Reciprocity – An Indirect Evolutionary Analysis*

Siegfried K. Berninghaus**

and

Christian Korth‡

and

Stefan Napel‡‡

Version: October 19, 2006

No. 147

Abstract This paper investigates agents who face a stylized pecuniary 'game of life' comprising the ultimatum game and the dictator game. Utility may but need not be attached to equity and reciprocity as formalized by Falk and Fischbacher (2006) but, critically, this social component of preferences cannot be conditioned on whether an ultimatum or a dictator game is played. Evolutionary fitness of agents is determined solely by material success. Under these conditions, a strong preference for reciprocity but little interest in equity as such evolves. Possible exogenous constraints that link reciprocity and equity concern imply long-run levels of both which depend on the relative frequency of ultimatum vs. dictator interaction in agents' multi-game environment.

Forthcoming in: *Journal of Evolutionary Economics*

Keywords: Reciprocity, evolutionary stability, fairness
JEL codes: C78, C90

---

*We thank two anonymous referees for constructive comments and J. Oechssler for helpful discussion.

**University of Karlsruhe, School of Economics and Business Engineering, D-76128 Karlsruhe, Germany, berninghaus@wiwi.uni-karlsruhe.de

‡Grindelhof 85, D-20146 Hamburg, Germany, korth@stanfordalumni.org

‡‡University of Hamburg, Department of Economics, D-20146 Hamburg, napel@econ.uni-hamburg.de – Financial support from the German Academic Exchange Service (DAAD) is gratefully acknowledged.

1. INTRODUCTION

In experimental investigations of the *ultimatum game* participants quite consistently offer 30–50% of an available monetary surplus as first-moving proposers. They reject offers of less than 20% as second-moving responders, which results in zero payoff for both players. Particularly the latter observation is hard to reconcile with the assumption that economic actors are rational maximizers of their monetary payoffs.[1] However, observations can be explained very well by including a consideration for fairness and reciprocity in players' preferences. This is also true regarding many other games for which experimental findings are puzzling from a monetary-payoff maximization point of view.

Neoclassical theory does not restrict preferences to be based only on monetary payoffs or to be strictly monotonic in them. But economic agents who are spiteful, enjoy a warm glow donating money to anonymous strangers, or feel it worthwhile to incur private costs to punish free-riders of public goods have not been the conventional assumption in economics. The award of 2002's Nobel prize to Daniel Kahneman and Vernon Smith is only one indicator that this is changing fast.

In adopting a more realistic view of homo oeconomicus, however, one needs to be careful to not jump to the other extreme, i. e. to count any observation of supposedly odd behavior as evidence that 'standard assumptions' are wrong and to suppose that human beings universally have the nicer-than-expected character exhibited in some laboratory experiments. The latter would be invalidated by many other experiments, e. g. on market games, in which participants' behavior is explained well by egoistic maximization of monetary rewards. Also, the tale – told in different versions before the advent of behavioral economics – that (an unspecified kind of) evolution would make economic agents behave *as if* they maximized payoffs in a world of scarce material resources in our view contains some grain of truth. The question is: Under which circumstances is behavior more of the standard homo oeconomicus-type, and which environments defined by which conditions induce human beings to act e. g. like homo reciprocans (cp. Fehr and Gächter 1998) or benevolent dictators?

This paper combines methods of evolutionary and behavioral game theory to address the above question in a simple but still powerful model. The analysis concentrates on a possible preference for reciprocity as formalized by Falk and Fischbacher (2006). The key departure from the literature is that we consider an environment consisting not of just one, but two distinct distribution tasks (or

[1] It is unlikely that stakes are just too small for people to bother: experiments in which the available surplus amounted to several monthly wages of participants (e. g. in Indonesia by Cameron 1999 or in the Slovak Republic by Slonim and Roth 1998) produce roughly the same results. Findings are also very robust concerning the subject pool. See, e. g. the large cross-cultural study in non-student populations by Henrich et al. (2001).

games) and impose restrictions on the degree to which social preferences can be conditioned on the particular task at hand. This serves as a first approximation of the complexities of the real 'game of life' and the observation that it has to be tackled with a limited set of social norms, moral rules and emotions pertaining to general classes of interaction rather than specific situations. To be concrete we study evolution of particular reciprocity-based preferences in a world where agents randomly face either the ultimatum or the dictator game.[2]

Positive (negative) *reciprocity* refers to the impulse or desire to be kind (unkind) to those who have been kind (unkind) to us.[3] Reciprocity is to be distinguished from simple altruism, i. e. unconditional generosity. While the narrow self-interest hypothesis in 'standard theory' fails to explain stylized facts of many experiments, the notion of reciprocity preference sheds a fairly consistent light on possible motivational forces behind a number of observations.

A related motive which seems to guide behavior of economic agents in situations of social exchange is *inequity aversion*.[4] Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) have provided prominent models of agents with preferences that exhibit inequity aversion. In these models, agents may increase their utility by sacrificing their own material payoff if by doing so their payoff is closer to their counterparts' payoffs. Preferences can depend on the entire payoff distribution, but crucially not on any intentions ascribed to other players.

Still, it is a key feature of the psychology of reciprocity that decisions to be kind or unkind to others are based not only on material consequences implied by other players' actions but also on the *intentions* attributed to these players. Agents who are motivated by reciprocity discriminate between players who take an (un)generous action by choice and those who are forced to do so.[5]

Prominent formalizations of reciprocity based on intentions have been given by Rabin (1993), Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006). Rabin was the first to adopt the framework

[2]What we in the following refer to as a multi-game environment may also be viewed as a recurrent version of a comprehensive single game, in which Nature has the first move and selects either an ultimatum or a dictator subgame.

[3]See Fehr and Gächter (1998) for a more detailed account.

[4]Inequity aversion even seems to be an important determinant of social networks: subjects in recent experiments form equilibrium networks involving nearly equal payoffs; they abstain from forming a network with a player designed to earn significantly less than others (e. g. Falk and Kosfeld 2003, Berninghaus et al. 2006).

[5]Experimental evidence is given, for example, by Falk et al. (2003) for four different mini ultimatum games. In each game the proposer had two choices, one of which always was to offer 20%. The alternatives were 0%, 50% or 80%, respectively. The rejection rate of the 20% offer was highest when the alternative was equal division. When the only alternative was to offer nothing to the second player, i. e. when the 20% offer revealed good intentions, the rejection rate was lowest but remained positive. The latter suggests that pure equity concern also plays a role.

of psychological games of Geanakoplos et al. (1989) to model reciprocity. He introduced so-called fairness games in which a reciprocity payoff is added to the material payoff of the players. The reciprocity payoff is calculated as the product of a kindness term and a reciprocity term. The kindness term is positive whenever a player feels treated well. Then he or she tries to make the reciprocity term positive, too, in order to increase his or her total utility payoff. This is achieved by being nice in return. Negative reciprocity is modelled analogously.

While Rabin's original model only applies to two-player normal form games, Charness and Rabin (2002, Appendix 1) consider reciprocity and parallel concern for social welfare in multiperson settings. They define a reciprocal-fairness equilibrium, which imposes homogeneity in players' preferences and does not entail sequential rationality. It is therefore unsuitable to analyze the stylized 'game of life' focused on in this paper. The model of Dufwenberg and Kirchsteiger (2004) incorporates sequential rationality in general n-person extensive form games. Its main restriction is that it captures intention-based reciprocity but not any direct equity or social welfare orientation.[6]

Falk and Fischbacher (2006), first, extend Rabin's approach to extensive form games of perfect information with finitely many stages and, second, propose a utility function that allows for both intention-based reciprocity *and* pure equity concern. Equilibrium calculations and parameter calibrations are quite complex but predictions of Falk and Fischbacher's hybrid model are very consistent with a broad range of experimental evidence. This and its comprehensiveness – nesting 'traditional' preferences, pure inequity aversion and pure intentional reciprocity – make it a very powerful tool for the joint analysis of different games and therefore the focus of this paper.

Our choice of Falk-Fischbacher preferences obviously entails a loss of generality. But permitting *all* possible preferences would, in fact, lead back to allowing unimpeded context-dependent specialization of social preferences. So any model that considers the evolution of a limited set of behavioral norms pertaining to distinct interaction types would have to impose similar restrictions.

We have selected combinations of ultimatum and dictator games as agents' environment because these rather simple games already capture two very fundamental forms of conflictual social and economic interaction, admittedly in a highly stylized fashion: in the first, individual success depends on mutual cooperation and is influenced by both agents' behavior. In the second, an agent either has no power at all, or full control over own and someone else's success. Bolton and Ockenfels

[6] Another restriction of psychological games more generally (hence also the Falk-Fischbacher model) is that players are assumed to have consistent second-order beliefs, i. e. to correctly anticipate not only others' strategy choices but even their respective beliefs about one's own choice. This contrasts with evidence, for example, on human overconfidence or wishful thinking.

