Available online at www.sciencedirect.com
ScienceDirect

Journal of Economic Theory 136 (2007) 197–216

JOURNAL OF Economic Theory
www.elsevier.com/locate/jet

# Tit for tat: Foundations of preferences for reciprocity in strategic settings

Uzi Segal[a], Joel Sobel[b,*]

[a]Department of Economics, Boston College, Chestnut Hill, MA 02467, USA
[b]Department of Economics, University of California, San Diego, La Jolla, CA 92093, USA

Received 19 November 2004; final version received 14 July 2006
Available online 7 September 2006

**Abstract**

This paper assumes that in addition to conventional preferences over outcomes, players in a strategic environment have preferences over strategies. It provides conditions under which a player's preferences over strategies can be represented as a weighted average of the utility from outcomes of the individual and his opponents. The weight one player places on an opponent's utility from outcomes depends on the players' joint 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 identifies restrictions that the theory places on rational behavior.

*JEL classification:* C72; D63

*Keywords:* Reciprocity; Game theory; Extended preferences; Representation theorems

## 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 the domains of individuals' preferences depend only on those aspects of an allocation that directly influence their material well-being.

* Corresponding author. Fax: +1 858 534 7040.
*E-mail addresses:* uzi.segal@bc.edu (U. Segal), jsobel@ucsd.edu (J. Sobel).

0022-0531/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jet.2006.07.003

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 relations defined on lotteries over outcomes. Our starting point is to assume in addition 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 $\sigma^1$, $\sigma^2$, and $\sigma^3$, and for all $\alpha \in (0,1]$, $\sigma^1 \succcurlyeq \sigma^2$ iff $\alpha\sigma^1 + (1-\alpha)\sigma^3 \succcurlyeq \alpha\sigma^2 + (1-\alpha)\sigma^3$. This, together with continuity and transitivity, implies that preferences over strategies can be represented by an expected utility functional. This utility does not have to agree with the expected utility from payoffs obtained when the player uses this strategy.

Section 2 presents the basic representation theorem. We show that in a fixed game $G$, and given that a strategy profile $\sigma^*$ describes the expected pattern of play, player $i$'s preferences over his own strategies $\sigma_i$, given that the rest of players are playing $\sigma^*_{-i}$, are represented by a utility function of the form:

$$u_i(\sigma_i, \sigma^*_{-i}) + \sum_{j \neq i} a^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i}).$$

The representation is a weighted sum of the players' utilities, where the weight player $i$ gives to player $j$'s utility, $a^j_{i,\sigma^*}$, depends on the entire strategy profile $\sigma^*$. This result is a consequence of a theorem due to Harsanyi [23]. The critical assumption is that if, given a fixed strategy profile of player $i$'s opponents, two of player $i$'s strategies lead to the same distribution of expected utility (from outcomes) for *all* players, then player $i$ is indifferent between these two strategies. The coefficients $a^j_{i,\sigma^*}$ represent the degree to which player $i$ is willing to take person $j$'s interests into consideration. In standard theory, $a^j_{i,\sigma^*} \equiv 0$ for $j \neq i$. Positive values of the coefficient suggest that player $i$ is willing to sacrifice his non-strategic payoff from outcomes in order to increase the payoff of player $j$. Negative values suggest a willingness to sacrifice non-strategic payoff in order to lower player $j$'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 sacrifices (lowering his utility from outcomes) to increase or decrease his opponent's payoff in the same strategic setting.

The analysis in Section 2 relaxes the assumption that a player's preferences over outcomes completely determine his preferences over strategies. Consequently, the theory permits a much wider range of preferences than standard theory. In Section 3 we ask whether the theory provides any restrictions at all. We place no restrictions on the functional form identified in Section 2 and show that there are outcomes that an outside observer, knowing the players' preferences over payoffs, but not their preferences over strategies, could rule out either on the basis of rationality or equilibrium behavior. We characterize dominance relationships when players have general preferences over strategies.

Section 4 shows how our model can be used to organize some experimental results in which players systematically exhibit behavior that is not consistent with standard game theory. Section 5 discusses a few examples that illustrate important features of our model. In particular, we argue that equilibrium outcomes depending on whether mixed strategy equilibria are viewed as equilibria in beliefs or the result of deliberate randomization.

Section 6.1 discusses the relationship between our model and psychological games. Psychological games were introduced by Geanakoplos et al. [21] to study strategic situations in which the beliefs that players have about their opponents enter independently into preferences. Rabin [27] and others have used psychological games to model attitudes towards fairness and reciprocity in games. Our approach amounts to a reformulation of an interesting class of psychological games. The representation theorem provides an axiomatic foundation for preferences that are linear in opponents' utilities over outcomes. By identifying our model with psychological games, we demonstrate that one cannot reproduce our results under standard assumptions.

A number of papers have proposed models of extended preferences designed to describe and organize experimental findings that are inconsistent with narrow notions of rationality. A comprehensive survey is provided in Sobel [29]. We describe some of these contributions in Section 6.2.

## 2. Representation theorems

This section introduces our model, states the basic axioms, and provides a representation theorem for preferences over strategies.

Let $X_i$ be the space of outcomes to player $i, i = 1, \dots, I$. Each player has preferences $\succsim_i^{\text{out}}$ over $\Delta(X_i)$, the space of lotteries over $X_i$. A game is a finite collection $\mathbf{s}_i = \{s_i^1, \dots, s_i^{n_i}\}$ of strategies for player $i, i = 1, \dots, I$, together with the payoff function $O : \prod_{j=1}^I \mathbf{s}_j \to \prod_{j=1}^I X_j$. Let $\sum_i$ be the space of mixed strategies of player $i$ and extend $O$ to be from $\prod_{j=1}^I \sum_j$ to $\prod_{j=1}^I \Delta(X_j)$. Throughout the paper, $\sum = \prod_{j=1}^I \sum_j$. [1]

Given a game, player $i$ has a complete and transitive preference relation over $\sum_i$. These preferences depend of course on $\sigma_{-i}$, the strategies of other players, and possibly also on $i$'s interpretation of these strategies or the "context" in which the game is being played. We assume that the context is summarized by a mixed strategy profile $\sigma^*$, which we interpret as a description of the conventional way in which the game is played. [2] It is within this context that players rank their available strategies. Formally, given $\sigma^* = (\sigma_i^*, \sigma_{-i}^*)$, player $i$ has preferences $\succsim_{i,\sigma^*}$ over $\sum_i$. The statement $\sigma_i \succ_{i,\sigma^*} \sigma_i'$ says the given the context $\sigma^*$, player $i$ would prefer to play $\sigma_i$ rather than $\sigma_i'$.

Preferences over outcomes do not depend on the strategic context and can be elicited (in principle) by observing choices over lotteries over outcomes. Similarly, an agent's preferences over strategies can be elicited by asking him to choose between lotteries over strategies. Once preferences are elicited, it is possible to check whether they satisfy the axioms we impose below. [3]

In this framework, two solution concepts are relevant.

**Definition 1.** A Nash equilibrium is a strategy profile $\sigma^*$ in which agent $i$'s strategy, $\sigma_i^*$, is maximal according to $\succsim_{i,\sigma^*}$.

In the definition of Nash equilibrium player $i$ selects $\sigma_i$, treating the context $\sigma^*$ as fixed. Player $i$ can vary his strategy, but potential deviations do not influence the context that he uses to judge

[1] Strategies ($\mathbf{s}_i$), strategy sets ($\sum_i$), outcome functions ($O$) and preferences over strategies (see below) can vary with the game. Our analysis always concentrates on a fixed game, however, so we suppress this dependence in our notation.

[2] Alternatively, for each $i$, $\sigma_{-i}^*$ represents player $i$'s beliefs about how his opponents play the game.

[3] Strictly speaking, we must maintain the self-interest assumption (stated below) in order to rule out the possibility that elicited preferences reveal indifference rather than strict preference.

his opponents' behavior. When interpreting how nice an opponent *j* is, player *i* should evaluate *j*'s action on the basis of what player *j* thinks *i* is going to do ($\sigma_i^*$), not on what he actually does ($\sigma_i$), as player *j* cannot see *i* deviating. Of course, in equilibrium *j*'s beliefs about *i*'s choice of strategy and *i*'s actual choice should agree.

As we discuss in Section 5 the interpretation of the equilibrium depends critically on whether one believes that players consciously randomize.[4] For this reason, we define equilibrium in beliefs.

