for example, if preferences are over payoff distributions, then the strategic environment may be represented by

$3,3$	$0,0$
$0,0$	$1,1$

(In this example, both players' utilities are functions of the joint income distribution, and are given by $u_i(x_i, x_j) = u_j(x_i, x_j) = \min\{x_i, x_j\}$).

Our approach does more than can be done by simply redefining preferences over outcomes. The following example demonstrates that, in contrast to standard game theory, if a strategy is a unique best response to every pure strategy, then it need not be a dominant strategy. Consider the following example.

15,30	9,10
20,20	10,20

The players are going to cook dinner together. Player i will bring the main course, either beef (U) or pheasant (D). Player j will bring the wine, either red (L) or white (R). Player i prefers red wine to white and pheasant to beef.

[7] The standard equilibrium outcome would continue to be an equilibrium.

Player j prefers to drink red wine with beef, but hates a beef-white wine menu. If the weight that player i gives to player j's utility when j brings red wine is sufficiently positive (greater than $\frac{1}{2}$), then the optimal response of player i to red wine is to supply beef. On the other hand, if player j brings white wine, player i will give j's utility a negative weight, and if it is sufficiently negative (that is, less than $-\frac{1}{10}$), he will "punish" her by making her eat beef with the wrong wine. Under standard analysis, this means that player i should always play U. However, if player j uses a non-degenerate mixed strategy, player i may give j's utility zero weight, and eat pheasant. If player j always places zero weight on her opponent's utility, then we get two Nash equilibria, one is up–left, the other is down for player i and a mixed strategy (say $\frac{1}{2} - \frac{1}{2}$) for player j.

It is easy to verify that such a set of equilibrium points cannot be obtained under standard game theory.

MAYBE IT IS BETTER TO EXTEND THE SET OF STRATEGIES AND ASSUME A VIOLATION OF THE REDUCTION OF COMPOUND LOTTERIES AXIOM (RCLA)? Suppose we consider all mixtures as pure strategies. That is, strategy s_α for player i is "play a lottery where with probability α you choose beef, and with probability $1-\alpha$ you choose pheasant." Similarly, strategy β for player j is to play a lottery where with probability t_β she will bring a bottle of white wine, and with probability $1-\beta$ she will bring a bottle of red wine. Of course, if player i is indifferent between the mixed strategy "s_α with probability p and $s_{\alpha'}$ with probability $1-p$" and the pure strategy $s_{p\alpha+(1-p)\alpha'}$, then this extra structure will make no difference. So assume that this last indifference is (sometimes) violated. In other words, assume that players' preferences violate RCLA.[8]

Permitting arbitrary violations of RCLA would provide an alternative explanation of the pheasant-beef example. There is no systematic theory of violations of RCLA that would account for the example, however. Indeed, although many experiments show widespread violations of the reduction of

[8]For models where decision makers violate the reduction of compound lotteries axiom, see Kreps and Porteus [32], or Segal [43]. As Kreps and Porteus show, violations of the reduction axiom do not imply violations of expected utility *within each stage*. Of course, preferences over two or more stages must violate expected utility. For an example where at each stage the preferences can be represented by the *same* expected utility functional, but not between the stages, see Segal [43, Exp. 1].

compound lotteries axiom (see references in [43]), all experiments that we know show nonindifference between

1. A lottery that yields with probability p_k a ticket to a lottery that yields x_k with probability 1, $k = 1, \ldots, m$.

2. A lottery that with probability 1 yields a ticket to a lottery that with probability p_k pays x_k, $k = 1, \ldots, m$.

