The Survival and Welfare Implications of Altruism When Preferences are Endogenous

Anders Poulsen
Department of Economics, The Aarhus School of Business
Silkeborgvej 2, DK-8000 Aarhus C, Denmark

April 7, 2003

Abstract

This paper is a contribution to the economic literature studying altruism. In a simple evolutionary model of endogenous preferences we show that individuals with altruistic preferences can survive. We also analyze the material welfare implications of altruism. Policies that promote altruism in the population can be detrimental to material welfare.

Keywords: Endogenous preferences; altruism; welfare; reciprocity; materialism; evolutionary stability.

JEL Classification: C70; D64.

1 Introduction

Several economists have studied the behavioral and welfare consequences of individuals having altruistic preferences. See, for example, Becker (1974, 1976), Bergstrom (1989), Bernheim and Stark (1988), Bernheim, Shleifer and Summers (1985), Bruce and Waldman (1990) and Hirshleifer (1977). In this paper we endogenize individuals' preferences: Individuals may differ in their preferences, and the population proportions of the preferences adjust over time, according to the material success the preferences confer on the individuals carrying them. See e.g. Bester and Güth (1998). We ask: Can individuals with altruistic preferences survive? If so, what are the material welfare implications of altruism?

In our model interaction is formalized as a Prisoner's Dilemma game and altruism is modeled in the simplest possible way: An altruist always co-operates. We show that even though such behavior can be exploited by selfish individuals it can still, on average, 'pay': Altruists perform well against each other, and their preference structure induces reciprocal individuals to treat them just as well, and, which is crucial, better than reciprocal individuals treat each other. This may more than outweigh the loss an altruist suffers when interacting with materialistic individuals. Then altruism pays and survives together with other preferences. The survival of altruism has been studied in other contexts. Bester and Güth (1998) studies the evolution of altruism in a dupoly market [see also Bolle (2000) and Possajennikov (2000)]. They show that altruism can completely dominate the population whenever actions are strategic complements. In our model this cannot happen and we instead get preference heterogeneity.[1] This difference occurs because of the way altruism is modeled: In our model, an altruist behaves the same way toward all opponents, while in Bester and Güth (1998) the altruist conditions his behavior on the opponent's preferences. In their model there are consequently two strategic effects at work when an altruist meets e.g. a selfish person: The altruistic person's preferences affect the selfish person's behavior and vice versa. In our model, only the first strategic effect is active and this implies that altruism can never dominate the population alone.

Given that altruism survives, we next consider the relationship between altruism and material welfare.[2] We show that policies that increase the proportion of altruists in the population can reduce welfare. Conversely, a policy that reduces the proportion of altruists can increase welfare. The reason is that a policy that promotes altruism also promotes materialism, since materialistic individuals benefit from interacting with altruists. Since the first (second) preference has a positive (negative) effect on the population's welfare, the total effect is ambiguous. We show that the total effect is sometimes positive and at other times negative, depending on parameter values. Thus the relationship between policy changes, their effects on the preference distribution and the overall effect on welfare is far from obvious. In Bester and Güth (1998), on the other hand, an increase in the players' degree of altruism always increases welfare. This is because when altruism survives in their model, it is the only preference that survives, so the negative effects, present in our model, are not present in their model.[3]

The rest of the paper is organized as follows: In Section 2 we consider the interaction between reciprocity, materialism and altruism. Section 3 analyzes the welfare implications of altruism. Section 4 concludes. All proofs are in the Appendix.

2 When does Altruism Survive?

The matrix below gives the payoffs for the classical Prisoner's Dilemma (PD) game:

	C	D
C	R	S
D	T	P

Each person has two actions, Co-operate (C) and Defect (D). 'R' stands for

[1] Other evolutionary approaches are studied in e.g. Bergstrom and Stark (1993) and Bergstrom (1995). See also the excellent survey in Bergstrom (2002).

[2] Whenever we use the word 'welfare' in this paper, we will mean the *material* welfare ('money').

[3] Bernheim and Stark (1988) show, in a non-evolutionary framework, that altruism can reduce welfare. In repeated interaction, for example, an increase in the degree of altruism can lower welfare because it makes the use of certain punishment strategies supporting co-operation less credible. Another model analyzing the relationship between changes in altruism and welfare is Milchtaich (2002). In his model all players have the same degree of altruism and preferences are not endogenized. The focus in that paper is on what happens when all players' common degree of altruism changes.

