Reciprocity, Materialism and Welfare: An Evolutionary Model

Anders Poulsen*
Department of Economics
The Aarhus School of Business
Fuglesangs Allé 20
8210 Aarhus V
Denmark
E-mail: aup@asb.dk

March 19, 2001

Abstract

This paper analyses preference evolution in a bargaining situation. We show that preferences for reciprocity, that sustain a conflict-free outcome, are viable if players have enough information about opponents' preferences. However, depending on the initial starting point, preference evolution can in general both enhance or reduce subjective and material welfare, relative to the situation where all players have the usual materialistic preferences.

Keywords: Preference evolution; reciprocity; altruism; materialism; subjective and material efficiency; bargaining; indirect evolutionary approach.

JEL Classification: C7

*I thank the seminar audience at Department of Economics, University of Aarhus and Gert Tinggaard Svendsen for helpful comments.

1 Introduction

In most economic models people maximise their personal material gains. While having served the economics profession well in many contexts, the assumption of such "materialistic" preferences is not always supported by experiments. There subjects often seem to be guided by concerns for "fairness" or "justice" and willing to punish opponents who violate such norms, even though they will never meet the opponent again and punishing means forgoing material gains.[1]

In this paper we allow for non-materialistic preferences and, using an evolutionary approach, we analyse what preferences are most likely to win the "struggle for survival". In particular, we wish to see whether the preferences that evolve are "good" for the players, both in material and subjective terms. We do this in the context of a simple bargaining situation, where two players are in dispute over a valuable resource and must simultaneously choose between being "moderate" or "aggressive". If both players are aggressive, there is a costly conflict, while mutual moderation avoids it. The players play the bargaining game only once. Thus we deliberately ignore the well-known repeated game effects that can bring about outcomes that are superior to those of the one-shot game (see Fudenberg and Maskin (1986)). Instead we are interested in whether preference evolution on its own can improve the situation for the players in the one-shot game.

To study the evolution of preferences, we employ the so-called indirect evolutionary approach: Players act rationally given their preferences, but over time their preferences can change. The basic idea is that if it pays better in material terms to be, say, "selfish" than "cooperative", people gradually adopt the former and abandon the latter preference.[2] As pointed out by Güth (1995), the indirect evolutionary approach can be interpreted as a generalisation of the standard models, where preferences are exogenous.

A player with *materialistic* preferences maximises her material payoffs. However, two materialistic players end up in conflict some of the time. Were both players dovish, each would be strictly better off, both subjectively and materially. A player with *reciprocal* preferences prefers to be moderate (aggressive) if the opponent is moderate (aggressive). That is, she is kind to those who is kind to her and unkind to those who is unkind to her. This is the behaviour that is often observed in experimental settings and, indeed, in everyday life. One key implication of reciprocity is that when two reciprocal players meet, and it is common knowledge that both are reciprocal, mutual moderation *can* be sustained as a Nash equilibrium.[3] As a consequence, two reciprocal players perform better in material terms than two Materialists.

However, there are also two other feasible preference types: *Altruistic* players, who strictly prefer to be moderate and *Bullies*, who strictly prefer to be aggressive. These

[1] See Section 5 below, where we review the literature.

[2] Huck (1997) provides a good survey of the basic ideas and methodological issues. A more formal treatment is Königstein and Müller (2000). The indirect approach differs from the "direct" evolutionary approach, where players are "programmed" to a certain behaviour and where preferences are consequently immaterial. See e.g. Weibull (1995) for an exposition of the direct approach.

[3] There is also an equilibrium where both players are aggressive, but we will assume (as do other papers in the literature) that they co-ordinate on the payoff-dominant equilibrium.

types should not be a priori excluded from the model. In fact, we show below that it makes a significant difference for preference evolution whether they are available or not. It is by allowing for these preference types that the paper is different from other contributions studying preference evolution in the literature (there are a number of other differences; see Section 5 below, where we review the literature).

It is important how much information a player has about her opponent's preferences. We follow Güth (1995) in assuming that a player learns the opponent's preferences with a certain exogenous probability and otherwise observes only the aggregate distribution of players over the preference types. The higher this probability, the easier it is to "read" the opponent's intentions. It is also important to remark that players are restricted to "truth-telling", i.e., a player cannot pretend to have other preferences than her true ones.

For expositional reasons, we first assume that players are either Reciprocators or Materialists. A reciprocal player performs well in a population of similarly minded players, since conflict is avoided. Moreover, if players have enough information about opponents' preferences, a mutant with materialistic preferences performs badly, since she invariably ends up in some conflict with the Reciprocators. Then reciprocity is a stable outcome. If, on the other hand, players have very little or no information, a reciprocal player is unable to distinguish between an opponent with the same preferences and a materialistic one. Then reciprocity cannot be sustained and there is always some conflict.

We then assume that a player can also be of the Altruist or Bully type. Reciprocity is now less stable than before and the reason is, ironically enough, the presence of the altruistic players. Moreover, if there initially are many Materialists and/or Bullies in the population, the Reciprocators die out. This happens because they get involved in too much costly conflict with their opponents, while the materialistic players avoid this. The population then ends up in a mixture of materialistic players and Bully players. The Bullies survive in this population because they are effectively "committed" to aggression and this gives them an advantage when meeting the Materialists.[4] When players have little or no information about opponents' preferences, the survival chances of materialistic players are improved, since the "commitment" of Bullies become useless when they cannot be properly recognised.