In other words,[9] the pure strategy $s_{0.5}$ must be indifferent in player i's preferences to the mixture $(s_1, \frac{1}{2}; s_0, \frac{1}{2})$. So even if we now have two pure-strategies equilibria (s_1, t_1) and $(s_0, t_{0.5})$, there is still the mixed strategy equilibrium where player i plays down, and player j plays left or right with probability $\frac{1}{2}$ each. As before, s_1 is the unique best response to both pure strategies of player j, but there exists an equilibrium in which player i uses his other pure strategy.[10]

ARE THERE SOLUTIONS TO GAMES THAT CANNOT BE EXPLAINED BY THIS MODEL? In other words, is this model at all restrictive? Yes. Consider a 2×2 game of two players, and suppose that the following are equilibrium points of this game (the pair (p_i, p_j) means that player k plays s_k^1 with probability p_k, $k = i, j$): $(1, 0.5)$, $(0, 0.5)$, $(0.5, 1)$, $(0.5, 0)$. It follows that if player i plays 0.5, then j's best response includes 0 and 1, hence by Axiom **IND**, it must also include 0.5. Likewise, if j plays 0.5, i's best response must include 0.5, hence $(0, 5, 0.5)$ is also an equilibrium.

IS THIS MODEL SENSITIVE TO THE CHOICE OF THE UTILITY FUNCTIONS? In standard models, where players care only about their own utility, taking a positive affine transformation of person i's utility will not change the nature of the game. Since eq. (1) involves utility levels of more than one player, will changing the vN–M utility index of a player change the nature of the game? This turns out to be one of the major obstacles in social choice theory, where one person's manipulation of utility may change the social optimum (see Weymark [47]). Despite its similarity to Harsanyi's utilitarian framework,

[9] The requirement that the last two are always indifferent is called time neutrality in Segal [43].

[10] If the preferences over outcomes are not linear in probabilities, then we can explain the example without our theory. For more on games with nonlinear utilities, see Crawford [14].

our model does not suffer from this problem. It is straightforward to check that if for $k = i, j$, $\widetilde{u}_k = \alpha_k u_k + \beta_k$ with $\alpha_k > 0$, then the utility function $u_i(\sigma) + a_{i,\sigma_j} u_j(\sigma)$ represents the same preferences as $\widetilde{u}_i(\sigma) + \widetilde{a}^j_{i,\sigma_j} \widetilde{u}_j(\sigma)$, where $\widetilde{a}^j_{i,\sigma_j} = \frac{\alpha_i}{\alpha_j} a_{i,\sigma_j}$.

CAN THIS MODEL BE EXTENDED TO MORE THAN TWO PLAYERS? Technically yes. Suppose there are N players. Let $\sigma = (\sigma_1, \ldots, \sigma_N)$, and let $\sigma_{-i} = (\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_N)$. By fixing the strategies of all players but i and j, we can use the structure of this paper to conclude that player i maximizes a function of the form

$$u_i(\sigma) + \sum_{k \neq i} a^k_{i,\sigma_{-i}} u_k(\sigma)$$

When there are more than two players, new issues arise involving how to define preferences over opponents' strategies. Probably the most difficult of them is how should player i evaluate person j's utility, when person j is nice to k but mean to ℓ. We do not deal with these issues here.

5 Related Literature

There is a large literature that documents instances in which agents reward kindness and punish nastiness in ways that are difficult to explain using conventional economic models of agents maximizing their material payoffs.[11] This section describes some of this literature and indicates, informally, the extent to which our approach is consistent with empirical findings.

[11] While this section discusses the relationship between the theoretical model and experiments conducted by economists, we should note that anthropologists have contributed several classic ethnographic studies describing exchange in non-market societies (Malinowski [35], Mauss [36], Sahlins [42], and Service [46]). These works have rich discuss reciprocal behavior in 'primitive' cultures. While economic theorists would be tempted to use repeated games to explain many of the observations because reciprocity arises in the context of repeated, non-anonymous interaction, some examples, in particular Malinowski's discovery that reciprocity occurs through extended and indirect chains, suggest an explanation based on non-selfish preferences would be less contrived. Gintis [23] reviews further evidence from other disciplines.

