IS ALTRUISM EVOLUTIONARILY STABLE?

Helmut Bester and Werner Güth*†

November 1994

Abstract

We develop an evolutionary approach to explain altruistic preferences. Given their preferences, individuals interact rationally with each other. By comparing the success of players with different preferences, we investigate whether evolution favors altruistic or selfish attitudes. The outcome depends on whether the individuals' interactions are strategic complements or substitutes. Altruism and self-interest are context dependent.

Keywords: Altruism, Evolutionary Stability, Endogenous Preferences, Strategic Complements and Substitutes; *JEL Classification No.:* A13, C72, D64

*CentER, Tilburg University, and Humboldt University, Berlin; respectively. The second author is grateful to the Humboldt Foundation for supporting his visit at CentER, where this research was completed.

†Mailing address: Helmut Bester, CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

1 Introduction

There is an abundance of observations showing that individuals do not always pursue self-interest: People risking their own life to rescue others, soldiers voluntarily going to war, the many forms of charity etc. These observations can be made consistent with standard economic theory by postulating utility functions that include the well- being of others in addition to the own one. Yet, this only rephrases the question of why individuals behave in this way. Instead of explaining altruistic behavior, one now has to explain why people sometimes have altruistic preferences. It is this question that we want to address.

We adopt an evolutionary approach to investigate whether altruism may have evolved in humans through a process of natural or cultural selection. Formally, the degree of altruism is expressed by a preference parameter describing how much an individual cares for the success of others. The range of possible parameters includes pure self-interest as the special case where an agent's objective is identical to his private success. In the interaction with others, each player rationally selects a strategy to maximize his preferences. As a result, in equilibrium each player's effective success depends on the altruistic attitudes of all the involved players. This allows us to compare the success of players with different preference parameters. In an evolutionary environment, players with a higher expected success are less likely to be eliminated. Since success is related to preferences, we can study the question of whether evolution favors altruistic or selfish attitudes. Altruism is said to be evolutionarily stable if it survives evolutionary selection.

Instead of studying directly the evolution of behavior, which is the usual approach in evolutionary biology[1] and in evolutionary game theory (see, for instance, Hammerstein and Selten (1994)), we consider rational behavior for given preferences. These preferences determine the players' behavior and their effective success via their effect on the outcome of strategic interactions. By assuming rational behavior and applying the concept of evolutionary stability (see Maynard Smith (1982)) to preferences rather than to

strategies, we endogenously determine preferences. Our approach thus offers a way of endogenizing individual objective functions, which neoclassical theory usually treats as exogenous.[2]

Our analysis of individual interactions yields two insights: First, a comparison of the interaction between altruists and the interaction between egoists reveals that the altruists achieve a higher material payoff than the egoists. This is so because altruistic preferences internalize some externalities in the game between the players. Second, when an altruist interacts with an egoist, the altruist's material payoff is lower than the egoist's payoff. This finding is in line with the conventional view that altruistic preferences reduce the individual's success, while at the same time increasing the opponent's success.

The second result is often used as an argument that altruism cannot possibly evolve by natural selection. Yet, this argument does not directly address evolutionary considerations. For the process of natural selection, the relevant question is whether an egoistic mutant facing a population of altruists is more successful than the altruists among themselves. Altruism will be evolutionarily stable if an egoist in the interaction with an altruist receives a lower material payoff than an altruist. In our model, this depends on the strategic dependence between the players. Altruism turns out to be evolutionarily stable only if the game exhibits strategic complementarities in the sense of Bulow *et. al.* (1985). This suggests that preferences may be context dependent. Situational factors may decide whether individuals are motivated by altruism or self-interest.

As Frank (1987, 1988) and Schelling (1978), our analysis emphasizes the strategic role of preferences and emotions. A player's preferences affect not only his own equilibrium behavior but also the behavior of his opponent. Depending on the type of interaction, this effect can be either beneficial or harmful for a player with altruistic preferences. As a result, natural selection favors altruism in the case of strategic complements but not in the case of strategic substitutes. The strategic role of preferences distinguishes our approach from alternative explanations of altruism that rely on 'kin selection' arguments.

These arguments show that evolution can sustain altruism between genetically linked individuals (see, e.g. Bergstrom and Stark (1993)).

The following section describes the interaction between individuals and defines their success resulting from their behavior. Section 3 studies the interaction between egoistic players and discusses efficiency implications. Altruistic preferences are introduced in section 4, where we also study the impact of preferences on the equilibrium outcome. Section 5 investigates the evolutionary stability of altruism. In section 6 we extend our conclusions to a more general framework. Finally, section 7 concludes and discusses extensions.

2 Success and Behavior

Consider a population whose members interact with each other in pairs. All members are identical and so the interaction between a pair of individuals is described by a symmetric game. In this game, one of the players is labelled as player 1 and the other is labelled as player 2. Player 1's choice of action is denoted by $x \geq 0$; player 2 chooses some action $y \geq 0$. Each player's *material payoff* or *evolutionary success* depends on the joint actions (x, y) according to

$$U_1(x, y) \equiv x(ky + m - x), \quad U_2(x, y) \equiv y(kx + m - y), \tag{1}$$

where the parameters k and m are assumed to satisfy the restriction

$$-1 < k < 1 \quad \text{and} \quad m > 0. \tag{2}$$

The function $U_i(\cdot)$ does not necessarily represent player i's *subjective utility* or his *preferences*. In the interaction with his opponent, player i seeks to maximize the utility $V_i(x, y)$. The function $V_i(\cdot)$, which may differ from $U_i(\cdot)$, will be defined in Section 4. The parametric specification of payoffs allows us to derive a closed form solution for the players' equilibrium success. Assumption (2) together with the specification of the utility function $V_i(\cdot)$ in Section 4 guarantees that the equilibrium of the game between the two individuals is unique. Also, the constraints $x \geq 0$ and $y \geq 0$ are never binding.

The specification of material payoffs in (1) is sufficiently general to illustrate the main arguments of our analysis; we consider more general payoff functions in section 6.