We then consider the players' material welfare. When players can be either Materialists or Reciprocators, they are always materially better off under perfect information in any stable outcome, relative to the benchmark where they are all Materialists. However, when all four preference types are allowed, the material welfare effects of preference evolution become ambiguous: In the Bully-Materialist population mix, material player welfare is lower than in the all-Materialist population, while players are best off when everybody are Reciprocators. Thus the paper shows that preference evolution can worsen the players' material well-being, relative to the situation where all players have materialistic preferences.

The implications of the model for the interpretation of observed experimental behaviour is quite simple: The subject population has Reciprocal and/or Altruistic pref-

[4]The insight that such a "commitment" can confer an advantage on a player in bargaining and other situations is due to Schelling (1960). See also Hirshleifer (1987).

erences. Moreover, the model informs us that we have been lucky: Were the initial conditions different, we would have observed a much inferior performance.

The rest of the paper is organised as follows. In Section 2, we set up the model. In Section 3, we analyse the case where players are either Materialists or Reciprocators. Then in Section 4 we allow for all feasible preference types. Section 5 relates the results to the existing literature. Finally, Section 6 concludes. Section 7 is the Appendix, which contains all proofs.

2 The Model

Consider two players who must divide a valuable surplus between them and who simultaneously choose between being "Hawkish" (H) and "Dovish" (D). The material payoffs (shares of the surplus) are shown below:

	D	H
D	1/2	1/4
H	3/4	0

Figure 1: The material payoffs in the bargaining game.

If the opponent is dovish, it pays a player to be hawkish, since that gives more than half of the surplus (a very good deal) and being dovish only gives one-half (a fair deal). However, if the opponent is hawkish, dovishness is better, since a poor deal is better than no deal. Two hawkish players waste the surplus completely, so they each perform strictly worse than two dovish players.[5] The game is a variant of the Hawk-Dove (or "Chicken") game, which has been used to study behaviour in a variety of economic, political, military and even biological contexts. See, for example, Hirshleifer (1982), Kahn (1965), Lipman (1986), Maynard-Smith (1982). Here we restrict ourselves to a bargaining interpretation.

In a 2×2 game a player can have 24 $(= 4!)$ different strict preference orderings over outcomes. However, in terms of the best replies the ordering induces, many of them are equivalent. Let $(i, j) \succ (m, n)$, with $i, j, m, n = D, H$, indicate that a player strictly prefers the outcome where she plays i and the opponent plays j to the outcome where she plays m and the opponent plays n. We consider the following four strict orderings:

Definition 1 :

A player with material *preferences (a* Materialist*) has the ranking* $(H, D) \succ (D, D) \succ (D, H) \succ (H, H)$.

[5]The results below hold for other material payoff numbers, as long as (i). playing H gives strictly more than one-half, but not all, of the surplus when the opponent plays D, (ii). H is strictly better than D when the opponent plays D and (iii). a mutual choice of D gives each player a strictly higher material payoff than when both players choose H.

A player with reciprocal *preferences (a* Reciprocator*) has the ranking* $(D,D) \succ (H,D) \succ (H,H) \succ (D,H)$.

A player with altruistic *preferences (an* Altruist*) has the ranking* $(D,D) \succ (H,D) \succ (D,H) \succ (H,H)$.

A player with bully *preferences (a* Bully*) has the ranking* $(H,D) \succ (D,D) \succ (H,H) \succ (D,H)$.

A Materialist maximises the material payoffs, i.e., she is the type normally assumed in economics. A Reciprocator maximises "fairness" or "justice": A dovish opponent is rewarded, since the Reciprocator is also dovish; and a hawkish opponent is punished, since the Reciprocator is hawkish, too. We may say that a person with reciprocal preferences adheres to a social norm of not exploiting dovish opponents and of punishing hawkish opponents. An Altruist strictly prefers to "help" the opponent, by raising the latter's material payoff, no matter what the opponent intends to do. Finally, a Bully may be said to be spiteful or even malicious, since she acts to lower the opponent's material payoff, no matter what the latter does.[6]

For convenience, we will say that a player's preferences gives her *type*. Consider now a large population of players, where a player is of one of the four types above. Let s_M, s_R, s_A and s_B denote the population shares of Materialists, Reciprocators, Altruists and Bullies, respectively. Then $s = (s_M, s_R, s_A, s_B)$, with $\sum_i s_i = 1$ and $s_i \geq 0$ for $i = M, R, A, B$, is the population state. At each point in time, players are randomly matched in pairs and then play a Nash equilibrium of the game with *subjective* payoffs. Let $V(i,j)$ denote the resulting *material* payoff to a player of type i when she is matched with a player of type j, where $i, j = M, R, A, B$ (we derive these payoffs below). Also, let $V(i,s)$ denote the expected material payoff to a player of type i, when the population state is s. Finally, $V(s,s) = \sum_i s_i V(i,s)$ is the material average payoff in the population state s.

We assume that the share of players of type i grows if and only if those players earn an expected material payoff that is higher than the average material payoff. Formally, the evolution of the population shares is governed by the Replicator Dynamic (Taylor and Jonker (1978)):

$$\dot{s}_i = s_i[V(i,s) - V(s,s)],$$

where $i = M, R, A, B$. Note that only the material performance matters for the growth of players of a given preference. However, preferences matter indirectly, since they (together with the entire population's preference configuration) determine a player's material performance.