Many papers have been written on the ultimatum game and its simpler relation, the dictator game.[12] In the ultimatum game, one player offers a division of a fixed surplus, and the other player can either accept or reject. If it is accepted, player one's offer determines material payoffs for both players. Otherwise, neither player receives anything. In the dictator game, the second player must accept the first player's offer. Experiments find that the first player offers a positive amount to the second player in dictator games (amounts are sensitive to how the game is framed), and that in ultimatum games the first player offers even more, with the second player rejecting relatively small offers. Informal notions of fairness appear to play a role in the experiments (with equal division given prominence that it would not receive in a conventional theory). If the second player can make a counterproposal after rejecting the initial offer, disadvantageous counteroffers (offers in which player two rejects player one's offer only to make a counteroffer that yields him a lower monetary payoff than the rejected offer) arise (Ochs and Roth [37]). If agents have preferences that are described by eq. (1), then disadvantageous counteroffers arise for the same reason that unfair offers are rejected in the ultimatum game: In response to a nasty strategy, a player puts negative weight on his opponent's material payoff and is therefore willing to sacrifice his own material payoff in order to reduce his opponent's material payoff.

In models of gift exchange,[13] the second agent to move rewards behavior that is kind and, to a greater extent, punishes unkind behavior. The moonlighting game of Abbink, Irlenbusch, and Renner [2] is a paradigmatic example. Player one first decides on a transfer to player two. The transfer can be positive or negative. Positive transfers to player two are tripled (for every \$1 player one donates, player two receives \$3). The second player observes the first player's decision and can either reward or punish the first player. Rewards to player one are tripled. However, in order to reduce player one's payoff by \$1, player two must sacrifice \$1. In the unique subgame perfect equilibrium for this game (when players are motivated solely by their material payoffs), player two neither rewards nor punishes player one (be-

[12] References include: Bolton [7], Bolton and Zwick [9], Camerer [11], Costa-Gomes and Zauner [13], Eckel and Grossman [16], Güth [24], Güth, Schmittberger, and Schwarze [25], Hoffman, McCabe, and Smith [29, 30], and Roth [40].

[13] Experimental studies include Abbink, Irlenbusch, and Renner [2], Berg, Dickhaut, and McCabe [5], Fehr and Gächter [17], and Fehr, Gächter, and Kirchsteiger [18].

cause both actions are costly, but yield no material gain), while player one takes the maximum possible amount from player two. In experiments, player one typically makes a positive transfer to player two. Player two tends to reward positive transfers and punish negative ones. The results are consistent with equilibrium behavior under the assumptions that player two places a negative weight on player one's material payoff when player one takes money from player two and places a positive weight on player one's material payoff when one gives money to two.

There are games in which experimental results are more consistent with the predictions of equilibrium behavior of selfish players.[14] Games that operate like markets or auctions tend to replicate conventional equilibrium predictions for two reasons. First, the predictions of market models remain valid if only a small number of participants in the model behave selfishly. Since experimental results confirm the existence of some individuals that behave selfishly, the outcomes are consistent with other market participants having a preference for reciprocity. Second, in some games the kindness ordering $\gg_i$ is likely to be degenerate. For example, in the best-shot game in which two players sequentially make contributions c_i and material payoffs are $f(\max(c_1, c_2)) - c_i$, for $f(\cdot)$ increasing and $f'(0) > 1$, the theoretical prediction that player one makes no contribution and player two's contribution solves: $\max_{c_2} f(c_2) - c_2$ is consistent with experimental findings. These findings are also consistent with a model in which agents have a preference for reciprocity. Partial contributions from player one do not influence player two's maximum payoff, so our theory would not predict that player two would sacrifice material payoff in order to punish player one if player one contributes nothing.

Interesting attempts to explain these findings using models of learning and bounded rationality[15] or cognitive psychology (for example, Jacobsen

[14] Andreoni [3], Andreoni, Brown, and Vesterlund [4], Harrison and Hirschleifer [26], Ledyard [33], and Prasnikar and Roth [38] perform and or describe some of these experiments.