By (1), each individual's success depends not only on his own action but also on the choice of the other player. If $k > 0$ the game exhibits *positive externalities* because a higher action by player i increases the success of player j. *Negative externalities* occur if $k < 0$. The simplest example is a production game with externalities, where x and y denote the players' effort or input decisions. Player 1's success can then be defined as the difference between his output, $x(ky+m)$, and his (quadratic) effort cost, x^2. Equivalently, in the presence of cost externalities, his output is mx and his cost is $x(ky - x)$. Positive production externalities may not only result from technological interdependence; they also occur when agents share the joint output of their individual production efforts or if their efforts contribute to the production of a public good. Negative cost externalities arise naturally when the players exploit a common resource.

Several authors have used evolutionary arguments to explain the market behavior of firms (see, e.g. Penrose (1952) and Winter (1971)). We can apply our approach to the standard models of oligopolistic competition by identifying a firm's success with its profits: Player 1 and 2 are engaged in a symmetric duopoly game with heterogeneous products and linear demand functions. In a Cournot market, the actions x and y would represent the firms' quantity choices. Their products are imperfect substitutes for all $k \in (-1, 0)$ and their prices are $m + ky - x$ and $m + kx - y$. In a Bertrand market, the firms would compete by choosing prices x and y, respectively. They face the demand functions $m + ky - x$ and $m + kx - y$ and their products are imperfect substitutes for all $k \in (0, 1)$. Thus, the payoffs in (1) can be interpreted as the Cournot or Bertrand profits in a symmetric duopoly market with zero production costs.

As a benchmark, we define a symmetric *optimum* by actions $(\hat{x}, \hat{y})$ that maximize the players' joint success, i.e.

$$(\hat{x}, \hat{y}) \in \underset{x,y}{\operatorname{argmax}}[U_1(x, y) + U_2(x, y)]. \tag{3}$$

Since $-1 < k < 1$, this optimum is well-defined and determined by the necessary and sufficient first-order conditions $x = (2ky + m)/2$ and $y = (2kx + m)/2$. Therefore,

$$\hat{x} = \hat{y} = \frac{m}{2-2k} \quad \text{and} \quad U_1(\hat{x}, \hat{y}) = U_2(\hat{x}, \hat{y}) = \frac{m^2}{4(1-k)}. \tag{4}$$

Typically, the presence of externalities prevents the players from reaching the outcome $(\hat{x}, \hat{y})$ through individual preference maximization. In the following two sections, we investigate the relation between altruistic preferences and this inefficiency.

3 Equilibrium with Egoistic Preferences

An egoistic player seeks to maximize his private success. He shows no concern for the success of his partner. That is, player i acts egoistically if his subjective utility satisfies $V_i(x, y) = U_i(x, y)$. Since the players interact non-cooperatively, each of them chooses his action taking the action of the other as given. This results in actions $(\tilde{x}, \tilde{y})$ that constitute a Nash equilibrium of the game. Accordingly,

$$\tilde{x} \in \underset{x}{\operatorname{argmax}} U_1(x, \tilde{y}), \quad \tilde{y} \in \underset{y}{\operatorname{argmax}} U_1(\tilde{x}, y). \tag{5}$$

The players' best-response functions are given by

$$\tilde{x} = 0.5(ky + m), \quad \tilde{y} = 0.5(kx + m). \tag{6}$$

To characterize the type of strategic interdependence, it will be useful to employ the terminology of Bulow *et. al.* (1985): For $k > 0$, the reaction functions are upwards sloping and the game exhibits *strategic complementarities*. If $k < 0$, the game exhibits *strategic substitutes* because the reaction functions are downwards sloping. For instance, the Cournot game discussed in Section 2 induces strategic substitutes, while the Bertrand game leads to strategic complements.

The players' reactions generate the following equilibrium actions and payoffs:

$$\tilde{x} = \tilde{y} = \frac{m}{2-k}, \quad U_1(\tilde{x}, \tilde{y}) = U_2(\tilde{x}, \tilde{y}) = \frac{m^2}{(2-k)^2}. \tag{7}$$

As a result $U_i(\tilde{x},\tilde{y}) < U_i(\hat{x},\hat{y}), i = 1,2$, unless $k = 0$. The reason is, of course, that with egoistic behavior each player ignores the impact of his action on the other player's success. This kind of externality explains why $\tilde{x} = \tilde{y} < \hat{x} = \hat{y}$ if $k > 0$, and $\tilde{x} = \tilde{y} > \hat{x} = \hat{y}$ if $k < 0$.

Usually, one defines evolutionary stability in terms of strategies rather than preferences.[3] The definition of an evolutionarily stable strategy implies immediately that this strategy constitutes a symmetric Nash equilibrium. In fact, if a symmetric Nash equilibrium is 'strict' in the sense that the players' best responses are unique, then the equilibrium strategy is also evolutionarily stable. By (6) the equilibrium $(\tilde{x},\tilde{y})$ is strict and, since $\tilde{x} = \tilde{y}$, it is symmetric. Therefore, only the strategy $\tilde{x}$ is evolutionarily stable. That is, only the egoistic behavior $\tilde{x} = \tilde{y}$ survives selection of the most successful strategy. To explain the evolution of altruism, one has to adopt an alternative method. This is done by our 'indirect' evolutionary approach, which applies the idea of evolutionary selection to preferences instead of strategies.

