J. theor. Biol. (1998) **194,** 561–574
Article No. jt980775

The Dynamics of Indirect Reciprocity

MARTIN A. NOWAK*‡ AND KARL SIGMUND†

**Institute for Advanced Study, Princeton, NJ* 08540, *U.S.A.* and
†*Institut für Mathematik, Universität Wien, Strudlhofg* 4, *A*-1090 *Wien, Austria*
and IIASA, Laxenburg

(*Received on* 8 *April* 1998, *Accepted in revised form on* 23 *June* 1998)

Richard Alexander has argued that moral systems derive from indirect reciprocity. We analyse a simple case of a model of indirect reciprocity based on image scoring. Discriminators provide help to those individuals who have provided help. Even if the help is never returned by the beneficiary, or by individuals who in turn have been helped by the beneficiary, discriminating altruism can be resistant against invasion by defectors. Indiscriminate altruists can invade by random drift, however, setting up a complex dynamical system. In certain situations, defectors can invade only if their invasion attempts are sufficiently rare. We also consider a model with incomplete information and obtain conditions for the stability of altruism which differ from Hamilton's rule by simply replacing relatedness with acquaintanceship.

© 1998 Academic Press

1. Introduction

Altruistic behaviour is usually explained by inclusive fitness, group advantage, or reciprocity. The idea of reciprocal altruism, which is essentially economic, was introduced by Trivers (1971): a donor may help a recipient if the cost (to the donor) is less than the benefit (to the recipient), and if the recipient is likely to return the favour. This principle was explored in many papers, we mention only Axelrod & Hamilton (1981), Axelrod (1984), Sugden (1986), Boyd & Lorberbaum (1987), May (1987), Sherratt & Roberts (1998), Lindgren (1991), Nowak & Sigmund (1992, 1993), Nowak *et al.* (1995), Sigmund (1995), Crowley (1996), Leimar (1997).

In his seminal paper of 1971, Trivers mentioned the further possibility of a "generalised altruism," where the return is directed towards a third party. "Individuals ... may respond to an altruistic act that benefits themselves by acting altruistically toward a third individual uninvolved in the initial interaction." Trivers goes on to say: "In a system of strong multiparty interactions, it is possible that in some situations individuals are selected to demonstrate generalised altruistic tendencies." This possibility is further stressed in Triver's book on *Social Evolution* (1985), where it is speculated that a sense of justice may evolve "in species such as ours in which a system of multi-party altruism may operate such that an individual does not necessarily receive reciprocal benefit from the individual aided but may receive the return from third parties."

Richard Alexander greatly extended this idea, and coined the term of "indirect reciprocity" (see

‡Author to whom correspondence should be addressed.

0022–5193/98/020561 + 14 $30.00/0

© 1998 Academic Press

Alexander, 1979 and 1987, and references quoted therein). In this case, one does not expect a return from the recipient (as with direct reciprocity), but from someone else. Co-operation is thereby channelled towards the cooperative members of the community. A donor provides help if the recipient is likely to help others (which is usually decided on the basis of experience, i.e. according to whether the potential recipient has helped others in the past). According to Alexander (1987), indirect reciprocity, which "involves reputation and status, and results in everyone in the group continually being assessed and reassessed," plays an essential role in human societies. Alexander argues (convincingly, to our mind) that systems of indirect reciprocity are the basis of moral systems.

The principles of direct reciprocity are usually studied by means of games (like the Prisoner's Dilemma) repeatedly played between the same two players. In this paper we investigate situations where the players engage in several rounds of the game, but with a negligible probability of ever encountering the same co-player again. This is, of course, an idealisation, and in human communities, both direct and indirect reciprocity occur together. In fact, Alexander stresses that "indirect reciprocity is a consequence of direct reciprocity occurring in the presence of others." But in order to better understand the mechanism of indirect reciprocity, we shall essentially eliminate direct reciprocity from our model.

In Nowak & Sigmund (1998), we analysed populations of individuals having the options to help one another or not. Following usual practice, we denote the benefit of the altruistic act to the recipient by b, the cost to the donor by c, and assume $c < b$. If the donor decides not to help, both individuals receive zero payoff. The payoff is in terms of incremental fitness.

Each player has an image score, s. The score of a potential donor increases by one unit if he or she performs the altruistic act; if not, it decreases by one unit. The image score of a recipient does not change. At birth, each individual has score 0. We consider strategies where potential donors decide to help according to the image score of the recipient. A strategy is given by an integer k: a player with strategy k provides help if and only if the image score of the potential recipient is at least k. Players who provide help must pay some cost, but they increase their score and are, therefore, more likely to receive help in the future. During their lifetime, individuals undergo several rounds of this interaction, either as donors or as recipients, but the possibility of meeting the same co-player again will be neglected in our model. (More precisely, we use random meeting of partners, which implies that the two players meet again with a small probability: but we could just as well exclude meeting twice, without changing the conclusions.) At the end of each generation, individuals leave offspring in proportion to their accumulated payoff, which inherit the stategy of their parent (we assume clonal reproduction, as is usual in evolutionary games, see Maynard Smith, 1982).