'Reward', 'S' for 'Sucker', 'T' for 'Temptation', and 'P' for 'Punishment', where $T > R > P > S$ and $R > (1/2)(S+T)$.[4] There is a single large population of individuals. At every time instant, individuals are randomly matched in pairs and play the PD game, where the payoffs S, P, R and T are interpreted as *money* payoffs. All decision makers act rationally given their preferences. The objects of evolutionary selection are preferences, not behavioral programs: Those preferences that give their users more than average money earnings proliferate in the population. See e.g. Bester and Güth (1998).

There are three feasible preferences: The *Reciprocator* preference (R), the *Materialist* preference (M) and the *Altruist* preference (A). A Materialist person's subjective payoffs are represented by the money payoffs. Such an individual always plays D. A Reciprocator prefers C (D) if the opponent plays C (D). The Altruist strictly prefers to co-operate and so always chooses C. Preferences are, as usual, assumed to be common knowledge. We use the term 'Altruist' because an individual who always chooses C can be interpreted as seeking to maximize a weighted average of his own and the opponents' material payoff, and where the weight assigned to the opponent's material payoff is sufficiently large.[5]

When two Reciprocators meet, both the (C, C) and the (D, D) outcome are Nash equilibria. There is also a symmetric mixed Nash equilibrium. It seems rather natural to assume that a Reciprocator prefers (C, C) to (D, D) and prefers (C, C) to the payoff in the mixed Nash equilibrium. However, we will not *a priori* exclude the (D, D) Nash equilibrium: It is a real possibility that each individual does not trust the other to co-operate and hence chooses to defect instead.[6] Various kinds of 'noise', or misunderstandings, may have the same effect: There is a risk that two Reciprocators get stuck in the (D, D) equilibrium. In order to formalize these considerations in the simplest possible way, we make the following assumption: When two Reciprocators meet, each plays C with probability λ and plays D with probability $1 - \lambda$, where $\lambda \in (0, 1)$. When λ is close to unity the two Reciprocators almost perfectly co-ordinate on (C, C); when λ is small, they mostly defect on each other. It is, from a formal point of view, in this respect that our model differs from the existing models of endogenous preferences in the Prisoner's Dilemma game, such as Fershtman and Weiss (1998), Guttman (2000), Ockenfels (1993) and Witt (1986). These models assume that reciprocal individuals can co-ordinate with probability one on their preferred equilibrium, i.e., they assume implicitly that $\lambda = 1$.

In the matrix below, entry (i, j) is the money payoff to an individual with preference i when matched with another individual with preference j, where $i, j = R, M, A$.[7] This

[4]This is a standard condition for the Prisoner's Dilemma game, see e.g. Rapoport and Chammah (1965) and Axelrod (1984).

[5]Formally, suppose the Altruist's subjective payoff function is $u(i,j) = \alpha\pi(i,j) + (1-\alpha)\pi(j,i)$, where $\pi(i,j)$ [$\pi(j,i)$] is the material payoff to the individual [opponent] when the individual chooses action i and the opponent chooses j, where $i, j = C, D$. Some simple calculations reveal that C is a strictly dominant strategy when $\alpha < max\{(R-S)/(T-S), (T-P)/(T-S)\}$. That is, the player assigns a sufficiently high weight to the opponent's payoff.

[6]Moreover, depending on the Reciprocator's exact subjective utility numbers, (D, D) may risk-dominate the (C, C) equilibrium, and so, according to the risk-dominance criterion (Harsanyi and Selten (1988)), this outcome should be selected.

[7]For example, entry (R, M) contains the payoff P because each player plays D; similarly, when a Reciprocator and an Altruist meet, each plays C.

matrix defines a symmetric 3×3 symmetric normal form game, to which we can apply well-known concepts in evolutionary game theory; see e.g. Weibull (1995).

	R	M	A
R	μ	P	R
M	P	P	T
A	R	S	R

Table 1: The money payoffs in the evolutionary game.