4 Altruistic Preferences

A player is altruistic when his preferences reflect some concern for the other player's success. We describe such preferences by

$$V_1(x,y) = \alpha U_1(x,y) + (1-\alpha)U_2(x,y), \quad V_2(x,y) = \beta U_2(x,y) + (1-\beta)U_1(x,y). \tag{8}$$

Accordingly, the concern that players 1 and 2 express for the other player's success is represented by the weights $1-\alpha$ and $1-\beta$, respectively. If $\alpha, \beta < 1$, the players are said to be *altruistic*. This formulation of altruism has been employed already by Edgeworth (1881, p. 53), who called the values $(1-\alpha)/\alpha$ and $(1-\beta)/\beta$ the 'coefficients of effective sympathy'. In what follows, we restrict these coefficients to lie in the unit interval by considering only values of α and β such that

$$1/2 \leq \alpha \leq 1, \quad 1/2 \leq \beta \leq 1. \tag{9}$$

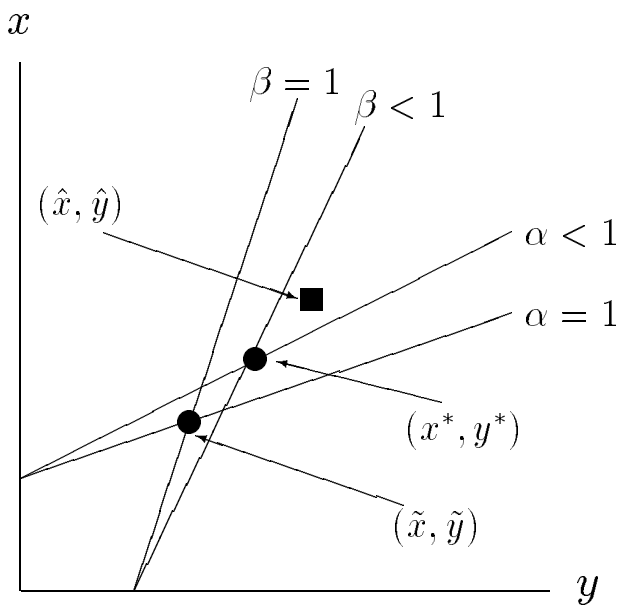

Figure 1: *Altruism and Equilibrium Behavior*

That is, each player is taken to weigh his own success at least as much as the opponent's success. We maintain the assumption of common knowledge to analyse the game between the $\alpha-$ and the $\beta-$player. This means, not only the material payoffs but also the preference parameters α and β are commonly known by the players. We will discuss the significance of this assumption later in this section and in section 7.

Altruism influences the strategic interactions between the players even though it does not directly affect their success as defined by (1). Altruism has an indirect impact on the players' success since their behavior depends on the parameters α and β. Each player seeks to maximize his subjective preferences so that

$$x^* \in \underset{x}{\operatorname{argmax}}\, V_1(x, y^*), \quad y^* \in \underset{y}{\operatorname{argmax}}\, V_2(x^*, y). \tag{10}$$

From the first-order conditions for preference maximization we can derive the players' best response functions:

$$x = \frac{ky + \alpha m}{2\alpha}, \quad y = \frac{kx + \beta m}{2\beta}. \tag{11}$$

Altruistic preferences shift a player's best response function upwards if $k > 0$, and downwards if $k < 0$. For $k > 0$, this is illustrated in Figure 1. The figure depicts player 1's best

response for given parameter values $\alpha = 1$ and $\alpha < 1$. Similarly, player 2's behavior is represented for two different values of β. The intuition for this effect is that an altruistic player internalizes, at least partially, the externality of his behavior on the other player's success. This induces him to select a higher action with positive externalities and a lower action with negative externalities.

The equilibrium of the game between two players with preference parameters α and β, respectively, is given by

$$x^*(\alpha, \beta) = \frac{\beta m(2\alpha + k)}{4\alpha\beta - k^2}, \quad y^*(\alpha, \beta) = \frac{\alpha m(2\beta + k)}{4\alpha\beta - k^2}. \tag{12}$$

In Figure 1 the equilibrium is determined as the intersection of the players' best response function. For $\alpha = \beta = 1$, this is the point $(\tilde{x}, \tilde{y})$. In a game between two altruists (x^*, y^*) is realized as $\alpha, \beta < 1$. As the figure illustrates, this outcome is closer to the optimum $(\hat{x}, \hat{y})$ than the egoistic equilibrium $(\tilde{x}, \tilde{y})$.

For what follows, it is important to notice that each player's preference parameter not only affects his own equilibrium behavior but also the opponent's choice of action. Since the players are engaged in a non-cooperative game, each of them bases his decision on his knowledge about the other player's attitudes. Following Schelling (1978, p. 229), who calls "a behavioral propensity [...] strategic if it influences others by affecting their expectations," we refer to the dependence of x^* on β and of y^* on α as the 'strategic' effect of altruism. This strategic effect is consistent with the psychological evidence that individuals do not act uniformly against other individuals; rather, they condition their own behavior on the attitudes of those with whom they interact.[4] This, of course, presumes that they can anticipate the attitudes of their opponent. Preferences can have a strategic effect only if, at least to some extent, they are communicated to the other player. In this sense, our analysis refers to environments where the players learn about each other before choosing their actions. In fact, as Frank (1987, 1988) argues, many observable physical symptoms may provide some indication of a person's affective condition. These symptoms include posture, the rate of respiration, the pitch and timbre of the voice, and facial muscle tone and expression. Also, he reports some experimental

evidence that, even on the basis of brief encounters involving strangers, subjects are adept at predicting the behavior of their opponent (see Frank (1988), ch. 7).

Using (12), we can determine the direction of the strategic effect. As $\partial x^*/\partial\beta < 0$ and $\partial y^*/\partial\alpha < 0$ for all $k \neq 0$, the opponent's equilibrium action will be the higher the more altruistically inclined a player is. In the case of positive externalities ($k > 0$), therefore, the strategic effect has a positive impact on the altruistic player's success. Effectively, the opponent will choose a more favorable action because he interacts with an altruist. If $k < 0$, however, the strategic effect turns out to be disadvantageous: Player i's altruism induces player j to choose a higher action and this reduces player j's success.

Can a population of altruistic players reach a higher level of success than egoistic players? We answer this question by considering the outcome of the interaction between two players with identical preference parameters. A comparison with the egoistic outcome and the symmetric optimum shows that, for all $1/2 < \alpha < 1$,

$$\tilde{x} = \tilde{y} < x^*(\alpha,\alpha) = y^*(\alpha,\alpha) < \hat{x} = \hat{y}, \quad \text{if} \quad k > 0; \tag{13}$$

$$\tilde{x} = \tilde{y} > x^*(\alpha,\alpha) = y^*(\alpha,\alpha) > \hat{x} = \hat{y}, \quad \text{if} \quad k < 0. \tag{14}$$

Altruism shifts the equilibrium outcome closer towards optimal behavior. In fact, in the extreme case $\alpha = 1/2$ the players' equilibrium actions become identical to the symmetric optimum. This has the following implication for the players' success.