In extensive computer simulations, Nowak & Sigmund (1998) showed that even for a very low number of rounds per generation, a cooperative regime based on indirect reciprocity can be stable. If one allows for mutations, then long-term cycling becomes likely. Populations of altruists discriminating according to the score of the recipient are undermined by indiscriminate altriusts. Then, unconditional defectors invade, until discriminating cooperators return, etc. We also extended the model so that individuals would only witness a fraction of the interactions in their community, and therefore have incomplete information about their co-player's score.

In this paper we shall study analytically a class of simple models for indirect reciprocity, based on two score values only, which we denote by G (for "good") and B (for "bad"). We obtain some of the cycling behaviour seen in the computer simulations. Furthermore, we show that the probability q that a player knows the score of another player must exceed c/b, if indirect reciprocity is to work. This is an intriguing parallel to Hamilton's rule, the cornerstone of the kin-selection approach to altruism (Hamilton, 1963). Hamilton's rule states that the coefficient of relatedness must exceed c/b. In this sense, indirect reciprocity differs from kin selection in replacing relatedness with acquaintance. If the average number of rounds per

lifetime exceeds $(bq + c)/(bp - c)$, then cooperation based on score discrimination is evolutionarily stable.

2. The Basic Model

For indirect reciprocity to work, some members of the group must assess the "score" of other members, and discriminately channel their assistance toward those with a higher score. Of course, the group may also contain members who do not discriminate, and either always give help, or never. We shall denote the frequency of the former by x_1, and that of the latter by x_2. By x_3, we denote the frequency of the discriminators. These individuals assess their group members and keep track of their "score". If they only remember the last round, they distinguish between those who have helped and thereby acquired score G, and those who have withheld assistance, and acquired score B. Discriminators help only G-players.

We shall now assume that each generation experiences a certain number of rounds of interactions. In each round, every player is both in the position of a donor and in the position of a recipient. (This simplifies the calculations without changing the basic results. In Nowak & Sigmund (1998) as well as in the last section of this paper, we assume that every player can be, with the same probability, a donor or a recipient.) In each of these roles, the player interacts with a randomly chosen co-player. If only few rounds occur, then the likelihood of meeting the same co-player twice is very small. The strategies which we consider take no account of this possibility.

In the first round, discriminators do not know the score of the potential recipient of their altruistic action. They have to rely on an *a priori* judgement, and assume with a certain "subjective" probability p that they are matched with a G-individual. If they help, they acquire G-status and become possible beneficiaries of other discriminators in the next round. We first consider the case $0 < p < 1$, and later the case $p = 1$.

With g_n we denote the frequency of G-players in round n (it is convenient to set $g_1 = p$, the discriminators' initial guess). Clearly

$$g_n = x_1 + g_{n-1}x_3 \tag{1}$$

for $n = 1, 2, \ldots$, so that by induction

$$g_n = \frac{x_1}{1 - x_3} + x_3^{n-1}(p - \frac{x_1}{1 - x_3}). \tag{2}$$

Hence g_n converges to $x_1/(x_1 + x_2)$, the percentage of cooperators among the indiscriminating players.

In order to compute the payoff, we have to monitor whether a recipient who meets a discriminating donor is perceived by the donor as a G-player. In the first round, this happens with probability p. From then on, it happens with probability 1 to the undiscriminate altruists (who have had occasion to prove their altruism), with probability 0 to the unconditional defectors (who are unmasked in the first round), and with probability g_{n-1} to the discriminators (since this is the probability that they have encountered a G-player and consequently provided help in the previous round).

In the first round, the payoff for an indiscriminate altruist is $-c + b(x_1 + px_3)$ (he always provides help, and he receives help from the indiscriminate altruists as well as from those discriminators who believe that he has label G). The payoff for unconditional defectors is similarly $b(x_1 + px_3)$ and that for discriminators is $-cp + b(x_1 + px_3)$. Obviously, if there is only one round, unconditional defectors win.

In the n-th round ($n > 1$), the indiscriminate altruists receive payoff $-c + b(x_1 + x_3)$, and unconditional defectors obtain bx_1. The proportion of G-scorers among the discriminators is g_{n-1} and their payoff is $-cg_n + b(x_1 + x_3)$. The other discriminators obtain $-cg_n + bx_1$. Adding up, we receive as the discriminators' payoff in the n-th round $-cg_n + b(x_1 + x_3g_{n-1})$, which by (1) is just $(b - c)g_n$.