(2000, p. 188) explicitly show how "many facets of behavior, over a wide class of games, can be deduced from [these] two ... most elementary games." In particular, knowledge about agents' respective dictator offers and lowest accepted ultimatum offers allows fairly accurate predictions of their behavior in other bargaining, market, and social dilemma games. Insights about the evolutionary (in)stability of social preferences in our quite specific set-up may therefore also have more general economic relevance.

We study the evolution of preferences among rational decision makers. Agents have preferences satisfying the usual axioms and select strategies with the goal to maximize *expected utility*. Preferences can but need not depend on an agent's individual material payoff alone; they may e. g. exhibit reciprocity in the way indicated above. However, it is assumed that (average) material payoffs ultimately determine the *fitness* of the agents having particular preferences and hence preferences themselves in an evolutionary process: Agents with preferences that earn material payoffs above (below) the average reproduce more (less) successfully, and their preferences' population share increases (decreases).

The considered agents make choices based on anticipated consequences, evaluating their options by preferences which evolved dependent on past success. Thus, as argued in more detail e. g. in Berninghaus et al. (2003) and Güth et al. (2003), an *indirect evolutionary* model such as ours combines the main elements of the traditional neoclassical approach (rational purposeful actions selected according to their anticipated consequences or the 'shadow of the future') and the direct evolutionary approach (agents carry out fixed behavioral programs which evolve based only on past success or the 'shadow of the past').

Throughout this paper agents are assumed to know their opponents' preferences before they interact. This allows them to anticipate their opponents' action and optimally respond to it, as formalized by the (subgame perfect) Nash equilibrium concept. These are somewhat restrictive but not overly critical assumptions of the indirect evolutionary approach: immediate play of an equilibrium could be replaced by an adaptive learning process which operates sufficiently faster than evolution; general perfect observability of preferences – a standard assumption in game theory – could be relaxed to only occasional bilateral encounters with complete information or the possibility that players collect and process information about their opponents at non-prohibitive costs which they compare to population-dependent benefits.

The following analysis is closely related to Güth and Napel (2006). There, equity-based preferences similar to those of Fehr and Schmidt (1999) are considered in the same stylized 'game of life'. Here, the multi-game indirect evolutionary analysis of Güth and Napel[7] is extended to the domain of psychological games.

[7] Cf. Stahl and Haruvy (2006) for an experimental study of multi-game environments. Also see

The remainder of the paper is organized as follows: The following section presents our model, which is analyzed in section 3 with focus on the question: Under which circumstances is a reciprocity-based notion of fairness evolutionary stable? Section 4 compares results to those obtained in the related model of Güth and Napel (2006). Section 5 concludes.

2. THE MODEL

Repeatedly, two agents who are randomly drawn from a single population are given the chance to create a surplus (the 'pie'), and subsequently to decide about its distribution. Taking the population to be large enough to rule out repeated-game effects, agents are assumed to act fully rationally according to commonly known preferences for the possible outcomes of the game presented to them. Utility is not restricted to own material payoff, but (average) material payoff alone defines an agent's fitness or reproductive success in a – not explicitly modelled – payoff-monotonic evolutionary process imposed on the population.[8] As expressed in a pointed way by Samuelson (2001, p. 226f): "Nature can thus mislead her agents, in that preferences and fitness can diverge, but cannot mislead herself, in that high fitness wins the day."

2.1. The Material World

A matched pair of agents randomly faces either of two different games: The ultimatum game or the dictator game. In the ultimatum game one of the two players is randomly (with probability 0.5) selected to be the proposer (role X), who proposes how to split the pie of one unit. The amount he or she offers is denoted by $c \in [0,1]$. The other player (the responder, role Y) decides whether to accept the proposed split or to reject it. The resulting material payoffs are $(\pi_X, \pi_Y) = (1-c, c)$ and $(\pi_X, \pi_Y) = (0,0)$, respectively.

In the dictator game one agent is similarly assigned to the proposer role and decides how to split the pie by offering $c \in [0,1]$ to the responder. In contrast to the ultimatum game the responder must accept, so resulting material payoffs are $(\pi_X, \pi_Y) = (1-c, c)$.

Which game the agents play is determined randomly, with exogenous probability $\lambda \in [0,1]$ for the ultimatum game and $1-\lambda$ for the dictator game. The game realization becomes common knowledge to the agents. In both events material payoffs (π_X, π_Y) determine reproductive success.

Poulsen and Poulsen (2006) for simultaneous preference evolution in several games.

[8] See Benaïm and Weibull (2003) for a concise overview of models with agents who learn or imitate rather than biologically reproduce which can be very closely approximated by evolutionary models coming from a purely biological background.

Differences between dictator game vs. ultimatum game very loosely resemble those between private interaction vs. anonymous market interaction: Success in the ultimatum game depends on mutual cooperation and is influenced by both agents' behavior; in the dictator game one player is a mere price taker without influence on terms of trade. It will be interesting to see if and how evolutionary stable preferences depend on the frequency or 'importance' of each type of interaction in agents' lives.

2.2. Agents' Preferences

Agents can have fairness preferences as defined by Falk and Fischbacher (2006). Details about the utility representation are discussed in the Appendix.[9] Basing reciprocity on intentions has a price in terms of model complexity even in simple ultimatum or dictator games. Paying it, however, cannot be avoided if one does not want to rule out a priori that intentions matter in the considered context. Most important for this paper is that the utility functions $u_i^X(\cdot)$ and $u_i^Y(\cdot)$ which represent preferences of a given agent i in the roles of proposer X or responder Y, respectively, have the free parameters $\rho_i^X, \varepsilon_i^X$ and $\rho_i^Y, \varepsilon_i^Y$. Reciprocity parameter $\rho_i^k \in \mathbb{R}_+$ describes how much weight agent i places on reciprocal behavior in role $k \in \{X, Y\}$. For $\rho_i^k = 0$ the considered agent is purely interested in his or her own material payoff, behaving as in most orthodox economic models. The equilibrium rejection probability of ultimatum game offers below 50% increases in ρ_i^Y (see section 3.1). The second parameter, $\varepsilon_i^k \in [0, 1]$, measures agent i's pure concern for an equitable outcome in role k. Agents without any pure equity concern have an entirely intention-driven notion of fairness: whenever no good or bad intentions can be attributed to the other player, the agent simply maximizes material payoff. In contrast, for agents with positive equity concern more equitable splits of the pie are more valuable even in the absence of intentions. This is best seen in the dictator game where there is no scope for reciprocation: An agent with $\varepsilon_i^X = 0$ offers $c = 0$ whereas an agent with $\varepsilon_i^X = 1$ may offer up to half of the pie.[10]

In principle, the intensities of reciprocity and equity concern could be different in ultimatum and dictator games and also in the proposer or responder roles of each game. This would correspond to a setting in which agents can have moral sentiments that are tailor-made to very specific economic situations.[11] Here, we prefer to focus

[9] In short, player i's utility at any terminal node t of the game tree is the sum of his or her material payoff associated with t and a reciprocity payoff which is calculated decision node by decision node and weighted by parameter ρ_i. For each of i's decision nodes n along the path to t, the products of a term measuring the (un)kindness of others to i which has lead to n (here, parameter ε_i may matter) and a term measuring i's reciprocated (un)kindness from choosing to move further towards t are added up.

[10] The actual offer depends on the pure equity concern *and* the reciprocity parameters.

[11] The possibility that agents can condition equity concern on the game and/or their role in it is

on social preferences that reflect the general character of a given agent, i. e. which are the same in both randomly assigned roles and randomly selected distribution tasks. Therefore, we will apply $\rho_i^X \equiv \rho_i^Y \equiv \rho_i$ and $\varepsilon_i^X \equiv \varepsilon_i^Y \equiv \varepsilon_i$ for agent i in his or her entire stylized 'game of life'.

Allowing preferences of the Falk-Fischbacher type (which nest material payoff maximization, purely intention-based reciprocity, and inequity aversion) but excluding other types clearly is a restriction. It entails some arbitrariness, but can be motivated by the former type's descriptive success and generality. Giving up *any* restriction in the spirit of Dekel et al. (2005) would go too far in our view: It would implicitly allow for moral discrimination between the components of any mixed habitat. One would thus force about a simple superimposition of preference types that are each tailor-made for a specific single-game environment.

2.3. Stability Concepts

The approach pursued here can formally be subsumed under classical (direct) evolutionary game theory.[12] In particular, preference evolution for a 2-player game Γ in which fitness is determined by strategy profiles $\mathbf{s} \in S^2$ can be regarded as evolution in a higher-level game $\hat{\Gamma}$ in which payoffs and fitness are (indirectly) determined by *preference profiles* $\hat{\mathbf{s}} \in \hat{S}^2$. Namely, $\hat{\Gamma}$'s 'strategy space' $\hat{S}$ is the set of feasible preferences over outcomes in Γ and its payoff function $\hat{\pi}$ is the composition $\pi \circ \mu$ of original payoff function π and a mapping μ from preference profiles in $\hat{S}^2$ to equilibrium strategy profiles of Γ.