Proposition 1: *Let $k \neq 0$. Then a population of altruistic players reaches a higher level of success than a population of egoists, i.e. $U_i(\tilde{x},\tilde{y}) < U_i(x^*(\alpha,\alpha), y^*(\alpha,\alpha))$, $i = 1,2$, if $\alpha < 1$.*

Proposition 1 shows that altruism produces more efficient outcomes.[5] Yet, this does not mean that an altruistically inclined actor is more successful than a player who acts on egoistic principles. Indeed, the conventional view is that altruistic behavior reduces the actor's success while enhancing the success of others. The following result, which

follows from Lemma 7 in the Appendix, confirms this intuition.

Proposition 2: *Let* $k \neq 0$. *Then in the interaction between two players, the more altruistically motivated player is less successful than his opponent. That is* $U_1(x^*(\alpha,\beta), y^*(\alpha,\beta)) < U_2(x^*(\alpha,\beta), y^*(\alpha,\beta))$, *for all* $\alpha < \beta$.

An altruist is willing to reduce his own success in order to increase the success of others. Therefore, one might conclude that self-interest has a higher survival value than altruism. Yet, Proposition 2 presents only one consideration that is important for evolutionary selection. As Proposition 1 indicates, a population consisting largely of altruists will perform better than a population of egoists. An egoist within a population of altruists may have a relatively low expected success because the altruists among themselves attain a higher level of success than the egoist against the altruists. In fact, even an altruist interacting with an egoist may have a higher success than an egoist who faces another egoist. In the following section we will address the issue of evolutionary preference selection by using the concept of evolutionary stability.

5 The Stability of Altruism

Can altruism emerge in an evolutionary process where only the most successful players survive? By Proposition 1, a population of altruists is more successful than a population of egoists. But this does not necessarily mean that altruism is evolutionarily stable. When an egoist invades a population of altruists and performs better than his opponents, then egoism will spread out and eliminate altruistic behavior in the process of evolution. Conversely, an altruist may successfully invade a population of egoists if he does better than the egoists against each other.

To study the evolutionary stability of altruism, we employ the 'indirect' evolutionary approach, which is schematically presented in Figure 2. In the previous section we studied the equilibrium behavior of two players with preference parameters α and β, respectively. This equilibrium determines each player's success. In an environment where

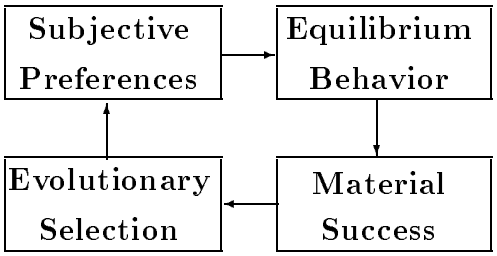

Figure 2: *Evolutionary Selection of Preferences*

evolutionary selection favors the more successful players, players with lower material payoffs will become extinct. In this way, preferences are selected for their capacity to generate material payoffs.

To complete our analysis, we investigate whether a monomorphic population of players with parameter α^* is immune against invading mutant players with a different preference parameter. In what follows, $R(\alpha, \beta)$ denotes a player's success when he has the altruism parameter α while his opponent has the parameter β. Since the interaction between these players results in the equilibrium $(x^*(\alpha, \beta), y^*(\alpha, \beta))$, we get

$$R(\alpha, \beta) \equiv U_1(x^*(\alpha, \beta), y^*(\alpha, \beta)). \tag{15}$$

The mutant space $M = [1/2, 1]$ is the set of all possible values for the parameters α and β. The function $R(\cdot)$ together with the set M defines a symmetric evolutionary game. This game allows us to study the evolutionary stability of a preference parameter by using the concept of evolutionarily stable strategies (ESS, see Maynard Smith (1982)).

Definition: A preference parameter $\alpha^* \in [1/2, 1]$ is called *evolutionarily stable* if

$$R(\alpha^*, \alpha^*) \geq R(\alpha, \alpha^*) \qquad \text{for all} \quad \alpha \in [1/2, 1]; \text{ and} \tag{16}$$

$$R(\alpha^*, \alpha) > R(\alpha, \alpha) \qquad \text{whenever} \quad R(\alpha^*, \alpha^*) = R(\alpha, \alpha^*). \tag{17}$$

These conditions capture the idea that a population with parameter α^* cannot be invaded by a small minority with deviant parameter α : According to the first require-

ment, an evolutionarily stable parameter α^* is a best reply against itself. Any $\alpha-$mutant invading a society of α^*-players cannot be more successful than the members of the society. If several parameters are equally successful, the second condition rules out that an alternative best reply $\alpha \neq \alpha^*$ can spread out in the population: Since α^* is better against α than α itself, α will be eliminated as soon as it becomes more frequent within the population.

The ESS-concept originates from biology and is based on the idea that higher success reflects an advantage in reproducing. In an economic context, of course, success is mostly identified with monetary payoffs. One can directly extend the idea of evolutionary stability to this context when monetary payoff is an important determinant for reproductive success. Indeed, some empirical evidence indicates that, over the more recent human history, individual wealth has been positively related to the number of surviving offspring.[6] For the economist, however, the social mechanisms of learning and imitation are probably more important than the genetic mechanism: Evolutionary selection occurs because successful behavioral attitudes tend to be imitated. Individual traits that yield lower payoffs will, therefore, be driven out by more successful traits. In this way, imitation may induce a process that resembles natural selection or the 'survival of the fittest'.[7]

We first apply the ESS-concept to the case $k > 0$, where the players' interactions exhibit strategic complementarities. By Lemmas 4 - 6 in the Appendix we get the following result.

Proposition 3: *Let $k > 0$. Then $\alpha^* = (2-k)/2$ is the unique evolutionarily stable preference parameter.*

As $\alpha^* < 1$, evolutionarily stable preferences must exhibit some degree of altruism. The level of altruism is positively related to the parameter k. Altruism becomes more important when the strategic interdependence between the players is relatively high. In fact, $k \to 1$ implies $\alpha^* \to 1/2$.