If we assume that there are exactly N rounds per generation, then the total payoff for indiscriminate altruists is

$$\hat{P}_1 = N[b(x_1 + x_3) - c] - (1 - p)bx_3, \tag{3}$$

that for defectors is

$$\hat{P}_2 = Nbx_1 + bpx_3, \tag{4}$$

and that for discriminators is

$$\hat{P}_3 = (b-c)(g_1 + \ldots + g_N) + b(x_1 + px_3 - p). \quad (5)$$

It is easy to check that

$$g_1 + \ldots + g_N = \left[\frac{1}{1-x_3}\right]\left[(g_1 - g_2)\frac{1-x_3^N}{1-x_3} + Nx_1\right], \quad (6)$$

so that

$$\hat{P}_3 = (p - px_3 - x_1)\left(-b + \frac{b-c}{1-x_3}\frac{1-x_3^N}{1-x_3}\right) + \frac{N(b-c)x_1}{1-x_3}. \quad (7)$$

It is well-known that the structure of a game is unchanged if the same function is subtracted from all payoff functions (see, e.g. Hofbauer & Sigmund, 1998). It turns out that it is most convenient to substract $\hat{P}_2$. We then obtain as normalised payoff values $P_i := \hat{P}_i - \hat{P}_2$, the values $P_2 = 0$,

$$P_1 = (N-1)bx_3 - Nc \quad (8)$$

and

$$P_3 = \frac{x_1}{x_1 + x_2}P_1 + \left(p - \frac{x_1}{x_1 + x_2}\right)\frac{bx_3 - c - (b-c)x_3^N}{1-x_3}. \quad (9)$$

For instance, if the game is stopped after the second round already, i.e. $N = 2$, then

$$P_1 = -2c + bx_3, \quad (10)$$

$$P_2 = 0, \quad \text{and} \quad P_3 = -cp - cx_1 + (b-c)px_3. \quad (11)$$

3. The Replicator Equation for a Constant Number of Rounds

This allows to investigate the evolution of the frequencies of the three types of players under the influence of selection. We can use either a discrete game dynamics monitoring the frequencies from generation to generation, or the continuous replicator dynamics (see Hofbauer & Sigmund, 1998)

$$\dot{x}_i = x_i(P_i - \overline{P}) \quad (12)$$

on the (invariant) simplex $S_3 = \{\mathbf{x} = (x_1, x_2, x_3) \in R^3 : x_i \geqslant 0, \ \Sigma x_i = 1\}$. Here, $\overline{P} = \Sigma x_i P_i$ is the average payoff in the population. We stick to the latter, somewhat more transparent dynamics, emphasising that it is obtained as a limiting case of the dynamics with discrete generations (see Hofbauer & Sigmund, 1998).

For simplicity, let us start with the case $N = 2$. If $b > 2c$, as we shall assume in the following, then there exists a unique fixed point $\hat{\mathbf{p}} = (p_1, p_2, p_3)$ in the interior of S_3, i.e. with all three types present. It is given by $P_1 = P_2 = P_3$, which yields (since $P_2 = 0$)

$$p_1 = p(1 - \frac{2c}{b}), \ p_2 = (1-p)(1 - \frac{2c}{b}), \ p_3 = \frac{2c}{b}. \quad (13)$$

This point is a center. Indeed, one checks by a straightforward computation that the Jacobian at $\hat{\mathbf{p}}$ has trace 0 and determinant $2c^2p(1-p)(1-2c/b)^2$. The eigenvalues are therefore purely imaginary.

On the boundary of the simplex S_3, we find five fixed points. In addition to the corners $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ (where only one type is present), we find two mixed equilibria, namely

$$\mathbf{F}_{23} = \left(0, \frac{b-2c}{b-c}, \frac{c}{b-c}\right) \quad (14)$$

and $\mathbf{F}_{13}$, which is obtained from $\mathbf{F}_{23}$ by exchanging the first and the second coordinate. In the absence of discriminators (i.e. on the edge $x_3 = 0$), the flow points from $\mathbf{e}_1$ to $\mathbf{e}_2$: defectors win. In the absence of defectors, i.e. for $x_2 = 0$, the flow on the edge $\mathbf{e}_1\mathbf{e}_3$ leads toward $\mathbf{F}_{13}$. In the absence of indiscriminate altruists, i.e. when $x_1 = 0$, the system on the edge $\mathbf{e}_2\mathbf{e}_3$ is bistable (see Fig. 1).

Since $(b-c)x_3 = c$ is an invariant line (along this line, one has $\dot{x}_3 = 0$), it follows that there exists an orbit in the interior of S_3 which points along this straight line from $\mathbf{F}_{13}$ (its α-limit) to $\mathbf{F}_{23}$ (its ω-limit). The boundary of the triangle

spanned by $\mathbf{e}_3$, $\mathbf{F}_{13}$ and $\mathbf{F}_{23}$ is a heteroclinic cycle: it consists of three saddle-points connected by three orbits.

