Journal of Behavioral and Experimental Economics 76 (2018) 68–75

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Interdependent preferences and endogenous reciprocity☆

José A. Carrasco[a], Rodrigo Harrison[b], Mauricio Villena[*,c]

[a] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibañez, 2640 Diagonal Las Torres, Santiago, Chile
[b] Instituto de Economía, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macúl, Santiago, Chile
[c] Dirección de Presupuestos, Ministerios de Hacienda, Chile, and Escuela de Negocios, Universidad Adolfo Ibañez, 2700 Diagonal Las Torres, Santiago, Chile

ARTICLE INFO

Keywords:
Reciprocity
Endogenous preferences
Asymmetric evolutionary game

JEL classification:
C72
A13

ABSTRACT

This paper employs an indirect approach to formally examine the evolutionary stability of interdependent preferences when players randomly engage in pairwise interactions. Following the model specification for altruism and spitefulness in experiments proposed by Levine (1998), we also explore the stability of reciprocity and reciprocal preferences. In particular, we study how individuals equipped with intrinsic preferences such as altruism, selfishness or spitefulness adjust their behavior depending on who they interact with. The key aspect of our method is that behavioral preferences are choice variables that optimally evolve, accounting for strategic interaction. Our model predicts that in a specific economic framework characterized by negative externalities and strategic substitutes, there is a continuum of evolutionary stable interdependent preference profiles: At least one player behaves spitefully, and at most one acts selfishly. The emergence of altruism as an evolutionarily stable preference crucially depends on how large the support for preferences is. When players have reciprocal preferences, altruism might arise even in meetings where one player is intrinsically spiteful, but not necessarily from the intrinsically altruistic player.

1. Introduction

Economic theory usually dictates that agents are self-interested and rational. However, the experimental literature (Güth et al., 1982; Isaac and Walker, 1988; Fehr and Schmidt, 1999; Charness and Rabin, 2002 among others) suggests that agents are better characterized as having *interdependent preferences* and being concerned about the payoffs of others.[1] Additional supporting evidence for this interpretation is also provided by Brandts and Solà (2001), Güth et al. (1998), Sadrieh and Schröder (2016) and Thunström et al. (2016). Furthermore, the seminal experimental work of Levine (1998) suggests that agents behave as if they have *reciprocal preferences*, a more specific type of interdependent preference. Agents with these preferences adjust the concern they express for others based on their perceptions of how they are being treated by their opponents. Of course, these perceptions do not have to remain unchanged. To wit, the behavior of players evolves, as they may perceive intentions differently based on who the opponent is. This experimental evidence inspired our work to analytically solve Levine's model to more accurately predict how preferences evolve.

In this paper, we employ an indirect approach to analytically explore the evolutionary stability of reciprocal preferences using the specification that Levine used to address experimental evidence. In our model, a large population of individuals are continuously and randomly matched in pairs. Players interact in a strategic environment that shows *negative externalities and strategic substitutes*. Matched players preferences are common knowledge; they determine players' choices, which in turn determine outcomes and payoffs. Behavior is guided by *adjusted utility* maximization, whereas the stability of preferences is driven exclusively by *material payoff* maximization. We follow (Levine, 1998) and assume that the adjusted utility functions are linear on both individual material payoffs. We use a quadratic monetary payoff specification, as this may represent many social dilemmas in which the individual choice that maximizes individual payoffs differs from the one that maximizes group payoffs. In addition, its tractability allows us to derive a closed form solution for optimal strategies, which is an appealing property.

☆ We thank the editor and two anonymous referees for many helpful comments and suggestions. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

* Corresponding author.
E-mail addresses: jose.carrasco@uai.cl (J.A. Carrasco), harrison@uc.cl (R. Harrison), mauricio.villena@uai.cl, MVillena@dipres.gob.cl (M. Villena).
URLS: http://www.tonocarrasco.com (J.A. Carrasco), http://www.rodrigoharrison.com (R. Harrison), http://www.mauriciovillena.com (M. Villena).

[1] From a theoretical perspective, Heifetz et al. (2007) show that self-interest payoff-maximization is suboptimal when agents interact in pairwise-meetings. That is, some degree of distortion — difference between agents' objectives and payoffs — is beneficial to each player and is also evolutionarily stable.

https://doi.org/10.1016/j.socec.2018.08.002
Received 26 August 2017; Received in revised form 10 August 2018; Accepted 15 August 2018
Available online 18 August 2018

So equipped, we first endogenize interdependent preferences instead of considering them as exogenously given. We find that when players *choices* are perfect substitutes, there is a continuum of evolutionarily stable interdependent preferences (Proposition 1). Altruism or selfishness might arise as optimally evolved preferences but only by one of the matched players. Instead, spiteful preferences might arise as evolved preferences by both players. That is, whenever one player behaves altruistically, his opponent always behaves spitefully. Otherwise, at least one player behaves spitefully and at most, one acts selfishly. More generally, *stable preferences act as substitutes*: When players strategically choose how to shape their preferences, they are downward sloping functions of their opponent's preferences. As one player behaves more altruistically (less spitefully), the other behaves more spitefully (less altruistically). Aside from these predictions, our model quantitatively predicts how altruism and spitefulness impact players payoffs.

We then turn to reciprocity to model reciprocal preferences. We use the linear approach proposed by Levine (1998) and distinguish the *intrinsic preferences* of each player and their *behavioral preferences*. We consider the former as genes that are acquired through genetic inheritance (Güth, 1995) and the latter as weighted averages of intrinsic values. Our key observation is that some notion of fairness or reciprocity shapes players' preferences. To wit, while players' preferences depend to some extent on their genetics, they will be able to adjust their behavior depending on who they interact with. Therefore, we aim to address contextual behavior without assuming that players' intrinsic preferences change. Instead, we let reciprocity evolve, which indirectly determines preferences and behavior. This is more consistent with recent empirical works, which suggest that culture (reciprocity), rather than genes (intrinsic preference), provides a greater scope for large-scale human evolution (Bell et al., 2009). Similarly, Boyd and Richerson (1988, 2006) argue that cultural adaptation — the ability to create non-genetic evolution — is what makes the human species different from others. This cultural adaptation process is much faster than genetic evolution and leads to highly adaptive behavior.

In this more specific version of the model, pairwise meetings occur between players from two large populations, each characterized by its own intrinsic preference parameters (players types). We propose that each population intrinsic preference parameter is *common knowledge*, (as in Gth and Napel, 2006; Menicucci and Sacco, 2009; Sethi and Somanathan, 2001). While this can be seen as a rather strong assumption, it can be noted that there is psychological evidence that several observable physical symptoms, such as posture, respiration, voice, and facial muscle tone and expression, can provide some indication of a person's disposition towards others (Frank, 1987; Frank, 1988). These physical symptoms act as signals that can condition people's behavior towards others. Alternatively, this perfect information assumption may be replaced with an assumption of sufficiently accurate pre-play signals or the intrinsic preferences information is available for both players at sufficiently small costs.[2]

For our tractable quadratic material payoff function, we compute the set of evolutionarily stable reciprocal preferences that would arise in the setting proposed by Levine (1998). Players adjust their behavior depending on who they interact with and sometimes, their behavioral preferences coincide with their intrinsic ones. More specifically, we find that in each meeting, at least one player acts reciprocally and that his concern about his opponent's material payoff depends on the intrinsic preferences of his opponent. Furthermore, we find that genes might restrict the induced behavioral preferences. When they do, *strong reciprocity* (when a player's behavioral preference equals his opponent's intrinsic value) together with *no reciprocity* (when a player's behavioral preference coincides with his own intrinsic value) both arise as an evolutionarily stable strategy profile. This occurs when both players are either too altruistic or too spiteful (Proposition 2).