**Definition 2.** The belief profile $\mu^* \in \sum$ forms an equilibrium in beliefs if $\mu_i^*(s_i^k) > 0$ implies that $s_i^k \succsim_{i,\mu^*} s_i$ for all $s_i \in \mathbf{s}_i$.

We interpret $\mu_{-i}^*$ as player *i*'s beliefs over his opponents' strategy choice. In an equilibrium in beliefs, each player believes his opponents place positive probability only on those pure strategies that are maximal according to their preferences over strategies. That is, in equilibrium, a player believes that his opponent only selects a strategy that is a best response to beliefs about how others play. Player *i* may have non-degenerate beliefs about the behavior of his opponents' strategy because *i* does not know what pure strategy other players will play. Nash equilibrium, taken literally, requires that the uncertainty be the result of conscious randomization. Equilibrium in beliefs allows the possibility that other players do not randomize, but player *i* is uncertain about which best responses they use.

We assume that the preferences $\succsim_{i,\sigma^*}$ satisfy the following axioms.

(C) Continuity: For every $\sigma_i \in \sum_i$ and $\sigma^* \in \sum$, the sets $\{(\sigma_i', \sigma^*) : \sigma_i' \succsim_{i,\sigma^*} \sigma_i\}$ and $\{(\sigma_i', \sigma^*) : \sigma_i \succsim_{i,\sigma^*} \sigma_i'\}$ are closed subsets of $\sum_i \times \sum$.

(IND) Independence: For every $\sigma_i, \sigma_i', \sigma_i'' \in \sum_i$, $\sigma^* \in \sum$, and $\alpha \in (0, 1]$, $\sigma_i \succsim_{i,\sigma^*} \sigma_i'$ iff $\alpha\sigma_i + (1-\alpha)\sigma_i'' \succsim_{i,\sigma^*} \alpha\sigma_i' + (1-\alpha)\sigma_i''$.

The following straightforward existence result, Lemma 1, follows from standard arguments.[5]

**Lemma 1.** *If, for a given game, all 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 $\succsim_i^{\text{out}}$ over lotteries of payoffs satisfy the assumptions of expected utility theory.

It follows by this axiom that there are vN-M utility functions $u_i : X_i \to \Re$ such that the preferences $\succsim_i^{\text{out}}$ over $\Delta(X_i)$, the set of lotteries over $X_i$, are represented by the expected value of the utility $u_i$ from their payoffs. These functions are unique up to affine transformations of the form $\zeta_i u_i + \eta_i$, with $\zeta_i > 0$. To simplify notation, denote by $u_i(\sigma)$ the expectation of the utility $u_i$ player *i* receives from the lottery $O_i(\sigma)$ ($O_i$ is the lottery person *i* receives from $O$). Let $u(\sigma) = (u_1(\sigma), \ldots, u_I(\sigma))$.

(SI) Self interest: Suppose that $O_j(\sigma_i, \sigma_{-i}^*) \sim_j^{\text{out}} O_j(\sigma_i', \sigma_{-i}^*)$ for all $j \neq i$. Then $\sigma_i \succsim_{i,\sigma^*} \sigma_i'$ if, and only if, $O_i(\sigma_i, \sigma_{-i}^*) \succsim_i^{\text{out}} O_i(\sigma_i', \sigma_{-i}^*)$.

[4] To our knowledge, the importance of the interpretation of mixed strategies in models of reciprocity has not received detailed attention in the literature. Dufwenberg and Kirchsteiger [12] are certainly aware of the issue. They interpret randomization in their model as a result of incomplete information about population behavior rather than individual choice.

[5] All proofs are contained in the Appendix.

In terms of utilities, this axiom can be expressed as

- Suppose that $u_j(\sigma_i, \sigma^*_{-i}) = u_j(\sigma'_i, \sigma^*_{-i})$ for all $j \neq i$. Then $\sigma_i \succcurlyeq_{i,\sigma^*} \sigma'_i$ if, and only if, $u_i(\sigma_i, \sigma^*_{-i}) \geqslant u_i(\sigma'_i, \sigma^*_{-i})$.

The self-interest axiom implies in particular,

$$\text{If } u(\sigma_i, \sigma^*_{-i}) = u(\sigma'_i, \sigma^*_{-i}) \quad \text{then } \sigma_i \sim_{i,\sigma^*} \sigma'_i. \tag{1}$$

The axioms are weaker than the assumptions on preferences in standard game theory and, as we show below in Fact 1, yield a correspondingly weaker representation theorem. Standard game theory does not assume that players have preferences over strategies, but it assumes that players act to maximize their utility over outcomes given the behavior of opponents. Preferences over outcomes induce preferences over strategies in a natural way: $\sigma_i \succcurlyeq_{i,\sigma^*} \sigma'_i$ iff $u_i(\sigma_i, \sigma^*_{-i}) \geqslant u_i(\sigma'_i, \sigma^*_{-i})$. The preferences over strategies derived from standard game theory therefore satisfy the continuity, independence, expected utility, and self-interest axioms. The continuity axiom is weaker in our framework because it assumes that preferences over strategies do not change too much in response to small changes in the context $\sigma^*$, while standard game theory assumes that these preferences are independent of $\sigma^*$. The standard self-interest axiom requires that player $i$'s preferences over strategies always agree with his preferences over outcomes. We require agreement only when other players are indifferent (relative to their preferences over outcomes) between him playing $\sigma'_i$ and $\sigma''_i$.[6]

The structure of the model so far resembles that of Harsanyi's social choice theory [23]. 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 obtain the following fact.

**Fact 1.** *Assume expected utility, continuity, independence, and condition* (1). *For any given choice of the utilities $u_1, \ldots, u_n$, the preferences $\succcurlyeq_{i,\sigma^*}$ over $\sum_i$ can be represented by*

$$V_{i,\sigma^*}(\sigma_i) = \sum_j a^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i}). \tag{2}$$

*Moreover, if the utility allocations induced by person i's strategies*

$$A_i(\sigma^*_{-i}) = \left\{ u(\sigma_i, \sigma^*_{-i}) : \sigma_i \in \sum\nolimits_i \right\}$$

*has non-empty interior in $\Re^I$ then the coefficients $a^j_{i,\sigma^*}$, $j = 1, \ldots, I$, are unique up to multiplication of all by the same positive number.*

If we replace condition (1) with the self-interest axiom we get the following further restriction on the representation function.

[6] There is a sense in which the self-interest axiom is restrictive. The premise of the axiom is that player $i$'s attitudes towards an opponent $j$ depend on $j$'s preferences over outcomes. The axiom essentially rules out an interdependence in which player $i$'s preferences over strategies depend on player $j$'s preferences over strategies.

**Theorem 1.** *Given the expected utility, continuity, independence, and self-interest axioms,* $a^i_{i,\sigma^*}$ *can be chosen to be equal to one. That is, the preferences* $\succcurlyeq_{i,\sigma^*}$ *can be represented by*

$$V_{i,\sigma^*}(\sigma_i) = u_i(\sigma_i, \sigma^*_{-i}) + \sum_{j \neq i} a^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i}). \tag{3}$$

*Moreover, if* $A_i(\sigma^*_{-i})$ *has non-empty interior in* $\Re^I$, *then the coefficients* $a^j_{i,\sigma^*}$, $j \neq i$, *are unique.*

If the preferences $\succcurlyeq_{i,\sigma^*}$ can be represented by (3) we say that player *i* has reciprocity preferences. Note that some (or even all) of the $a^j_{i,\sigma^*}$, $j \neq i$, may be negative.

Theorem 1 states conditions under which the coefficients $a^j_{i,\sigma^*}$, $j \neq i$, are uniquely determined by the utility functions over outcomes. Since our basic assumptions are restrictions on material preferences, not utility functions, one could ask whether these coefficients are uniquely determined by preferences. In our framework, the answer is no. To see this, suppose that we write $\tilde{u}_k = \zeta_k u_k + \eta_k$ with $\zeta_k > 0$. It follows that the utility function $u_i(\sigma_i, \sigma^*_{-i}) + \sum_{j \neq i} a^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i})$ represents the same preferences as

$$\tilde{u}_i(\sigma_i, \sigma^*_{-i}) + \sum_{j \neq i} \tilde{a}^j_{i,\sigma^*} \tilde{u}_j(\sigma_i, \sigma^*_{-i}),$$

where $\tilde{a}^j_{i,\sigma^*} = \zeta_i a^j_{i,\sigma^*} / \zeta_j$.