In the following, we will not explicitly model a dynamic evolutionary process. Our goal is to identify preferences $(\rho, \varepsilon) \equiv \hat{s}$ that are stable at least in the sense of being a *neutrally stable strategy* (NSS) in game $\hat{\Gamma}$. So considering payoffs $\hat{\pi}$ defined by the outcome of the respective (unique) subgame perfect equilibrium (SPE) of the underlying material game, we call preferences $\hat{s}$ *stable* if and only if for all $\hat{s}' \in \hat{S}$

$$\hat{\pi}(\hat{s}, \hat{s}) \geq \hat{\pi}(\hat{s}', \hat{s}) \tag{1}$$

and, moreover, whenever $\hat{\pi}(\hat{s}, \hat{s}) = \hat{\pi}(\hat{s}', \hat{s})$ then

$$\hat{\pi}(\hat{s}', \hat{s}') \leq \hat{\pi}(\hat{s}, \hat{s}'). \tag{2}$$

This is equivalent to preferences (ρ, ε) satisfying (2) for all (ρ', ε') in a neighborhood of (ρ, ε) (see e. g. Weibull 1995, Prop. 2.7).

explored in Güth and Napel (2006). If real economic agents' fair behavior is adequately described by a utility function with one or a few fairness-related parameters at all, the same parameters should in our view be valid for more than a very special class of games. Admittedly, to expect them to be valid in *all* games would be too much.

[12] The opposite is also true: Models that study the direct evolution of behavior can be regarded as the special case of preference evolution where feasible preferences are restricted to those making distinct strategies strictly dominant.

For finite strategy spaces, close links between static stability concepts such as NSS and stationary points of various dynamic evolutionary processes exist. In particular, a NSS corresponds to a population state which satisfies *(Lyapunov) dynamic stability* under the well-known replicator dynamics, i. e. a small group of invading mutants cannot spread (e. g. Weibull 1995, ch. 3). Unfortunately, even stronger concepts like *evolutionary stable strategy* (ESS), which replaces (2) by a strict inequality and implies actual repelling of a small invasion, do not guarantee that an arbitrary initial preference distribution converges. Static concepts like ESS and NSS are nevertheless a focal prediction for long-run evolution and have been the benchmark of most related investigations.[13]

As links between static stability concepts like NSS and actual evolutionary dynamics are much harder to pin down in the continuous case[14] (and one can argue that the world is fundamentally discrete anyhow, at least at the quantum level), we analyze evolution of preferences on a finite grid. In particular, we consider preference parameters $(\rho, \varepsilon) \in \hat{S} \equiv \{0, \frac{1}{n}, \frac{2}{n}, \ldots, \bar{P}\} \times \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ given an arbitrarily small grid size $1/n$ ($n \in \mathbb{N}$), and a large upper bound $0 < \bar{P} \in \mathbb{N}$.

3. EVOLUTIONARY ANALYSIS

First, ultimatum and dictator games will be studied in isolation, corresponding to the boundary cases $\lambda = 1$ and $\lambda = 0$. Then, the mixed environment consisting of both games will be analyzed. When agents A and B interact, we will in the following write ρ_X and ε_X (ρ_Y and ε_Y) for the preference parameters of the agent who is assigned to role X (role Y). Recall that we assume agents to have complete and perfect information when they interact.

3.1. The Ultimatum Game

Equilibrium play[15] results in *acceptance probability*

$$p^*(c) = \begin{cases} \min\left\{1, \frac{c}{\rho_Y \cdot (1-2c)(1-c)}\right\} & \text{if} \quad c < \frac{1}{2} \\ 1 & \text{if} \quad c \geq \frac{1}{2} \end{cases} \tag{3}$$

for offer c if $\rho_Y \neq 0$, and $p^* \equiv 1$ if $\rho_Y = 0$. So an offer of half of the pie or more

[13] A recent exception is Possajennikov (2005), who explicitly studies a dynamic process.

[14] We are aware of no general sufficient condition for dynamic stability if a straightforward definition of 'closeness' of two population states is applied. Oechssler and Riedel (2002) obtain a sufficient condition regarding the natural weak topology in the special case of *doubly symmetric games*. General sufficient conditions exist for the much less appealing variational norm (Oechssler and Riedel 2001).

[15] The derivations of the equilibrium for the ultimatum and dictator games are sketched in the appendix.

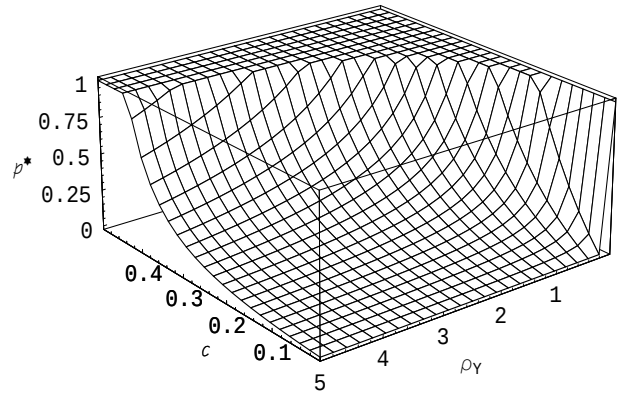

FIG. 1 Acceptance probability p^* of agent with reciprocity parameter ρ_Y for offer c in the ultimatum game

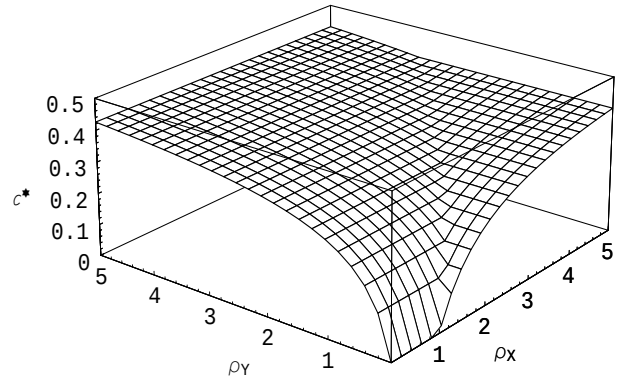

FIG. 2 Split c^* proposed by an agent with reciprocity parameter ρ_X to one with reciprocity parameter ρ_Y in the ultimatum game

is always accepted for sure. This is not the case for lower offers, for which the acceptance probability is decreasing in ρ_Y as shown in figure 1.

The share c^* offered by the proposer X to responder Y in equilibrium is given by

$$c^* = \max\left\{\frac{3\rho_Y + 1 - \sqrt{1 + 6\rho_Y + \rho_Y^2}}{4\rho_Y}, \frac{1}{2}\cdot(1 - \frac{1}{\rho_X})\right\} \tag{4}$$

for $\rho_X, \rho_Y \neq 0$. For $\rho_X > \rho_Y = 0$ and $\rho_Y > \rho_X = 0$, c^* is (4)'s limit for $\rho_Y \downarrow 0$ and $\rho_X \downarrow 0$, respectively, while for $\rho_X = \rho_Y = 0$ one obtains $c^* = 0$ (corresponding to 'traditional' preferences). Note that preference parameter ε has no effect in the ultimatum game.

Offer c^* can result from two distinct motives, reflected by the two terms in (4). The first term depends only on the responder's reciprocal inclination ρ_Y and is the

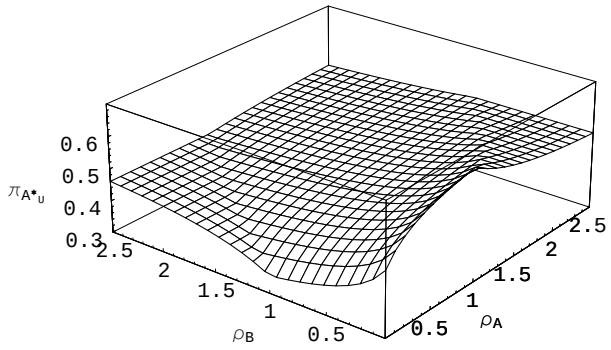

FIG. 3 Expected payoff $\pi^*_{A_U}$ of agent A with reciprocity parameter ρ_A in the ultimatum game when matched against an agent B with reciprocity parameter ρ_B

maximizer of the proposer's expected utility in case that the responder's concern for reciprocity is 'binding' – dominating a relatively weak reciprocity concern of the proposer. The associated share of the pie is just high enough to ensure acceptance ($p = 1$). The second expression depends only on the proposer's ρ_X and reflects how much the proposer would *voluntarily* offer in order to maximize utility in the light of his or her intrinsic concern for a fair outcome. As either ρ_X or ρ_Y grow large, c^* approaches 1/2 (see figure 2).