Using the classification of phase portraits of the replicator equation due to Zeeman (1980) and Bomze (1983), we can conclude that the fixed point $\hat{\mathbf{p}}$ is surrounded by closed orbits filling the afore-mentioned triangle (in Bomze's notation, we obtain phase portrait 13). The time-averages of these orbits all converge toward the point $\mathbf{p}$. In the remaining part of the simplex S_3, all orbits converge to $\mathbf{e}_2$. If the frequency of discriminators x_3 is less than $c/(b-c)$, therefore, then defectors take over. If not, then the frequencies of the three types oscillate periodically. We note however that this situation is not persistent: a sequence of random fluctuations can lead to larger and larger oscillations, and finally cause the system to cross the separatrix line $(b-c)x_3 = c$ and end up with a regime of all-out defection.

We mention without proof that if there are $N > 2$ rounds, nothing much changes. The unique fixed point $\hat{\mathbf{p}}$ in the interior of S_3 has now the coordinates

$$p_1 = p(1-p_3),\ p_2 = (1-p)(1-p_3),$$

$$p_3 = \frac{Nc}{(N-1)b}. \tag{15}$$

(The third equation follows because $P_1 = 0$, the first because for this value of p_1 one has $p - p_1/(1-p_3) = 0$. Again, the eigenvalues at $\hat{\mathbf{p}}$ are pure imaginary; this fixed point is a center surrounded by periodic orbits. The points $\mathbf{F}_{13}$ and $\mathbf{F}_{23}$ now satisfy

$$x_3 + \ldots + x_3^{N-1} = c/(b-c) \tag{16}$$

(the equation for $\mathbf{F}_{23}$ is given by $P_3 = 0$, that for $\mathbf{F}_{13}$ by $P_1 = P_3$.) We note that the solution of (16) satisfies $x_3 > c/b$.

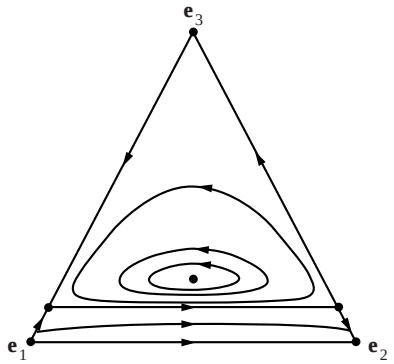


FIG. 1. Phase portrait of the model described in Section 3. There are three strategies: cooperators, defectors and discriminators (corresponding to the three corner fixed points $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$, respectively). Discriminators help G-players (see text). In the first round, they help other individuals with a fixed probability, p We assume the game is played for two rounds; the payoff values are given by eqns (10, 11). In the absence of discriminators, $x_3 = 0$, defectors win. In the absence of defectors, $x_2 = 0$, a stable equilibrium between cooperators and discriminators is reached. In the absence of cooperators, $x_1 = 0$, there is an unstable equilibrium between defectors and discriminators. If all three strategies are present, there is a separatrix connecting the two boundary equilibria on the edges. If the initial frequency of discriminators is below a critical value, then defectors will win. If it is above this critical value, then we obtain neutral oscillations around a center.

4. The Prejudice p as an Evolutionary Variable

So far we have treated p, the prejudice of the discriminator, as a parameter. But p may well be an evolutionary variable. So let us consider a model where, in addition to the types used so far, with frequencies x_1, x_2 and x_3, we have another type of discriminator with a prejudice $\rho \neq p$. The frequency of this new type is denoted by x_4 (with $\Sigma x_i = 1$). Again we can describe the payoffs of the different types of players in the different rounds. In the first round, all players receive (as recipients) the payoff $b(x_1 + px_3 + \rho x_4)$, which we neglect henceforth, since it is the same for all; as donors, indiscriminate altruists pay $-c$, unconditional defectors 0, and the two types of discriminators $-cp$ and $-c\rho$, respectively. In the first round, it pays to have as low an opinion as possible concerning the score of the unknown partner. From then onward, the score is always G for the indiscriminate altruists, and never G for the unconditional defectors. The two types of discriminators have score G, in the second round, with probability p and ρ, respectively. It follows that in the second round, the frequency of G-players is $g_2 = x_1 + px_3 + \rho x_4$. For the n-th round, with $n > 2$, the frequency g_n of G-players satisfies the recurrence relation

$$g_n = x_1 + (x_3 + x_4)g_{n-1}, \tag{17}$$

In the second round, the payoff for p-discriminators is therefore given by $-cg_2 + b(x_1 + p(x_3 + x_4))$, and that for ρ-discriminators by $-cg_2 + b(x_1 + \rho(x_3 + x_4))$. In the n-th round ($n > 2$) the payoff is $-cg_n + bg_{n-1}$ for both types of discriminators. If there are altogether two rounds or more per generation, then the total payoff for the p-discriminators differs from that of the ρ-discriminators by $(p - \rho)(-c + b(x_3 + x_4)$. By the quotient rule for the replicator dynamics (see Hofbauer & Sigmund, 1998) it follows that