In the sequel we fix the vNM utilities $u_1, \ldots, u_n$. The values of the coefficients $a^j_{i,\sigma^*}$ are irrelevant (beyond their sign). However, *changes* in their values are relevant, for example, when $\sigma^*$ changes.

## 3. Predictions with extended preferences

This paper expands the set of allowable preferences in strategic settings. Since we impose fewer restrictions on preferences than standard theory, it is obvious that the model permits more predictions. In this section, we investigate the restrictions on behavior placed by the model. In order to do so, we ask two questions: What restrictions does rationality place on individual behavior in our model? Given these restrictions, what kinds of outcomes would an observer expect to see? A complete theory would predict a single strategy profile for every game. A weak theory would place no restrictions at all on observations. Standard game theory falls somewhere between these extremes. We will show that our theory also falls between the extremes, but is weaker than standard game theory.

There is an extensive literature in standard game theory that takes a decision-theoretic view of behavior in games. At one extreme is the approach associated with Bernheim [6] and Pearce [26]. They demonstrate that common knowledge of rationality leads to the conclusion that players will only use rationalizable strategies.[7] Other authors have justified equilibrium predictions under more restrictive assumptions about beliefs. Aumann [2] demonstrates that common knowledge of rationality and a common prior assumption lead to correlated equilibria. Aumann's model assumes that there is common uncertainty about the state of the world, and the state of the world includes the strategies to be played in the game. Our decision-theoretic model of preferences

[7] In two-player games, rationalizable strategies are precisely those that survive iterative deletion of strictly dominated strategies.

over strategies makes no assumption about the structure of preferences when there is (individual) uncertainty about the context. The profile $\sigma^*$ must be fixed in order to deduce the representation (3). Consequently, the approach of Aumann [2] goes beyond the scope of this paper.

There is an alternative epistemic approach that is consistent with our model. Aumann and Brandenberger [3] show, for two-player games, if there is mutual knowledge of the payoff functions, rationality, and conjectures over opponent's strategies, then the conjectures constitute a Nash equilibrium. [8] The essential difference between these results is that Aumann and Brandenberger [3] assume that players share common beliefs about strategies. The assumption that players have common beliefs about strategies is natural in our setting, where these shared beliefs become the "context" used to determine preferences over strategies. Hence, Aumann and Brandenberger's framework is an appropriate starting point for our model. With that understanding that conjectures over strategies determine strategic context, Aumann and Brandenberger's result extends to our framework without change. Consequently, there are epistemic foundations for looking at Nash equilibria as the set of predictions in our model.

We next seek an understanding of the size of the set of Nash equilibria. In standard games, the set of Nash equilibria is typically a small set of the feasible strategy profiles. Not surprisingly, the set is larger in our setting. The main result of the section provides a characterization.

We take the perspective of an outside observer who wishes to make predictions given a strategic setting. It is natural to assume that the observer knows the players' preferences over outcomes. Standard analysis makes this assumption and it is difficult to see how one can make predictions about strategic behavior without it. In standard analysis, once preferences over outcomes are known, preferences over strategies are determined. In our approach, it is possible for an outside observer to know preferences over outcomes without knowing preferences over strategies. We ask, therefore, what would an outside observer view as the set of possible outcomes of a game under the assumption that players have mutual knowledge of payoff functions, rationality, and conjectures over opponent's strategies when the observer knows the strategy sets, the material preferences, and that preferences over strategies are characterized by a function of the form (3). That is, we seek to describe the set of possible Nash equilibria when preferences over strategies obey our axioms, preferences over outcomes are known, but the weights $a_{i,\bullet}$ are unknown.

We find that the observer thinks that rational players can play precisely those strategies that are, in a sense we make precise, undominated. We begin the investigation by looking at which strategies can be rational responses to beliefs. Strategies that are never best responses will not be played in equilibrium.

We say that $\sigma_i$ is a best response to beliefs $\sigma_j$ if there exists an $a$ such that

$$u_i(\sigma_i, \sigma_j) + a u_j(\sigma_i, \sigma_j) \geqslant u_i(\sigma_i', \sigma_j) + a u_j(\sigma_i', \sigma_j) \tag{4}$$

for all $\sigma_i' \in \sum_i$. $\sigma_i$ is a best response if there exists $\sigma_j$ and $a$ for which (4) holds.

Our first result, Lemma 2, provides necessary and sufficient conditions on utilities over outcomes for a mixed strategy $\sigma_i$ to be a best response to some strategy of player $j$. Theorem 2, generalizing results from standard game theory, relates best responses to undominated strategies: in two-player games a strategy is undominated if, and only if, it is a best response to some mixed strategy choice of the opponent. Theorem 3 characterizes the set of strategy profiles that can be Nash equilibria for some preferences over strategies.

[8] Slightly stronger conditions are needed to obtain the same conclusion in $n$-player games.

Recall that we say that player $i$ has reciprocity preferences if $\succcurlyeq_{i,\sigma^*}$ can be represented by

$$V_{i,\sigma^*}(\sigma_i) = u_i(\sigma_i, \sigma^*_{-i}) + a^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i}).$$

**Lemma 2.** *There exist reciprocity preferences for player i such that $\sigma_i$ is a best response to $\sigma_j$ if, and only if, $\{u(\sigma'_i, \sigma_j) : \sigma'_i \in \sum_i\} \cap \{(w_i, u_j(\sigma)) : w_i > u_i(\sigma)\} = \emptyset$.*

Notice that the sets compared in the statement of Lemma 2 are subsets of $\Re^2$ representing pairs of utilities. In words, Lemma 2 states that a mixed strategy $\sigma_i$ for player $i$ can be a best response to his opponent's strategy $\sigma_j$ for some supporting weights $a_{i,\sigma^*}$ unless given his opponent's strategy, player $i$ can improve his payoff without changing player $j$'s payoff. In standard game theory, whether $\sigma_i$ is a best response does not depend on $u_j$.

**Remark 1.** Lemma 2 requires that the set of pure strategies be finite. For example, assume that $\sum_i = [-1, 1]$, $\sum_j = \{0\}$, and that $u(x, 0) = (x, -x^2)$. In this game, player $j$ is a dummy. It is clear that

$$\{u(x, 0) : x \in [-1, 1]\} \cap \{(w_i, 0) : w_i > 0\} = \emptyset$$

so the condition in Lemma 2 is satisfied. On the other hand, there is no $a$ such that $x = 0$ solves $\max_{x\in[-1,1]} x - ax^2$. That is, there is no $a$ that makes $x_i = 0$ a best response (to $x_j = 0$). The proof of Lemma 2 fails because it is only possible to separate the disjoint sets $\{u(x, 0) : x \in [-1, 1]\}$ and $\{(w, 0) : w > 0\}$ with a horizontal line. We could salvage the conclusion of Lemma 2 if we permitted player $i$'s preferences over strategies to place zero weight on his utility from outcomes.

Consider now the notion of dominance. With no restrictions on $a_{i,\sigma^*}$, the appropriate notion of dominance is:

**Definition 3.** $\sigma'_i \in \sum_i$ strictly dominates $\sigma_i \in \sum_i$ if for all $\sigma_j \in \sum_j$, $u_i(\sigma'_i, \sigma_j) > u_i(\sigma_i, \sigma_j)$ and $u_j(\sigma'_i, \sigma_j) = u_j(\sigma_i, \sigma_j)$. If there does not exist a strategy $\sigma'_i$ that strictly dominates $\sigma_i$, then we say that $\sigma_i$ is undominated.

In order for a strategy of player $i$ to be dominated in our setting, there must be another strategy that provides, for each strategy of player $j$, a higher payoff from outcomes to player $i$ and the same payoff from outcomes to player $j$.

The next lemma generalizes the connection between dominance (a decision-theoretic implication of rationality) and best responding (a strategic implication of rationality) that is familiar from standard game theory. Dominated strategies are those strategies that are never best responses.

**Theorem 2.** *The strategy $\sigma_i \in \sum_i$ is undominated if, and only if, there exist reciprocity preferences for player i and $\sigma_j \in \sum_j$ such that $\sigma_i$ is a best response to $\sigma_j$.*

Just as in standard game theory, rational players will assume that their opponents never play strictly dominated strategies. Once strictly dominated strategies are deleted, it may be possible to delete further strategies. In this way, one can develop a theory of rationalizability for games with reciprocal preferences that is completely analogous to the standard theory.