The equilibrium offer is the maximum of both expressions. This means that if a selfish proposer plays against a reciprocal responder, the offer is increasing in ρ_Y. If the responder's concern for reciprocity is low, the offer depends on the fairness concern of the proposer, hence it is increasing in ρ_X. If both players are selfish, the offer is close to zero.

Agent A's role in interaction with agent B is random, with probability 0.5 for each. The fitness and expected material payoff $\pi^*_{A_U}$ of agent A with preference parameters (ρ_A, ε_A) who is matched with agent B to play the ultimatum game is the average of the payoffs he or she receives in both roles. This turns out to be:

$$\pi^*_{A_U}(\rho_A,\rho_B) = \begin{cases} \frac{5\rho_A+3-\sqrt{1+6\rho_A+\rho_A^2}}{8\rho_A} & \text{if} \quad \rho_A \geq \frac{-2\rho_B}{\rho_B+1-\sqrt{1+6\rho_B+\rho_B^2}} \quad \text{(a)} \\ \frac{3\rho_B-3+\sqrt{1+6\rho_B+\rho_B^2}}{8\rho_B} & \text{if} \quad \rho_A \leq \frac{\rho_B(\rho_B-1)}{1+\rho_B} \quad \text{(b)} \\ \frac{3\rho_A+1-\sqrt{1+6\rho_A+\rho_A^2}}{8\rho_A} & \\ \quad + \frac{\rho_B-1+\sqrt{1+6\rho_B+\rho_B^2}}{8\rho_B} & \text{if} \quad \frac{-2\rho_B}{\rho_B+1-\sqrt{1+6\rho_B+\rho_B^2}} > \rho_A > \frac{\rho_B(\rho_B-1)}{1+\rho_B}. \quad \text{(c)} \end{cases} \tag{5}$$

Figure 3 illustrates this. Expected payoff depends on both agents' concerns for reciprocity, captured by ρ_A and ρ_B.

There are three cases. In case (a), ρ_A is large compared to ρ_B and thus behavior in *both* possible role realizations is determined by agent A's strong concern for fairness. Agent A's payoff is strictly decreasing in ρ_A, i. e. whenever (a) applies, agents with minimal reciprocity concern (ρ_A at the boundary to case (c)) are fittest.

Case (b) applies if ρ_A is small compared to ρ_B. Then, behavior in both roles is determined by ρ_B. Agent A's payoff is constant in ρ_A, i. e. evolutionary pressures only apply to the agent with greater concern for reciprocity (the smaller ρ_B, the fitter).

In the intermediate case (c), ρ_A and ρ_B do not differ too much. This is the only relevant case for NSS-based stability analysis since it concerns *symmetric* Nash equilibrium profiles. The proposer offers a split that maximizes his or her own material payoff subject to the binding constraint imposed by the rejection behavior of the responder, i. e. c^* depends only on the responder's reciprocity parameter. It follows from (5) that agent A's (B's) payoff is strictly increasing in ρ_A (ρ_B). Restricting attention to parameter ρ, the unique symmetric Nash equilibrium of preference game $\hat{\Gamma}$ is thus (ρ^*, ρ^*) with $\rho^* = \bar{P}$, i. e. the highest possible level (see Sec. 2.3).[16] Replicator dynamics – and, in fact, all weakly payoff positive selection dynamics – started from a fully mixed population state must therefore bring about ρ^* if they converge (Weibull 1995, Prop. 4.11(c)).

Since agents' pure concern for an equitable outcome, ε, has no effect on payoffs we can conclude that all preferences

$$\hat{s} \in \hat{T}_U \equiv \{(\bar{P}, \varepsilon) \in \hat{S} \colon \varepsilon \in [0, 1]\}$$

are stable in the ultimatum game played in isolation. This corresponds to an approximately *equal split* (exactly equal for $\bar{P} \to \infty$), which is offered for *strategic reasons*, i. e. to prevent the responder from rejecting and not because of the proposer's intrinsic motivation. This finding is broadly consistent with various other evolutionary and behavioral investigations of ultimatum bargaining: Huck and Oechssler (1999) consider indirect evolution in a 2×2-version of the ultimatum game and find that a preference for punishing unfair proposers survives and induces equitable offers (even under anonymous interaction). Direct evolutionary studies of the ultimatum game include Gale et al. (1995), Nowak et al. (2000), and Nowak and Page (2002). They respectively demonstrate the persistence of fair offers given relatively more noise in the responder than the proposer population (or none at all), when proposers have access to observations of past responder behavior, and when at least some fraction of agents are empathetic in the sense of always offering at least what they themselves would accept. The aspiration-based satisficing model

[16] It is even strict and thus corresponds to an ESS. – Preferences with $\rho = \infty$ would correspond to maximization of the so-called reciprocity payoff regardless of own material payoff (cf. (13) in the Appendix).

investigated by Napel (2003) entails fair ultimatum offers linked to parameters that reflect player characteristics such as stubbornness and capriciousness. Reinforcement learning models proposed e. g. by Roth and Erev (1995) also predict rather equal offers coupled with rejection of unequal ones, at least in the intermediate run.

3.2. The Dictator Game

The split c^* offered by the proposer X in the dictator game in equilibrium depends only on his or her *own* reciprocity parameter ρ_X and pure outcome-concern parameter ε_X:

$$c^* = \max\left\{0, \frac{1}{2} \cdot (1 - \frac{1}{\varepsilon_X \rho_X})\right\}. \tag{6}$$

A proposer offers a positive amount if $\varepsilon_X \rho_X > 1$, which means that he or she is reasonably concerned about reciprocity in the unintentional case. No matter how large $\varepsilon_X \rho_X$ is, the offer is never greater than half. Equation (6) corresponds to the second term in (4), i. e. the possible 'voluntary offer' in the ultimatum game, where $\varepsilon_X \rho_X$ replaces ρ_X. Here, the receiver has no choice but to accept and therefore the outcome is 'unintentional' in the sense of Falk and Fischbacher (2006); this is reflected by a 'discounting' of reciprocity parameter ρ_X by pure equity concern parameter ε_X in the dictator game. A given agent always offers weakly *less* in the dictator game than in the ultimatum game, because $0 \leq \varepsilon_X \leq 1$.

Agent A is assigned the proposer role in a dictator game with agent B with probability 0.5. His or her expected material payoff $\pi^*_{A_D}$ in equilibrium is then given by:

$$\pi^*_{A_D}(\rho_A, \varepsilon_A, \rho_B, \varepsilon_B) = \begin{cases} \frac{1}{2} & \text{if} \quad \varepsilon_A \rho_A \leq 1 \text{ and } \varepsilon_B \rho_B \leq 1 \quad \text{(I)} \\ \frac{1}{4} + \frac{1}{4\varepsilon_A \rho_A} & \text{if} \quad \varepsilon_A \rho_A > 1 \text{ and } \varepsilon_B \rho_B \leq 1 \quad \text{(II)} \\ \frac{3}{4} - \frac{1}{4\varepsilon_B \rho_B} & \text{if} \quad \varepsilon_A \rho_A \leq 1 \text{ and } \varepsilon_B \rho_B > 1 \quad \text{(III)} \\ \frac{1}{2} + \frac{1}{4\varepsilon_A \rho_A} - \frac{1}{4\varepsilon_B \rho_B} & \text{if} \quad \varepsilon_A \rho_A > 1 \text{ and } \varepsilon_B \rho_B > 1. \quad \text{(IV)} \end{cases} \tag{7}$$

We have four cases, illustrated by figure 4. In cases (II) and (III) either of the two agents offers a positive amount to the other agent when he or she is in the role of proposer. The expected payoff for such 'generous' behavior is smaller than in case (I), and strictly decreasing in the agent's parameters ρ and ε. Cases (II) and (III) are hence unstable.

In case (IV) *both* agents as proposers share the pie with the receiver; their payoff need not be smaller than in case (I). Still, material payoff is decreasing in each agent's individual parameters ρ and ε. The smaller they are, the fitter is the

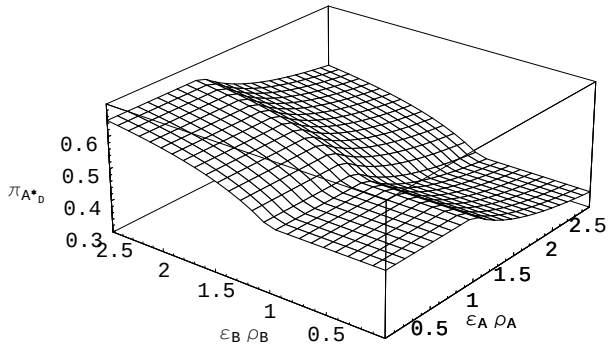

FIG. 4 Expected payoff $\pi^*_{A_D}$ of agent A with parameters ε_A and ρ_A in the dictator game when matched against an agent B with parameters ε_B and ρ_B

agent carrying these preferences. In particular, no preference profile corresponding to (IV) forms a Nash equilibrium (let alone NSS) of the preference game.