The emergence of altruism as an evolved preference crucially depends on how intrinsic preferences are restricted. Our specification includes an extended support $[-\infty, 1]$ that includes models of altruistic and egoistic behavior, such as those in Bester and Güth (1998), in which intrinsic values are restricted to be in [0,1], as well as models of spitefulness, as suggested by Bolle (2000) and Possajennikov (2000). We find that if intrinsic values are only allowed to be in $[-1, 1]$, as in Levine (1998), then altruism only arises in meetings between altruistic players (Proposition 3 and Proposition 4). However, with an extended support for spite, altruism may also arise in meetings between an altruistic player and a highly spiteful player. More surprisingly, in these meetings, altruism may arise as an evolved behavioral preference by the intrinsically spiteful player, and when this happens, the intrinsically altruistic player may have spiteful preferences. In fact, in these meetings, multiple more natural combinations of preferences might also arise as evolutionarily stable. These include both players behaving spitefully or each player behaving according to their intrinsic value. That is, the spiteful player behaving spitefully and the altruistic behaving altruistically.

LITERATURE REVIEW: There are as many ways to model interdependent preferences as there are players concerns other than their own payoffs (Sobel, 2005). Fehr and Schmidt (1999) propose a model of inequality aversion, when players avoid inequitable outcomes, whereas (Güth and Napel, 2006) further explore the evolution of inequality aversion. Charness and Rabin (2002) and Alger and Weibull (2013) study interdependent preferences when players have a concern for efficiency. Kokesen et al. (2000) set up a model for negatively interdependent preferences. They focus on players behaving either selfishly or spitefully, and they consider conditions under which those who behave spitefully earn higher material payoffs than those behaving selfishly. In an alternative approach, Dekel et al. (2007) focus on the stability of outcomes and show that when preferences are observable, only efficient outcomes are stable. That is, the efficiency of outcomes is necessary for the stability of preferences. Herold and Kuzmics (2009) extends this result accounting for preferences that may depend on the opponent's preferences. Unlike us, they consider exogenously specified preferences that do not account for players' optimizing behavior or strategic interaction.[3]

Our paper contributes to the extensive game theoretic literature on reciprocal preferences (Sethi and Somanathan, 2001; Sethi and Somanathan, 2003; Levine, 1998). In the reciprocity games literature, it is widely accepted that reciprocity is driven by perceived kindness. This raises the question of whether agents consider intentions, consequences or both when they evaluate kindness and at the moment of reciprocation. Rabin (1993) develops a theory of fairness equilibria, where a player's reciprocity is exclusively driven by the underlying belief about his opponent's intention. In Falk and Fischbacher (2006), reciprocity is not only based on intentions but also driven by the observed consequences of actions. Our focus on reciprocity and altruism considers exclusively intentions — summarized in players types — rather than consequences. As in Levine's static setting, observed choices change as reciprocity and preferences evolve (and not the other way around); our reciprocity coefficients cannot depend on observed consequences of actions. This approach greatly improves (Rabin, 1993) tractability by replacing beliefs about intentions with beliefs about the players' intrinsic altruism (types). Similar to our study, (Sethi and Somanathan, 2001) conduct an evolutionary analysis of reciprocity

[2] While there is a line of research that assumes unobservable preferences (Ok and Vega-Redondo, 2001; Ely and Yilankaya, 2001; Dekel et al., 2007), these attempts have recently been criticized by Gamba (2011, 2013), as they rely on the assumption that a Bayesian Nash equilibrium is played at any state of the evolutionary dynamics. This assumption favors selfish preferences. By adopting a weaker solution concept — the self-confirming equilibrium — (Gamba, 2013) shows that altruistic preferences can be evolutionarily stable, even under unobservable preferences.

[3] Gamba (2013) explores the evolution of altruistic preferences in the centipede game.

using a slight variation of Levine's specification to model preferences. Unlike this paper, they consider only two types of players, materialists and reciprocators, and provide sufficient conditions for stable preferences. Furthermore, all reciprocators are altruists and have the same intrinsic preferences. That is, meetings between reciprocators translates into a symmetric game. We model reciprocal preferences more generally and do not require that reciprocators have the same intrinsic preferences; therefore, interactions always occur between heterogeneous players. More importantly, they do not endogenize reciprocity or preferences, which we see as a key feature of our model.

We present the model in Section 2 and offer theoretical predictions for evolutionarily stable interdependent preferences in Section 3. We then explore the reciprocal preferences case in Section 4. Finally, we present our conclusions in Section 5. Brief and instructive proofs are in the text, and lengthier ones are provided in the appendix.

2. The model

A large population of individuals are continuously and randomly matched in pairs. We label matched players as *player i* and *player j*, with $i, j \in \{1, 2\}$. In a meeting, they independently choose quantities $x_i, x_j \in \mathbb{R}_+$. Conditional on his opponent's choice x_j, player i receives a *material payoff* $\pi_i(x_i, x_j) = x_i(1 - x_i - x_j)$. That is, the strategic environment allows for *negative externalities and strategic substitutes*.

Preferences are *interdependent* and players care about the material payoff of their opponents. Player i perceives an *adjusted utility* of $u_i(x_i, x_j) = \pi_i(x_i, x_j) + \beta_i\pi_j(x_j, x_i)$, where $\beta_i \leq 1$ is his *behavioral preference* coefficient that summarizes his concern about his opponent's material payoff.[4] We consider that player *i behaves altruistically, selfishly or spitefully* if $\beta_i > 0$, $\beta_i = 0$ or $\beta_i < 0$, respectively. Players choose quantities to maximize their adjusted utility. Preferences evolve as players pursue their individual material payoffs that in general will diverge from their adjusted utility.

In Section 4, we explore a more specific model in which players have *reciprocal preferences*. Here, we distinguish a player's behavioral preference from his *intrinsic preference* coefficient, denoted by $a_i \leq 1$. Meetings occur between players from two large populations of individuals. Each population is characterized by its own intrinsic preference parameter; therefore, (a_1, a_2) fully describes two different populations, and each one includes players that share the same intrinsic value. We refer to player i as *intrinsically altruistic, selfish or spiteful* if $a_i > 0$, $a_i = 0$ or $a_i < 0$, respectively. In each meeting, the intrinsic values are common knowledge. We assume that players share the same notion of reciprocity, which apply symmetrically according to $\beta_{ij}(\lambda_i) = (a_i + \lambda_i a_j)/(1 + \lambda_i)$, where $\lambda_i \geq 0$ is a *reciprocity coefficient* that weights players intrinsic values. Evolutionary stable reciprocal preferences arise as players exclusively pursue their material payoff by adjusting how much reciprocity to exert.

3. Evolutionarily stable interdependent preferences

We first explore interdependent preferences that are evolutionarily stable. Ultimately, we wish to predict the behavior of the players and understand how much concern they express for others. First, let preferences be given by β_i for player i. In a meeting, the players simultaneously maximize their adjusted utility by choosing quantities; player i chooses $x_i \geq 0$.[5] We now solve for the Nash equilibria in this two player sub-game.

Player i's best response is $x_i(x_j) = (1 - x_j(1 + \beta_i))/2$. When $\beta_1 = \beta_2 = 1$, the best responses perfectly overlap, as each player maximizes the joint payoffs. Equilibrium quantities obey $x_1^* + x_2^* = 1/2$. Otherwise, the unique equilibrium is described by:

$$x_i^* = \frac{1 - \beta_i}{4 - (1 + \beta_i)(1 + \beta_j)} \tag{1}$$

Optimized material payoffs are:

$$\pi_i(x_i^*, x_j^*) = \frac{(1 - \beta_i)(1 - \beta_i\beta_j)}{(4 - (1 + \beta_i)(1 + \beta_j))^2} \tag{2}$$

Although (1) and (2) allow for negative quantities and payoffs,[6] as preferences optimally evolve, we prove that both quantities and payoffs always remain positive. To show this formally, we define $\mathcal{B} = \{\beta_1, \beta_2 \leq 1 \text{ and } \beta_1 \cdot \beta_2 \leq 1\}$ as the set of interdependent preferences that yield positive quantities and payoffs[7], and we show that evolutionarily stable preferences obey $(\beta_1, \beta_2) \in \mathcal{B}$.

To study the evolution of preferences, we now let behavioral preferences coefficients act as choice variables. Preferences evolve as players pursue individual material payoffs — instead of adjusted utility —which represent evolutionary success in our evolutionary game theoretical framework. For a fixed pure strategy β_j, player i maximizes $\pi_i(x_i^*, x_j^*)$ by choosing $\beta_i \leq 1$. Using Eq. (2), both players' FOCs coincide at:

$$1 + \beta_1 + \beta_2 = 3\beta_1\beta_2 \tag{3}$$

which fully describes the set of evolutionarily stable interdependent preferences.