However, rather than studying the Replicator Dynamic itself, we shall limit ourselves to studying the end-points of the dynamic process. That is, we wish to find those population states that are stationary and (asymptotically) stable for the Repli-

[6]The results would not change if we interchanged (H,H) with (H,D) in the Reciprocator's ranking. The same is true if we interchange (H,D) and (D,H) in the Altruists's ranking and, finally, if we interchange (H,H) and (D,D) in the Bully's ranking. This is because the players' best reply structures remain unchanged.

cator Dynamic. To do this, we use the following concepts from the evolutionary game theory literature:

Definition 2 *(Maynard-Smith (1973,1982)):*

A population state s is an Evolutionarily Stable Strategy (ESS) *if (i). $V(s', s) \leq V(s, s)$ for all s' and (ii). If $V(s', s) = V(s, s)$, then $V(s, s') > V(s', s')$.*

A population state s is a Neutrally Stable Strategy (NSS) *if (i). $V(s', s) \leq V(s, s)$ for all s' and (ii). If $V(s', s) = V(s, s)$, then $V(s, s') \geq V(s', s')$.*

The definition says that for a population state to be an ESS, it must, first, be a Nash equilibrium (part (i)). The interpretation is that if mutants, with behaviour corresponding to s', enter the population, they must not perform better than the incumbents, whose behaviour corresponds to s. This ensures that the mutants cannot spread in the population. Second, if the mutants perform exactly as well as the incumbents, the incumbent players outperform the mutants against similar mutants (part (ii)). The two conditions ensure that the mutants disappears again. The NSS definition differs only in the slightly weaker second condition: The mutants may not be driven out, but will not spread. Clearly, an ESS is an NSS, while the converse may or may not be true.

The ESS and NSS concepts are very useful to check the stability properties of a population state: If a population state s is an ESS, then s is (locally) asymptotically stable for the Replicator Dynamic (Taylor and Jonker (1978)). If s is an NSS, then s is Lyapunov stable for the Replicator Dynamic (Thomas (1985) and Bomze and Weibull (1995)). In other words, if a population state s is an ESS, we converge to s when close enough and, once there, we stay there. If s is an NSS, no exogenous shock can lead away from s. However, unlike asymptotic stability, there may be no "pull" towards the population state. The reader may consult e.g. Weibull (1995) for details.

3 Materialists and Reciprocators

We start by assuming that a player is either a Materialist or a Reciprocator. We also initially assume that players are perfectly informed about opponents' preferences.

When two Materialists meet, the game with subjective payoffs has three Nash equilibria: Two asymmetric pure equilibria, (D,H) and (H,D), and a mixed Nash equilibrium. We assume that players cannot condition their choice on the player position (and have no correlation devices). Thus two Materialists play the mixed Nash equilibrium.[7] Since this equilibrium involves some conflict, the material payoff to each player is strictly below one-half (the expected material payoff to each player is a convex combination of the material payoffs 1/2 and 1/4.) Moreover, the outcome is inefficient: Each player would be subjectively strictly better off had they each played D.

[7]We will assume that the mixing probabilities in this equilibrium are the same as those of the mixed equilibrium of the material payoffs game in Figure 1 above. It follows that each Materialist plays Dove with probability 1/2.

Consider next a meeting between a Reciprocator and a Materialist. In this case there is a unique and mixed Nash equilibrium, with each player receiving a material payoff strictly below one-half. Again this outcome is inefficient, since each player would be subjectively strictly better off under mutual dovishness.

Suppose finally two Reciprocators meet. There are then two strict Nash equilibria, (D,D) and (H,H). There is also a mixed Nash equilibrium, but we ignore it here and focus on the strict equilibria. In the (D,D) equilibrium, both the material and the subjective payoff to each player is strictly higher than in the (H,H) equilibrium. Moreover, the former equilbirium is efficient, while the latter is not. We make the following assumption:

Assumption 1 *When two Reciprocators meet, they co-ordinate on the (D,D) Nash equilibrium.*

We will comment below on how important this assumption of perfect coordination is.[8]

We then obtain the following material payoff matrix for the evolutionary game.

	M	R
M	$V(M,M)$	$V(M,R)$
R	$V(R,M)$	$1/2$

Figure 2: The evolutionary game with Materialists (M) and Reciprocators (R).

From this matrix we get the following result:

Proposition 1 *Suppose players have perfect information about opponents' preferences. Then the population consisting of only Reciprocators is an ESS. The all-Materialist population is an ESS if and only if* $V(R,M) < V(M,M)$.

Proposition 1 is good news for reciprocity: If all players have reciprocal preferences, no mutant can displace the population. And if the initial population state is close, we end up in the all-reciprocal population.[9] However, materialism can also be an ESS, if a Materialist performs strictly better than a Reciprocator against another Materialist. Intuitively, this can happen because a Reciprocator, when she meets a Materialist and there is the resulting conflict, lowers not only the Materialist's material payoff, but also her own material payoff. And she may lower her material payoff so much that she performs worse against a Materialist than another Materialist.

Assume now that a player knows only the overall distribution of player types, i.e., the population state, s. This is her prior belief about any opponent's type. Then, with some exogenous probability, she sobserves the opponent's type, which is then her posterior belief. With complementary probability, she observes nothing and her

[8] A similar assumption is made in other papers, e.g. Guttman (2000).

[9] Note that the Reciprocator strategy will remain an ESS even if two Reciprocators sometimes do not manage to co-ordinate on (D,D), as long as this does not happen too often.

posterior equals her prior belief. This is the set-up developed in Güth (1995) and we use it to model a situation where there is always *some* chance that a player learns something correct about her opponent.