[15] Evolutionary models (for example, Gale, Binmore, and Samuelson [21]) and learning models (for example, Roth and Erev [41]) are useful ways to understand some of the experimental results. These models provide useful explanations of the failure to play the subgame perfect equilibrium in ultimatum games, but are less able to explain why agents make disadvantageous counterproposals in multi-period bargaining games. Abbink, Bolton, Sadrieh, and Tang [1] describe experiments that suggest fairness considerations better explain outcomes in experimental ultimatum games than learning models.

and Sadrieh [31]) contribute ideas that complement our approach. Most related to our approach, however, are models that assume equilibrium behavior of optimizing agents, but relax the assumption that agents seek to maximize their material utility. Bolton and Ockenfels [8],[16] Fehr and Schmidt [19], Levine [34], and Rabin [39] introduce models of this kind. The main contribution of our paper is that it provides an axiomatic foundation for using extended preferences of this sort in strategic settings; the other papers provide no formal justification for the functional forms that they use. The different papers provide similar, but distinct, predictions. It may be useful to contrast the approaches. Bolton and Ockenfels [8] and Fehr and Schmidt [19] present models in which agents have preferences that exhibit inequality aversion. In these models, agents are willing to sacrifice their own material payoff if by doing so they obtain a payoff that is closer to (some measure of what) other agents receive. In contrast to our approach, therefore, one player's preferences do not depend on the intentions of his opponents. Provided that it is possible to identify the comparison group to which an individual compares his payoff, it should be possible to distinguish the predictions between these theories and ours.[17]

The intentions of other players are important in Rabin's [39] model. He uses the theory of psychological games (Geanakoplos, Pearce, and Stacchetti [22]) to allow beliefs about an opponent's intentions to determine an equilibrium. Our approach demonstrates that intentions can be included in a game-theoretic analysis without using psychological games. Rabin makes restrictive assumptions about the weight that one player places on his opponent's utility.[18] Under these assumptions, he is able to derive some general properties of the fairness equilibria that he studies. These properties would not hold universally in our model.

Rabin explicitly assumes that an agent cares about his opponent's material payoff only as a response to intentions. For this reason, his approach

[16] Bolton [7] introduces a related approach.

[17] Experiments conducted by Blount [10] and Charness [12] finds evidence for reciprocal behavior in situations where first-move in ultimatum and gift-exchange games are random. These results suggest that the desire to reciprocate is not simply a response to an opponent's intentions.

[18] The functional forms Rabin chooses to describe fairness reduce his model's ability to explain some observations. Hausman [28] argues that Rabin's approach does not provide a satisfactory prediction in gift-exchange models.

can be distinguished from the inequality aversion models of Bolton and Ockenfels [8] and Fehr and Schmidt [19] or from our approach, which places few restrictions on the weight placed on opponent's utility. Nothing prevents a combination of the two approaches, however.

Rabin's use of psychological games permits his model to include one qualitative phenomena that would not arise using our approach. His paper provides an example of a game in which a player could use the same strategy in two distinct strict psychological equilibria. This cannot happen using conventional Nash equilibrium. In his analysis of the battle of the sexes, there are two strict equilibria in which j plays right. What is peculiar about this is that up is a strict best response to right in one situation, but down is a strict best response to right in another. This can happen for Rabin because in psychological games expectations matter. So, if i thinks that j is playing right because j thinks that i is playing down, then i thinks that j is being nice, and is willing to be nice (and play down). If i thinks that j is playing right because j thinks that i is playing up, then i thinks that j is being nasty, and is willing to be nasty (and play up).

In Levine's [34] model, the weight a player places on his opponents' material payoffs depends on what he thinks opponents' preferences are. Levine's players are inclined to make material sacrifices to benefit agents they believe to be altruistic and to harm agents they believe to be spiteful. These preferences do not depend on the behavior of opponents — unlike our approach the weight placed on opponents' strategies does not depend on the opponents' strategy choice. It should therefore be possible to identify the preferences of the agents in Levine's model using information from behavior in non-strategic settings.