Proposition 1. *Any $\beta_1, \beta_2 \leq 1$ obeying(3) are evolutionarily stable interdependent preferences. They satisfy $(\beta_1, \beta_2) \in \mathcal{B}$ and so $x_i^* \geq 0$ and $\pi_i(x_i^*, x_j^*) \geq 0$ for all $i \in \{1, 2\}$.*

Proof. We first compute player i's best response against a pure strategy $\beta_j < 1$. We let $\beta_i \leq 1$ in the extended reals. As the constraints are linear, the constraint qualification is met and, thus, any optimum solves the Kuhn-Tucker FOC. The unique interior candidate is $\beta_i = (1 + \beta_j)/(3\beta_j - 1)$, by Eq. (3). Because $\beta_i - 1 = 2(1 - \beta_j)/(3\beta_j - 1) \leq 0$ and since the SOC is $-(1 - 3\beta_j)^4/64(1 - \beta_j)^5$, it is the unique local maximum only for $\beta_j < 1/3$. Payoffs are $1/8(1 - \beta_j) > 0$. For global optimality, we check that payoffs obey $1/8(1 - \beta_j) - \pi_i(x_i^*, x_j^*) = (1 + \beta_i + \beta_j - 3\beta_i\beta_j)^2/8(1 - \beta_j)(4 - (1 + \beta_i)(1 + \beta_j))^2 > 0$, by Eq. (2). To wit, $\beta_i = (1 + \beta_j)/(3\beta_j - 1)$ is the unique global maximum when $\beta_j < 1/3$. When $1 > \beta_j \geq 1/3$, then the optimum lies on one corner. At $\beta_i = 1$, we obtain $\pi_i(x_i^*, x_j^*) = 0$. Using Eq. (2), when $1 > \beta_j \geq 1/3$, we obtain $(1 - \beta_i)(1 - \beta_i\beta_j) > 0$; therefore, $\pi_i(x_i^*, x_j^*) > 0$. Then, $\beta_i = -\infty$ is optimal. For $\beta_j = 1$, $\beta_i = 1$ is optimal if $x_i^* \geq 1/2$; otherwise, any $\beta_i < 1$ is optimal. Best responses intersect at (3), which yields a continuum of strict equilibria. We now verify that $(\beta_1, \beta_2) \in \mathcal{B}$. Exploiting (3), we check that in any of such equilibria, $x_i^* = 1/2(1 - \beta_j) \geq 0$, $x_j^* = (1 - 3\beta_j)/4(1 - \beta_j) \geq 0$ and $x_1^* + x_2^* = 3/4 < 1$. According to Selten (1980), any strict Nash equilibrium is evolutionarily stable. □

Observe that $\beta_1 = \beta_2 = 1$ is not a Nash equilibrium; therefore, this socially efficient preference profile cannot be evolutionarily stable.[8]

[4] The assumption that $\beta_i \leq 1$ precludes the *"after you" problem* of altruism (See Collard, 1978).

[5] Our payoff specification represents social dilemmas in which individual choices are inefficient. The simplest example is a common resource game in which players costlessly exploit a unit of the resource, each at the effort level x_i. Other examples are standard games of oligopolistic competition, where x_i represents firm i's production choice.

[6] If $\beta_i = -1$ and $\beta_j = -3$, then $x_i^* = 1/2$, $x_j^* = 1$, $\pi_i(x_i^*, x_j^*) = -1/4$ and $\pi_j(x_j^*, x_i^*) = -1/2$. If $\beta_i = -2$ and $\beta_j = -1$, then $x_i^* = 3/4$, $x_j^* = 1/2$, $\pi_i(x_i^*, x_j^*) = -3/16$ and $\pi_j(x_j^*, x_i^*) = -1/8$.

[7] As $\beta_1, \beta_2 \leq 1$, exploiting (1) and (2) we require $4 \geq (1 + \beta_1)(1 + \beta_2)$ for $x_1^*, x_2^* \geq 0$ and $\beta_1\beta_2 \leq 1$ for positive prices and payoffs. But if $\beta_1\beta_2 \leq 1$, then $(1 + \beta_1)(1 + \beta_2) \leq 4$.

[8] If it is, then $x_1^* + x_2^* = 1/2$; therefore, $\pi_1(x_1^*, x_2^*) = x_1^*/2$ and $\pi_2(x_2^*, x_1^*) = x_2^*/2$. For $\beta_2 = 1 > \beta_1$, we obtain$\pi_1(x_1^*, x_2^*) = 1/4$, and for $\beta_1 = 1 > \beta_2$, we obtain $\pi_2(x_2^*, x_1^*) = 1/4$, by Eq. (2). Equilibrium dictates $x_1^* \geq 1/2$ and $x_2^* \geq 1/2$. A contradiction.

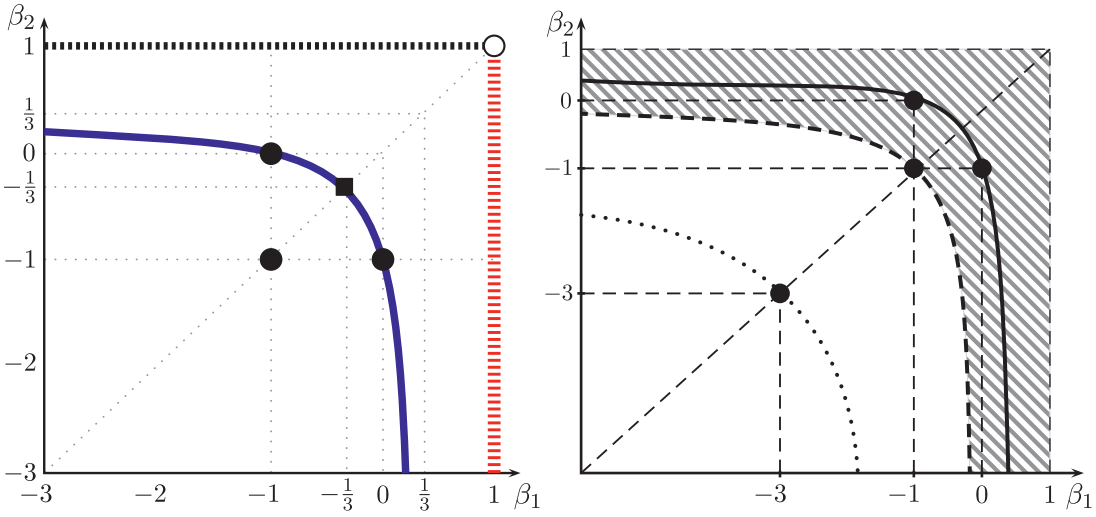

Fig. 1. Evolutionarily stable preferences. On the left, player 1's best response is the vertically-dashed line and the solid curve. If $1 > \beta_2 \geq 1/3$, then $\beta_1 = -\infty$ is optimal. Player 2's best response is the horizontally-dashed line and the solid curve. If $1 > \beta_1 \geq 1/3$, then $\beta_2 = -\infty$ is optimal. $\beta_1 = \beta_2 = 1$ is not an equilibrium. Best responses intersect at the solid curve (3), which describes a continuum of evolutionarily stable preferences. The right panel depicts the set $\mathcal{B}$. The solid curve is (3), which is contained in $\mathcal{B}$. Profiles between the dotted and dashed curves induce $x_1^*, x_2^* \geq 0$ but $x_1^* + x_2^* > 1$. Below the dotted curve, $x_1^* + x_2^* \leq 1$ but $x_1^*, x_2^* < 0$.

More generally, any preference profile inducing the efficient outcome requires $x_1^* + x_2^* = 1/2$ and so $(1 - \beta_1)(1 - \beta_2) = 0$,by (1). By Proposition 1, none of them are equilibrium profiles.