When two players are matched, there are four possible information situations: Both players observe the opponent's preference (this was analysed in the previous section); only one of the players observe the opponent's preference (and vice versa); and, finally, neither player observes the opponent's preferences. We assume that Nature decides between these four possibilities with probabilites α, β, γ and $1-(\alpha+\beta+\gamma)$, respectively. In this symmetric context, Nature cannot condition her choice on players' identities; we therefore have $\beta = \gamma$. We also assume that Nature's choice is revealed to the players before they choose their strategies. Thus it is common knowledge what the information situation is.

Consider first the population state consisting of only reciprocating players. Each Reciprocator earns expected material payoff equal to one-half. Can a mutant with materialistic preferences invade this population? With probability α, the mutant is recognised and gets the material payoff $V(M,R)$, as before. With probability β, the Reciprocator learns that she faces a Materialist, while the Materialist observes only the aggregate distribution of preference types. This again gives material payoff $V(M,R)$ to the mutant. If the Reciprocator does not observe the type of the materialistic mutant, but the mutant learns she faces a Reciprocator, the Reciprocator plays D and the mutant plays H. This is clearly the most favourable situation for the mutant. If neither player observes the oponent's type, one gets the same outcome. Thus, at the all-Reciprocator population state, the expected material payoff to a materialistic mutant is $V(M,s) = (\alpha+\beta)V(M,R) + (1-\alpha-\beta)(3/4)$. It follows that when $V(M,s) < 1/2$, the Reciprocator population is an ESS.

Using a similar argument for the all-Materialist population gives the following proposition:

Proposition 2 *Suppose players have imperfect information about opponents' preferences, but still have some information: $\alpha+\beta > 0$. Then*

(a). The all-Reciprocator population is an ESS when $\alpha+\beta > \frac{1/4}{3/4-V(M,R)}$.

(b). The all-Materialist population is an ESS if $V(R,M) < V(M,M)$.

This proposition generalises Proposition 1 to the case of imperfect type information: The all-reciprocating population remains an ESS as long as it is *sufficiently* likely that a hostile mutant is recognised, such that the resulting conflict lowers the mutant's material expected payoff enough to prevent it from invading. Similarly, the Materialist population remains an ESS under imperfect information only if a Materialist outperforms a Reciprocator against other Materialists.

Finally, we consider the case where a player receives no specific information about her opponent's preferences at all:

Proposition 3 *Suppose players have no specific information about their opponents' preferences, i.e., $\alpha = \beta = 0$. Then the Reciprocator population is unstable and the all-Materialist population is an NSS (and may be an ESS).*

The all-Reciprocator population is unstable because the Reciprocators are dovish against all opponents, also against a materialistic mutant, who exploits this by playing H. Such a mutant performs better than the incumbents and invades.

4 Allowing for all four preference types

The previous analysis was restrictive in that the Altruist and the Bully preference type were not available. Making these types feasible gives the following 4×4 matrix:

	M	R	A	B
M	$V(M,M)$	$V(M,R)$	3/4	1/4
R	$V(R,M)$	1/2	1/2	0
A	1/4	1/2	1/2	1/4
B	3/4	0	3/4	0

Figure 3: Material payoffs in the evolutionary game. M: Materialist, R: Reciprocator, A: Altruist, B: Bully.

Note that neither the Bully or the Altruist strategy is dominated. In particular, facing a Materialist, the Bully strategy performs better in material terms than any other strategy.

A study of this matrix yields the following proposition, which is proven in the Appendix.

Proposition 4 *Suppose players have perfect information. Then*

(a). The population of only Reciprocators is an NSS. The same is true for any population state with only Reciprocators and Altruists, as long as the population share of Altruists, s_A, is sufficiently small. Precisely, the condition is $s_A < [1/2 - V(M,R)]/[3/4 - V(M,R)]$.

(b). The population state s^* consisting of only Materialists and Bullies and where the share of Bullies is $s_B^* = [3/4 - V(M,M)]/[1 - V(M,M)]$ is an ESS.

The bad news is now part (b): There is an ESS where no Reciprocators are present at all, but only Bullies and Materialists.[10] Even though there is conflict whenever two Bullies or two Reciprocators meet, preferences are optimally adapted: In an encounter with a Bully, a best material reply is to be dovish, which is what the Materialist prefers. Conversely, a best reply when meeting a Materialist is to be hawkish, which is what the Bully-type prefers. Moreover, no reciprocal (or altruistic) mutant can invade the population: Relative to a materialistic player a reciprocal mutant ends up in total conflict with a Bully and is not compensated enough by meeting the Materialists. In other words, preference evolution has gone into a "trap", where conflict persists.

Note that the Bully players survive in the ESS because each is, by virtue of her strict H-preference, "committed" to gettting the lion's share of the surplus. A Materialist is therefore dovish when meeting a Bully and consequently a population of Materialists is invaded by a Bully mutant. This implies that the all-Materialist population is now unstable.[11]

[10] This is true, irrespective of whether $V(R,M) > V(M,M)$ or not, which was important in Propoposition 1 above.