We have

$$\mu = \lambda^2 R + \lambda(1-\lambda)(S+T) + (1-\lambda)^2 P. \tag{1}$$

We first notice that the assumption $R > (1/2)(S+T)$ implies $\mu < R$ for all $\lambda \in (0,1)$. This means that an Altruist can invade an all-Reciprocator population. Moreover, letting $x = (x_R, x_M, x_A)$ denote the distribution of preferences in the population, we obtain the result that altruism survives with the two other preferences:

Proposition 1 *Suppose $P < (1/2)(S+T)$ and $\lambda \in (0,1)$. There is a unique interior equilibrium population, x^*, given below, where all three preferences are represented. This equilibrium is a center for the Replicator Dynamic.*

$$x^* = (x_R^*, x_M^*, x_A^*) = \left(\frac{(P-S)(T-R)}{D}, \frac{(R-\mu)(T-R)}{D}, \frac{(\mu-P)(P-S)}{D} \right), \tag{2}$$

where $D = (\mu - P - R)[P + R - (S+T)] + PR - TS$.

Proof: In the Appendix.

Altruism survives because a player with such preferences can induce a reciprocally minded person to treat the altruist better than any other individual is treated; the Reciprocator is induced to co-operate with probability one with the Altruist.

We have illustrated Proposition 1 in the figure below, which shows the phase portrait for the Replicator Dynamic. We have set $S = -1$, $P = 0$, $R = 1$, $T = 2$ and $\lambda = 1/2$, such that $\mu = 1/2$. The vertex labeled i is the population where all individuals are of preference type i, where $i = R, M, A$. The equilibrium is $x^* = (1/2, 1/4, 1/4)$.

Let us now consider the case where $(1/2)(S+T) < P$.

Proposition 2 *Suppose $(1/2)(S+T) < P$ and $\lambda \in (0, \lambda')$. The Materialist preference is then the unique ESS and the unique asymptotically stable population for the Replicator Dynamic. If $\lambda \in (\lambda', 1)$, Proposition 1 continues to hold.*

$$\lambda' = \frac{2P - (S+T)}{P + R - (S+T)}. \tag{3}$$

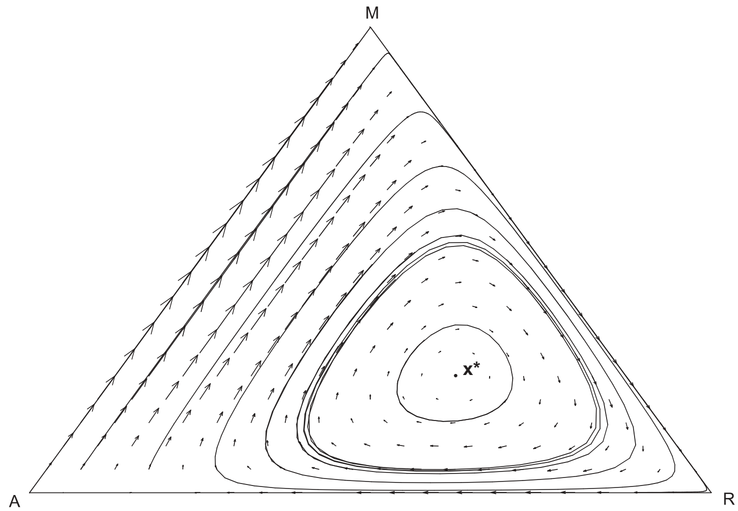

Figure 1: Illustration of Proposition 1. Equilibrium proportions: $x_R^* = 1/2$, $x_M^* = 1/4$ and $x_A^* = 1/4$.

Proof: In the Appendix.

Materialism is the only preference that survives when two conditions hold: The Reciprocators must get a very low payoff whenever they fail to co-ordinate on C; and they must be sufficiently likely to realize the inferior (D, D) outcome. When this is the case a Materialist earns strictly higher expected payoff than a Reciprocator when all three preferences are present; this drives out the Reciprocator preference, leaving only the Materialist and the Altruist preference, from which the Materialist emerges as the victor. Altruism disappears because it can no longer, so to speak, 'hide behind' the Reciprocators.