$$(x_3/x_4)^{\cdot} = (x_3/x_4)(p - \rho)(-c + b(x_3 + x_4)). \quad (18)$$

If the total frequency $x_3 + x_4$ of discriminators is sufficiently high (namely larger than c/b), then (18) shows that the ratio x_3/x_4 increases if and only if $p > \rho$. In particular, in a population where the p-discriminating type is established and defectors have gone to extinction, or are on their way to vanish (which means, as we have seen, that x_3 is larger than c/b), then the ρ-discriminating type can invade and take over, if and only if $\rho > p$. Thus we can conclude that if indirect reciprocity works at all, then it favours those discriminators having larger p-values i.e. with a more positive prejudice in favour of an unknown partner. This leads to a trait-substitution sequence in the sense of Metz *et al*. (1992): mutations introducing larger and larger p-values will successively take over under the influence of selection. The p-value will therefore grow, as an evolutionary variable, until it approaches its maximal value 1. We shall therefore restrict our attention to the limiting case $p = 1$.

From now on, a discriminator is a player who, in the first round, gives help, and from then on helps recipients with G-score only. (The first help can be viewed as an entrance fee to the club of G-players.) It should be stressed that discriminators are not Tit For Tat players. Tit For Tat is a very successful strategy for the iterated Prisoner's Dilemma, and consists in cooperating in the first round, and from then on doing whatever the co-player did in the previous round. Tit For Tat strategists base their decisions on their own previous experience with the co-player, whereas discriminators use the experience of others. Pollock & Dugatkin (1992), in their interesting paper on reputation, described this strategy as "observer TFT".

It should also be mentioned that this discriminator strategy is related to, but different from the so-called T_1-strategy in the book by Robert Sugden on *The Evolution of Rights, Cooperation and Welfare* (1986). The T_1-strategy is based on the concept of *good standing*. Every player is born with a good standing, and keeps it as long as he extends help to other players with good standing. If he does not, he loses his good standing. Sugden argues that such a strategy can work as a basis for an insurance principle within the population (in each round of his game, a randomly chosen player needs help, and all other players can contribute to it). We stress that a player can keep his good standing by refusing to help someone of bad standing, whereas in our model, he would lose his G-score whenever he refuses help, even if the potential recipient is a B-scorer. Sugden's T_1 strategy is more sophisticated, but like Contrite Tit For Tat, another strategy based on standing, it is vulnerable to errors in perception (see also Boerlijst *et al*. 1997).

5. Pyrrhic Victories, or the Advantage of Rarely Showing Up

If we denote the frequency of discriminators by x_3, again, then the payoffs for indiscriminate altruists, unconditional defectors and discriminators are, in the first round, given by $-c + b(x_1 + x_3)$, $b(x_1 + x_3)$, and $-c + b(x_1 + x_3)$, respectively, and in the following rounds by $-c + b(x_1 + x_3)$, bx_1 resp. $-cg_n + b(x_1 + x_3 g_{n-1}) = (b - c)g_n$ where g_n is, as before, the frequency of G-players in round n and $g_1 = 1$. We now have by (2):

$$g_n = [x_1 + x_3^{n-1}x_2]/(x_1 + x_2). \quad (19)$$

If there are exactly N rounds (with $N > 1$), then the total payoffs $\hat{P}_1$, $\hat{P}_2$ and $\hat{P}_3$ of indiscriminate altruists, unconditional defectors and discriminators, respectively, are given by

$$\hat{P}_1 = N[-c + b(x_1 + x_3)], \quad (20)$$

$$\hat{P}_2 = Nbx_1 + bx_3, \quad (21)$$

$$\hat{P}_3 = (b - c)(g_1 + g_2 + \ldots + g_N) - bx_2 \quad (22)$$

which yields

$$\hat{P}_3 = N(b-c) + x_2\left[-b + \frac{b-c}{1-x_3}(1 + x_3 + \ldots + x_3^{N-1} - N)\right]. \quad (23)$$

Normalising such that $P_2 = 0$, this yields

$$P_1 = -Nc + (N-1)bx_3 \quad (24)$$

and

$$P_3 = P_1 + x_2[(N-1) \times b + \frac{b-c}{1-x_3}(1 + \ldots + x_3^{N-1} - N)]$$

and hence

$$P_3 = P_1 + x_2[(N-1)(c - bx_3) + (b-c)x_3(1 + \ldots + x_3^{N-2})]/(1-x_3). \quad (25)$$