Why can an egoistic mutant with $\alpha > \alpha^*$ not invade a population of α^*-individuals? Actually, by Proposition 2, such a mutant has a higher success than the member of the population with whom he interacts. Yet, the low payoff of an α^*-player against an invading mutant is less important for evolutionary considerations. For the members of the population the likelihood of interacting with the mutant is small. Mostly they interact with each other and so, by Proposition 1, their expected level of success is relatively high. The parameter $\alpha^* < 1$ is evolutionarily stable because in the game between a pair of α^*- individuals each of them gets a larger material payoff than the egoist against an α^*-opponent.

Also, the uniqueness of α^* in Proposition 3 implies that a population of egoists is vulnerable against invasion by altruistic agents. This is so because the interaction between the egoists results in low payoffs. If an altruist enters a population of egoists, his payoff will be lower than the one of his partner. Nonetheless, he is still more successful than all the other egoists who have an egoist as their partner. This happens because preferences have a strategic effect: As it was shown in the foregoing section, the altruist induces his opponent to increase his action level. In the case of positive externalities, this is beneficial and so he will succeed in invading a society of egoists.

The last argument indicates that the sign of k may be important for the evolution of altruism. In fact, the following result reveals that egoistic preferences are the unique evolutionary outcome in the case of strategic substitutes.

Proposition 4: *Let $k < 0$. Then $\alpha^* = 1$ is the unique evolutionarily stable preference parameter.*

The proof of this statement follows immediately from Lemmas 4 - 6 in the Appendix. Propositions 3 and 4 reveal that the survival of altruism depends on the environment. It is consistent with evolutionary stability that individuals behave altruistically in some

situations and egoistically in others. Our simple example of strategic interactions demonstrates this by the dependence of α^* upon the parameter k. When $k < 0$, a population of egoists will defeat entry of altruism. On average the egoists will be more successful than an invading mutant. In fact, an altruistic entrant will suffer for two reasons. First, his choice of action does not aim at maximizing private success. Second, as $k < 0$, the strategic effect of his attitude turns out to be harmful. His egoistic opponent will choose a higher action level when facing an altruist. In the presence of negative externalities this lowers the altruist's payoff.

In terms of the examples discussed in section 2, our analysis shows that altruism emerges in the presence of positive production externalities, in the case of output sharing, or in a Bertrand market. Self-interest is stable in an environment with negative production externalities, with common resource exploitation, and in a Cournot market. The strategic effect of preferences explains why altruism is evolutionarily stable for $k > 0$, whereas egoism is evolutionarily stable for $k < 0$. When altruism induces a harmful reaction by the other player, one is better off by egoistically maximizing private success. Altruism may emerge only if its strategic effect is beneficial. In this situation, the evolutionarily stable parameter α^* is determined by the following tradeoff. On the one hand, the altruist reduces his success by choosing an action that reflects some concern for the other player's success; on the other hand, his attitude causes a favorable reaction by the other player. The latter effect becomes more important for larger values of k. Therefore, α^* and k are negatively related.

6 A Generalization

In the previous sections we have employed the parametric specification of material payoffs in equation (1) to study the evolutionary stability of altruism. For this specification, the equilibrium defined by (10) is uniquely determined so that the function $R(\cdot, \cdot)$ in (15) is unambiguously defined. Moreover, there is a unique preference parameter α^* satisfying conditions (16) and (17) for evolutionary stability. For general success functions,

this may no longer be the case. Even when the evolutionary game in section 5 is well-defined, it may happen that the parameter α^* is not unique or that an evolutionarily stable parameter does not exist. Nonetheless, we can extend our main conclusions to a more general framework.

In this section, we will generalize the specification of material payoffs. As before, we consider a monomorphic population of players who interact pairwise. The game between any pair of players, say player 1 and player 2, determines their evolutionary success. This is represented by the functions $U_1(x,y)$ and $U_2(x,y)$, where x and y are the actions chosen by player 1 and 2, respectively. The game is symmetric in the sense that $U_1(x,y) = U_2(y,x)$. By symmetry, the payoff of a strategy is independent of whether the player acts in the role of player 1 or player 2. We assume $U_1(\cdot,\cdot)$ to be strictly concave and twice differentiable.

To characterize the interaction between the players, we extend the terminology in the previous sections to the more general case. Let the signs of $\partial U_1(x,y)/\partial y$ and $\partial^2 U_1(x,y)/\partial x \partial y$ be constant for all $(x,y) \geq 0$. The game is said to exhibit *positive externalities* if $\partial U_1(x,y)/\partial y > 0$ and *negative externalities* if $\partial U_1(x,y)/\partial y < 0$. The players face *strategic complementarities* if $\partial^2 U_1(x,y)/\partial x \partial y > 0$ and *strategic substitutes* if $\partial^2 U_1(x,y)/\partial x \partial y < 0$. We focus on situations where $\partial U_1(x,y)/\partial y \neq 0$ and $\partial^2 U_1(x,y)/\partial x \partial y \neq 0$ for all $(x,y) \geq 0$.

Whenever two individuals interact with each other, each player i seeks to maximize his subjective utility $V_i(x,y)$, as defined by (8). An equilibrium is a pair of actions, $(x^*(\alpha,\beta), y^*(\alpha,\beta))$, that satisfies condition (10). As long as $\partial U_1(0,y^*)/\partial x$ is sufficiently large, the equilibrium actions satisfy $x^* > 0$ and $y^* > 0$ so that they can be derived from the first order conditions

$$\frac{\partial V_1(x^*(\alpha,\beta), y^*(\alpha,\beta))}{\partial x} = 0, \quad \frac{\partial V_2(x^*(\alpha,\beta), y^*(\alpha,\beta))}{\partial y} = 0. \tag{18}$$

To ensure that $R(\alpha,\beta)$ in (15) is well-defined, we assume that (18) has a unique solution $(x^*(\alpha,\beta), y^*(\alpha,\beta))$.[8]