We may also ask what happens when Reciprocators can perfectly co-ordinate on their preferred (C, C) equilibrium, i.e., $\lambda = 1$. In this limiting case it can be shown that our equilibrium above collapses and no population with all three preference types can be stable. Indeed, only populations with Reciprocators and some, but not too many, Altruists, are stable. However, as mentioned earlier, we regard the case $\lambda = 1$ as a somewhat extreme case of our analysis.

3 Altruism and Material Welfare

In the evolutionarily stable populations with altruism (described in Proposition 1), what is the relationship between the equilibrium proportion of altruists and the equilibrium average material payoff? Do the two always move in the same direction? To consider changes in the equilibrium proportions, we parameterize the payoff in the simplest possible way, using a single parameter, a: $S = -a$, $P = 0$, $R = 1$ and $T = 2a$, with $a > 1/2$ and $a < 2$; the latter ensures $R > (1/2)(S + T)$. One interpretation is that a measures the *severity* of the social dilemma: The larger is a, the larger the money payoffs from defecting, for any choice of the opponent.

The matrix for the evolutionary game is

	R	M	A
R	μ	0	1
M	0	0	$2a$
A	1	$-a$	1

Our restriction $a \in (1/2, 2)$ implies $P < (1/2)(S+T)$ for all $\lambda \in (0,1)$, so Proposition 1 applies. We now have

$$x^* = (x_R^*, x_M^*, x_A^*) = \left(\frac{a(2a-1)}{D}, \frac{(1-\mu)(2a-1)}{D}, \frac{\mu a}{D} \right),$$

where $D = (1-a)\mu - 1 + a + 2a^2$ and $\mu = \lambda^2 + \lambda(1-\lambda)a$. In the evolutionary equilibrium all three preference type's expected payoffs are equal, and hence equals the average payoff. Thus we may use the Materialist type's expected payoff to write the equilibrium average payoff, denoted π, as

$$\pi = 2ax_A^*. \tag{4}$$

Along any orbit (other than x^* itself), the equilibrium average payoff π fluctuates endlessly, too. However, it can be shown that in the long-run average equilibrium payoff in the population along any orbit is in our case given by the average payoff π in the equilibrium x^*. From (4) we can see how an increase in a affects the equilibrium average payoff: The increase in a gives a Materialist more whenever he meets an Altruist, but there will be fewer of them. Thus the total effect of an increase in a on equilibrium average payoff is not a priori obvious. We obtain the following result:

Proposition 3 *For any $\lambda \in (0,1)$, there are severities, a, such that: (i). Increasing the severity a* reduces *the equilibrium proportion of altruism and* increases *equilibrium average material payoff. (ii). Reducing the severity a* raises *the equilibrium proportion of altruism and* reduces *equilibrium average material payoff.*

The proof of the result is obtained from inspecting the graphs of x_A^* and π, plotted as functions of a and λ. These are shown in the top and middle figures in Figure 2 below (next page). We see in Figure 2 that for any fixed $\lambda \in (0,1)$ x_A^* is monotonically decreasing in a. Moreover, for any fixed $\lambda \in (0,1)$, π first falls, reaches a minimum in the interior of the interval $(1/2, 2)$ and then increases. Combining these two observations gives the proposition above.

To explain and interpret this result, we consider a specific example, where we fix $\lambda = 3/4$. The bottom diagram in Figure 2 shows how the average payoff, and the equilibrium proportions, vary with a. For a small (close to one-half), there are almost only Altruists in the population. As a grows, the expected payoff to Materialists and Reciprocators increase. The expected payoff to the Altruists, on the other hand, falls, since they are hurt more when meeting the Materialists. The proportion of Reciprocators grows monotonically; the proportion of Materialists grows initially, too. The