Our result in Proposition 1 extends the set of evolutionarily stable preferences proposed by Bester and Güth (1998) and Possajennikov (2000), since we do not impose symmetry. Interestingly, this symmetric game yields many — in most cases, asymmetric — evolutionarily stable interdependent preferences. The only symmetric profile is $\beta_1 = \beta_2 = -1/3$, which has been proven evolutionarily stable by Possajennikov (2000).

Unlike Levine (1998), we allow $\beta_i \leq -1$. If we restrict $\beta_i \in [-1, 1]$, then only spiteful or selfish preferences are evolutionarily stable, as depicted in the left panel of Fig. 1. More specifically, in each meeting, at least one player behaves spitefully and at most, one acts selfishly. Surprisingly, in our setting with a larger support for preferences, altruism might arise as optimally evolved behavior by only one of the players. In particular, whenever one player behaves altruistically, his opponent behaves spitefully.

We finish this section by offering a simple comparative static exercise: suppose player j behaves more altruistic. For $(\beta_1, \beta_2) \in \mathcal{B}$ and $\beta_j < 1$, we obtain:[9]

$$\frac{\partial \pi_i(x_i^*, x_j^*)}{\partial \beta_j} = \frac{(1 - \beta_i)(2 - \beta_i\beta_j(1 + \beta_i) - \beta_i(1 - \beta_i))}{(4 - (1 + \beta_i)(1 + \beta_j))^3} \geq 0$$

That is, someone who behaves more altruistically is willing to increase his opponents' material payoff. He achieves this by reducing x_j^* enough so that $x_1^* + x_2^*$ falls and $\pi_i(x_i^*, x_j^*)$ rises, regardless of how x_i^* changes.[10] The monotonicity of $\pi_j(x_j^*, x_i^*)$ in β_j is ambiguous in general.

If $\beta_i = -1$, then $\pi_j(x_j^*, x_i^*) = \left(1 - \beta_j^2\right)/16$; therefore, $\partial \pi_j(x_j^*, x_i^*)/\partial \beta_j = -\beta_j/8$.

Evolutionarily stable preferences obey $d\beta_i/d\beta_j = -4/(3\beta_j - 1)^2 < 0$, by Eq. (3). To wit, stable interdependent preferences act as substitutes: As player j behaves more altruistically (less spitefully), player i behaves more spitefully (less altruistically). We exploit this to explore how behavior optimally adjusts as player j behaves more altruistically. Recall that $x_i^* = 1/2(1 - \beta_j) \geq 0$, $x_j^* = (1 - 3\beta_j)/4(1 - \beta_j) \geq 0$ and $x_1^* + x_2^* = 3/4$, by Proposition 1. Then, as β_j rises, x_j^* falls, and x_i^* rises but $x_1^* + x_2^*$ remains unchanged. The game becomes a constant sum game with $\pi_1(x_1^*, x_2^*) + \pi_2(x_2^*, x_1^*) = 3/16$. By behaving more altruistically, player j increases his opponents' material payoff $\pi_i(x_i^*, x_j^*) = 1/8(1 - \beta_j)$. That is, "someone who behaves more altruistically is willing to reduce his own material payoff in order to increase his opponents", as stated by Bester and Güth (1998). In the extreme case, as $\beta_j \to 1/3$ and $\beta_i \to -\infty$, then $\pi_i(x_i^*, x_j^*) \to 3/16$ and $\pi_j(x_j^*, x_i^*) \to 0$. That is, player j totally reduces his payoffs in favor of player i, who earns 3/16.

4. Reciprocity and reciprocal preferences

We now use the indirect evolutionary approach to study reciprocity and explore evolutionarily stable reciprocal preferences.

Our key underlying observation is that individuals are better characterized by an intrinsic preference coefficient — like genes, which are acquired through genetic inheritance — yet they are able to adjust their behavior depending on who they interact with. That is, preferences evolve as reciprocity adjusts in each meeting, but intrinsic values do not vary at all. This setting allows us not only to predict how players will behave but also to answer a more specific question regarding *which* players will behave altruistically, spitefully or selfishly. For instance, would an intrinsically spiteful player ever behave altruistically? Would an intrinsically altruistic player ever behave spitefully?

To explore how reciprocity and preferences evolve, we now let pairwise meetings occur between players from two large populations of individuals. Each population $i \in \{1, 2\}$ is fully characterized by an intrinsic preference parameter $a_i \leq 1$. We assume $a_1 \neq a_2$, since otherwise, neither reciprocity nor preferences will evolve. We assume that each population intrinsic preference parameter — that summarizes beliefs about intentions —is common knowledge.

As in Levine (1998), we assume that players share the same notion

[9] To see this, observe that $1 - \beta_i \geq 0$ and that for (the denominator is positive. To wit, it suffices to show that the numerator is positive. For $\beta_i < 0$, observe that $\beta_i\beta_j(1 + \beta_i)$ rises in β_j only for $\beta_i \leq -1$, in which case $\beta_i\beta_j(1 + \beta_i) \leq \beta_i(1 + \beta_i)$; therefore, $\beta_i\beta_j(1 + \beta_i) + \beta_i(1 - \beta_i) \leq 2\beta_i < 2$. For $\beta_i \geq -1$, we obtain $\beta_i\beta_j(1 + \beta_i) \leq (1 + \beta_i)$; therefore, $\beta_i\beta_j(1 + \beta_i) + \beta_i(1 - \beta_i) \leq 2 - (1 - \beta_i)^2 \leq 2$.

[10] Eq. (1)can be easily used to deduce that $\partial x_j^*/\partial \beta_j = -2(1 - \beta_i)/(4 - (1 + \beta_i)(1 + \beta_j))^2 \leq 0$ and $\partial x_i^*/\partial \beta_j = (1 + \beta_i)(1 - \beta_i)/(4 - (1 + \beta_i)(1 + \beta_j))^2$ so that x_i^* rises in β_j iff $\beta_i \geq -1$. Altogether, $\partial(x_1^* + x_2^*)/\partial \beta_j = -(1 - \beta_i)^2/(4 - (1 + \beta_i)(1 + \beta_j))^2 \leq 0$.

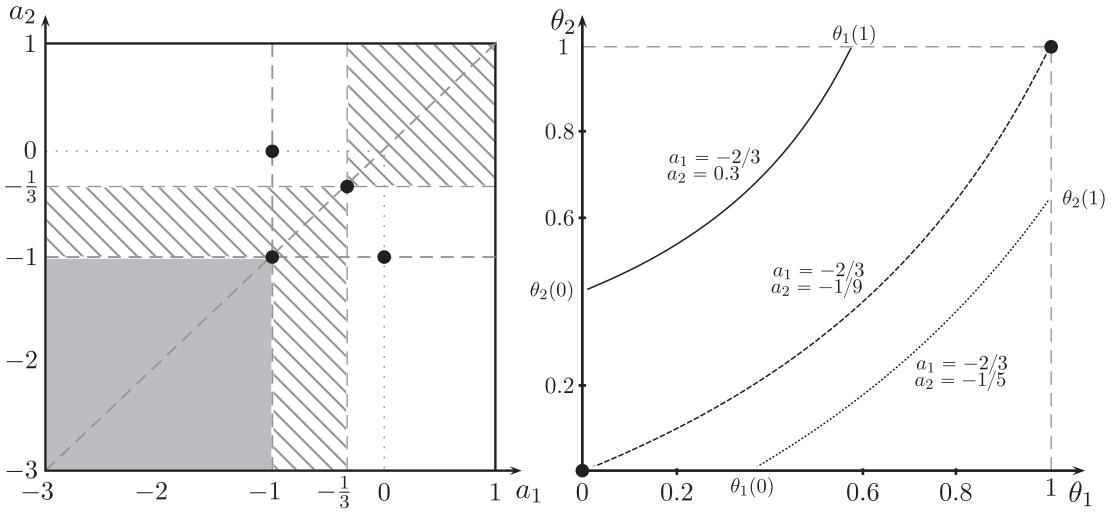

Fig. 2. Reciprocity. On the left, the intrinsic values space is partitioned. In the upper (lower) dashed region, the more (less) altruistic player exerts strong reciprocity; the other is not reciprocal. In the gray region, $\beta_{12}, \beta_{21} \notin \mathcal{B}$. In the white region, $(1 + 3a_1)(1 + 3a_2) < 0$, for each (a_1, a_2), there are infinite evolutionarily stable preferences that satisfy (3). On the right, the set of evolutionarily stable reciprocity weights are strategic complements. For each (a_1, a_2), there are infinite stable reciprocity weights that satisfy (5).

of reciprocity that apply symmetrically according to the linear specification[11]:

$$\beta_{ij}(\lambda_i) = \frac{a_i + \lambda_i a_j}{1 + \lambda_i} \tag{4}$$

Preferences are reciprocal in the sense that if someone is an altruist, others will be nicer to him than if he is spiteful or selfish.[12] In other words, players reward kindness and punish spitefulness by adjusting their reciprocity. For instance, for a spiteful person, reciprocity dictates that he should want an altruistic opponent to have a higher payoff than if he did not have a concern for reciprocity.