We now turn to Nash equilibrium. Lemma 2 has an immediate consequence.

**Theorem 3.** *The strategy profile $\sigma = (\sigma_i, \sigma_j)$ can be a Nash equilibrium for some reciprocity preferences if, and only if,*

(1) $\{u(\sigma'_i, \sigma_j) : \sigma'_i \in \sum_i\} \cap \{(w_i, u_j(\sigma)) : w_i > u_i(\sigma)\} = \emptyset$; *and*
(2) $\{u(\sigma_i, \sigma'_j) : \sigma'_j \in \sum_j\} \cap \{(u_i(\sigma), w_j) : w_j > u_j(\sigma)\} = \emptyset$.

In standard game theory, 2 × 2 games with generic preferences over outcomes have at most two pure-strategy equilibria. By contrast, Theorem 3 states that an undominated strategy profile $\sigma$ can be a Nash equilibrium of the game with an appropriate choice of weights. It follows that for generic preferences over outcomes, all pure strategy combinations of a 2 × 2 game with generic preferences over outcomes can be Nash equilibria.[9]

**Remark 2.** Let $S^*$ be the set of pairs of pure strategies for which the two conditions of Theorem 3 hold. Then, unless we put some restrictions on the functions $a^j_{i,\bullet}$ and $a^i_{j,\bullet}$, we can define these functions in such a way that all elements of $S^*$ become Nash equilibria.

The following example demonstrates that we cannot guarantee that there exist functions $a^j_{i,\bullet}$ and $a^i_{j,\bullet}$ such that all *mixed* strategies that satisfy the two conditions of Theorem 3 are Nash equilibria. That is, Remark 2 does not extend to the entire set of mixed strategies.

**Example 1.** Consider the game:

| | $s_j^1$ | $s_j^2$ |
|---|---|---|
| $s_i^1$ | 0, 0 | 2, 2 |
| $s_i^2$ | 1, 2 | 1, 1 |

Observe that condition (1) of Theorem 3 holds when $\sigma_j \neq (\frac{1}{3}, \frac{2}{3})$ or $\sigma = ((1, 0), (\frac{1}{3}, \frac{2}{3}))$. Similarly, Condition (2) holds when $\sigma_i \neq (0, 1)$ or when $\sigma = ((0, 1), (1, 0))$. Hence, the set of mixed strategies for which the conditions in Theorem 3 hold is neither open nor closed. It follows from the continuity axiom, however, that the set of Nash equilibria of a game must be closed. Therefore it is not possible to find functions $a^j_{i,\bullet}$ and $a^i_{j,\bullet}$ such that the entire set of mixed strategies for which the conditions in Theorem 3 hold are Nash equilibria.

**Remark 3.** Straightforward generalizations of Lemma 2 and Theorems 2 and 3 exist for games with more than two players. In this case the preferences $\succcurlyeq_{i,\sigma}$ can be represented by

$$V_{i,\sigma}(\sigma'_i) = u_i(\sigma'_i, \sigma_{-i}) + \sum_{j \neq i} a^j_{i,\sigma} u_j(\sigma'_i, \sigma_{-i}). \tag{5}$$

[9] There are non-generic 2 × 2 games with unique Nash equilibrium. For example, no matter how players weigh their opponent's payoff from outcomes, (2, 2) is the only Nash equilibrium of the game:

| | $s_j^1$ | $s_j^2$ |
|---|---|---|
| $s_i^1$ | 1, 1 | 1, 2 |
| $s_i^2$ | 2, 1 | 2, 2 |

As in the case of standard game theory, the relationship between dominance and the best-response property needs to be modified when there are more than two players.[10] Theorem 2 must be modified to read: the strategy $\sigma_i \in \sum_i$ is undominated if, and only if, there exists preferences satisfying Eq. (5) such that $\sigma_i$ is a best response to a (possibly correlated) distribution over $\sum_{-i}$.

The next example illustrates the results of this section in a 3 × 3 game.

**Example 2.** Consider the game:

| | $s_j^1$ | $s_j^2$ | $s_j^3$ |
|---|---|---|---|
| $s_i^1$ | $1, 1$ | $2, 0$ | $0, 3$ |
| $s_i^2$ | $2, 2$ | $2, -1$ | $2, 3$ |
| $s_i^3$ | $0, -3$ | $5, 4$ | $1, 3$ |

We first discuss dominance. The strategy $s_i^1$ is strictly dominated (by a mixture of $\frac{4}{5}$ of $s_i^2$ and $\frac{1}{5}$ of $s_i^3$), hence player *i* will not play $s_i^1$ in any equilibrium. None of player *j*'s strategies are dominated (even after $s_i^1$ is deleted). Therefore, the remaining strategies of the players are rationalizable.[11]

Next, we identify the possible pure-strategy Nash equilibria of the game. We noted that $s_i^1$ cannot be used in any Nash equilibrium because it is strictly dominated. $(s_i^3, s_j^3)$ cannot be NE because player *i* can play $s_i^2$ and $(s_i^2, s_j^2)$ and $(s_i^2, s_j^1)$ cannot be NE because player *j* can play $s_j^3$. In fact, $(s_i^2, s_j^3)$ *must* be a NE of this game, since for each player playing anything else will reduce his own outcome without changing that of his opponent.

In standard game theory (that is, $a \equiv 0$), the pair $(s_i^3, s_j^1)$ is not a NE, therefore it does not *have* to be NE when the reciprocity preferences are permitted. However, it *may* be a NE when such preferences are assumed. To support this equilibrium, let $a_{i,(s_i^3,s_j^1)} = -1$ and $a_{j,(s_i^3,s_j^1)} = -7$. Finally, is standard game theory $(s_i^3, s_j^2)$ *must* be a NE, while here it may not (for example, when $a_{i,(s_i^3,s_j^2)} = -1$).

A full analysis of mixed strategy NE is tedious, and without restrictions on the functions *a*, many such equilibria may exist. None place positive probability on $s_i^1$. Another strategy will never be a part of a NE. Let $\sigma_j^* = \frac{1}{2}s_j^1 + \frac{1}{2}s_j^2$. Then *i*'s response to $\sigma_j^*$ must be $s_i^3$. However, $\sigma_j^*$ cannot be the optimal response to $s_i^3$. To see why, observe that unless $a_{j,(s_i^3,\sigma_j^*)} = -\frac{7}{5}$, either $s_j^1$ or $s_j^2$ are better than $\sigma_j^*$ when $(s_i^3, \sigma_j^*)$ is played. But when $a_{j,(s_i^3,\sigma_j^*)} = -\frac{7}{5}$, $s_j^3$ is even better.

## 4. Reciprocity

There is considerable evidence that behavior in games is not consistent with the joint assumptions of equilibrium behavior and maximization of monetary payoffs. Some widely replicated findings from the literature include: proposers typically do not demand all of the surplus in dictator games; first movers in ultimatum game experiments rarely demand the entire surplus (and when

[10] A textbook treatment of this case for standard game theory appears in Fudenberg and Tirole [20, p. 52].

[11] This means that any mixture of $s_i^2$ and $s_i^3$ is a best response to a probability distribution over player *j*'s strategies and any mixed strategy of player *j* is a best response to a probability distribution over $s_i^2$ and $s_i^3$.

they do, their proposals are often rejected); experimental subjects contribute positive amounts in public goods games; and players frequently repay transfers and punish greed in trust games. For more detailed reviews of this literature see, for example, Camerer [9], Fehr and Schmidt [18], Ledyard [24], Roth [28], and Sobel [29].

Within the framework of optimizing agents and equilibrium behavior, there are two broad approaches to these anomalies. The first approach, introduced and developed in papers by Bolton and Ockenfels [7] and Fehr and Schmidt [17], assumes that the preferences of agents depend non-trivially on the distribution of material payoffs. Levine [25] proposes a variation of this approach. In his model, individuals differ in their intrinsic "niceness" and people care more about the material payoffs of nice people. [12] While Bolton and Ockenfels [7] and Fehr and Schmidt [17] provide evidence that some simple parametric forms of interdependent preferences account for some of the experimental findings, there is strong evidence that preferences over outcomes are not fixed, but vary with the game. We discuss below an example that demonstrates the need to go beyond interdependent preferences.