For parameter constellations pertaining to case (I) both agents are only weakly interested in fairness; both offer $c^* = 0$ as proposers. The case applies as long as the product of the parameters for reciprocity ρ and the pure equity-concern ε is smaller or equal to unity for both agents. All such parameter constellations are behaviorally equivalent. It follows from (7) that any

$$\hat{s} \in \hat{T}_D \equiv \{(\rho, \varepsilon) \in \hat{S} \colon \varepsilon\rho \leq 1\}$$

is a NSS of preference game $\hat{\Gamma}$, and there exist no other stable preferences.[17] Moreover, $\hat{T}_D$ comprises all Nash equilibria of $\hat{\Gamma}$. So, if any given payoff monotone selection dynamic converges from an interior point to some $\hat{s}^* \in \hat{S}$, then $\hat{s}^* \in \hat{T}_D$.

We can thus expect agents to offer zero (i. e. to be indistinguishable from selfish payoff maximizers) in the dictator game in the long run. They may have a preference for fairness to a degree that does not affect behavior.

3.3. The Mixed Environment

Now consider the stylized 'game of life' in which agents randomly get to interact in either an ultimatum or a dictator game setting. The payoff π^*_A of agent A facing agent B given the probability $\lambda \in [0, 1]$ to play the ultimatum game and probability

[17] In fact, $\hat{T}_D$ forms an *evolutionary stable set* (ES set): every $\hat{s} \in \hat{T}_D$ is locally superior to preference strategies outside $\hat{T}_D$ and not inferior to others inside $\hat{T}_D$; this implies asymptotic stability. The same holds for $\hat{T}_U$ in the ultimatum game. See, e. g., Weibull (1995, sec. 3.5.4) for details.

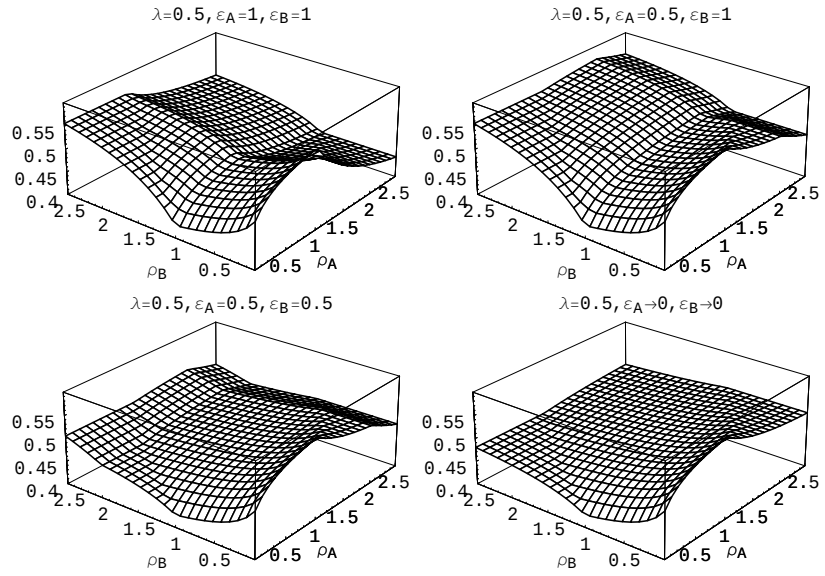

FIG. 5 Expected payoff π_A^* of agent A with parameters ε_A and ρ_A in the 'game of life' when matched against an agent B with parameters ε_B and ρ_B for $\lambda = 0.5$

$1-\lambda$ to play the dictator game is

$$\pi_A^*(\rho_A, \varepsilon_A, \rho_B, \varepsilon_B, \lambda) = \lambda \cdot \pi_{A_U}^*(\rho_A, \rho_B) + (1-\lambda) \cdot \pi_{A_D}^*(\rho_A, \varepsilon_A, \rho_B, \varepsilon_B). \qquad (8)$$

This expected material payoff function is illustrated for $\lambda = 1/2$ and several values of ε_A and ε_B in figure 5.[18] Total expected payoff is increasing in payoff $\pi_{A_U}^*(\cdot)$ from the ultimatum game and the payoff $\pi_{A_D}^*(\cdot)$ from the dictator game. Hence evolution would favor preferences that maximize $\pi_{A_U}^*(\cdot)$ and $\pi_{A_D}^*(\cdot)$ at the same time if this were possible.

Indeed, this is possible for the Falk-Fischbacher preferences considered here: The maximum payoff in the dictator game is reached for *any* parameter constellation with $\rho \cdot \varepsilon \leq 1$. The pure equity concern parameter ε can always totally compensate a high reciprocity concern ρ by being less or equal to $1/\rho$. So, the 'optimal', material payoff-maximizing preferences in the dictator game can be reached for any level of ρ. The stable preferences are hence all $\hat{s} \in T_U \cap T_D = \{(\bar{P}, \varepsilon) \in \hat{S} \colon \varepsilon \leq 1/\bar{P}\}$, which also correspond to the Nash equilibria of preference game $\hat{\Gamma}$ if $\lambda \in (0,1)$.

This implies that agents behave very fairly whenever their counterpart can reciprocate, but show no pure concern for equity. The evolutionary prediction is thus: approximately equal splits in the ultimatum game and full appropriation of the pie

[18] Note the saddle points (1,1) and (2,2) in the cases of $\varepsilon_A = \varepsilon_B = 1$ and $\varepsilon_A = \varepsilon_B = 0.5$, respectively. A saddle point – corresponding to an ESS *if* ε were exogenously fixed (or is restricted as in section 3.4) – does not exist for $\varepsilon_A = \varepsilon_B \approx 0$.

in the dictator game, i. e. a superimposition of the stable outcomes in the two components of our stylized 'game of life'. With large ρ and small ε agents reciprocate strongly in situations that imply intentionality but do not care about fairness when it comes to situations where intentions play no role. So from our evolutionary model's point of view, fair behavior should really be a matter of intention-based reciprocity rather than a general concern for equitable payoff distributions.

3.4. Restricted Parameters in the Mixed Environment

With two free parameters and two games it may not be surprising (though it is not trivial either) that evolution brings about 'optimal' behavior for isolated dictator and ultimatum game environments also in the combined habitat: Our agents are unable to have different sets of parameters for reciprocity and pure equity concern in ultimatum and dictator game, respectively, i. e. they cannot discriminate directly between the games. However, they succeed to *indirectly* discriminate based on distinctive features of the games (here: intentionality).

It can be questioned whether the degrees of freedom in the social component of our preferences in reality match the number of different classes of social interaction.[19] Humans seem to adjust behavior to a specific situation only to some extent. It therefore seems interesting and reasonable to limit nature's freedom in shaping agents' preferences by restricting the possible range of the pure outcome-concern parameter ε which applies in unintentional situations.

One practical possibility is to impose an *exogenous lower bound* $\varepsilon_l > 1/\bar{P}$ (chosen as a multiple of grid size $1/n$) for parameter ε, so that $\varepsilon_l \leq \varepsilon \leq 1$ is required instead of $0 \leq \varepsilon \leq 1$. This situation differs from the above in that a strong concern for reciprocity cannot completely be blanked out in the dictator game by a low ε. So even accounting for indirect discrimination possibilities, an agent's behavior in dictator realizations of the 'game of life' is thus no longer independent from that in ultimatum realizations.

Whenever $\rho \geq 1/\varepsilon_l$, the dictator game induces downward pressure on parameter ε: the smaller ε, the less is given away in the dictator role and hence the greater is own payoff. At the same time, the strategic reaction by ultimatum proposers to responders' reciprocity concern puts persistent upward pressure on parameter ρ. This has the following environment-dependent implications for stable preferences.

For λ not too big, the dictator game is important enough in agents' lives to make 'play' of $(1/\varepsilon_l, \varepsilon_l)$ the unique (strict) Nash equilibrium of preference game $\hat{\Gamma}$, which corresponds to an ESS. A payoff positive selection dynamic started at a fully mixed population state can only converge to $(1/\varepsilon_l, \varepsilon_l)$. This results in a zero offer

[19] Biological costs of discrimination by a given agent's (global) preferences would play the central role in a theoretical investigation of this relationship.