The functional forms used by Bolton and Ockenfels [8], Fehr and Schmidt [19], and Rabin [39] all permit an explicit comparison between one player's payoffs and those of his opponents'. In this way, these models incorporate informal notion of fairness into their analyses. Some kind of fairness seems necessary to explain the prominence of particular distributions (for example, equal division) across experimental studies. Our approach does not restrict the way in which relative material payoffs influence a^G_{i,σ_j}. Hence, we cannot explain the prominence of equal-division outcomes in ultimatum games without making further restrictions to our theory.[19]

[19] The linear specification of inequality aversion used in Fehr and Schmidt [19] lacks the

Appendix A: A Numerical Example

This section provides an example to demonstrate the consistency of our maintained assumptions. Let a_{i,σ_j} represent the relation $\succeq_i^{opp}$ over Σ_j, such that

1. $\forall \sigma_j$, $a_{i,\sigma_j} \in [-1, 1]$, and
2. $\exists \sigma_j$ such that $a_{i,\sigma_j} = 0$.

For example, let $\sigma_j^1 \succeq_i^{opp} \sigma_j^2$ iff $\max_{\sigma_i} u_i(\sigma_i, \sigma_j^1) \geqslant \max_{\sigma_i} u_i(\sigma_i, \sigma_j^2)$, and let

$$a_{i,\sigma_j^*} = \frac{2 \exp\left(\max\limits_{\sigma_i} u_i(\sigma_i, \sigma_j^*)\right)}{\exp\left(\max\limits_{\sigma_j} \max\limits_{\sigma_i} u_i(\sigma_i, \sigma_j)\right)} - 1$$

Lemma 3 *Let $\sigma_i^1 \succeq_{i,\sigma_j} \sigma_i^2$ iff*

$$u_i(\sigma_i^1, \sigma_j) + a_{i,\sigma_j} u_j(\sigma_i^1, \sigma_j) \geqslant u_i(\sigma_i^2, \sigma_j) + a_{i,\sigma_j} u_j(\sigma_i^2, \sigma_j)$$

where $a_{i,\cdot}$ represent $\succeq_i^{opp}$ on Σ_j. Then the Reciprocal Altruism axiom is satisfied.

Proof Conditions (a)–(c) of the axiom imply:

(a) For strategies as in the axiom,

(i) $u_i(\sigma_i^1, \sigma_j^1) = u_i(\sigma_i^2, \sigma_j^2)$.
(ii) $u_j(\sigma_i^1, \sigma_j^1) = u_j(\sigma_i^2, \sigma_j^2)$.
(iii) $u_i(\bar{\sigma}_i^1, \sigma_j^1) = u_i(\bar{\sigma}_i^2 \sigma_j^2)$.
(iv) $u_j(\bar{\sigma}_i^1, \sigma_j^1) = u_j(\bar{\sigma}_i^2, \sigma_j^2)$.

(b) $a_{i,\sigma_j^2} > a_{i,\sigma_j^1}$.

(c) $u_i(\sigma_i^1, \sigma_j^1) + a_{i,\sigma_j^1} u_j(\sigma_i^1, \sigma_j^1) = u_i(\bar{\sigma}_i^1, \sigma_j^1) + a_{i,\sigma_j^1} u_j(\bar{\sigma}_i^1, \sigma_j^1)$.

---

ability to describe the dispersion of offers typically observed in ultimatum games even when players are assumed to have heterogeneous preferences, but non-linear specifications are flexible enough to fit the data.

We want to show that $\sigma_i^2 \succeq_{i,\sigma_j^2} \bar{\sigma}_i^2$ iff $u_j(\sigma_i^2, \sigma_j^2) \geqslant u_j(\bar{\sigma}_i^2, \sigma_j^2)$. By the definition of $\succeq_{i,\sigma_j^2}$ and by (a-i) and (a-iii),

$$\begin{aligned}
&\sigma_i^2 \succeq_{i,\sigma_j^2} \bar{\sigma}_i^2 \iff \\
&u_i(\sigma_i^2, \sigma_j^2) + a_{i,\sigma_j^2} u_j(\sigma_i^2, \sigma_j^2) \geqslant u_i(\bar{\sigma}_i^2, \sigma_j^2) + a_{i,\sigma_j^2} u_j(\bar{\sigma}_i^2, \sigma_j^2) \iff \\
&u_i(\sigma_i^1, \sigma_j^1) + a_{i,\sigma_j^2} u_j(\sigma_i^2, \sigma_j^2) \geqslant u_i(\bar{\sigma}_i^1, \sigma_j^1) + a_{i,\sigma_j^2} u_j(\bar{\sigma}_i^2, \sigma_j^2)
\end{aligned}$$