Intrinsic and behavioral preferences only coincide when a player's reciprocity coefficient is zero. *Reciprocity* arises when $\lambda_i \neq 0$, and *strong reciprocity* when $\lambda_i = \infty$. Essentially, an individual shows reciprocity when the concern that he expresses for his opponent's payoff depends on his opponent's intrinsic value. Let $\theta_i = \lambda_i/(1 + \lambda_i)$ be player i's *reciprocity weight*; therefore, $\beta_{ij}(\theta_i) = a_i + \theta_i(a_j - a_i)$. Observe that despite the fact we do not impose restrictions on $\lambda_i \geq 0$, the reciprocity weights are naturally constrained in $\theta_i \in [0, 1]$.[13] Clearly, reciprocity arises when $\theta_i > 0$, and strong reciprocity arises when $\theta_i = 1$.

We now let the reciprocity coefficients evolve for fixed intrinsic values. This differs from Levine's original interpretation where the reciprocity coefficient is a fixed parameter. Unlike Section 3, in every pairwise meeting, preferences will now be constrained to be in $[\min(a_1, a_2), \max(a_1, a_2)]$. In this refined problem, the corner solutions (i.e strong reciprocity) arise more often, which reduces the set of evolutionarily stable preferences found in Section 3. Proposition 2 addresses the strong reciprocity cases.

Proposition 2. *If $a_j \geq -1$ and $(a_i - a_j)(3a_j + 1) \geq 0$ then $\theta_i^* = 1$ and $\theta_j^* = 0$ is the unique evolutionary stable equilibrium strategy profile. The induced evolutionarily stable reciprocal preference is $\beta_{ij} = \beta_{ji} = a_j$.*

This result shows that strong reciprocity together with no reciprocity might be an evolutionarily stable strategy profile.[14] This occurs if players are either altruistic or spiteful enough, depicted as the dashed region on the left panel of Fig. 2 if $\min(a_1, a_2) \geq -1/3$ or $\max(a_1, a_2) \leq -1/3$. In the first (second) case, the more (less) altruistic player exerts strong reciprocity, whereas the other is not reciprocal. Regardless, in each case, the induced evolutionarily stable preferences are the same.[15] If $\min(a_1, a_2) \geq -1/3$, then the induced evolutionarily stable preference is $\beta_{12} = \beta_{21} = \min(a_1, a_2)$; therefore, both players might behave spitefully, altruistically or selfishly. Observe that players would like to behave less altruistically according to Eq. (3) — as in Section 3 — but their behavior is constrained by their genes. To wit, if altruism arises, it does so at its lower possible value. If $\max(a_1, a_2) \leq -1/3$, then the induced evolutionarily stable preference is $\beta_{12} = \beta_{21} = \max(a_1, a_2)$; therefore, players always behave spitefully. In this case, players moderate their behavior and end up behaving less spitefully than what they actually are. They would like to behave even less spitefully, but they are constrained by their genes; therefore, spite arises at the lowest possible value.

Any other evolutionarily stable preferences obey (3) as in Section 3. This "one equation-two unknowns" problem yields multiple reciprocal preferences profiles. Using Eq. (4) in Eq. (3), the reciprocity weights obey:

$$\theta_i(\theta_j) = \frac{1}{3(a_j - a_i)}\left(1 - 3a_i + \frac{4}{3\beta_{ji}(\theta_j) - 1}\right) \tag{5}$$

Clearly, as $d\theta_i/d\theta_j = 4/(3\beta_j(\theta_j) - 1)^2 > 0$, the weight that each player assigns to his opponent's intrinsic altruism is a strategic complement: *in a game where actions are strategic substitutes, preferences are also strategic substitutes, but reciprocity weights are strategic complements.* The following result characterizes the interior pure strategy equilibria that are evolutionarily stable.

[11] Sethi and Somanathan (2001) use a slightly different specification for preferences. In their paper, preferences obey $\beta_{ij}(\lambda_i) = (a_i + \lambda_i(a_j - a_i))/(1 + \lambda_i)$. Furthermore, they consider only two types of players: materialists (with $a_i = \lambda_i = 0$) and homogeneous reciprocators who are intrinsically altruistic (with $a_j = \alpha \in (0, 1)$ and $\lambda_j > 0$).

[12] As stated in Levine (1998), the model *can be regarded as incorporating fairness, not in the sense that players have a particular target they consider "fair", but in the sense that they are willing to be more altruistic to an opponent who is more altruistic toward them.*

[13] This extends Levine's values; he lets $\lambda_i \in [0, 1]$.

[14] One might think that the fact that players weight intrinsic values somehow "forces" reciprocity.

[15] If we allow material payoffs to be negative, then Proposition 2 results easily extend changing the initial requirement from $a_j \geq -1$ to $a_j \geq -3$. This guarantees that equilibrium quantities are still positive. That is, in the gray region of the left panel of Fig. 2, the less intrinsically altruistic player exerts strong reciprocity, whereas the other is not reciprocal. The unique evolutionary stable preferences will be $\beta_{ij} = \beta_{ji} = \max(a_1, a_2)$.

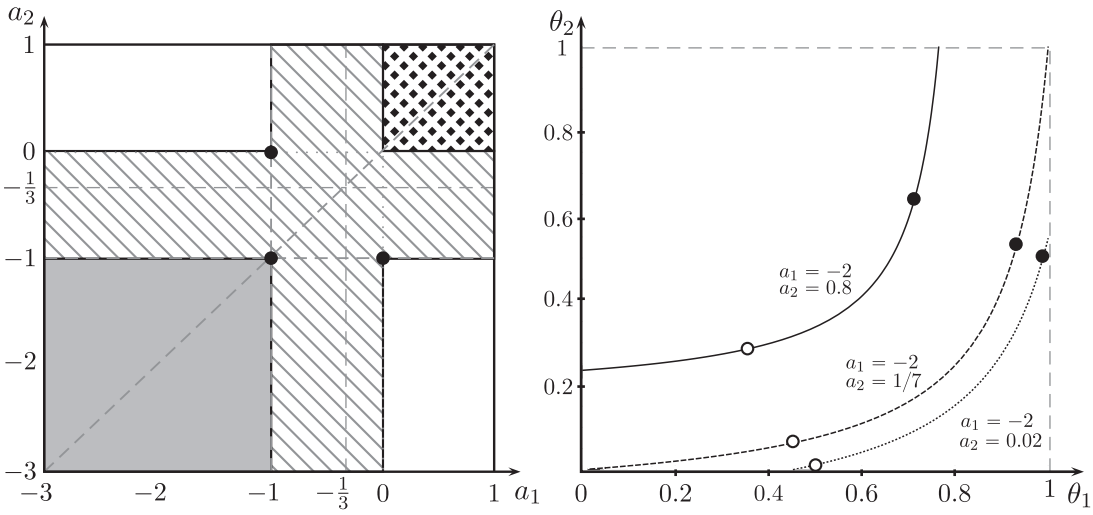

Fig. 3. Evolutionarily Stable Reciprocal Preferences. On the left, a partition of the intrinsic values space is partitioned. If a_1, $a_2 > 0$, then β_{12}, $\beta_{21} > 0$, and if either $a_1 \in [-1, 0]$ or $a_2 \in [-1, 0]$, we obtain β_{12}, $\beta_{21} \leq 0$. In the gray region, β_{12}, $\beta_{21} \notin \mathcal{B}$. In the white regions, $0 < a_j \leq 1$ and $a_i < -1$; for each (a_1, a_2), there are infinite evolutionarily stable reciprocity weights that satisfy (5). The right side shows the set of evolutionarily stable reciprocity weights and their induced preferences: When $\theta_2^* > \overline{\theta}_{21}$ (black circles), then $\beta_{12} > 0$ and $\beta_{21} < -1$. When $\theta_2^* < \underline{\theta}_{21}$ (white circles), then $\beta_{12} < -1$ and $\beta_{21} > 0$. Otherwise, $-1 \leq \beta_{12}, \beta_{21} \leq 0$.