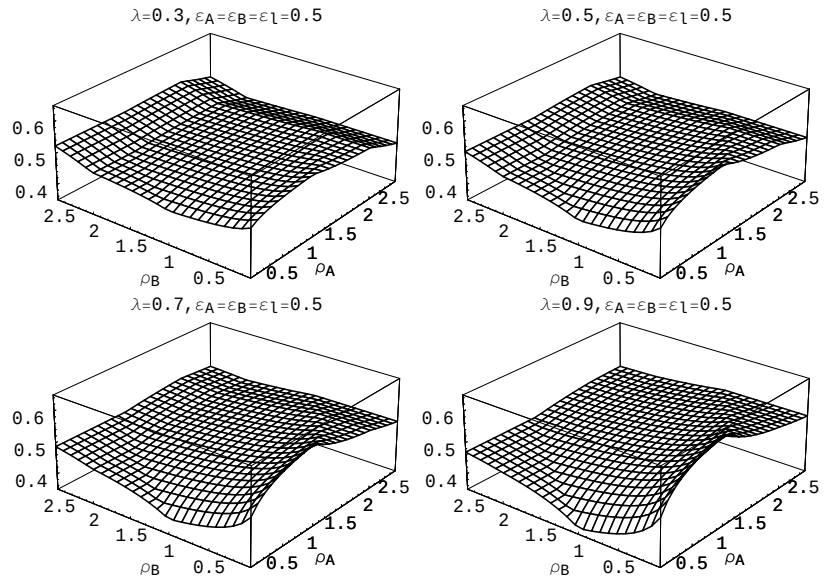

FIG. 6 Expected payoff π_A^* of agent A with parameters $\varepsilon_A = \varepsilon_l = 0.5$ and ρ_A in the 'game of life' when matched against an agent B with parameters $\varepsilon_B = \varepsilon_l = 0.5$ and ρ_B for several values of λ

in the dictator game and

$$c^* = \frac{1 + 3\frac{1}{\varepsilon_l} - \sqrt{1 + 6\frac{1}{\varepsilon_l} + \frac{1}{\varepsilon_l^2}}}{4\frac{1}{\varepsilon_l}} \tag{9}$$

in the ultimatum game. For example, a value of $\varepsilon_l = 0.25$ implies a split of $(\pi_X, \pi_Y) \approx (0.59, 0.41)$ or $\varepsilon_l = 0.125$ implies $(\pi_X, \pi_Y) \approx (0.55, 0.45)$ in the ultimatum game in the long run. The imposed restriction thus slightly reduces equitableness in ultimatum offers without consequences for selfish behavior in the dictator game.

The influence of parameter λ on the payoffs is illustrated by figure 6. The saddle point at $(\rho_A, \rho_B) = (2, 2)$, corresponding to stable preferences with $\rho^* = 1/\varepsilon_l$, is barely visible for $\lambda = 0.7$ and no longer exists for $\lambda = 0.9$. In fact, there is a *critical level* of λ above which $(1/\varepsilon_l, \varepsilon_l)$ is no longer stable, and instead $(\bar{P}, \varepsilon_l)$ becomes the evolutionary prediction. This critical level can be calculated by analyzing expected payoff for values of $\rho > 1/\varepsilon_l$; if it is lower (higher) than for $\rho = 1/\varepsilon_l$ the latter combination is (is not) stable. Constellations $\rho > 1/\varepsilon_l$ belong to case (c) in the ultimatum game and case (IV) in the dictator game (involving positive offers by

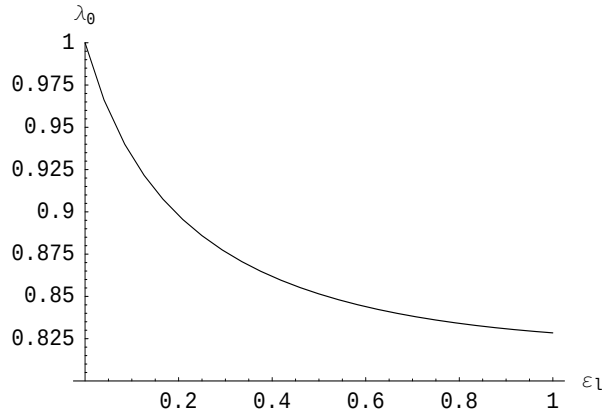

FIG. 7 The critical value $\lambda_0(\varepsilon_l)$ at which $\rho = 1/\varepsilon_l$ ceases to be an evolutionary stable point

both agents). The material payoff of agent A in the 'game of life' is therefore

$$\begin{aligned}\pi_A^*(\rho_A,\varepsilon_A,\rho_B,\varepsilon_B,\lambda) &= \lambda\left[\frac{3\rho_A+1-\sqrt{1+6\rho_A+\rho_A^2}}{8\rho_A}+\frac{\rho_B-1+\sqrt{1+6\rho_B+\rho_B^2}}{8\rho_B}\right]\\ &\quad+(1-\lambda)\left[\frac{1}{2}+\frac{1}{4\varepsilon_A\rho_A}-\frac{1}{4\varepsilon_B\rho_B}\right]. \qquad (10)\end{aligned}$$

The marginal change of agent A's payoff from greater ρ_A is

$$\frac{\partial\pi_A^*(\rho_A,\varepsilon_A,\rho_B,\varepsilon_B,\lambda)}{\partial\rho_A}=\frac{\varepsilon_A\lambda(1+3\rho_A)+(2\lambda-2-\rho_A\lambda)\sqrt{1+6\rho_A+\rho_A^2}}{8\varepsilon_A\rho_A\sqrt{1+6\rho_A+\rho_A^2}} \qquad (11)$$

and can be analyzed for $\rho_A > 1/\varepsilon_l$ and $\varepsilon_A = \varepsilon_l$ in order to determine at which critical level λ_0 the combination of selfish dictator offers and (moderately) fair ultimatum offers starts to yield smaller payoff than $(\bar{P},\varepsilon_l)$. The marginal effect of a ρ_A-increase on A's payoff is negative for values of λ up to the critical level

$$\lambda_0(\varepsilon_l)=\frac{2\sqrt{1+\frac{6}{\varepsilon_l}+\frac{1}{\varepsilon_l^2}}}{3+\varepsilon_l+(2-\varepsilon_l)\sqrt{1+\frac{6}{\varepsilon_l}+\frac{1}{\varepsilon_l^2}}} \qquad (12)$$

and then positive. So for $\lambda > \lambda_0(\varepsilon_l)$, $(\bar{P},\varepsilon_l)$ becomes the unique stable preference.[20] As illustrated in figure 7, $\lambda_0(\varepsilon_l)$ is strictly decreasing in the exogenous lower bound on pure equity concern ε_l.[21]

So when the ultimatum game is played very frequently compared to the dictator game, stable preferences involve reciprocity parameter $\rho^* = \bar{P}$ and pure outcome concern $\varepsilon^* = \varepsilon_l$. This implies approximately equal splits in *both* games (becoming exactly equal for $\bar{P} \to \infty$). Loosely speaking, if agents interact in comparatively many 'private' situations captured by the ultimatum game compared to few

[20] For $\lambda = \lambda_0(\varepsilon_l)$ all preferences $\hat{s} \in \{(\rho,\varepsilon_l) \in \hat{S}\colon 1/\varepsilon_l \le \rho \le \bar{P}\}$ are NSS.

[21] It falls from $\lambda_0 = 1$ for $\varepsilon_l \to 0$ to a minimum of $\lambda_0 = 2\sqrt{2}-2 \approx 0.838$ for $\varepsilon_l = 1$.

'market' situations reflected by the dictator game, they develop a strong notion of fairness which involves equity concern but is primarily driven by preference for reciprocity.

4. COMPARISON WITH GÜTH AND NAPEL (2006)

Recall that the crucial difference between this model and that of Güth and Napel (2006) is the considered class of preferences. Agents' sense of fairness is limited to inequity aversion in the paper of Güth and Napel, while we allow intention-based reciprocity.

Results of the separate analysis of ultimatum and dictator game are qualitatively the same. In the *ultimatum game* agents in the role of the responder benefit from the Fehr-Schmidt type of inequity aversion as well as from reciprocation, because the proposer anticipates that small offers are rejected with positive probability. In the long run a rather equitable or even equal split is reached (this depends on the precise utility specification).

In the *dictator game*, in contrast, both inequity aversion and reciprocal behavior are detrimental to an agent's average success, as it leads him or her to voluntarily give part of the pie away. In Güth and Napel's investigation as well as this one, agents can be inequity averse to any degree that fails to actually affect proposer behavior. In the analysis by Güth and Napel, this upper bound to stable inequity aversion of dictators is driving results for the mixed environment. Either the long-run level of inequity aversion in the ultimatum game is below this bound, resulting in no behaviorally relevant interaction between the two games, or it is above. In the latter case, the given composition of the stylized 'game of life' determines a stable level that, loosely speaking, balances marginal evolutionary benefits and costs of inequity aversion. In our analysis the situation is complicated by the presence of two free parameters instead of one. The original model of Falk and Fischbacher (2006) allows for both: Strong reciprocation and no inequity aversion in unintentional cases at all. If the possible discrepancy between these two aspects of fairness is not restricted exogenously, agents behave in both games just as if they were independent. This resembles Güth and Napel's case in which agents can directly condition the social component of their utility function – and hence fair behavior – on the game at hand. Here, such a moral discrimination between games is achieved *indirectly*. However, when the 'discounting' of fairness in the *unintentional* case is limited, our long-run outcome resembles the one of Güth and Napel's case of game-independent equity aversion.