Let us consider first the case $N = 2$ of two rounds only. In this case, we have

$$P_1 = -2c + bx_3, \quad (26)$$

and

$$P_3 = P_1 + cx_2, \quad (27)$$

It follows immediately that the replicator equation admits no interior fixed point. The edge $\mathbf{e}_1\mathbf{e}_3$ consists of fixed points: in the absence of unconditional defectors, both types do equally well. Along the edge $x_3 = 0$, the flow points from $\mathbf{e}_1$ to $\mathbf{e}_2$. On the edge $\mathbf{e}_2\mathbf{e}_3$, there exists a fixed point $\mathbf{F}_{23}$, with $x_3 = c/(b-c)$. The restriction to this edge is bistable: in a competition between unconditional defectors and discriminators, discriminators win if and only if their initial frequency is larger than $c/(b-c)$. Since the average payoff $\overline{P}$ is equal to $x_1P_1 + x_3(P_1 + cx_2)$, it follows that at $\mathbf{F}_{23}$, the transversal eigenvalue $\dot{x}_1/x_1$ is given by $c(2c-b)/b-c$, which is negative. Hence $\mathbf{F}_{23}$ is saturated.

Along the fixed point edge $\mathbf{e}_1\mathbf{e}_3$, the transversal eigenvalue $\dot{x}_2/x_2$ is equal to $-\overline{P}$ (i.e. to $2c - bx_3$). If we denote by F the point with $x_2 = 0$ and $\dot{x}_2/x_2 = 0$, i.e. with $x_3 = 2c/b$, then the points on the edge between $\mathbf{e}_3$ and $\mathbf{F}$ are saturated, and hence ω-limits of orbits in the interior of S_3, whereas all points on the segment between $\mathbf{F}$ and $\mathbf{e}_1$ are α-limits. If $(b-c)x_3 = c$ then $\dot{x}_3 = 0$ in the interior of S_3. It follows that the line l given by $x_3 = c/(b-c)$ is invariant. It corresponds to an orbit whose ω-limit is the saddle point $\mathbf{F}_{23}$ and whose α-limit, which we denote by $\mathbf{F}_{13}$, has coordinates $x_2 = 0$ and $x_3 = b/(b-c)$. This separatrix l divides the interior of the simplex S_3 into two regions. In one region, all orbits converge toward $\mathbf{e}_2$. In the other region, all orbits lead from the fixed point edge $\mathbf{e}_1\mathbf{e}_3$ back to that edge; their α-limit is between $\mathbf{F}_{13}$ and $\mathbf{F}$, their ω-limit between $\mathbf{F}$ and $\mathbf{e}_3$; they surround $\mathbf{F}$ (see Fig. 2.). The equation also admits an invariant of motion: $x_1x_3^{-2}[-c + (b-c)x_3]$ (courtesy of Josef Hofbauer).

The interplay between the three strategies leads to a fascinating long-term dynamics. Depending on the initial condition, selection leads either toward a homogeneous regime of all-out defectors, or to a mixture of

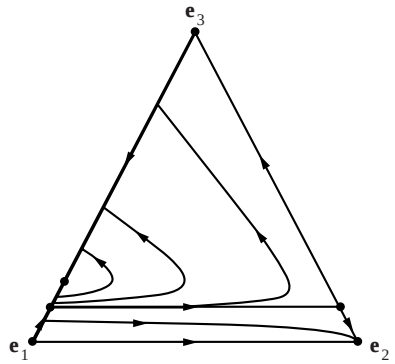


FIG. 2. Phase portrait of the model described in Section 5 [eqns (26–27)]. As for Fig. 1, we consider cooperators, defectors and discriminators, but this time discriminators always help in the first round ($p = 1$). Again there is a separatrix connecting two fixed points on the edges $\mathbf{e}_1\mathbf{e}_3$ and $\mathbf{e}_2\mathbf{e}_3$, but there is no fixed point in the interior of the simplex. Instead the whole edge $\mathbf{e}_1\mathbf{e}_3$ consists of fixed points, some of which are stable against invasion by defectors, while others are not. The overall dynamics of the system are as follows. Imagine a mixture between cooperators and discriminators. There is random drift along the edge $\mathbf{e}_1\mathbf{e}_3$. If there are sufficiently many discriminators then defectors cannot invade. There are two threshold levels of discriminators. If the frequency drops below the first value then defectors can invade, but will go extinct again leaving the system in a state with a higher frequency of discriminators. If the discriminator frequency fluctuates below the second value, then defectors can invade and take over. Hence, if defectors appear too often they cannot win. They only win when showing up rarely. This seems to be an interesting example for a more general, counter-intuitive principle where a mutation can only win if rare.