Subtract (c) from the last inequality to obtain, together with (a-ii) and (a-iv), that

$$[a_{i,\sigma_j^2} - a_{i,\sigma_j^1}] u_j(\sigma_i^2, \sigma_j^2) \geqslant [a_{i,\sigma_j^2} - a_{i,\sigma_j^1}] u_j(\bar{\sigma}_i^2, \sigma_j^2)$$

Which is equivalent to

$$[a_{i,\sigma_j^2} - a_{i,\sigma_j^1}][u_j(\sigma_i^2, \sigma_j^2) - u_j(\bar{\sigma}_i^2, \sigma_j^2)] \geqslant 0$$

It is therefore sufficient to prove that $a_{i,\sigma_j^2} > a_{i,\sigma_j^1}$, which follows by (b). ■

References

[1] Abbink, K., G. Bolton, A. Sadrieh, and F.-F. Tang, 1998: "Adaptive Learning versus Punishment in Ultimatum Bargaining," Bonn.

[2] Abbink, K., B. Irlenbusch, and E. Renner, 1997: "The moonlighting game," Bonn.

[3] Andreoni, J., 1995: "Cooperation in public goods experiments: Kindness or confusion?" *American Economic Review* 85:891–904.

[4] Andreoni, J., P. Brown, and L. Vesterlund, 1997: "Fairness, selfishness and selfish fairness: Experiments on games with unequal equilibrium payoffs," Wisconsin.

[5] Berg, J., J. Dickhaut, and K. McCabe, 1995: "Trust, reciprocity and social history," *Games and Economic Behavior* 10:122–142.

[6] Border, K., 1985: "More on Harsanyi's utilitarian cardinal welfare function," *Social Choice and Welfare* 1:279–281.

[7] Bolton, G., 1991: "A comparative model of bargaining: Theory and evidence," *American Economic Review* 81:1096–1135.

[8] Bolton, G. and A. Ockenfels, 1997: "ERC: A theory of equity, reciprocity and competition," Penn State.

[9] Bolton, G. and R. Zwick, 1995: "Anonymity versus punishment in ultimatum bargaining," *Games and Economic Behavior* 10:95–121.

[10] Blount, S., 1995: "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," 63:131–144.

[11] Camerer, C., 1997: "Progress in behavioral game theory," *Journal of Economic Perspectives* 11:267–288.

[12] Charness, G. 1996: "Attribution and reciprocity in a labor market: An experimental investigation," UC, Berkeley.

[13] Costa-Gomes, M. and K. G. Zauner, 1998: "Ultimatum Bargaining Behavior in Israel, Japan, the United States, and Yugoslavia: A Social Utility Analysis," AGSM Working Paper.

[14] Crawford, V., 1990: "Equilibrium without independence," *Journal of Economic Theory* 50:127–154.

[15] Debreu, G., 1960: "Topological methods in cardinal utility theory," in K.J. Arrow, S. Karlin, and P. Suppes: *Mathematical Methods in the Social Sciences*. Stanford University Press, Stanford.

[16] Eckel, C. and P. Grossman, 1996: "Altruism in anonymous dictator games," *Games and Economic Behavior* 16:181–191.

[17] Fehr, E. and S. Gächter, 1998: "Reciprocity and economics: The economic implications of homo reciprocans," *European Economic Review* 42:845–859.

[18] Fehr, E., S. Gächter, and G. Kirchsteiger, 1997: "Reciprocity as a contract enforcement device: Experimental evidence," *Econometrica* 65:833–860.