It is a common result of either analysis that a high share of ultimatum games affects evolutionary stable preferences. Coincidentally, the minimal level at which the frequency of ultimatum games affects the outcome is similar in both studies, about

80% in Güth and Napel (2006) and around 85% for a wide range of constellations here. So, on the one hand, the explicit consideration of intention-based reciprocity broadly speaking produces results similar to the analytically much simpler case of purely payoff-oriented inequity aversion. On the other hand, we find it interesting that in a hybrid setting the former really is the dominant force.

5. CONCLUDING REMARKS

In general, psychological fairness models explain experimental results better than more tractable equity-based models, which in turn perform better than the default assumption of monetary payoff maximization. This success comes at the expense of more free parameters, which have to be defended against accusations of ad hoc 'game fitting'. Since parameters *can* be such that an agent, in fact, has traditional materialistic preferences, indirect evolutionary analysis provides a useful theoretical benchmark. If under plausible modelling assumptions evolutionary forces select parameters that imply non-degenerate social preferences, this corroborates post-experimental econometric estimations.

The relevance of fairness preference can be expected a priori to vary with the decision situations at hand even if these are restricted to simple distribution tasks. Our analysis suggests that preference for fair behavior has sound evolutionary reasons, but – in line with experimental observations – more likely plays a significant role in games with punishment opportunities such as the ultimatum game than in what is basically a one-player decision problem like the dictator game. If Nature permits players to condition the social component of their preferences on different games (as investigated by Güth and Napel 2006) or at least indirectly on differences in games, the same agents can exhibit a pronounced sense of fairness in one type of social interaction while they are entirely selfish in another one. This is true even though preferences are assumed to evolve simultaneously in a multi-game environment.

If Nature imposes physical or psychological restrictions on the variance in agents' social attitudes across different games and different dimensions of fairness, the precise composition of the stylized 'game of life' faced by agents has an impact. When agents most of the time face the more fairness-conducive ultimatum game, they will eventually perceive it in their interest to be generous even as dictators.

We have studied environmental and psychological determinants of material-payoff maximization vs. social preferences in a particularly simple two-game environment. In our view, this is an improvement compared to the usual analysis of preferences in a single game. It highlights the importance of possibly unconscious links between behavioral modes in different classes of interaction. Whether human agents indeed face a binding restriction in their ability to discriminate between the

fairness implications of similar play in ultimatum and dictator games, reflected by the reciprocity and pure equity concern parameters ρ and ε in our model, is an empirical question. Experimental evidence on the ultimatum game is consistent with both the unrestricted and restricted evolution of ρ and ε; evidence on the dictator game supports the restricted view (although not the equal splits that would be implied for $\bar{P} \to \infty$).[22]

Finally, it seems desirable to us to extend our stylized 'game of life' to more than only ultimatum and dictator games. First, an environment with more than two component games would create a natural endogenous restriction on the two parameters of the considered Falk-Fischbacher reciprocal preferences – preventing the full 'specialization' observed in section 3.3. Second, reciprocal preferences are proposed as explanations for empirical observations of a considerable range of games, including e. g. trust and gift-exchange games, public good games, and variants of the ultimatum game such as the best-shot game. They should hence prove to be evolutionary stable in an environment including at least these games.

APPENDIX

Falk-Fischbacher Preferences

Falk and Fischbacher (2006) use the framework of psychological games (Geanakoplos et al. 1989) to model reciprocity. They consider extensive games of complete and perfect information with a finite set of decision nodes $\mathcal{N}$ and terminal nodes $\mathcal{T}$. A *reciprocity payoff* is added to the *material payoff* of the players. Total utility of player $i \in \{X, Y\}$ in a terminal node $t \in \mathcal{T}$ is

$$U_i(t) \equiv \pi_i(t) + \rho_i \sum_{\substack{n \in \mathcal{N}_i \\ n \to t}} \varphi_j(n) \cdot \sigma_i(n, t, b_j, c_i) \tag{13}$$

where $\pi_i(t)$ denotes the material payoff; the reciprocity payoff is accumulated over all of i's decision nodes $n \in \mathcal{N}_i \subset \mathcal{N}$ on the path to t and weighted with parameter ρ_i. It is based on player i's belief $b_j \in S_j$ about the strategy choice a_j of j, and player i's second-order belief $c_i \in S_i$ about what he believes player j believes he is choosing (i. e., c_i is player i's belief about b_i).

The *kindness term*

$$\varphi_j(n) \equiv \vartheta_j(n) \Delta_j(n).$$

reflects the kindness player i experiences from player j's expected actions at n. It is positive (negative) if player j is considered as kind (unkind). Its two determinants crucially depend on comparability of the levels of players' monetary payoffs, $\pi_i(\cdot)$

[22] Dictator offers in Güth and Napel (2006) continuously rise from zero to moderate positive levels (assuming that marginal disutility of inequality is increasing).

and $\pi_j(\cdot)$.[23] The factor

$$\Delta_j(n) \equiv \pi_i(n, b_j, c_i) - \pi_j(n, b_j, c_i)$$

is positive if player i expects to experience a higher payoff than his opponent. Abbreviating $(\pi_i^0, \pi_j^0) \equiv (\pi_i(n, b_j, c_i), \pi_j(n, b_j, c_i))$, the *intention factor*

$$\vartheta_j(n) \equiv \max\{\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) | (\tilde{\pi}_i, \tilde{\pi}_j) \in \Pi_i(n)\}.$$

tries to measure the intentionality player i attributes to player j's anticipated actions and is calculated in two steps:

First, (π_i^0, π_j^0) is compared with all payoff alternatives $(\tilde{\pi}_i, \tilde{\pi}_j)$ to (π_i^0, π_j^0) – collected in the set $\Pi_i(n)$ – which player j could have chosen. Every such comparison is summarized by a real number $\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) \in [0, 1]$, where a value of 1 signifies full intentionality:

$$\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) \equiv \begin{cases} 1 & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i < \pi_i^0 & \text{(a)} \\ \varepsilon_i & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i \geq \pi_i^0 & \text{(b)} \\ 1 & \text{if } \pi_i^0 < \pi_j^0, \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i \leq \tilde{\pi}_j & \text{(c)} \\ \max(1 - \frac{\tilde{\pi}_i - \tilde{\pi}_j}{\pi_j^0 - \pi_i^0}, \varepsilon_i) & \text{if } \pi_i^0 < \pi_j^0, \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i > \tilde{\pi}_j & \text{(d)} \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \tilde{\pi}_i < \pi_i^0. & \text{(f)} \end{cases}$$

For example, in both cases (a) and (b) player i receives a higher payoff than player j, but only when j could actually have left him with smaller payoff as in (a), this is interpreted as fully (here: well) intentioned. In case (b), player j's action is not perceived as particularly generous and the intention factor is only $\Omega = \varepsilon_i$, corresponding to an individual *pure outcome-concern parameter* $0 \leq \varepsilon_i \leq 1$.[24] In case (d), player i is worse off than j but j could have improved player i's situation only by becoming the worse-off player herself. Player i's evaluation of j's kindness then depends on the rate of transformation between both players' payoffs.

Second, the maximum of all these comparison values is taken to be the *overall intentionality* player i associates with (π_i^0, π_j^0). So if player j has at least one alternative, where she could give more to player i without becoming the worse-off player herself or less whilst staying disadvantaged, then (π_i^0, π_j^0) is considered as fully intentional.

[23] Equity considerations based on an evaluation of material payoffs according to, for example, agent-specific need (related, e. g., to different wealth levels and other asymmetries) are thus not compatible with the model.

[24] One can model behavior which is purely intention-driven (as in Rabin 1993 and Dufwenberg and Kirchsteiger 2004) by $\varepsilon_i = 0$, or purely outcome-oriented (as in Fehr and Schmidt 1999 and Bolton and Ockenfels 2000) by $\varepsilon_i = 1$.

The remaining component of (13) is the *reciprocation term*

$$\sigma_i(n,t,b_j,c_i) \equiv \pi_j(s(n,t),c_i,b_j) - \pi_j(n,c_i,b_j).$$

with $s(n,t)$ denoting n's successor on the path to t. The first (second) summand is player j's expected payoff after (before) player i's move. $\sigma_i(\cdot)$ captures the *alteration* of player j's *expected payoff* implied by player i's move towards t.