discriminators and indiscriminate altruists (with no unconditional defectors). In such a mixture, no type has a selective advantage. Random drift takes over, and the mixture fluctuates along the $\mathbf{e}_1\mathbf{e}_3$-edge. From time to time, mutation can also introduce unconditional defectors. If such an invasion is attempted when the state lies between $\mathbf{e}_3$ and $\mathbf{F}$, it is promptly repelled. If it occurs while the state is between $\mathbf{F}_{13}$ and $\mathbf{e}_1$, then it succeeds and defectors take over. But if the invasion attempt occurs while the state lies between $\mathbf{F}_{13}$ and $\mathbf{F}$, then it knows a transient success only; the frequency of defectors increases at first, but then the proportion of discriminators grows at the expense of the indiscriminate altruists, and causes the defectors to vanish. The end result of this failed invasion attempt is, as before, a mixture of discriminators and indiscriminate altruists, but now with a much higher amount of discriminators, so that now it is able to stop any invasion attempt by defectors in the bud. Somewhat related examples of successful invasions which are ultimately self-defeating (Pyrrhic victories, so to speak) can be found in Mylius *et al*. (1998) where strategies are studied which are invadible yet unbeatable.

Of course, random drift can slowly lead the state back into the threatened zone. But if invasions by defectors occur frequently enough, these invasions will be attempted while the state is between $\mathbf{F}_{13}$ and $\mathbf{F}$, and hence the state will be led back into the invasion-proof zone. It is only if the frequency of invasion attempts by defectors is low that random drift along the fixed point edge $\mathbf{e}_1\mathbf{e}_3$ can lead the state across the "gap" between $\mathbf{F}_{13}$ and $\mathbf{F}$ (whose width is $c(b-2c)/b(b-c)$). In this case the state enters into the segment between $\mathbf{F}$ and $\mathbf{e}_1$ where an invasion by defectors knows an irreversible success. Thus we see a remarkable phenomenon: a mutant that can succeed only if it occurs rather rarely!

Essentially the same situation holds when there are $N > 2$ rounds. The point $\mathbf{F}_{23}$ now has a coordinate x_3 which is given as the solution of the equation

$$x_2 + x_3^2 + \ldots + x_3^{N-1} = \frac{c}{b-c} \tag{28}$$

(see (16)). This is a value which, with increasing N, shifts from $c/(b-c)$ towards c/b. The point $\mathbf{F}$ has a coordinate x_2 given by $Nc/(N-1)b$. This is simply the limit of the interior fixed point $\hat{\mathbf{p}}$ given by (15), if p converges to 1.

This cycle of invasions is related to a phenomenon found in the numerical simulations by Nowak & Sigmund (1998), which are based on a more sophisticated model of indirect reciprocity where scores can take all integer values (see Fig. 3).

6. Random Numbers of Rounds

Let us now assume, not a constant number of rounds per generation, but rather a constant probability w for a further round. The total number of rounds per generation is then a geometrically distributed random variable with mean value $1/(1-w)$. The payoffs are of the form $A_1 + wA_2 + w^2A_3 + \ldots$, where A_n is the payoff in the n-th round. Then, by using the first paragraph of Section 5,

$$\hat{P}_1 = \frac{1}{1-w}\,[-c + b(x_1 + x_3)], \tag{29}$$

$$\hat{P}_2 = \frac{1}{1-w}\,bx_1 + bx_3 = \frac{b(x_1 + x_3) - wbx_3}{1-w}, \tag{30}$$

$$\hat{P}_3 = (b-c)(g_1 + wg_2 + w^2g_3 + \ldots) - bx_2. \tag{31}$$

Writing $g := g_1 + wg_2 + w^2g_3 + \ldots$, we see that $g = 1 + w(x_1 + g_1x_3) + w^2(x_1 + g_2x_3) + \ldots$, and hence that

$$g = 1 + \frac{wx_1}{1-w} + x_3wg. \tag{32}$$

Therefore

$$g = \frac{1 - w + wx_1}{(1-w)(1-wx_3)} \tag{33}$$

and thus

$$\hat{P}_3 = -bx_2 + \frac{(b-c)(1-w+wx_1)}{(1-w)(1-wx_3)} \tag{34}$$

It is convenient again to normalise the payoff values such that $P_2 = 0$. In this case