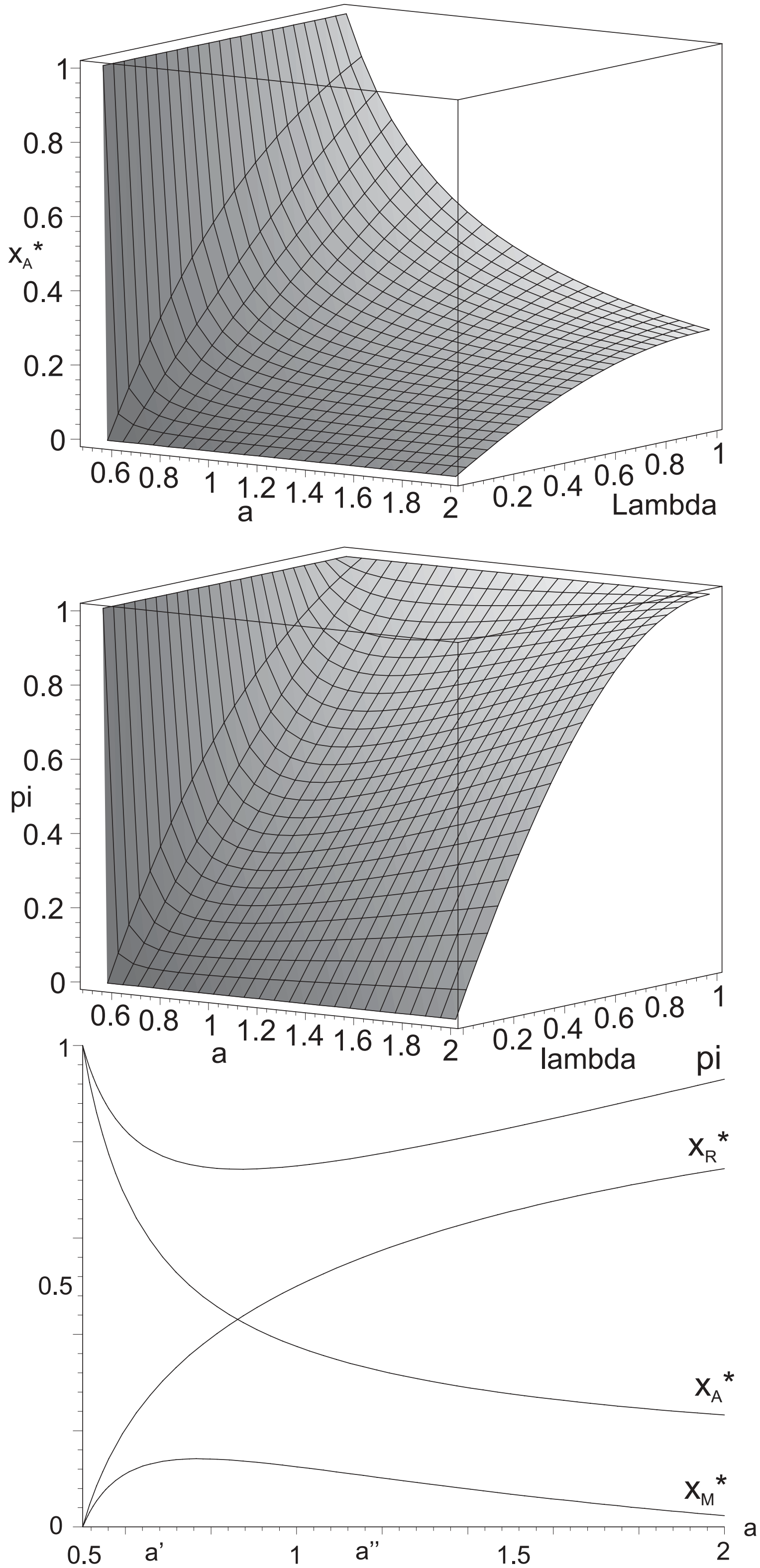

Figure 2: Top: The equilibrium proportion of Altruists, x_A^*, as a function of a and λ. Middle: The equilibrium average payoff, π, as a function of a and λ. Bottom: The equilibrium average payoff and the equilibrium proportions, x_i^*, where $i = M, R, A$, given as functions of the parameter a; $\lambda = 3/4$.

reason is that the Materialists 'prey' on the Altruists, and as long as the payoff T from preying increases and there are many altruists, the expected payoff of a Materialist is high. Thus the proportion of Materialists increase. The proportion of Altrusts fall. All this implies that the average payoff π falls. At the point a', a policy that raises the proportion of Altruists by lowering a also lowers the proportion of Materialists, and hence it raises average payoff.