The other approach considers models in which preferences over outcomes can reflect some form of reciprocity. Rabin [27] developed a model of reciprocity based on Geanakoplos et al.'s [21] theory of psychological games. Charness and Rabin [10] and Falk and Fischbacher [14] build on Rabin's model to provide alternative models of strategic behavior in which intentions matter and players have interdependent preferences over outcomes. Dufwenberg and Kirchsteiger [12] and Falk and Fischbacher [14] adapt Rabin's model to extensive-form games. [13] Cox et al. [11] describe a model that combines distributional preferences with concerns about status and reciprocity.

To get an idea of the importance of assuming preferences depend on more than outcomes, consider two variations of the ultimatum game. In the first ("three-split") variation, one player can propose that he receives 80%, 50%, or 20% of a fixed surplus and the second player can either accept or reject the proposal. In the second ("two-split") variation, only first player must choose between the two unequal proposals. In both cases, if individuals have preferences that depend only on (and are strictly increasing in) their own material payoffs, then the unique subgame-perfect equilibrium outcome is for the proposer to ask for 80% of the surplus and for the responder to agree. Experiments of Falk et al. [13] confirm this prediction for the two-split game, but find that when a 50–50 proposal is feasible, the proposer frequently makes it (and some responders reject offers of only 20% of the surplus). To explain these results, it is not enough to assume that players have preferences that depend on the entire distribution of payoffs. In one case, the responder rejects an 80–20 division; in the other, he accepts it. [14] The experimental results are intuitive and are inconsistent with equilibrium models in which preferences depend only on the distribution of monetary payoff. They suggest that preferences do depend on more than the final division of the surplus. It is plausible that the responder will view an unequal division of the surplus as "unfair" when the equal split is available, but not when there are only two proposals available. If unfair

[12] Gul and Pesendorfer [22] provide foundations for the models of interdependent preferences and an axiomatization of the utility function used by Levine [25].

[13] Battigalli and Dufwenberg [4] extend the theory of psychological games to extensive form.

[14] Levine [25] assumes that agents are uncertain about their opponent's preferences. In his framework, it is possible that the first player would use the equal split, if available, to convey information about preferences. Therefore, under some conditions, responses to the 80–20 division may depend on whether the equal-split proposal is available. These differences would disappear after players learned their opponent's preferences.

offers lead to a large enough (in absolute value) weight on the payoff of the proposer in the utility function of the responder, then these proposals will be rejected.

It is clear that it is possible to find preferences that satisfy our assumptions in which unequal offers are rejected in the three-split game but not in the two-split game. It is worth discussing a potential specification in some detail, however, because the functional form proposed in Rabin's [27] model implies that the responder will accept an offer of 20 in the two-split game if and only if he accepts it in the three-split game.

The weight placed on opponent's utility in Rabin's model is the product of a term that summarizes the fairness of the expected outcome and a scaling factor. The specific functional form used by Rabin does not distinguish between the two- and three-split ultimatum game because, roughly, in both cases an equal division is treated as the benchmark for a fair outcome. [15] Consequently, if it is not an equilibrium for the responder to accept a 20% share when the proposer has three splits available, it will also not be an equilibrium for the responder to accept this share when the proposer has only two choices available. Dufwenberg and Kirchsteiger [12] specification also has this property. On the other hand, in Falk and Fischbacher's [14] model it is possible for the second mover to accept a small share in the two-split game but not in the three-split game and the status and reciprocity parameters can be adjusted to permit the responder's reaction to the (80, 20) split to depend on the existence of the (50, 50) offer in the model of Cox et al. [11].

The simple way to modify the model is to maintain Rabin's assumption that whether a weight is positive or negative depends on some measure of fairness, but to refine the notion of fairness to distinguish between the two situations. We will now propose simple weights. We concentrate on the weights used by the responder and assume (simplifying Rabin's approach), that the weight depends only on the expected play of the proposer. Modifying our general notation to the example, let $a^G(O)$ denote the weight that the responder places on the proposer's material payoff in game $G$ (where $G$ is either II in the two-split game or III in the three-split game), where $O = 80, 50$, or 20 is the fraction of the surplus that the proposer offers to the responder. We propose that

$$a^G(O) = \lambda \frac{O - F^G}{80}, \tag{6}$$

where $F^G$ is a fair outcome for the game $G$ and $\lambda > 0$ is a normalization.

It is natural to assume that $F^{\text{III}} = 50$ since the utility possibility set of the three-split game is symmetric and equal division is the symmetric efficient outcome. With this assumption, it follows from Eq. (6) that $a^G(20) < 0$, so that if $\lambda$ is large enough, it is optimal for the responder to reject a low offer. On the other hand, unless public randomization is available (to convexify the set of feasible payoffs), it is reasonable to assume that $F^{\text{II}} < 50$. A full theory of fairness in this situation would take into account the set of feasible distributions and the first-mover's advantageous position. Provided that $20 \leqslant F^{\text{II}} < F^{\text{III}}$ it will be the case (for some choices of $\lambda$) that the (80, 20) split is an equilibrium in the two-split game but not in the three-split game. [16] This specification of the coefficient $a^G$ leads refutable predictions: if an agent rejects an offer of 20 in the two-split ultimatum game, then he will reject the same offer in the three-split game. In general, a refinement of our theory that specifies how $a^G$ changes with the game $G$ could lead to refutable hypotheses of behavior over families of games.

[15] More precisely, Rabin's fairness measure looks at deviations from the average between the largest and smallest Pareto-efficient material payoffs available given the players' beliefs. When the responder is expected to accept all offers, then this average is 50% of the surplus whether or not an equal split is available.

[16] The result would be clearest if we set $F^{\text{II}} = 20$ so that the second player will accept all offers in equilibrium in the two-split game.

Two simple modifications generalize Eq. (6) to arbitrary games. First, one should replace $O$ by the maximum payoff that the player could obtain given his opponent's strategy. Second, one should normalize by dividing by the range of possible payoffs in the game. If $u^h_j(\sigma_i) = \max_{s_j \in S_j} u_j(s_j, \sigma_i)$, $\bar{u}_j = \max_{s \in S_i \times S_j} u_j(s)$, and $\underline{u}_j = \min_{s \in S_i \times S_j} u_j(s)$, then

$$a(\sigma_i) = \begin{cases} \lambda \dfrac{u^h_j(\sigma_i) - F^G}{\bar{u}_j - \underline{u}_j} & \text{if } \bar{u}_j - \underline{u}_j > 0, \\ 0 & \text{if } \bar{u}_j - \underline{u}_j = 0. \end{cases} \tag{7}$$

A more complete theory would specify a fair outcome ($F^G$) and a normalization. Charness and Rabin [10], Dufwenberg and Kirchsteiger [12], Falk and Fischbacher [14], and Rabin [27] all propose specific functional forms for $\lambda$ and $F^G$. Unlike our simple model, the values of the normalization and fair outcome depend not only on the game, but on the context ($\sigma^*$).

## 5. Examples

In this section we present some examples to illustrate important ways in which our approach differs from standard game theory. The following example shows that in our model, Definitions 1 and 2 do not have to lead to the same analysis.

**Example 3.** Consider the game:

| | AM | PM |
|---|---|---|
| AM | 10, 10 | 0, 0 |
| PM | 0, 0 | 10, 10 |
| All | 7, 10 | 7, 10 |

Column is a plumber and Row is a homeowner with a leaky faucet. Column can come Monday in the AM or in the PM, while Row can cancel appointments and arrange to be at home Monday in the AM, in the PM, or all day. The plumber earns 10 if she coordinates with the homeowner, but nothing otherwise. The homeowner receives a payoff of 10 if he can meet the plumber and only cancel half of his appointments; he receives 7 if he stays home all day; and he receives 0 if he fails to coordinate with the plumber. If players' preferences over strategies agreed with their preferences over outcomes, then the game has three equilibrium outcomes: (AM, AM), (PM, PM), and a continuum of equilibria in which the homeowner stays home all day and the plumber places probability of at least 0.3 on each pure strategy.

Now assume that the homeowner has preferences over strategies that lead him to put a positive weight on the plumber's payoff in response to nice behavior (apparent coordination) and a negative weight in response to nasty behavior. (We assume that the plumber cares only about her own payoffs.) Clearly, the two pure-strategy equilibria will continue to be equilibria. But if we assume that randomization by the plumber is purposeful, then the homeowner may well think that a plumber who randomizes equally between AM and PM is nasty, because this behavior minimizes the probability of coordination. With a sufficiently negative weight on the plumber's payoffs, the homeowner may prefer to play either AM or PM rather than to stay at home all day. Consequently, there may be an equilibrium in which both the homeowner and the plumber randomize equally between AM and PM, while (All, $\frac{1}{2}$AM + $\frac{1}{2}$PM) is no longer anequilibrium.