The evolutionary success of preferences depends on their impact on equilibrium behavior. By differentiating (18), we obtain for $\alpha \approx \beta$ that[9]

$$\text{sign}\left[\frac{\mathrm{d}x^*(\alpha,\beta)}{\mathrm{d}\alpha}\right] = \text{sign}\left[-\frac{\partial U_1(x^*,y^*)}{\partial y}\right] \qquad \text{and} \tag{19}$$

$$\text{sign}\left[\frac{\mathrm{d}y^*(\alpha,\beta)}{\mathrm{d}\alpha}\right] = \text{sign}\left[-\frac{\partial U_1(x^*,y^*)}{\partial y}\frac{\partial^2 U_1(x^*,y^*)}{\partial x \partial y}\right]. \tag{20}$$

The effect described by (19) has the same intuition as in the more special case studied before. The more altruistic player 1 is, i.e. the lower α is, the more he tends to internalize the externality of his action upon the other player's utility. Therefore, α and x^* are negatively related in games with positive externalities and positively related in games with negative externalities. Equation (20) generalizes the strategic effect discussed in the previous sections. As before, altruism induces the opponent to select a higher action when both the externalities and the strategic interdependence between the players' actions have the same sign. But, in the more general situation considered here, altruism may also reduce the other player's equilibrium action. This happens in games with strategic complements when the externalities are negative and in games with strategic substitutes when the externalities are positive.

How does the strategic effect influence a player's success? Consider a game with strategic complements. If the game has positive externalities, then altruism induces the opponent to choose a higher action level. Clearly, this effect is beneficial for the altruist's success since, in the presence of positive externalities, raising the other player's action level increases his own success. Similarly, in a game with strategic complements and negative externalities altruism reduces the opponent's action level. Again, the altruist gains from the strategic effect. In games with strategic substitutes this conclusion is reversed. For instance, in games with strategic substitutes and negative externalities, altruism increases the other player's action. This is harmful because it creates a negative externality. In summary, the strategic effect of altruism on the other player's behavior is beneficial in games with strategic complements and harmful in games with strategic substitutes.

Our previous analysis suggests that altruism survives evolutionary selection only if it is associated with a beneficial strategic effect. The following result, which is proved in the Appendix, extends this conclusion to the more general environment.

Proposition 5: *The preference parameter $\alpha^* = 1$ is evolutionarily stable only if the interactions between individuals are characterized by strategic substitutes. A parameter $\alpha^* < 1$ is evolutionarily stable only if these interactions involve strategic complements.*

This result is weaker than Propositions 3 and 4 because it does not establish the existence or uniqueness of an evolutionarily stable preference parameter. Nonetheless, it shows that pure egoism cannot evolve in an environment with strategic complementarities. In such an environment only some form of altruism has the potential to survive evolutionary selection. Conversely, altruism will not survive the invasion of egoistic mutants when interactions exhibit strategic substitutes. The evolution of altruism requires an environment of strategic complements. Altogether, one should not expect evolution to result in a society where individuals *always* either pursue pure self-interest or care for the well-being of others. Instead, evolutionary arguments suggest that these attitudes will be contingent on the strategic interdependence between individual behaviors.

7 Conclusions

Unlike other evolutionary studies of altruistic behavior in strategic interaction, our indirect evolutionary approach does not deny rational decision making. In principle it allows for any hypotheses specifying how stimuli, e.g. preferences, influence behavior. A process of natural or cultural selection then determines which stimuli emerge in the course of evolution. Our study employs the usual rationality assumptions of game theory to endogenize preferences, which neoclassical theory typically treats as exogenous. In this sense, the indirect evolutionary approach generalizes neoclassical theory.

The most important finding of our study is that evolutionarily stable altruism depends on the type of strategic interaction, as expressed by the signs of the derivatives of material payoffs $U_i(x, y)$. Although in our context altruism always produces more efficient outcomes, it is evolutionarily stable only if it induces the interaction partner to respond favorably. As the evolution of preferences depends on this strategic effect, one may expect altruism to mitigate inefficiencies only when interactions can be characterized as strategic complements.

Another requirement for the evolution of altruism is related to the individuals' information about preferences. Our analysis employs the usual common knowledge assumption of game theory, which implies that the preference parameters α and β are commonly known. To illustrate the possible impact of incomplete information, consider a monomorphic population of altruists with parameter $\alpha < 1$. If now an egoistic mutant appears, each altruist will consider the probability of interacting with the mutant as negligible. Under incomplete information, the egoist will be treated as an altruist and he will earn a higher material payoff than his altruistic encounter. As result, altruism will become vulnerable against egoistic mutants.

Our analysis, therefore, suggests that altruism is more likely to emerge in societies where individuals are not anonymous. For instance, altruism may be restricted to relatives and close friends. In contrast with the kin-selection selection argument, in our framework this happens not because family members are genetically linked but because they are better informed about each other. Nonetheless, even when preferences are not directly known, altruism may evolve if there are signals that indicate a person's attitude. In addition to the physical symptoms mentioned by Frank (1987, 1988), for instance donations to charities might signal altrustic preferences. An egoist is less willing than an altruist to donate. If imitation is too costly for the egoist, donations can become a credible signal of altruism.

Appendix

Lemma 1: *Let* $\partial R(\alpha^*, \alpha^*)/\partial\alpha = 0$. *Then either* $\alpha^* = -k/2$ *or* $\alpha^* = (2-k)/2$.

Proof: By definition,

$$R(\alpha, \beta) = \frac{\beta m^2 (k + 2\alpha)[k^2(\alpha - 1) + \beta k(2\alpha - 1) + 2\alpha\beta]}{(k^2 - 4\alpha\beta)^2}. \tag{21}$$

Therefore, $\partial R(\alpha, \beta)/\partial\alpha = 0$ is equivalent to

$$\alpha = [4\beta + k(2-k)]/[4(\beta + k)]. \tag{22}$$

Setting $\alpha = \beta = \alpha^*$ and solving the resulting quadratic equation for α^* leads to the two solutions stated in the Lemma. Q.E.D.

Lemma 2: *The parameter* $\bar{\alpha} = -k/2$ *is not evolutionarily stable.*

Proof: Since $R(\alpha, -k/2) = m^2/4$, $\bar{\alpha} = -k/2$ satisfies the first requirement of stability. The second requirement, $R(\bar{\alpha}, \alpha) > R(\alpha, \alpha)$ is equivalent to the condition $\alpha(k+1) < k$. As $\alpha(k+1) > 0$, this implies $k > 0$. But then $\bar{\alpha} < 0$, a contradiction. Q.E.D.