However, as a keeps increasing, the proportion of Altruists become so low that even though the payoff to a Materialist from meeting an Altruist increases, the Materialists' overall expected payoff fall below the average. Then the proportion of Materialists falls, too. Then the average payoff π starts to increase. At the point a'', a policy that lowers a will raise *both* the proportion of Altruists and Materialists. However, the negative effect from the increased presence of the materialists outweigh the positive effect from the Altruists. Thus average payoff falls (part (ii) of the proposition). Conversely, starting at the same point and increasing a will reduce altruism and materialism and now raise material welfare (part (i))).

4 Summary and Conclusion

We analyze the survival ability of altruism in an evolutionary model where preferences are endogenous. Altruism can 'pay off', and survive. The reason is that altruists perform well against each other and that they, via a strategic preference effect (Schelling (1978)) induce reciprocal individuals to treat them well, and better than how reciprocal individuals treat each other. However, policies that aim at increasing altruism can actually reduce material welfare, and vice versa. The reason is that a policy that increases altruism, which on its own would increase welfare, can give more materialism in society. Since the latter is detrimental to welfare, the total effect can be positive or negative.

The observation in Bester and Güth (1998), that common knowledge of preferences is important for their analysis, also applies here: If opponents' preferences were completely unknown, there could be no strategic effect from preferences and altruism is not likely to survive. Based on the analysis in papers allowing for imperfect information about preferences, see e.g. Frank (1987) and Güth (1995), we conjecture, however, that our results will hold as long as individuals' preferences are *sufficiently* observable.

5 Appendix

Proof of Proposition 1: First, it is straightforward, using (1), to verify the following: $P < (1/2)(S+T)$ implies $P < \mu$ for all $\lambda \in (0,1)$. This, in conjunction with the fact that $\mu < R$, implies

$$P < \mu < R \tag{5}$$

for all $\lambda \in (0,1)$. This will be used below.

Our proof builds on the results in Bomze (1983). Bomze exploits the fact that there is a close relationship between the Lotka-Volterra Dynamic and the Replicator Dynamic: If (p, q) is a fixed point for the Lotka-Volterra dynamic, then

$$x^* = (x_R^*, x_M^*, x_A^*) = (1/(1+p+q), p/(1+p+q), q/(1+p+q)) \tag{6}$$

is a fixed point for the Replicator Dynamic. Moreover, results about the stability of one system hold for the other system (Hofbauer (1981)). We refer the reader to Bomze (1983) for details. See also Hofbauer and Sigmund (1998).

We may, instead of the matrix in the main text, study the equivalent matrix, obtained by deleting the number μ (P) [R] from all entries the first (second) [third] column:

	A	R	M
R	0	0	0
M	$P-\mu$	0	$T-R$
A	$R-\mu$	$S-P$	0

Or, in abbreviated form,

	A	R	M
A	0	0	0
R	α	β	γ
M	δ	ϵ	θ

According to Bomze (1983), Proposition 6, part (ii), if the quantities $\beta\theta-\gamma\epsilon$, $\alpha\epsilon-\beta\delta$ and $\gamma\delta-\alpha\theta$ all have the same sign, then the Lotka-Volterra dynamic has a unique fixed point, given by $p = \frac{\gamma\delta-\alpha\theta}{\beta\theta-\gamma\epsilon}$ and $q = \frac{\alpha\epsilon-\beta\delta}{\beta\theta-\gamma\epsilon}$. We compute $\beta\theta-\gamma\epsilon = 0-(T-R)(S-P) > 0$, $\alpha\epsilon-\beta\delta = (P-\mu)(S-P) - 0 > 0$ and $\gamma\delta-\alpha\theta = (T-R)(R-\mu) - 0 > 0$. The last two expressions are signed using (5) above. Hence there is a unique fixed point, (p, q), where $p > 0$ and $q > 0$ and

$$p = \frac{(T-R)(R-\mu)}{(T-R)(P-S)} = \frac{R-\mu}{P-S} \quad \text{and} \quad q = \frac{(P-\mu)(S-P)}{(T-R)(P-S)} = \frac{\mu-P}{T-R}.$$