The above analysis depends on the interpretation of mixed strategies, because how the homeowner interprets the plumber's behavior determines the weight that he puts on the plumber's payoffs over outcomes. In many applications, it is appropriate to treat the homeowner as if he is matched against a population of plumbers, some with a tendency to come in the morning, others with a tendency to come in the afternoon. If the homeowner does not attribute his uncertainty to a deliberate strategy of the plumber, then it is reasonable to assume that he places a non-negative weight on the plumber's payoff. In this case, however, there will be an equilibrium in beliefs in which the homeowner always stays at home (and she believes with probability greater than 0.3 that the plumber will come at any time).

In standard game theory preferences over strategies are independent of interpretation of mixed strategies. In this example we suggest that different interpretations of mixed strategies lead to a different notions of what is a sensible preference over strategies, which in turn leads to different predictions. As this paper makes minimal assumptions on preferences over strategies, it is completely consistent with our approach for preferences over strategies to be independent of the interpretation of mixed strategies. It is, of course, also consistent with our approach for preferences over strategies to be different depending on the interpretation of mixed strategies.

In standard game theory, if player $i$ strictly prefers $s_i^1$ to $s_i^2$ for every pure strategy selected by $j$, then $s_i^2$ is strictly dominated and will not be used in any equilibrium. In Example 4 there is a reasonable specification of preferences over strategies in which this relationship does not hold.

**Example 4.** The players are going to eat dinner together. Row will bring the main course, either beef or pheasant, and Column will bring the wine, either red or white. Row prefers red wine to white and pheasant to beef, while Column prefers to drink red wine with beef, but hates a beef, white-wine menu:

| | Red | White |
|---|---|---|
| Beef | 15, 30 | 9, 10 |
| Pheasant | 20, 20 | 10, 20 |

If the weight that the row player gives to the column player's utility when Column brings red wine is sufficiently positive (greater than $\frac{1}{2}$), then the optimal response of Row to red wine is to supply beef. On the other hand, if Column brings white wine, Row may give Column'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. In standard game theory, if it is a strict best response for Row to bring beef no matter what Column's pure strategy choice is, then pheasant is a dominated strategy and will not be used with positive probability in any equilibrium. This is not so in our model. There can be an equilibrium in which Row brings pheasant and Column randomizes between red and white. Specifically, assume that Column brings each wine with probability $\frac{1}{2}$. Column will receive the expected payoff of 20 from outcomes no matter what Row does. Consequently, Row's best response is to bring pheasant, the strategy that maximizes his payoff over outcomes. Assuming that Column places no weight on her opponent's payoffs, the game has two Nash equilibrium outcomes, one in which the meal consists of beef and red wine, the other in which the main course is pheasant while the choice of wine is uncertain. The same conclusion holds if Row believes that Column is not purposefully randomizing, but is uncertain about what Column is going to do.

With general preferences over strategies, this game can have an equilibrium in which Row offers pheasant and Column randomizes, (Pheasant, 50–50). At the same time, Row prefers to bring beef in response to either choice of wine. In standard game theory, however, if it is a strict best response for Row to bring beef no matter what Column's pure strategy choice, then pheasant is a dominated strategy and will not be used with positive probability in any equilibrium.

The next example demonstrates that permitting preferences over strategies to change with the game creates possibilities that do not exist in standard game theory. We describe a situation in which a strictly dominated strategy for one player becomes an equilibrium strategy when the opponent is given a new strategy.

**Example 5.** Consider the following $2 \times 1$ game:

| | Nice |
|---|---|
| Transfer | 10, 30 |
| Keep | 20, 10 |

This game is a decision problem for the row player. The row player starts with \$20 and the column player starts with \$10. Row can either transfer \$10 to Column or keep the entire \$20 for himself. If Row does give \$10 to Column, Column also receives \$10 from a third party. Selfish row players (and even players with preferences for equitable outcomes) will keep the money.

Now imagine that Column could (at a personal cost of \$5) take Row's money. This creates a game in which the monetary payoffs are

| | Nice | Greedy |
|---|---|---|
| Transfer | 10, 30 | 0, 35 |
| Keep | 20, 10 | 0, 25 |

(If Column plays her $G$ strategy, then Row gets nothing, while Column gets her original \$10 plus Row's \$20 minus the \$5 she spends to take Row's money. She receives an additional \$10 from the third party when Row plays $T$.) $(T, N)$ could be a strict Nash equilibrium of this game when players have preferences over strategies. If the players anticipate that the outcome of the game is $(T, N)$, then Row has reason to believe that Column is being nice to him—she is forgoing the opportunity to take his money. Consequently Row may place a sufficiently large positive weight on Column's payoff (greater than $\frac{1}{2}$) to guarantee that $T$ is a best response to $N$. On the other hand, a column player who places weight more than $\frac{1}{2}$ on Row's payoffs will respond to $T$ by playing $N$. It follows that $(T, N)$ can be an equilibrium of the $2 \times 2$ game even though it is not an equilibrium of the $2 \times 1$ game. That is, by deleting the strategy $G$ one eliminates the $(T, N)$ equilibrium. This could not happen in standard game theory where it is not possible to eliminate a Nash equilibrium by deleting a strategy that is not used with positive probability in the equilibrium.[17] The example is a simplification of models of gift exchange that give rise to similar qualitative behavior in experiments.[18]

[17] Stated generally, if $\sigma$ is a Nash equilibrium of a game $G$ and one obtains the game $G'$ by deleting strategies in $G$ that are not in the support of $\sigma_i$ for all $i$, then $\sigma$ is a Nash equilibrium of $G'$. The same result holds for standard refinements.

[18] Experimental studies include Abbink et al. [1], Berg et al. [5], Fehr and Gächter [15], and Fehr et al. [16].

## 6. Related approaches

In this section, we compare our approach to related literature.

### *6.1. Relationship to psychological games*

Geanakoplos, Pearce, and Stacchetti (GPS) [21] introduced the concept of psychological games. Other authors have used their formulation to model preferences for fairness and reciprocity in strategic settings. In this section, we discuss the connection between GPS's work and our paper. We argue that the results in Section 2 provide a reformulation and representation theorem for psychological games.

Psychological games have players, strategies, and preferences. They differ from standard games in that preferences are defined on the product space of outcomes and "collectively coherent" beliefs.[19] Permitting preferences to depend on beliefs makes it possible to use psychological games to model intentions. A psychological equilibrium consists of a strategy profile and collectively coherent beliefs with the property that the equilibrium strategy is common knowledge[20] and, given beliefs, each player's strategy choice is a best response. GPS's [21] central contribution is to examine the implications of preferences in games that depend on more than the outcomes of the game. To this extent, our approach is identical. In our model, the extended preferences depend on a strategy profile (how the game is expected to be played), which we have called "context" and denoted by $\sigma^*$. Common knowledge that players use a particular strategy profile identifies a hierarchy of beliefs (in psychological games) and a context (in our model). Equilibrium requires both the standard best-response property, but also that the additional argument of utility functions be the one determined by the putative equilibrium strategy profile.

Our context $\sigma^*$ truncates the hierarchy of beliefs after just one level. Consequently, the domain of preferences in GPS [21] is much larger than in our model. This difference is not significant for our theory or in the existing applications. Theoretically, there are no barriers to extending our representation theorem to models in which "context" is an element of an arbitrary space (that is, it is not limited to strategy sets or the space of collectively coherent beliefs). Provided that the continuity axiom holds when arbitrary hierarchies of beliefs replace context ($\sigma^*$), the general topological conditions provided in Border's [8] treatment of Harsanyi's [23] theorem still hold so the results in Section 2 still hold. That is, one can view Fact 1 as a representation theorem that applies to a class of games that includes an interesting class of psychological games. Our results do not apply to all psychological games because there is no reason for the preferences over strategies and beliefs in GPS to satisfy our assumptions. In particular, the separation of material and non-material utility derived in Theorem 1 does not apply to all psychological games.