[11] Similar results has been established in the evolutionary game theory literature (See Banerjee and

There is still some good news (part (a)), but it is less good than before: Reciprocal behaviour can still survive, but its survival ability is lower than in the previous section, where a Reciprocator performed *strictly* better than any other player type in a population of reciprocal players. Now the new Altruistic type performs exactly as well: When an Altruist meets a Reciprocator, the outcome is (D,D), just as when two Reciprocators meet. Moreover, the two types perform exactly as well against another Altruist, so in any mix of Reciprocators and Altruists each type is equally well off. This is why the all-reciprocal population is an NSS, but no longer an ESS.[12] However, an altruistic player is easy prey for a Materialist, so the population mix is stable only as long as there are sufficiently few Altruists. Any small perturbation of the population that would violate the inequality in the proposition above would allow Materialists to enter the population and the population will be taken away from the Reciprocator-Altruist population.

Another way to see that it is not the Bullies who threaten reciprocity, but altruism, is to assume that the altruistic type is not feasible. Then the Reciprocator strategy would be an ESS again (since it would be a strict Nash equilibrium of the reduced 3×3 game).[13] The important thing is that reciprocal behaviour would be immune to any mutant only if they also inflicted some loss on the Altruists. However, in this model, where players act rationally given their preferences, the optimal policy for a Reciprocator is, intuitively speaking, to "forgive" the Altruist and not "punish" her for her unwilingness to punish other people. Thus, preference evolution cannot solve the problem of "who punishes the non-punishers".

Let us now consider the players' material welfare under perfect information. If only the Materialistic and the Reciprocal type are available, as in the previous section, then material payoffs in any stable outcome is always at least as high as when only the Materialist type was available. In this sense preference evolution improves the material well-being of the players by "injecting" into the population a preference for reciprocity that allows players to sustain behaviour with higher material payoffs. Suppose then all four preference types are feasible. Then players are materially best off in the Reciprocator-Altruist NSS. Moreover, they are materially strictly better off in the all-Materialist population than in the Bully-Materialist ESS (the reader is referred to the Appendix for a proof). In other words, allowing players to have other preference orderings than the materialistic one can increase the players' material welfare or reduce it.[14]

Weibull (1994) and Sethi (1996)), but those papers use the direct evolutionary approach, i.e., the Bullies are directly "programmed" to "irrational" behaviour. Here, in contrast, the Bullies are fully rational, but have preferences that, at first sight, seem irrational.

[12]Note also that, unlike the previous section, reciprocal behaviour is no longer robust to small perturbations of their ability to co-ordinate on the (D,D) equilibrium: If reciprocal players sometimes mis-coordinated, they would perform strictly worse than the Altruists in any Reciprocator-Altruist mix.

[13]A similar phenomenon has been observed in evolutionary models of the Prisoner's Dilemma. There the "Tit-For-Tat" strategy (TFT), who co-operates conditionally (see Axelrod (1984)) and the strategy that co-operates unconditionally perform equally well in a population mix of the two strategies. The TFT-strategy is an ESS only when the latter strategy is excluded. See, for example, the survey in Sethi (2001).

[14]Banerjee and Weibull (1994) show the same in a framework where "optimising" players interact with "programmed" players. Sethi (1996) reaches similar results in his direct evolutionary analysis of a Prisoner's Dilemma game with punishment.

Finally, as regards imperfect type information, results similar to Propositions 2 and 3 hold for the four-strategy case. In particular, when players have no information about opponents' preferences at all, the all-Materialist population is an NSS, i.e., stable. The reason is that since a Bully is no longer recognised by a Materialist, the Bully cannot "convince" the Materialist that the latter should be dovish. Thus the Bully no longer performs better than a Materialist and there is no ESS as under perfect information. The all-Reciprocator population is unstable for the same reasons that were given in the previous section. Instead the only stable population is the one consisting of only Materialists.[15]

5 Related literature

Some good surveys of the importance of reciprocity are Fehr and Gächter (1998), Fehr and Tyran (1997) and Sethi and Somanathan (2001). The latter also surveys the literature using the direct evolutionary approach. Many economists have interpreted reciprocity as means to overcome inefficiences caused by opportunism, asymmetric information and incomplete contracts. See, for example, Arrow (1971), Ben-Ner and Putterman (2000), Bowles (1998), Fehr and Gächter (1998), Güth and Kliemt (1994), Hirshleifer (1987) and Sugden (1986).

There is much experimental evidence on the role of reciprocity. In the Ultimatum game, Responder subjects reject low offers (i.e., bad behaviour is punished) and accept sufficiently high offers (good behaviour is rewarded). See Güth et. al. (1982) and Roth (1995). In the Dictator Game, players donate significant amounts. See Forsythe, Horowitz, Savin and Sefton (1994). Another context where reciprocity seems important is in public goods experiments, where players often reach contribution levels that are substantally higher than those predicted by the theoretical model based on materialistic preferences. See, for example, Fehr and Gächter (1998) and Hoffman, McCabe and Smith (1998). The many Prisoner's Dilemma experiments show that players cooperate, even in one-shot simultaneous move games, where players know they will never meet each other again. See, for example, Cooper et. al. (1996). In the context of the labor market and the employer-employee relationship, several papers have documented the role of reciprocity. See Agell and Lundborg (1995), Fehr, Kirchsteiger and Riedl (1993), Fehr and Gächter (1998)), Fehr and Schmidt (2000) and Kirchler, Fehr and Evans (1996).

There are several contributions in the literature that study theoretically what behaviour preference evolution can sustain. In the following I highlight some of the differences and similarities between these models and the one studied here.