Proposition 3. *For* $(1 + 3a_1)(1 + 3a_2) < 0$, *the evolutionarily stable reciprocity weights obey one of the following three cases:*

(A) *If* $1 + a_1 + a_2 = 3a_1a_2$, *then* $\theta_2^* \in [0, 1]$ *and* $\theta_1^* = \theta_1(\theta_2^*)$, *where* $\theta_1(0) = 0$ *and* $\theta_1(1) = 1$.
(B) *If* $(1 + a_1 + a_2 - 3a_1a_2)(3a_1 + 1) < 0$, *then* $\theta_2^* \in [\theta_2(0), 1]$, *and* $\theta_1^* = \theta_1(\theta_2^*)$, *where* $0 < \theta_2(0)$, $\theta_1^*(1) < 1$.
(C) *If* $(1 + a_1 + a_2 - 3a_1a_2)(3a_2 + 1) < 0$, *then* $\theta_1^* \in [\theta_1(0), 1]$, *and* $\theta_2^* = \theta_2(\theta_1^*)$, *where* $0 < \theta_1(0)$, $\theta_2^*(1) < 1$.

Observe that Propositions 2 and 3 fully partition the relevant set of intrinsic preferences, as depicted on the left panel of Fig. 2. If $(1 + 3a_1)(1 + 3a_2) < 0$, then the induced preferences obey (3); therefore, players could end up behaving spitefully, selfishly or altruistically, as described in Section 2. The following Proposition characterizes the set of evolutionarily stable reciprocal preferences that arise in each specific meeting.

Proposition 4. *The evolutionarily stable preferences are:*

- *If* a_1, $a_2 > 0$, *then* β_{12}, $\beta_{21} > 0$.
- *If either* $a_1 \in [-1, 0]$ *or* $a_2 \in [-1, 0]$, *then* β_{12}, $\beta_{21} \leq 0$.
- *If* $0 < a_j \leq 1$ *and* $a_i < -1$, *then there exists* $0 < \underline{\theta}_{ji} < \overline{\theta}_{ji} < 1$ *such that: If* $\theta_j^* > \overline{\theta}_{ji}$, *then* $\beta_{ij} > 0$ *and* $\beta_{ji} < -1$. *If* $\theta_j^* < \underline{\theta}_{ji}$, *then* $\beta_{ij} < -1$ *and* $\beta_{ji} > 0$. *If* $\theta_j^* \in [\underline{\theta}_{ji}, \overline{\theta}_{ji}]$, *then* $-1 \leq \beta_{ij}, \beta_{ji} \leq 0$.

Players adjust their behavior depending on who they interact with. Regardless, two altruistic (spiteful) players always end up behaving altruistically (spitefully). In fact, by exploiting Proposition 4, this is the only way for altruism to arise as evolved preferences if intrinsic values are restricted to be in $[-1, 1]$, as in Levine (1998). In any other case with this restricted support, at least one player behaves spitefully and at most, one acts selfishly. Notably, this implies that an altruist will never behave altruistically when meeting a spiteful or a selfish player.

With an extended support for intrinsic preferences, allowing for $a_i < -1$, additional combinations of stable behavioral preferences might arise in meetings between an altruistic and a highly spiteful player. Not only does at least one player behave spitefully, but at most one acts selfishly. In particular, if $0 < a_j \leq 1$ and $a_i < -1$, depicted as the white areas on the left panel of Fig. 3, we also determine that either the spiteful player behaves altruistically and the altruistic one behaves spitefully, or the spiteful player behaves spitefully and the altruistic one behaves altruistically. That is, altruism might arise even in meetings where one player is spiteful. More surprisingly, altruism might not arise from the intrinsically altruistic player and spitefulness might not arise from the intrinsically spiteful player. In any case, altruistic or selfish preferences never arise as evolutionary stable preferences for both players.

5. Conclusions

Levine's experimental work suggests that agents are better characterized as having reciprocal preferences. We used an indirect evolutionary approach to formally examine the evolutionary stability of interdependent preferences as well as the stability of reciprocity to explore how it shapes individual preferences and behavior. Unlike previous papers, we let reciprocity and preferences optimally evolve, and we do not take them as given. This strategic aspect of the preferences is the source of many surprising predictions. In a specific economic context characterized by negative externalities and strategic substitutes, players act reciprocally, leading to evolutionary stable preferences that differ from intrinsic ones. This implies that a players' concern for his opponent's payoff depends on his perception of the opponent's intention: the intrinsic preference of his opponent. In fact, strong reciprocity may be evolutionarily stable, entailing that an individual's concern for the success of others depends only on his opponent's intrinsic preference.

In particular, in some cases, reciprocity will induce evolutionarily stable behavioral preferences such that an intrinsically spiteful player and an altruistic player will behave spitefully in equilibrium. In other cases, the behavioral preference coefficients are asymmetric. Surprisingly, this includes stable preference profiles such that a spiteful player behaves altruistically, while his altruistic opponent behaves spitefully.

Future extensions include replicating our analytical framework for more general economic environments characterized by positive externalities and strategic complements and the experimental testing of the hypotheses inferred from this theoretical model, which could provide empirical evidence for the evolution of reciprocity and individual preferences. Another interesting topic for future research could be to consider a Bayesian game, where there is incomplete information on the other players' intrinsic preferences. Finally, a natural extension would be to combine the framework proposed by Sethi and Somanathan (2001) with ours to explore how reciprocity shapes preferences when agents are matched in larger groups.

Appendix

Proof of Proposition 2.. Let player i and j belong to populations with intrinsic values a_i and $a_j \neq a_i$, respectively. WLOG consider player i's maximization taking as given a_j and the pure strategy θ_j; therefore, β_j. As the constraints $\theta_i \geq 0$ and $\theta_i \leq 1$ are linear, the constrained qualification is met, and the Kuhn-Tucker FOC is necessary. We set up the Lagrangean $\mathcal{L} = \pi_i(x_i^*, x_j^*) + \underline{\gamma}_i\theta_i + \overline{\gamma}_i(1-\theta_i)$, where $\underline{\gamma}_i, \overline{\gamma}_i \geq 0$ are the constraints multipliers. The FOCs are:

$$\frac{(1-\beta_j)(1+\beta_i+\beta_j-3\beta_i\beta_j)}{(4-(1+\beta_i)(1+\beta_j))^3} = \frac{\underline{\gamma}_i - \overline{\gamma}_i}{a_j - a_i} \tag{A.1}$$

with $\underline{\gamma}_i\theta_i = 0$ and $\overline{\gamma}_i(1-\theta_i) = 0$. When $\theta_i = 1$ and $\theta_j = 0$, then $\beta_j = \beta_j = a_j$; therefore, Eq. (A.1) and (*A.j*) yield $(a_i - a_j)(1-a_j)^2(3a_j+1) = \overline{\gamma}_i(4-(1+a_j)^2)^3 = \underline{\gamma}_j(4-(1+a_j)^2)^3$, which clearly holds if $a_j = 1$. However, $\beta_i = \beta_j = 1$ is not an equilibrium. Otherwise, optimization dictates $(a_i - a_j)(3a_j+1)/(a_j+3)^3(1-a_j) = \overline{\gamma}_i = \underline{\gamma}_j \geq 0$. We restrict to $\beta_1, \beta_2 \in \mathcal{B}$; therefore, $a_j \geq -1$ and $(a_i - a_j)(3a_j+1) \geq 0$.