### *6.2. Comparison to other papers*

Rabin [27] introduced the idea of using psychological games to study fairness and reciprocity. Our paper does two things that Rabin does not do. We provide a representation theorem for preferences over strategies, thereby giving an axiomatic foundation to Rabin's approach and we

[19] An individual's belief hierarchy is coherent if the marginal distribution of a belief of order $k + 1$ is equal to the corresponding belief of order $k$. A profile of belief hierarchies—one for each player—is collectively coherent if it is common knowledge that all beliefs are coherent.

[20] This condition determines the hierarchy of beliefs for each player.

have general results on dominance. In addition, the discussion in Section 4 suggests that the specific functional form proposed by Rabin is limited.

Several recent papers [10,14] combine models of context-dependent strategic preferences with preferences over outcomes that depend non-trivially on the distribution of material payoffs. Our approach posits two different preference relationships, making the distinction between strategic and non-strategic preferences clear. The first relationship, preferences over outcomes, is non-strategic. In principle, one should be able to identify these preferences by examining the behavior of individuals in decision-theoretical problems. To highlight the distinction, consider the following. Suppose given a strategy profile, player one could generate any of the *monetary* payoffs: $(3, 4, 0)$, $(3, 0, 4)$ and $(3, 2, 2)$ by using his first, second, or third pure strategy, respectively. If these monetary payoffs were (material) utilities and player 1 was indifferent between his first two strategy choices, then by (3) he must also be indifferent between all three. One might think that this means that representation (3) prevents a player from exhibiting preferences over the distribution of payoffs. This is false. If player 1 prefers the monetary payoffs $(3, 2, 2)$ to the other profiles (because it treats his opponents equally), then these preferences must be exhibited by his preference over outcomes. If this player's preferences over outcomes exhibited inequity aversion (as in [17,7]) then according to his material preferences $(3, 2, 2)$ would be strictly preferred to $(3, 4, 0)$ and $(3, 0, 4)$. The example in Section 4 demonstrates that interdependent preferences alone are not sufficient to describe experimental regularities.

## Acknowledgment

We are grateful to Miguel Costa-Gomes, Vincent Crawford, Rachel Croson, Jürgen Eichberger, Armin Falk, Drew Fudenberg, David Levine, Mark Machina, Joe Ostroy, Matthew Rabin, Ariel Rubinstein, Ennio Stacchetti, Dan Vincent, Joel Watson, Bill Zame, referees, and an Associate Editor for their comments. We thank SSHRCC and NSF for financial support.

## Appendix A. Proofs

**Proof of Fact 1.** For the linear form, see Border [8] or Fishburn [19]. Uniqueness is similar to standard uniqueness results of the vN&M utility function. □

**Proof of Lemma 1.** For each $\sigma^* \in \sum$ define the correspondence $\Phi : \sum \to \sum$ by

$$\Phi_i(\sigma^*) = \left\{ \sigma_i \in \sum_i : \sigma_i \succcurlyeq_{i,\sigma^*} \sigma'_i \text{ for all } \sigma'_i \in \sum_i \right\}.$$

$\Phi(\cdot)$ satisfies all of the conditions of Kakutani's Theorem (convexity follows from the independence axiom). Consequently, $\Phi(\cdot)$ has a fixed point, $\mu^*$. It follows from the construction that if $\mu^*_i(s^k_i) > 0$, then $s^k_i \succcurlyeq_{i,\sigma^*} \sigma'_i$ for all $\sigma'_i \in \sum_i$. □

**Proof of Theorem 1.** For a given $\sigma^* = (\sigma^*_i, \sigma^*_{-i})$, suppose first that $A_i(\sigma^*_{-i})$ has non-empty interior in $\Re^I$. Let $\sigma_i, \sigma'_i \in \sum_i$ such that $u_i(\sigma'_i, \sigma^*_{-i}) > u_i(\sigma_i, \sigma^*_{-i})$ and $u_{-i}(\sigma'_i, \sigma^*_{-i}) = u_{-i}(\sigma_i, \sigma^*_{-i})$. Hence, by axiom SI, $\sigma'_i \succ_{i,\sigma^*} \sigma_i$. By Eq. (2), $a^i_{i,\sigma^*} u_i(\sigma'_i, \sigma^*_{-i}) > a^i_{i,\sigma^*} u_i(\sigma_i, \sigma^*_{-i})$ (since $u_{-i}(\sigma'_i, \sigma^*_{-i}) = u_{-i}(\sigma_i, \sigma^*_{-i})$). It follows from $u_i(\sigma'_i, \sigma^*_{-i}) > u_i(\sigma_i, \sigma^*_{-i})$ that $a^i_{i,\sigma^*} > 0$. Multiply all the coefficients by $1/a^i_{i,\sigma^*}$ to get the desired result.

Now suppose that $A_i(\sigma^*_{-i})$ has an empty interior in $\Re^I$.

1. If there exist $u_{-i}$ and $u_i \neq u'_i$ such that $(u_i, u_{-i})$ and $(u'_i, u_{-i}) \in A_i(\sigma^*_{-i})$, then, as in the first part of the proof axiom SI implies that $a^i_{i,\sigma^*} > 0$ and all coefficients can be multiplied by $1/a^i_{i,\sigma^*}$.
2. If $u_i$ is constant for all elements of $A_i(\sigma^*_{-i})$, then $a^i_{i,\sigma^*}$ can be any number, in particular, it can be one.
3. Otherwise, since $A_i(\sigma^*_{-i})$ is convex and has an empty interior, it is contained in a set of the form $\{u : k_i u_i + \sum_{j\neq i} k_j u_j = C\}$ with $k_i \neq 0$ (since $u_i$ is a function of $u_{-i}$ on $A_i(\sigma^*_{-i})$). That is, there exists $C'$ and $k'_j$ for $j \neq i$ such that on $A_i(\sigma^*_{-i})$ we can write

$$u_i = C' - \sum_{j\neq i} k'_j u_j. \tag{A.1}$$

Define new coefficients $\bar{a}^i_{i,\sigma^*} = 1$ and $\bar{a}^j_{i,\sigma^*} = a^j_{i,\sigma^*} + (1 - a^i_{i,\sigma^*})k'_j$. It follows that

$$\begin{aligned}
&a^i_{i,\sigma^*} u_i(\sigma_i, \sigma^*_{-i}) + \sum_{j\neq i} a^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i}) \\
&= a^i_{i,\sigma^*} C' + \sum_{j\neq i} (a^j_{i,\sigma^*} - a^i_{i,\sigma^*} k'_j) u_j(\sigma_i, \sigma^*_{-i}) \\
&= a^i_{i,\sigma^*} C' + \sum_{j\neq i} (\bar{a}^j_{i,\sigma^*} - k'_j) u_j(\sigma_i, \sigma^*_{-i}) \\
&= (a^i_{i,\sigma^*} - 1)C' + \bar{a}^i_{i,\sigma^*} u_i(\sigma_i, \sigma^*_{-i}) + \sum_{j\neq i} \bar{a}^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i}),
\end{aligned}$$

where the first and last equations follow from (A.1) and the second equation comes from the definition of $\bar{a}^i_{i,\sigma^*}$. It follows from Eq. (2) that $\succcurlyeq_{i,\sigma^*}$ can be represented using the coefficients $\bar{a}^j_{i,\sigma^*}$. □

**Proof of Lemma 2.** If the condition in the lemma does not hold, then there exists $\sigma'_i \in \sum_i$ such that $u_i(\sigma'_i, \sigma_j) > u_i(\sigma)$ and $u_j(\sigma'_i, \sigma_j) = u_j(\sigma)$. Hence, regardless of the value of $a^j_{i,\sigma}$,

$$u_i(\sigma'_i, \sigma_j) + a^j_{i,\sigma} u_j(\sigma'_i, \sigma_j) > u_i(\sigma) + a^j_{i,\sigma} u_j(\sigma)$$

so $\sigma_i$ is not a best response to $\sigma_j$.