Possajennikov (2000) considers preference evolution in general symmetric 2×2 games, including the Hawk-Dove game, which our material stage game is an instance of. Like us, he assumes that two Reciprocators (he calls them "conformists") coordinate on the equilibrium with the highest material payoff. Similarly, two Materialists

[15]Since the Materialists play a mixed Nash equilibrium where the material payoffs from being dove and hawk are exactly balanced, any other mutant perform exactly as well as a Materialist. Hence the mutant can enter, but not displace, the all-Materialist population. This implies that the all-Materialist population is stable. A similar result is derived in Banerjee and Weibull (1994).

(or, in his terminology, "non-conformists"), play the mixed Nash equilibrium. The key difference between his approach and ours is that he also allows for the preference type who is indifferent between outcomes. This type is quite successful in his model, since it can be led to choose the optimal action in each encounter with other types. However, we find it rather difficult to interpret such a preference in the context of the paper. Consequently, we restrict attention to strict preference orderings. Another difference is that Possajennikov only studies the evolutionary stability of population states where all players use the same strategy ('monomorphic' population states).

Güth and Yaari (1992) study a game where players have an opportunity to punish an opponent's "unfair" behaviour, after a surplus has been divided. Assuming perfect type information, a preference for such punishment will evolve. This happens because a player who is willing to punish is treated fairly and, moreover, she exploits an opponent who is not himself willing to punish. One difference between their model and the present is that their punishers benefit when "punishing" an unfair opponent. This, in turn, implies that the state with materialists is unstable in their model. In this model, on the other hand, a Reciprocator can, when she meets a Materialist, lower her own material payoff so much that she performs worse than the materialist against another materialist. Then the all-Materialist population state is stable. Another difference is that their Reciprocators are "meaner" than ours: The former prefer to exploit an opponent who is fair but unwilling to punish, rather than being fair themselves (and to not punish). This happens because preferences in their model are specified such that a player cannot decide to "forgive" a player who is fair but unwilling to punish. Thus the problem of "who punishes the non-punishers" does not arise in their model, unlike the present, where altruism destabilised reciprocity. Güth (1995) generalises the analysis of Güth and Yaari (1992) to the case of imperfect type information. In this paper we followed his modelling approach.

Güth and Kliemt (1994) analyse a "game of trust", where a first-mover can either "stay out" or trust a second-mover and the second-mover can either reciprocate or exploit the first-mover's trust. When all players have materialistic preferences, the second-mover will exploit the first-mover's trust and the outcome is inefficient. They show that when second-movers can evolve a preference for reciprocation and if the first-mover can distinguish between trust-worthy and untrustworthy second-movers, an efficient outcome can evolve. They also show that when there is no specific information, but first-movers can, at a cost, learn the second-mover's type, both exploiting and reciprocal types may co-exist.

Fershtman and Weiss (1998), Guttman (2000) and Ockenfels (1993) analyse evolution of preferences in the Prisoner's Dilemma game. These papers assume more specific forms of preferences than the present paper. In Guttman (2000), for example, two preference types are feasible (roughly corresponding to our Reciprocators and Materialists). The conclusions of this and the other papers would be somewhat different if all the four types analysed in this paper were feasible. The reader is referred to Poulsen (2001b).

Other papers using the indirect approach are Bester and Güth (1998), who show that altruism can be evolutionarily stable; Bowles and Gintis (1998), who study a situation where players may develop a preference for punishing free riders; Fehr and Schmidt (1999) emphasise the role of inequity aversion as explaining many observed

empirical regularities; Huck (1997, 1998) study the interplay between preferences and the design of legal institutions; Huck and Oechssler (1999) study fairness preferences in the Ultimatum Game; Königstein (1998) show that efficiency can evolve in a set-up with bargaining and production; Poulsen (2001a) show that Responders can, under certain circumstances, develop preferences that allow them to rationally refuse low offers. Finally, papers who analyse preference evolution in other set-ups are Ely and Yilankaya (1997), Kockesen, Ok and Sethi (2000a,b), Ok and Vega-Redondo (2000) and Sethi and Somanathan (2000).

6 Conclusion

Experimental research has shown that reciprocity is an important determinant of many kinds of behaviour in many economic and social situations. This has raised some doubt about the validity of the assumption that players always have the materialistic preferences that are typically assumed in economics. In this paper we constructed a simple model where players were allowed to have different kinds of preferences. By allowing the various preference types to "compete for survival", we avoided the methodological pitfall of simply assuming that players are always "nice".

Two players with reciprocal preferences share the surplus without any conflict and perform strictly better than two Materialists. We showed that reciprocity is stable whenever Reciprocators inflict a loss on players of other types.

This result was quite optimistic: Preference evolution would improve players' material welfare through the evolution of reciprocity. However, preference evolution need not "inject" only reciprocity into an all-materialistic population. It may also develop a taste for blind aggression (the Bully type) or for pure altruism (Altruism). We therefore made these preferences available to the players, in addition to the materialistic and reciprocal types. Our results then became less optimistic. First, the Altruists tended to destabilise reciprocity, because the Reciprocators treated the Altruists as they treated other Reciprocators. Moreover, very aggressive players and Materialists could end up co-existing and no reciprocal player could invade such a population. Preference evolution could therefore improve or worsen the situation for the players, relative to the situation where all players were Materialists. What matters here is the initial starting point of the dynamic process of preference change.

An important insight of the paper is therefore that preference evolution need not always lead to materially better outcomes for the players. It is possible that it leads into a "trap", where preferences are optimally co-adapted, but where the outcome for players is both subjectively and materially worse than when all players have materialistc preferences.