We may verify that when these expressions are used in (6), we get exactly the solutions above and in the main text. Bomze shows furthermore that if $\beta p + \theta q = 0$, then (p, q) is a center for the Lotka-Volterra system. Since $\beta = \theta = 0$, we have $\beta p + \theta q = 0$. We may therefore conclude that (p, q) is a center for the Lotka-Volterra system. This, in turn, allows us to conclude that our equilibrium x^*, given in (2), is a center for the Replicator Dynamic. Finally, in order to verify that $0 < x_i^* < 1$ for $i = R, M, A$, we may use the fact that $p > 0$, $q > 0$ and (6). ■

Proof of Proposition 2: It is not difficult to see, using (1), that $(1/2)(S+T) < P$ and $\lambda \in (0, \lambda')$ is equivalent to $\mu < P$. Whenever the latter holds, the M preference is a strict Nash equilibrium and hence an ESS. If $\lambda \in (\lambda', 1)$, then $\mu > P$ and Proposition 1 applies. ■

6 References

Axelrod, R. (1984): The Evolution of Co-operation, New York: Basic Books.

Becker, G. (1974): "A Theory of Social Interactions", Journal of Political Economy, 82, 1063-1093.

Becker, G. (1976): "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology", Journal of Economic Literature, 14 (3), 817-26.

Bergstrom, T. (1989): "A Fresh Look at the Rotten Kid Theorem–and Other Household Mysteries", Journal of Political Economy, 97 (5), 1138-59.

Bergstrom, T. (1995): "On the evolution of altruistic ethical rules for siblings", American Economic Review, 85, 58-81.

Bergstrom, T. (2002): "Evolution of Social Behavior: Individual and Group Selection", Journal of Economic Perspectives, 16 (2), 67-88.

Bergstrom, T. and Stark, O. (1993): "How altruism can prevail in an evolutionary environment", American Economic Review (Papers and Proceedings), 83, 149-155.

Bernheim, D.: and Stark, O. (1988): "Altruism Within the Family Reconsidered: Do Nice Guys Finish Last?", American Economic Review, 78 (5), 1034-1045.

Bernheim, D., Shleifer, A. and Summers, L. (1985): "The Manipulative Bequest Motive", Journal of Political Economy, 93, 1045-1076.

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

Bolle, F. (2000): "Ia altruism evolutionarily stable? And envy and malevolence?", Journal of Economic Behavior and Organization, 42, 131-133.

Bomze, I. (1983): "Lotka-Volterra Equation and Replicator Dynamics: A Two-Dimensional Classification", Biological Cybernetics, 48, 201-211.

Bruce, N. and Waldman, M. (1990): "The rotten-kid theorem meets the Samaritan's dilemma", Quarterly Journal of Economics, 105, 155-165.

Frank, R. (1987): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" American Economic Review, 77, 593-604.

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

Harsanyi, J. and Selten, R. (1988): A General Theory of Equilibrium Selection in Games, MIT Press.

Hirshleifer, J. (1977): "Shakespeare vs. Becker on Altruism: The Importance of Having the Last Word", Journal of Economic Literature, 15, 500-502

Hofbauer, J. (1981): "On the occurrence of limit cycles in the Volterra-Lotka equation", Nonlinear Analysis, Theory, Methods and Applications, 5, p. 1003-1007.

Hofbauer, J. and Sigmund, K. (1998): Evolutionary Games and Population Dynamics, Cambridge University Press.

Milchtaich, I. (2002): "How Does Selfishness Affect Well-Being?", Department of Economics, Bar-Ilan University, Israel.

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

Possajennikov, A. (2000): "On the evolutionary stability of altruistic and spiteful preferences", Journal of Economic Behavior and Organization, 42, 125-129.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics", American Economic Review, 83, 1281-1302.

Rapoport, A. and Chammah, A. (1965): Prisoner's Dilemma: A Study in Conflict and Cooperation, Ann-Arbor, the University of Michigan Press.

Schelling, T. (1978): "Altruism, Meanness, and Other Potentially Strategic Behaviors", American Economic Review (Papers and Proceedings), 68, 229-230.

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

Wilson, E. (1975): Sociobiology, Harvard University Press.

Witt, U. (1986): "Evolution and Stability of Cooperation Without Enforceable Contracts", Kyklos, 38, 245-265.