For uniqueness, we argue by contradiction, letting a_i, $a_j \leq 1$, $a_j \geq -1$, $(a_i - a_j)(3a_j+1) \geq 0$ and $\theta_i < 1$ or $\theta_j > 0$ or both. Observe that $4 \geq (1+\beta_i)(1+\beta_j)$ and $1+\beta_i+\beta_j-3\beta_i\beta_j \geq 0$ when $\beta_1, \beta_2 \in \mathcal{B}$.

If $\theta_i < 1$ and $\theta_j > 0$, then $\overline{\gamma}_i = \underline{\gamma}_j = 0$; therefore, $(1-\beta_j)(a_j - a_i)(1+\beta_i+\beta_j-3\beta_i\beta_j) \geq 0$, by Eq. (A.1) and $(1-\beta_i)(a_j - a_i)(1+\beta_i+\beta_j-3\beta_i\beta_j) \geq 0$, by (*A.j*). If $\beta_j = 1 > \beta_i$, then $a_j \geq a_i$, by (*A.j*); therefore, $\theta_j = (a_j-1)/(a_j-a_i) \leq 0$, which is a contradiction. By the same logic, we discard $\beta_i = 1 > \beta_j$. For $\beta_i, \beta_j < 1$, if $a_i > a_j$, then $a_j \geq -1/3$ and $\beta_i, \beta_j > -1/3$; therefore, $1+\beta_i+\beta_j-3\beta_i\beta_j > 0$, which is a contradiction. If $a_j > a_i$, then $a_j \leq -1/3$ and $\beta_i, \beta_j < -1/3$; therefore, $1+\beta_i+\beta_j-3\beta_i\beta_j < 0$, which is a contradiction.

If $\theta_i < 1$ and $\theta_j = 0$, then $\overline{\gamma}_i = \overline{\gamma}_j = 0$, $\beta_j = a_j$, $a_j < \beta_i \leq a_i$ if $a_i > a_j$ and $a_i \leq \beta_i < a_j$, and if $a_j > a_i$, so $(1-a_j)(a_j - a_i)(1+\beta_i+a_j-3\beta_i a_j) \geq 0$ by Eq. (A.1) and $(1-\beta_i)(a_i - a_j)(1+\beta_i+a_j-3\beta_i a_j) \geq 0$, by (*A.j*). If $a_j = 1$, then $(1-\beta_i)^2(a_i - 1) \geq 0$, by (*A.j*), which is a contradiction. If $\beta_i = 1$, then $(1-a_j)^2(a_j - a_i) \geq 0$, by Eq. (A.1). If $a_j = 1$, then $\beta_j = 1$, which is a contradiction. Otherwise, if $a_j \geq a_i$, then $a_j = 1 > \beta_i$, which is a contradiction. For β_i, $a_j < 1$, if $a_i > a_j$, then $\beta_j = a_j \geq -1/3$; therefore, $-1/3 < \beta_i \leq 1$. To wit, $1+\beta_i+a_j-3\beta_i a_j > 0$, which is a contradiction. If $a_j > a_i$, then $\beta_j = a_j \leq -1/3$, $\beta_i < -1/3$; therefore, $1+\beta_i+\beta_j-3\beta_i\beta_j < 0$, which is a contradiction.

If $\theta_i = 1$ and $\theta_j > 0$, then $\underline{\gamma}_i = \underline{\gamma}_j = 0$, $\beta_i = a_j$, $a_j < \beta_j \leq a_i$ if $a_i > a_j$ and $a_i \leq \beta_j < a_j$ if $a_j > a_i$; therefore, $(1-\beta_j)(1+a_j+\beta_j-3\beta_j a_j)(a_i - a_j) \geq 0$ by Eq. (A.1) and $(1-a')(1+a'+\beta'-3\beta' a')(a'-a) \geq 0$, by (*A.j*). If $a_j = 1$, then $(1-\beta_j)^2(a_i - 1) \geq 0$, by Eq. (A.1), which is a contradiction. If $\beta_j = 1$, then $(1-a_j)^2(a_j - a_i) \geq 0$, by (*A*.2). If $a_j = 1$, then $\beta_i = 1$, which is a contradiction. Otherwise, if $a_j \geq a_i$, then $a_j = 1 > \beta_j$, which is a contradiction. For β_j, $a_j < 1$, both players' FOCs dictate $1+\beta_i+a_j-3\beta_i a_j = 0$, by Eq. (A.1). If $a_i > a_j$, then $a_j \geq -1/3$; therefore, $-1/3 < \beta_j < 1$. To wit, $1+\beta_i+a_j-3\beta_i a_j > 0$, which is a contradiction. If $a_j > a_i$, then $a_j \leq -1/3$, $\beta_j < -1/3$; therefore, $1+\beta_i+\beta_j-3\beta_i\beta_j < 0$, which is a contradiction.

Altogether, if $a_j \geq -1$ and $(a_i - a_j)(3a_j+1) \geq 0$, then only $\theta_i = 1$ and $\theta_j = 0$ solves the necessary Kuhn-Tucker FOC. Finally, as the set [0,1] is compact, the functions $\pi_i(x_i^*, x_j^*)$ and $\pi_j(x_j^*, x_i^*)$ attain a maximum at θ_i^* in [0,1] and at θ_j^* in [0,1], for any $\theta_j \in [0,1]$ and $\theta_i \in [0,1]$, respectively. To wit, $\theta_i^* = 1$ and $\theta_j^* = 0$ is the unique strict Nash equilibrium and the unique evolutionary stable strategy profile. □

Proof of Proposition 3.. We find all $(\theta_i, \theta_j) \in [0,1]^2$ that satisfy (5), which are analogous for the other player. For (*A*): $\theta_i^*(0) = (1+a_i+a_j-3a_i a_j)/(a_j - a_i)(3a_j - 1)$ and $\theta_i^*(1) = (1+3a_i)(1-a_i)/(a_j - a_i)(3a_i - 1)$. If $1+a_i+a_j = 3a_i a_j$, then $\theta_i^*(0) = 0$ and $\theta_i^*(1) = 1$, to wit, $\theta_i^* = \theta_j^* = 0$ and $\theta_i^* = \theta_j^* = 1$ are equilibrium profiles. As (5) rises in θ_j, then any $\theta_j^* \in (0,1)$ and $\theta_i^* = \theta_i(\theta_j^*)$ are also equilibrium profiles. For (*B*): As $\theta_j^*(0) = (1+a_i+a_j-3a_i a_j)/(a_i - a_j)(3a_i - 1) = 1 - \theta_i^*(1)$, then $0 < \theta_i^*(1) < 1 \Rightarrow 0 < \theta_j^*(0) < 1$. This rules out $\theta_j^* = 1$ with $\theta_i^* \in (0,1)$ and $\theta_j^* = 0$ and $\theta_i^* \in (0,1)$ as profiles when multiple equilibria exists. Solving by cases, we find that $\theta_i^*(1) > 0$ iff $a_j < a_i \in (-1/3, 1/3)$, $a_j > a_i \in (1/3, 1)$ or $a_j > a_i$ and $a_j < -1/3$. By the same token, $0 < \theta_i^*(1) < 1$ iff $(a_j - a_i)(1+a_i+a_j-3a_i a_j)(3a_i+1) < 0$. Since $1+a_i+a_j = 3a_i a_j$ at $a_i = a_j$, $0 < \theta_j^*(0)$, $\theta_i^*(1) < 1$ iff $(1+a_i+a_j-3a_i a_j)(3a_i+1) < 0$. Finally, as $\theta_i(\theta_j^*(0)) = 0$, then $\theta_i^* = 0$, $\theta_j^* = \theta_j^*(0)$ and $\theta_j^* = 1$, $\theta_i^* = \theta_i^*(1)$ are equilibrium profiles. Since (5) is increasing, any interior profile is also an equilibrium profile. Furthermore, all these equilibrium profiles are strict Nash equilibria, since they induce preferences obeying (3). According to Selten (1980), they are all evolutionarily stable. □

Proof of Proposition 4.. The first statement is obvious, as preferences are weighted averages of positive numbers. This holds true for the second statement if a_1, $a_2 < 0$. If $a_j = 0$ and $a_i \leq -1$, then $\beta_{ij} = a_i(1-\theta_i)$ and $\beta_{ji} = a_i\theta_j$. Reciprocity obeys either *A* or *C* by Proposition 3; therefore, β_{ij}, $\beta_{ji} \leq 0$. For $0 \leq a_j \leq 1$ and $-1/3 \leq a_i < 0$ we obtain $\theta_j^* = 1$, $\theta_i^* = 0$ and $\beta_{ij} = \beta_{ji} = a_i < 0$, by Proposition 2. For $0 \leq a_j \leq 1$ and $-1 \leq a_i < -1/3$, as $\beta_{ji} \geq -1$ we obtain $\beta_{ij} = (1+\beta_{ji})/(3\beta_{ji} - 1) \leq 0$. Then, as $\beta_{ij} \geq -1$, we obtain $\beta_{ji} \leq 0$.