7 Appendix

Proof of Proposition 1: Since $V(M,R) < V(R,R) = 1/2$, the R-strategy is a strict Nash equilibrium and so an ESS (see e.g. Weibull (1995)). If $V(R,M) < V(M,M)$, the

M-strategy is an ESS for the same reason. If, on the other hand, $V(R,M) > V(M,M)$, the M-strategy is not even a Nash equilibrium. If $V(R,M) = V(M,M)$, the M-strategy again fails to be an ESS (and an NSS), since $V(M,R) < V(R,R)$.

Proof of Proposition 2: A proof is available from the author upon request.

Proof of Proposition 3: A proof is available from the author upon request.

Proof of Proposition 4: (a). Let s be any population state with carrier in {R, A}. Then $V(s,s) = 1/2$ and $V(s',s) \leq V(s,s)$ for all s' when $s_A \leq [1/2 - V(M,R)]/[3/4 - V(M,R)]$. Then (s,s) is a symmetric Nash equilibrium and part (i) of the NSS definition is satisfied. When the inequality holds strictly, one has $V(s',s) = V(s,s)$ only for those s' with carrier in {R,D}. Moreover, since for these s' one has $V(s,s') = V(s',s') = 1/2$, s satisfies the second part of the NSS definition as well. (b). Let s be a population state with carrier {M,H}. One computes that (s,s) is a symmetric Nash equilibrium when the population share of Bullies satisfies $s^*_H = [3/4 - V(M,M)]/[1 - V(M,M)]$. To verify the second part of the ESS definition, note that $V(s',s^*) = V(s^*,s^*)$ only for those s' with carrier in {M,B}. To verify that $V(s^*,s') > V(s',s')$ for all such s', one may compute $V(s^*,s') - V(s',s') = (s^*_H - s'_H)[V(H,s') - V(M,s')]$. The right hand side can be simplified to $(s^*_H - s'_H)[3/4 - V(M,M) - s'_H(1 - V(M,M))]$ or to $(s^*_H - s'_H)^2[1 - V(M,M)]$. This expression is strictly positive for all $s'_H \neq s^*_H$. Thus s^* is an ESS.

We finally prove the statement that we made in Section 4 about the material payoffs in the various outcomes. In the Reciprocator-Altruist mix everybody earns one-half. In the all-Materialist population the mixed Nash equilibrium of the material payoffs game is played. This gives an expected material payoff equal to $V(M,M) = 3/8 < 1/2$. Finally, in the Bully-Materialist ESS, each player earns expected material payoff $(1 - s^*_H)(3/8) + s^*_H(1/4) < 3/8$, since $s^*_H > 0$. This ranking of material payoffs holds also for the more general material payoff specification outlined in Footnote 4 above.

8 References

Agell, J. and Lundborg, P. (1995): "Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms", Scandinavian Journal of Economics, 97, 295-307.

Arrow, K. (1971): "Political end economic evaluation of social effects and externalities", 3-25 in Intrilligator, M. (ed.): Frontiers of Quantitative Economics, North-Holland Publishing Company.

Axelrod, R. (1984): The evolution of cooperation, New York: Basic Books.

Banerjee, A. and Weibull, J. (1995): Evolutionary selection and rational players. In Kirman, A. and Salmon, M. (eds): Learning and Rationality in Economics. Oxford: Blackwell, 343-363.

Ben-Ner, A. and Putterman, L. (2000): "On some implications of evolutionary psychology for the study of preferences and institutions", Journal of Economic Behavior

and Organization, 43, 91-99.

Bester, H. and Güth, W. (1998): "Is Altruism Evolutionarily Stable?", Journal of Economic Behavior and Organization 34, 193-209.

Bomze, I. and Weibull, J. (1995): "Does neutral stability imply Lyapunov stability", Games and Economic Behavior, 11, 173-192.

Bowles, S. (1998): "Endogenous Preferences: The Cultural Consequences of Markets and other Economic Institutions", Journal of Economic Literature, 56, 75-111.

Bowles, S. and Gintis, H. (1998): "The Evolution of Strong Reciprocity", Santa Fe Institute Working Paper 98-08-073E.

Cooper, R., Dejong, D., Forsythe, R. and Ross, T. (1996): "Cooperation without reputation: Experimental Evidence from Prisoner's Dilemma Games", Games and Economic Behavior, 12, 187-218.

Ely, J. and Yilankaya, O. (1997): "Nash equilibrium and the Evolution of Preferences", Department of Economics, Northwestern University.

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993): "Does fairness prevent market clearing? An experimental investigation", The Quarterly Journal of Economics, 437-459.

Fehr, E. and Gächter, S. (1998): "Reciprocity and economics: The economic implications of *Homo Reciprocans*", European Economic Review, 42, 845-859.

Fehr, E. and Schmidt, K. (1999): "A Theory of Fairness, Competition and Cooperation", Quarterly Journal of Economics, 114, 817-868.

Fehr, E. and Schmidt, K. (2000): "Fairness, incentives and contractual choices", European Economic Review, 44, 1057-1068.

Fehr, E. and Tyran, J. (1997): Institutions and Reciprocal Fairness. Nordic Journal of Political Economy 23, 133-144.

Fershtman, C. and Weiss, Y. (1998): "Why do we care what others think about us?", 133-151 in Ben-Ner, A. and Putterman, L. (eds.): Economics, Values and Organization, Cambridge University Press.