Equilibrium in the Ultimatum Game

Falk and Fischbacher (2006) call a subgame perfect psychological Nash equilibrium a *reciprocity equilibrium*. For a proof of existence in the considered class of games and more details on the following equilibrium derivations see their paper. To confirm that equations (3) and (4) in section 3.1 describe the unique reciprocity equilibrium of the ultimatum game, first, let p' denote the proposer's belief about acceptance probability p and let p'' denote the responder's belief about p'. Let $\vartheta_X(c)$ be the intentionality factor at the decision node after player X's choice of c. The responder's utility is

$$U_{Y_A} = c + \rho_Y \cdot \vartheta_X(c)\, p'' \,[c-(1-c)] \cdot [(1-c) - p''(1-c)]$$

in case she accepts the offer and

$$U_{Y_R} = \rho_Y \cdot \vartheta_X(c)\, p'' \,[c-(1-c))] \cdot [0 - p''(1-c)]$$

if she rejects. The former is greater for $c \geq \frac{1}{2}$, implying acceptance. For $c < \frac{1}{2}$ define (by setting $U_{Y_A} = U_{Y_R}$)

$$p''_{crit} = \frac{c}{\rho_Y \vartheta_X(c)(1-2c)(1-c)}. \tag{14}$$

Note that $p'' > p''_{crit}$ would ask for $p = p'' = 0$ in contradiction to $p''_{crit} \geq 0$ (for $c < \frac{1}{2}$). Hence either $p'' < p''_{crit}$ so that optimal responder behavior in equilibrium (involving consistent beliefs) requires $p = p'' = 1$, or we have $p'' = p''_{crit}$ so that $p = p'' = p''_{crit}$ is optimal for consistent beliefs. Optimal responder behavior can thus be summarized by $p^*(c) = \min\{1, p''_{crit}\}$. If $c < \frac{1}{2}$, then player Y is disadvantaged and player X's move is considered as fully intentional because $c = \frac{1}{2}$ would lead to a higher payoff for player Y without making X the worse-off player. Therefore, $\vartheta_X(c) = 1$ in (14).

The expected utility of a proposer with a correct belief p' determined by $p^*(\cdot)$ is

$$U_X = p^*(c) \cdot (1-c) + \rho_X \cdot \vartheta_Y(c)\, p^*(c'')\,(1-2c'') \cdot [p^*(c)c - p^*(c'')c''] \tag{15}$$

where c'' denotes player X's second-order beliefs.[25] In equilibrium we must have $c \leq \frac{1}{2}$ because U_X is decreasing in c for $c'' \geq \frac{1}{2}$. For $c > 0$ the responder's move $p^*(\cdot)$ is fully intentional as she has the option of rejecting the offer, which leads to a smaller payoff for the proposer; so $\vartheta_Y = 1$.

Now define $c_0 = \frac{1+3\rho_Y - \sqrt{1+6\rho_Y+\rho_Y^2}}{4\rho_Y}$ as the smallest c such that $p^*(c) = 1$. U_X is increasing in c for $c < c_0$; so a rational proposer must choose $c^* \geq c_0$, implying acceptance with probability 1. Setting $p^* = 1$ and $c \geq c_0$ in (15), one obtains

$$U_X = (1 - c) + \rho_X \cdot (1 - 2c'') \cdot (c - c'')$$

and

$$\frac{\partial U_X}{\partial c} = -1 + \rho_X \cdot (1 - 2c'').$$

So U_X is decreasing in c for

$$c'' > c''_{crit} = \frac{1}{2}\left(1 - \frac{1}{\rho_X}\right)$$

and increasing for $c'' < c''_{crit}$. First, consider $c''_{crit} < c_0$: Since in equilibrium $c = c''$, we get $c''_{crit} < c_0 \leq c = c''$ and U_X is decreasing in c. Then, the optimal proposal is $c^* = c_0 (= \max(c_0, c''_{crit}))$. Second, consider $c''_{crit} \geq c_0$: If $c'' > c''_{crit}$, U_X is decreasing in c and therefore c would have to be chosen equal to c_0 which is, however, incompatible with $c = c''$ because $c'' > c''_{crit} \geq c_0 = c$. If $c'' < c''_{crit}$, U_X is increasing in c and therefore c is chosen equal to 1 which is also incompatible with $c = c''$ because $c'' < c''_{crit} < \frac{1}{2} < 1 = c$. Therefore $c^* = c'' = c''_{crit} (= \max(c_0, c''_{crit}))$.

Equilibrium in the Dictator Game

The key difference to the ultimatum game is that acceptance is not intentional in the dictator game. So the intentionality factor at the proposer's decision node equals ε_X. Then

$$U_X = (1 - c) + \rho_X \varepsilon_X (1 - c'' - c'')c$$

and the first order condition yields $c''_{crit} = \frac{1}{2}\left(1 - \frac{1}{\rho_X \varepsilon_X}\right)$.

REFERENCES

Benaïm, M. and J. W. Weibull (2003). Deterministic approximation of stochastic evolution in games. *Econometrica 71*(3), 873–903.

Berninghaus, S., K.-M. Ehrhart, and M. Ott (2006). A network experiment in continuous time: The influence of link costs. *Experimental Economics 9*(3), 237–251.

[25] One can think of c'' as an offer that player X conjectures Y's response to be based on in order to evaluate her kindness. This kindness renders a particular reciprocation and corresponding actual offer c optimal, which must coincide with c'' in equilibrium.

Berninghaus, S. K., W. Güth, and H. Kliemt (2003). From teleology to evolution – Bridging the gap between rationality and adaptation in social explanation. *Journal of Evolutionary Economics 13*(4), 385–410.

Bolton, G. E. and A. Ockenfels (2000). ERC – A theory of equity, reciprocity and competition. *American Economic Review 90*(1), 166–193.

Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry 37*(1), 47–59.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*, 817–869.

Dekel, E., J. C. Ely, and O. Yilankaya (2005). Evolution of preferences. Mimeo, Northwestern University and University of British Columbia. [http://www.faculty.econ.northwestern.edu/faculty/ely/evlprf.pdf].

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games and Economic Behavior 47*, 268–298.

Falk, A., E. Fehr, and U. Fischbacher (2003). On the nature of fair behavior. *Economic Inquiry 41*(1), 20–26.

Falk, A. and U. Fischbacher (2006). A theory of reciprocity. *Games and Economic Behavior 54*(2), 293–315.

Falk, A. and M. Kosfeld (2003). It's all about connections: Evidence on network formation. IEW Working Paper 146, University of Zürich.

Fehr, E. and S. Gächter (1998). Reciprocity and economics. the economic implications of homo reciprocans. *European Economic Review 42*, 845–859.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*(3), 817–868.

Gale, J., K. G. Binmore, and L. Samuelson (1995). Learning to be imperfect: The ultimatum game. *Games and Economic Behavior 8*(1), 56–90.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological games and sequential rationality. *Games and Economic Behavior 1*, 60–79.

Güth, W., H. Kliemt, and S. Napel (2003). Wie du mir, so ich dir! – Evolutionäre Modellierungen. In M. Held, G. Kubon-Gilke, and R. Sturn (Eds.), *Jahrbuch Normative und Institutionelle Grundfragen der Ökonomik, Band 2: Experimentelle Ökonomik*, pp. 113–139. Marburg: Metropolis-Verlag.

Güth, W. and S. Napel (2006). Inequality aversion in a variety of games – An indirect evolutionary analysis. *Economic Journal (forthcoming)*.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review 91*(2), 73–78.

Huck, S. and J. Oechssler (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior 28*(1), 13–24.

Napel, S. (2003). Aspiration adaptation in the ultimatum minigame. *Games and Economic Behavior 43*(1), 86–106.

Nowak, M. A. and K. M. Page (2002). Empathy leads to fairness. *Bulletin of Mathematical Biology 64*, 1101–1116.

Nowak, M. A., K. M. Page, and K. Sigmund (2000). Fairness versus reason in the ultimatum game. *Science 289*(5485), 1773–1775.

Oechssler, J. and F. Riedel (2001). Evolutionary dynamics on infinite strategy spaces. *Economic Theory 17*, 141–162.

Oechssler, J. and F. Riedel (2002). On the dynamic foundation of evolutionary stability in continuous models. *Journal of Economic Theory 107*, 223–252.

Possajennikov, A. (2005). Cooperation and competition: Learning of strategies and evolution of preferences in prisoners' dilemma and hawk-dove games. *International Game Theory Review 7*, 443–459.

Poulsen, A. and O. Poulsen (2006). Endogenous preferences and social dilemma institutions. *Journal of Institutional and Theoretical Economics (forthcoming)*.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review 83*(5), 1281–1302.

Roth, A. E. and I. Erev (1995). Learning in extensive-form games: Experimental data and simple dynamic models of the intermediate term. *Games and Economic Behavior 8*(1), 164–212.

Samuelson, L. (2001). Introduction to the evolution of preferences. *Journal of Economic Theory 97*, 225–230.

Slonim, R. and A. E. Roth (1998). Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica 66*(3), 569–596.

Stahl, D. O. and E. Haruvy (2006). Level-n bounded rationality in two-player two-stage games. *Journal of Economic Behavior and Organization (forthcoming)*.

Weibull, J. W. (1995). *Evolutionary Game Theory*. Cambridge, MA: MIT Press.