Suppose now that the condition of the lemma is satisfied. It follows that the convex subsets of $\Re^2$, $\{u(\sigma'_i, \sigma_j) : \sigma'_i \in \sum_i\}$ and $\{(w, u_j(\sigma)) : w > u_i(\sigma)\}$ are disjoint. Therefore, there exists a separating line between them. That is, there exists $(p, q) \neq (0, 0)$ such that for all $w \geqslant u_i(\sigma)$ and for all $\sigma'_j \in \sum_j$,

$$pw + qu_j(\sigma) \geqslant pu_i(\sigma'_i, \sigma_j) + qu_j(\sigma'_i, \sigma_j). \tag{A.2}$$

Since inequality (A.2) holds for all $w \geqslant u_i(\sigma)$, it must be that $p \geqslant 0$. Further, since $\{u(\sigma'_i, \sigma_j) : \sigma'_i \in \sum_i\}$ is a polygon, we can take $p > 0$. Now set $a^j_{i,\sigma} = \frac{q}{p}$. It follows from inequality (A.2) that for all $\sigma'_j \in \sum_j$,

$$u_i(\sigma) + a^j_{i,\sigma} u_j(\sigma) \geqslant u_i(\sigma'_i, \sigma_j) + a^j_{i,\sigma} u_j(\sigma'_i, \sigma_j). \tag{A.3}$$

Inequality (A.3) means that if player $i$ has reciprocity preferences, then $\sigma_i$ is a best response to $\sigma_j$. □

**Proof of Theorem 2.** It is clear that if $\sigma_i'$ strictly dominates $\sigma_i$, then $\sigma_i$ can never be a best response to any probability distribution over player $j$'s pure strategies. To prove the converse, define for each $\sigma_j \in \sum_j$, $S(\sigma_j) = \{(u_i(\sigma_i', s_j^1), \ldots, u_i(\sigma_i', s_j^{n_j})) : \sigma_i' \in \sum_i\}$ and $u_j(\sigma_i', \sigma_j) = u_j(\sigma)\}$. Let $v = (u_i(\sigma_i, s_j^1), \ldots, u_i(\sigma_i, s_j^{n_j}))$. Clearly, $v \in S(\sigma_j)$, so $S(\sigma_j)$ is non-empty. Also, $S(\sigma_j)$ is convex. Let $\Psi(\sigma_j)$ be the set of all $p$ that satisfy $p = (p_1, \ldots, p_{n_j})$, $p \geqslant 0$, and $\sum_{k=1}^{n_j} p_k = 1$ such that

$$u_i(\sigma_i, p) = \sum_{k=1}^{n_j} p_k v_k \geqslant \sum_{k=1}^{n_j} p_k u_i(\sigma_i', s_j^k) = u_i(\sigma_i', p) \tag{A.4}$$

for all $\sigma_i' \in \sum_i$ such that $u_j(\sigma_i', \sigma_j) = u_j(\sigma)$. If $\sigma_i$ is undominated, then $v$ is undominated in $S(\sigma_j)$. It follows that $\Psi(\sigma_j)$ is non-empty (it contains the normal to a hyperplane that separates $S(\sigma_j)$ from $v$). It is straightforward to verify that $\Psi(\cdot)$ is an upper hemi-continuous, convex-valued correspondence from the simplex to itself. Hence it has a fixed point, $\sigma_j^*$. It follows from (A.4) with $p = \sigma_j^*$ that $u_i(\sigma_i, \sigma_j^*) \geqslant u_i(\sigma_i', \sigma_j^*)$ for all $\sigma_i' \in \sum_i$ such that $u_j(\sigma_i', \sigma_j^*) = u_j(\sigma_i, \sigma_j^*)$. Therefore, $\{u(\sigma_i', \sigma_j^*) : \sigma_i' \in \sum_i\} \cap \{(w, u_j(\sigma_i, \sigma_j^*)) : w > u_i(\sigma_i, \sigma_j^*)\} = \emptyset$ holds. From Lemma 2, we see that there exists $a$ such that $\sigma_i$ is a best response to $\sigma_j^*$. □

**Proof of Theorem 3.** It follows from Lemma 2 that $\sigma_i$ is a best response to $\sigma_j$ for some $a_{i,\sigma}^j$ if, and only if, the first condition in the lemma holds. The second condition is necessary and sufficient for $\sigma_j$ to be a best response to $\sigma_i$ for some $a_{j,\sigma}^i$. □

## References

[1] K. Abbink, B. Irlenbusch, E. Renner, The moonlighting game—an experimental study on reciprocity and retribution, J. Econ. Behav. Organ. 42 (2) (2000) 265–277.

[2] R.J. Aumann, Correlated equilibrium as an expression of Bayesian rationality, Econometrica 55 (1) (1987) 1–18.

[3] R.J. Aumann, A. Brandenberger, Epistemic conditions for Nash equilibrium, Econometrica 63 (5) (1995) 1161–1180.

[4] P. Battigalli, M. Dufwenberg, Dynamic psychological games, Technical Report, University of Arizona, September 2005.

[5] J. Berg, J. Dickhaut, K. McCabe, Trust, reciprocity and social history, Games Econ. Behav. 10 (1995) 122–144.

[6] B.D. Bernheim, Rationalizable strategic behavior, Econometrica 52 (4) (1984) 1007–1028.

[7] G. Bolton, A. Ockenfels, ERC: a theory of equity, reciprocity and competition, Amer. Econ. Rev. 90 (2000) 166–193.

[8] K.C. Border, More on Harsanyi's utilitarian cardinal welfare function, Soc. Choice Welfare 1 (4) (1985) 279–281.

[9] C. Camerer, Behavioral Game Theory, Princeton University Press, Princeton, NJ, 2003.

[10] G. Charness, M. Rabin, Understanding social preferences with simple tests, Quart. J. Econ. 117 (3) (2002) 817–869.

[11] J.C. Cox, D. Friedman, S. Gjerstad, A tractable model of reciprocity and fairness, Technical Report, UCSC, 2005. URL: ⟨http://econ.ucsc.edu/faculty/dan/Tractable.pdf⟩.

[12] M. Dufwenberg, G. Kirchsteiger, A theory of sequential reciprocity, Games Econ. Behav. 47 (2) (2004) 268–298.

[13] A. Falk, E. Fehr, U. Fischbacher, Testing theories of fairness—intentions matter, Working Paper 63, University of Zurich, September 2000.

[14] A. Falk, U. Fischbacher, A theory of reciprocity, Games Econ. Behav. 54 (2) (2006) 293–315.

[15] E. Fehr, S. Gächter, Reciprocity and economics: the economic implications of homo reciprocans, Europ. Econ. Rev. 42 (1998) 845–859.

[16] E. Fehr, S. Gächter, G. Kirchsteiger, Reciprocity as a contract enforcement device: experimental evidence, Econometrica 65 (1997) 833–860.

[17] E. Fehr, K. Schmidt, A theory of fairness, competition, and cooperation, Quart. J. Econ. 114 (1999) 817–868.

[18] E. Fehr, K. Schmidt, Theories of fairness and reciprocity—evidence and economic applications, in: M. Dewatripont, L.P. Hansen, S.J. Turnovsky (Eds.), Advances in Economics and Econometrics: Eighth World Congress, vol. 1, Cambridge University Press, Cambridge, MA, 2003, pp. 208–257, (Chapter 6).

[19] P.C. Fishburn, On Harsanyi's utilitarian cardinal welfare theorem, Theory Dec. 17 (1) (1984) 21–28.

[20] D. Fudenberg, J. Tirole, Game Theory, MIT Press, Cambridge, MA, 1991.

[21] J. Geanakoplos, D. Pearce, E. Stacchetti, Psychological games and sequential rationality, Games and Econ. Behav. 1 (1989) 60–79.

[22] F. Gul, W. Pesendorfer, The canonical type space for interdependent preferences, Technical Report, Princeton University, November 2005.

[23] J.C. Harsanyi, Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility, J. Polit. Economy 63 (4) (1955) 309–321.

[24] J.O. Ledyard, Public goods: a survey of experimental research, in: J. Kagel, A. Roth (Eds.), Handbook of Experimental Economics, Princeton University Press, Princeton, NJ, 1995, pp. 111–194, (Chapter 2).

[25] D. Levine, Modelling altruism and spitefulness in game experiments, Rev. Econ. Dyn. 1 (1998) 593–622.

[26] D.G. Pearce, Rationalizable strategic behavior and the problem of perfection, Econometrica 52 (4) (1984) 1029–1050.

[27] M. Rabin, Incorporating fairness into game theory, Amer. Econ. Rev. 83 (5) (1993) 1281–1302.

[28] A.E. Roth, Bargaining experiments, in: J. Kagel, A. Roth (Eds.), Handbook of Experimental Economics, Princeton University Press, Princeton, NJ, 1995, pp. 253–348.

[29] J. Sobel, Interdependent preferences and reciprocity, J. Econ. Lit. 43 (2) (2005) 396–440.