Forsythe, R., Horowitz, J., Savin, N. and Sefton, M. (1994): "Replicability, fairness and pay in experiments with simple bargaining games", Games and Economic Behavior, 6, 347-369.

Fudenberg, D. and Maskin, E. (1986): "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information", Econometrica, 54, 533-554.

Guttman, J. M. (2000): "On the evolutionary stability of preferences for reciprocity", European Journal of Political Economy, 16, 31-50.

Güth, W.: (1995): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives", International Journal of Game Theory, 24, 323-344.

Güth, W. and Kliemt, H. (1994): "Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes", Metroeconomica, 45,

155-187.

Güth, W., Schmittberger, R. and Schwarz, B. (1982): "An experimental analysis of ultimatum bargaining", Journal of Economic Behavior and Organization, 3, 367-388.

Güth, W. and Yaari, M.: (1992): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game". In Witt, Ulrich (ed.): Explaining Process and Change - Approaches to Evolutionary Economics, Ann Arbor, MI: University of Michigan Press.

Hirshleifer, J. (1977): "Economics from a Biological Viewpoint", Journal of Law and economics, 20, 1-52.

Hirshleifer, J. (1987): "The Emotions as Guarantors of Promises and Threats", 307 -, in Dupre, J. (ed): The Latest on the Best, Cambridge: MIT Press.

Hoffman, E., McCabe, K. and Smith, V. (1998): "Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology", Economic Inquiry, 36, 335-352.

Huck, S. (1997): "Institutions and Preferences: An Evolutionary Perspective", Journal of Institutional and Theoretical Economics, vol. 153, 771-779.

Huck, S. (1998): "Trust, Treason and Trials: An Example of how the Evolution of Preferences can be driven by legal Institutions", Journal of Law, Economics and Organization, 14, 44-60.

Huck, S. and Oechssler, J. (1999): "The Indirect Evolutionary Approach to Explaining Fair Allocations", Games and Economic Behavior, 28, 13-24.

Kahn, H. (1965): On Escalation: Metaphors and Scenarios, Greenwood Press Reprint, New York.

Kirchler, E., Fehr, E. and Evans, R. (1996): "Social exchange in the labor market: Reciprocity and trust versus egoistic money maximization", Journal of Economic Psychology, 17, 313-341

Kockesen, L., Ok, E. and Sethi, R. (2000a): "Evolution of Interdependent Preferences in Aggregative Games", Games and Economic Behavior, 31, 303-310.

Kockesen, L., Ok, E. and Sethi, R. (2000b): "The Strategic Advantage of Negatively Interdependant Preferences", Journal of Economic Theory, 92, 274-299.

Königstein, M. (1998): "Efficiency and Evolution of Social Preferences and Prosocial Behavior", Discussion Paper 90, SFB 373, Humboldt-Universitt zu Berlin.

Königstein, M. and Müller, W. (2000): "Combining rational choice and evolutionary dynamics: The indirect evolutionary approach", Metroeconomica, 51, 235-256.

Lipman, B. (1986): "Cooperation between egoists in Prisoner's Dilemma and Chicken games", Public Choice, 51, 513-331.

Martinez Coll, J. C. and Hirshleifer, J. (1991): "The Limits of Reciprocity", Rationality and Society, 3, 35-64.

Maynard-Smith, J. and Price, G.R. (1973): "The Logic of Animal Conflict", Na-

ture, 246, 15-18.

Maynard Smith, J. (1982): Evolution and the Theory of Games, Cambridge University Press.

Ockenfels, P. (1993): "Cooperation in prisoner's dilemma", European Journal of Political Economy, 9, 567-579.

Ok, E. and Vega-Redondo, F. (2000): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario", Journal of Economic Theory,

Possajennikov, A. (2000): "Evolution of Preferences in 2×2 Symmetric Games", Working paper $00 - 03$, Department of Economics, Dortmund University.

Poulsen, A. (2001a): "Evolutionary Selection in the Ultimatum Game", mimeo, Department of Economics, The Aarhus School of Business.

Poulsen, A. (2001b): "A Note on Preference Evolution in the Prisoner's Dilemma", mimeo, Department of Economics, The Aarhus School of Business.

Robson, A. J. (1990): "Efficiency in evolutionary games: Darwin, Nash and the secret handshake", Journal of Theoretical Biology, 144, 379-396.

Roth, A. (1995): Bargaining experiments. In Kagel, J. and Roth, A. (eds): Handbook of Experimental Economics, 253-348, Princeton University Press.

Schelling, T. (1960): The Strategy of Conflict, Harvard University Press, Cambridge.

Sethi, R. (1996): "Evolutionary stability and social norms", Journal of Economic Behavior and Organization, 29, 113-140.

Sethi, R. and Somanathan, E. (2000): "Preference Evolution and Reciprocity", Journal of Economic Theory, forthcoming.

Sethi, R. and Somanathan, E. (2001): "The Evolution of Reciprocity: A Survey", unpublished manuscript.

Sugden, R. (1986): The Economics of Rights, Cooperation and Welfare, Basil Blackwell, Oxford.

Taylor, P.D. and Jonker, L.B. (1978): "Evolutionarily Stable Strategies and Game Dynamics", Mathematical Biosciences, 40, 145-156.

Thomas, B. (1985): "On Evolutionarily Stable Sets", Journal of Mathematical Biology, 22, 105-15.

Weibull, J. (1995): Evolutionary Game Theory, MIT Press.