Next, take $0 < a_j \leq 1$ and $a_i < -1$; therefore, reciprocity obeys *A*, *B* or *C*, by Proposition 3. Observe that $\beta_{ij} > 0$ iff $\beta_{ji} < -1$, which occurs iff $\theta_j^* > (1+a_j)/(a_j - a_i) = \overline{\theta}_{ji}$. Equivalently, $\beta_{ij} < -1$ iff $\beta_{ji} > 0$, which occurs iff $\theta_j^* > a_j/(a_j - a_i) = \underline{\theta}_{ji}$. As $\overline{\theta}_{ji} - 1 = (1+a_i)/(a_j - a_i) < 0$ we obtain $0 < \underline{\theta}_{ji} < \overline{\theta}_{ji} < 1$. We now show that the best response in (5) intersects with $\underline{\theta}_{ji}$ and $\overline{\theta}_{ji}$ when reciprocity obeys *A*, *B* or *C*. It suffices to show that $\underline{\theta}_{ji} > \theta_j(0)$ when reciprocity obeys *A* or *B* and that $\overline{\theta}_{ji} < \theta_j(1)$ when reciprocity obeys *C*. In the first case, $\underline{\theta}_{ji} - \theta_j(0) = (1+a_i)/(a_j - a_i)(3a_i - 1) > 0$; in the second case, $\overline{\theta}_{ji} - \theta_j(1) = 4a_j/(a_j - a_i)(3a_j - 1) < 0$. Finally, $\beta_{ij} \leq 0$ iff $\beta_{ji} \geq -1$, which occurs iff $\theta_j \leq \overline{\theta}_{ji}$, and $\beta_{ij} \geq -1$ iff $\beta_{ji} \leq 0$ which occurs iff $\theta_j \geq \underline{\theta}_{ji}$. Altogether, if $\theta_j^* \in [\underline{\theta}_{ji}, \overline{\theta}_{ji}]$, we obtain $-1 \leq \beta_{ij}, \beta_{ji} \leq 0$. □

References

Alger, I., Weibull, J.W., 2013. Homo moralis preference evolution under incomplete information and assortative matching. Econometrica 81 (6), 2269–2302. https://doi.org/10.3982/ECTA10637.

Bell, A.V., Richerson, P.J., McElreath, R., 2009. Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. Proc. Natl. Acad. Sci. 106 (42), 17671–17674.

Bester, H., Güth, W., 1998. Is altruism evolutionarily stable? J. Econ. Behav. Organ. 34 (2), 193–209. https://doi.org/10.1016/S0167-2681(97)00060-7.

Bolle, F., 2000. Is altruism evolutionarily stable? And envy and malevolence?; Remarks on Bester and Güth. Journal of Economic Behavior and Organization 42 (1), 131–133.

Boyd, R., Richerson, P.J., 1988. Culture and the Evolutionary Process. University of Chicago Press.

Boyd, R., Richerson, P.J., 2006. Culture and the evolution of the human social instincts. Roots of Human Sociality. pp. 453–477.

Brandts, J., Solà, C., 2001. Reference points and negative reciprocity in simple sequential games. Games Econ. Behav. 36 (2), 138–157.

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

Collard, D.A., 1978. Altruism and Economy: A Study in Non-selfish Economics. Oxford University Press.

Dekel, E., Ely, J.C., Yilankaya, O., 2007. Evolution of preferences. Rev. Econ. Stud. 74 (3),

685–704.
Ely, J.C., Yilankaya, O., 2001. Nash equilibrium and the evolution of preferences. J. Econ. Theory 97 (2), 255–272.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav 54 (2), 293–315. https://doi.org/10.1016/j.geb.2005.03.001.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114 (3), 817–868.
Frank, R.H., 1987. If homo economicus could choose his own utility function, would he want one with a conscience? Am. Econ. Rev. 77 (4), 593–604.
Frank, R.H., 1988. Passions Within Reason: The Strategic Role of the Emotions. WW Norton & Co.
Gamba, A., 2011. On the Evolution of Preferences. Technical Report. Friedrich- Schiller-University Jena, Max-Planck-Institute of Economics.
Gamba, A., 2013. Learning and evolution of altruistic preferences in the centipede game. J. Econ. Behav. Organ. 85 (Suppl C), 112–117. https://doi.org/10.1016/j.jebo.2012.11.009. Financial Sector Performance and Risk
Güth, W., 1995. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. Int. J. Game Theory 24 (4), 323–344.
Güth, W., Huck, S., Müller, W., 1998. The Relevance of Equal Splits: On a Behavioral Discontinuity in Ultimatum Games. Technical Report. Discussion Papers, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes
Güth, W., Napel, S., 2006. Inequality aversion in a variety of games an indirect evolutionary analysis*. Econ. J. 116 (514), 1037–1056. https://doi.org/10.1111/j.1468-0297.2006.01122.x.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3 (4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7.
Heifetz, A., Shannon, C., Spiegel, Y., 2007. What to maximize if you must. J. Econ. Theory 133 (1), 31–57. https://doi.org/10.1016/j.jet.2005.05.013.
Herold, F., Kuzmics, C., 2009. Evolutionary stability of discrimination under observability. Games Econ. Behav. 67 (2), 542–551.
Isaac, R.M., Walker, J.M., 1988. Group size effects in public goods provision: the voluntary contributions mechanism. Q. J. Econ. 103 (1), 179–199.
Kokesen, L., Ok, E.A., Sethi, R., 2000. The strategic advantage of negatively interdependent preferences. J. Econ. Theory 92 (2), 274–299. https://doi.org/10.1006/jeth.1999.2587.
Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dyn. 1 (3), 593–622. https://doi.org/10.1006/redy.1998.0023.
Menicucci, D., Sacco, P.L., 2009. Evolutionary selection of socially sensitive preferences in random matching environments. J. Math. Sociol. 33 (4), 241–276.
Ok, E.A., Vega-Redondo, F., 2001. On the evolution of individualistic preferences: an incomplete information scenario. J. Econ. Theory 97 (2), 231–254.
Possajennikov, A., 2000. On the evolutionary stability of altruistic and spiteful preferences. J. Econ. Behav. Organ. 42 (1), 125–129. https://doi.org/10.1016/S0167-2681(00)00078-0.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83 (5), 1281–1302.
Sadrieh, A., Schröder, M., 2016. Materialistic, pro-social, anti-social, or mixed–a within-subject examination of self-and other-regarding preferences. J. Behav. Exp. Econ. 63, 114–124.
Selten, R., 1980. A note on evolutionarily stable strategies in asymmetric animal conflicts. J. Theor. Biol. 84 (1), 93–101. https://doi.org/10.1016/S0022-5193(80)81038-1.
Sethi, R., Somanathan, E., 2001. Preference evolution and reciprocity. J. Econ. Theory 97 (2), 273–297.
Sethi, R., Somanathan, E., 2003. Understanding reciprocity. J. Econ. Behav. Organ. 50 (1), 1–27.
Sobel, J., 2005. Interdependent preferences and reciprocity. J. Econ. Lit. 43 (2), 392–436.
Thunström, L., Cherry, T.L., McEvoy, D.M., Shogren, J.F., 2016. Endogenous context in a dictator game. J. Behav. Exp. Econ. 65, 117–120.