Lemma 3: *The parameter* $\bar{\alpha} = 1/2$ *is not evolutionarily stable.*

Proof: Straightforward calculations show that for $\alpha > \bar{\alpha} = 1/2$ the requirement $R(\bar{\alpha}, \bar{\alpha}) \geq R(\alpha, \bar{\alpha})$ is equivalent to

$$[k^2 + k(2\alpha - 1) - 1]/[k - 1] \leq 0. \tag{23}$$

As $k < 1$ this is equivalent to $k^2 + k(2\alpha - 1) \geq 1$. If $k > 0$, this condition cannot hold for α close enough to $1/2$. If $k < 0$, then (23) holds for $\alpha = 1$ only if $k^2 + k \geq 1$. But for $-1 < k < 0$ one cannot have $k^2 + k \geq 1$. This proves that $\bar{\alpha} = 1/2$ does not satisfy the first requirement of evolutionary stability. Q.E.D.

Lemma 4: *Let* α^* *be evolutionarily stable. Then either* $\alpha^* = 1$ *or* $\alpha^* = (2-k)/2$.

Proof: The statement simply follows from the fact that by the first requirement of evolutionary stability one must have $\partial R(\alpha^*, \alpha^*)/\partial\alpha = 0$ whenever $1/2 < \alpha^* < 1$. By Lemma

1, this equality has exactly two solutions,$-k/2$ and $(2-k)/2$. Lemmas 2 and 3 eliminate the possibility that $\alpha = -k/2$ or $\alpha = 1/2$ are evolutionary stable. This leaves only the two values $\alpha^* = 1$ and $\alpha^* = (2-k)/2$ as candidates for evolutionary stability. Q.E.D.

Lemma 5: *The parameter $\alpha^* = (2-k)/2$ is evolutionarily stable if and only if $k > 0$.*

Proof: Note that, by assumption (2), $\alpha^* \in [1/2, 1]$ if and only if $k \geq 0$. Straightforward calculations show that for $\alpha^* = (2-k)/2$ the inequality $R(\alpha^*, \alpha^*) \geq R(\alpha, \alpha^*)$ is equivalent to

$$[k^2 - 4][k - 2(1-\alpha)]^2/[k-1] \geq 0. \tag{24}$$

By assumption (2) this inequality is always satisfied. The inequality also shows that $R(\alpha^*, \alpha^*) > R(\alpha, \alpha^*)$ for $\alpha \neq \alpha^*$. This proves that also the second requirement of evolutionary stability is satisfied. Q.E.D.

Lemma 6: *The parameter $\alpha^* = 1$ is evolutionarily stable if and only if $k < 0$.*

Proof: Straightforward calculations show that for $\alpha < \alpha^* = 1$ the requirement $R(\alpha^*, \alpha^*) \geq R(\alpha, \alpha^*)$ is equivalent to

$$k^3 - 2k^2(1-\alpha) - 4\alpha k + 4(1-\alpha) \geq 0. \tag{25}$$

For $k \in (0, 1)$ this condition does not hold for α close enough to unity. But for $k \in (-1, 0]$ it holds for all $\alpha \in [1/2, 1]$ so that the first requirement of evolutionary stability is satisfied. Indeed, since the strict inequality holds in (25) for $\alpha < 1$, one has $R(\alpha^*, \alpha^*) > R(\alpha, \alpha^*)$. Therefore, also the second requirement of evolutionary stability is satisfied. Q.E.D.

Lemma 7: *$U_1(x^*(\alpha, \beta), y^*(\alpha, \beta)) < U_2(x^*(\alpha, \beta), y^*(\alpha, \beta))$, for all $\alpha < \beta, k \neq 0$.*

Proof: By symmetry of the functions $U_1(\cdot), U_2(\cdot)$ and by definition of $R(\alpha, \beta)$, the statement of the Lemma is equivalent to $R(\alpha, \beta) < R(\beta, \alpha), \alpha < \beta$. Using the expression for $R(\cdot)$ from Lemma 1, this is equivalent to $k^2(k + \alpha + \beta) > 0$. By (2) and (9), this inequality is always satisfied if $k \neq 0$. Q.E.D.

Proof of Proposition 5: By (15) one has

$$\frac{\mathrm{d}R(\alpha,\beta)}{\mathrm{d}\alpha} = \frac{\partial U_1(x^*,y^*)}{\partial x}\frac{\mathrm{d}x^*(\alpha,\beta)}{\mathrm{d}\alpha} + \frac{\partial U_1(x^*,y^*)}{\partial y}\frac{\mathrm{d}y^*(\alpha,\beta)}{\mathrm{d}\alpha}. \tag{26}$$

Suppose that the game exhibits strategic complements and that $\alpha^* = 1$. Then $\alpha^* = 1$ together with (8) and (18) implies $\partial U_1(x^*,y^*)/\partial x = 0$. This in combination with $\partial^2 U_1(x,y)/\partial x \partial y > 0$, and (20) implies $\mathrm{d}R(\alpha^*,\alpha^*)/\mathrm{d}\alpha < 0$. Thus for some $\alpha < 1$ in the neighborhood of α^* one gets $R(\alpha^*,\alpha^*) < R(\alpha,\alpha^*)$, a contradiction to requirement (16). This proves that $\alpha^* = 1$ only if the game exhibits strategic substitutes.

Now suppose that the game exhibits strategic substitutes and that $\alpha^* < 1$. Then (8) and (18) imply $\partial U_1(x^*,y^*)/\partial x = -(1-\alpha^*)/\alpha^* \cdot \partial U_2(x^*,y^*)/\partial x$. By (19) and symmetry of $U_i(\cdot)$, this yields $\partial U_1(x^*,y^*)/\partial x \cdot \mathrm{d}x^*/\mathrm{d}\alpha > 0$. Similarly, $\partial^2 U_1(x,y)/\partial x \partial y < 0$ and (20) imply $\partial U_1(x^*,y^*)/\partial y \cdot \mathrm{d}y^*/\mathrm{d}\alpha > 0$. Therefore, $\mathrm{d}R(\alpha^*,\alpha^*)/\mathrm{d}\alpha > 0$. Thus for some $\alpha > \alpha^*$ in the neighborhood of α^* one gets $R(\alpha^*,\alpha^*) < R(\alpha,\alpha^*)$, a contradiction to requirement (16). This proves that $\alpha^* < 1$ only if the game exhibits strategic complements. Q.E.D.