[19] Fehr, E. and K. M. Schmidt, 1998: "A theory of fairness, competition, and cooperation," Zürich.

[20] Fishburn, P.C., 1984: "On Harsanyi's utilitarian cardinal welfare theorem," *Theory and Decision* 17:21–28.

[21] Gale, J., K. Binmore, and L. Samuelson, 1995: "Learning to be imperfect: The ultimatum game," *Games and Economic Behavior* 8:56–90.

[22] Geanakoplos, J., D. Pearce, and E. Stacchetti, 1989, "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1, 60–79.

[23] Gintis, H., 1998: "The Individual in Economic Theory," U. Mass.

[24] Güth, W., 1995: "On ultimatum bargaining experiments — a personal review," *Journal of Economic Behavior and Organization* 27:329–344.

[25] Güth, W., R. Schmittberger, and B. Schwarze, 1982: "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior and Organization* 3:367–388.

[26] Harrison, G. and J. Hirschleifer, 1989: "An experimental evaluation of the weakest link, best shot models of public goods," *Journal of Political Economy* 97:201–225.

[27] Harsanyi, J. C., 1955: "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility," *Journal of Political Economy* 63:309–321.

[28] Hausman, D. 1998, "From simple game forms to games: The incredible journey," U. Wisconsin

[29] Hoffman, E., K. McCabe, and V. Smith, 1995: "Ultimatum and dictator games," *Journal of Economic Perspectives* 9:236–239.

[30] Hoffman, E., K. McCabe, and V. Smith, 1996: "Social distance and other-regarding behavior in dictator games," *American Economic Review* 86:653–660.

[31] Jacobsen, E. and A. Sadrieh, 1996: "Experimental Proof for the Motivational Importance of Reciprocity," Bonn.

[32] Kreps, D.M. and E.L. Porteus, 1978: "Temporal resolution of uncertainty and dynamic choice theory," *Econometrica* 46:185–200.

[33] Ledyard, J., 1995: "Public goods: A survey of experimental research," in J. Kagel and A. Roth (eds.), *Handbook of Experimental Economics*, Princeton: Princeton University Press.

[34] Levine, D. 1998, "Modelling Altruism and Spitefulness in Game Experiments," *Review of Economic Dynamics* 1:593–622.

[35] Malinowski, B., 1961 (originally, 1922): *Argonauts of the Western Pacific*, New York: Dutton.

[36] Mauss, M., 1990 (originally, 1925): *The Gift: The Form and Reason for Exchange in Archaic Societies*, London: Routledge.

[37] Ochs, J. and A. Roth, 1989: "An experimental study of sequential bargaining," *American Economic Review* 78:355–384.

[38] Prasnikar, V. and A. Roth, 1992: "Considerations of fairness and strategy: Experimental data from sequential games," *Quarterly Journal of Economics* 79:355–384.

[39] Rabin, M., 1993, "Incorporating fairness into game theory," *American Economic Review* 83:1281-1302.

[40] Roth, A., 1995: "Bargaining experiments," in J. Kagel and A. Roth (eds.), *Handbook of Experimental Economics*, Princeton: Princeton University Press.

[41] Roth, A. and I. Erev, 1995: "Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior* 8:164–212.

[42] Sahlins, M., 1968: *Tribesmen*, Englewood Cliffs, N.J.: Prentice-Hall.

[43] Segal, U., 1990: "Two-stage lotteries without the reduction axiom," *Econometrica* 58:349–377.

[44] Segal, U., 1992: "Additively separable representations on non-convex sets," *Journal of Economic Theory* 56:89–99.

[45] Segal, U. and J. Sobel, 1999: "Max, min, and sum," mimeo.

[46] Service, E., 1966: *The Hunters*, Englewood Cliffs, N.J.: Prentice-Hall.

[47] Weymark, J.A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In J. Elster and J.E. Roemer (eds.): *Interpersonal Comparisons of Well-Being*, ch. 8. Cambridge: Cambridge University Press.