Footnotes

1. Note that also in evolutionary biology one often considers the assumption of genetically determined behavior as questionable (see van Lawick-Goodall (1974)). Higher developed species like mammals live in such a complex and stochastic environment that a genetically determined reaction behavior to all circumstances appears to be impossible.

2. The exceptions include Becker (1976), Frank (1987) and, more recently, Güth and Yaari (1992), Güth and Kliemt (1994), Hanson and Stuart (1990), Rabin (1993), Rogers (1994), and Waldman (1994).

3. A strategy x^s is evolutionarily stable if (i) $U_1(x^s, x^s) \geq U_1(y, x^s)$ for all y; and (ii) $U_1(x^s, y) > U_1(y, y)$ whenever $U_1(x^s, x^s) = U_1(y, x^s)$.

4. For a brief presentation of some evidence, see Rabin (1993) who incorporates these facts by deriving a 'psychological game' from basic 'material games'.

5. A setting in which altruism induces inefficient behavior is studied by Lindbeck and Weibull (1988). For a discussion of the efficiency aspects of altruism, see also Friedman (1988).

6. See, e.g., Chagnon and Irons (1979) or Boyer (1989).

7. See, e.g., Mailath (1992) and Selten (1991) for a discussion. Björnerstedt and Weibull (1994) show that population dynamics based on imitation may be closely related to biological dynamics.

8. Friedman (1986, p.42ff) presents conditions guaranteeing a unique equilibrium.

9. In the derivation of (19) and (20) we use the symmetry of the game and the fact that $V_i(\cdot)$ is strictly concave.

References

Becker, Gary S., "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology," *Journal of Economic Literature* 14(3), September 1976, 817-826.

Bergstrom, Theodore C. and Stark, Oded, "How Altruism Can Prevail in an Evolutionary Environment," *American Economic Review* (Papers and Proceedings) 83(2), May 1993, 149-155.

Björnerstedt, Jonas and Weibull, Jörgen, "Nash Equilibrium and Evolution by Imitation," mimeo, Dept. of Economics, Stockholm University, February 1994.

Boyer, George, "Malthus Was Right After All: Poor Relief and Birth Rates in Southeastern England," *Journal of Political Economy* 97(1), February 1989, 93-114.

Bulow, Jeremy I., Geanakoplos, John D., and Paul D. Klemperer, "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy* 93(3), June 1985, 488-511.

Chagnon, Napoleon A. and Irons, William, eds. *Evolutionary Biology and Human Social Behavior,* North Scituate, MA: Duxbury, 1979.

Edgeworth, Francis Y., *Mathematical Physics,* London: P. Kegan, 1881,

Frank, Robert H., "If Homo Economicus Could Choose His Own Utility Function, Would He Want One With a Conscience?," *American Economic Review* 77(4), September 1987, 593-604.

Frank, Robert H., *Passions Within Reason,* New York: W. W. Norton, 1988.

Friedman, David D., "Does Altruism Produce Efficient Outcomes? Marshall versus Kaldor," *Journal of Legal Studies* 17(1), January 1988, 1-13.

Friedman, James W., *Game Theory with Applications to Economics,* New York - Oxford: Oxford University Press, 1986.

Güth, Werner and Yaari, Menahem, "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," in *Explaining Process and Change - Approaches to Evolutionary Economics,* edited by Ulrich Witt, Ann Arbor: The University of Michigan Press, 1992, 23-34.

Güth, Werner and Kliemt, Hartmut "Competition or Cooperation: On the Evolu-

tionary Economics of Trust," *Metroeconomica* 45(3), October 1994, 155-187.

Hammerstein, Peter and Selten, Reinhard, "Game Theory and Evolutionary Biology," Chapter 28 in *Handbook of Game Theory with Economic Applications, Vol. 2,* edited by Robert J. Aumann and Sergiu Hart, Amsterdam: Elsevier, 1994, 929-993.

Hanson, Ingemar and Stuart, Charles, "Malthusian Selection of Preferences," *American Economic Review* 80(3), June 1990, 529-544.

Lindbeck, Assar and Weibull, Jörgen, "Altruism and Time Consistency: The Economics of Fait Accompli," *Journal of Political Economy* 96(6), December 1988, 1165-1182.

Mailath, George J., "Introduction: Symposium on Evolutionary Game Theory," *Journal of Economic Theory* 57(2), August 1992, 259-277.

Maynard Smith, John, *Evolution and the Theory of Games,* Cambridge: Cambridge University Press, 1982.

Penrose, Edith, "Biological Analogies in the Theory of the Firm," *American Economic Review* 42(5), December 1952, 804-819.

Rabin, Matthew, "Incorporating Fairness into Game Theory and Economics," *American Economic Review* 83(5), December 1993, 1281-1302.

Rogers, Alan R., "Evolution of Time Preference by Natural Selection," *American Economic Review* 84(3), June 1994, 460-481.

Selten, Reinhard, "Evolution, Learning, and Economic Behavior," *Games and Economic Behavior* 3(1), February 1991, 3-24.

Schelling, Thomas C., "Altruism, Meanness, and Other Potentially Strategic Behaviors," *American Economic Review* (Papers and Proceedings) 68(2), May 1987, 229-230.

van Lawick-Goodall, Jane, *In the Shadow of Man,* London: Fontana, 1974.

Waldman, Michael, "Systematic Errors and the Theory of Natural Selection," *American Economic Review* 84(3), June 1994, 482-497.

Winter, Sidney G., "Satisficing, Selection, and the Innovating Remnant," *Quarterly Journal of Economics* 85(2), May 1971, 237-261.