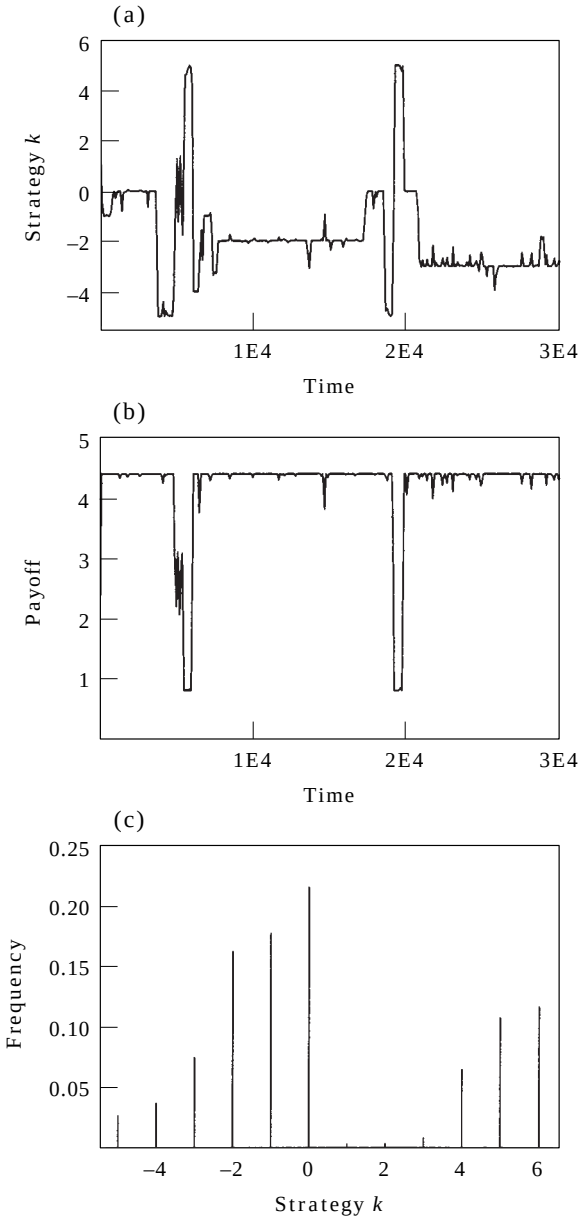
$$P_1 = \frac{wbx_3 - c}{1-w} \tag{35}$$

and

$$P_3 = \frac{(b-c)(1-w+wx_1)}{(1-w)(1-wx_3)} - bx_2 - bx_3 - \frac{bx_1}{1-w} \quad (36)$$

which yields

$$P_3 = \frac{(1-w+wx_1)}{1-w}\left(\frac{b-c}{1-wx_3} - b\right), \quad (37)$$



and thus finally

$$P_3 = \frac{1-w+wx_1}{1-wx_3}P_1. \quad (38)$$

In contrast to the case of a fixed number of rounds, we now obtain a line l of fixed points in the interior of S_3, given by $x_3 = c/wb$ (we assume from now on that $w > c/b$). The edge $\mathbf{e}_1\mathbf{e}_3$ consists of fixed points too (see Fig. 4). On the edge $\mathbf{e}_1\mathbf{e}_2$ the flow leads towards $\mathbf{e}_2$, and on the edge $\mathbf{e}_2\mathbf{e}_3$ we have a bistable competition, with threshold point $\mathbf{F}_{23}$ given by the intersection with the fixed point line l. This line l acts as separatrix. It divides S_3 into two regions, in one region the ratio x_1/x_2 decreases and in the other it increases. All orbits in the former region converge to $\mathbf{e}_2$ and lead to a population of unconditional defectors; in the other region, all orbits converge to the fixed point edge, and hence lead to a mixture of discriminators and indiscriminate altruists. Random fluctuations along the fixed point edge will eventually lead to the region where defectors can invade.

7. An Analogy with the Prisoner's Dilemma Game

Although the dynamics of indirect reciprocity given by (29)–(32) is based on a model which is

FIG. 3. Cycling behaviour in the model by Nowak & Sigmund (1998). Donor–recipient pairs are formed at random. The score of a newborn is 0, it increases by one unit whenever the individual provides help and decreases by one unit if the individual refuses to help. A strategy is given by an integer j. An individual with strategy j provides help to all potential recipients with score at least j. Players with strategy $j = 0$ can be viewed as discriminate altruists. Players with a low j (for instance $j = -3$) are *de facto* indiscriminate altruists, because they help every co-player; indeed, if players experience only two or three rounds per lifetime, there will be no players with score less than -3. Players with a high j (for instance $j = 4$), on the other hand, are defectors; they will never provide help. Numerical simulations show how populations of discriminate altruists are eventually undermined by indiscriminate altruists (the average j-value drops), that defectors cash in (the average j-value sharply increases) and that this brings discriminators to the fore again (the average j-value drops back to 0): (a) the average j-value of the population; (b) the average payoff per individual, per generation; (c) frequency distribution of strategies sampled over many generations ($t = 10^7$). Parameter values: $b = 1$, $c = 0.1$ (to avoid negative payoffs we add 0.1 in each interaction); $m = 300$ rounds per generation.

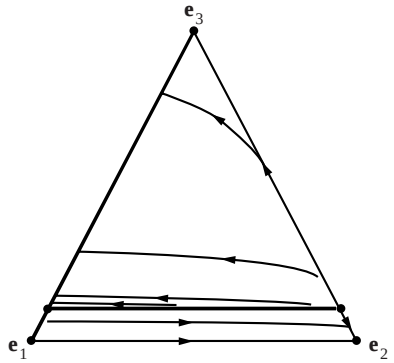


FIG. 4. Phase portrait of the model described in Section 6 [eqns (29–31)]. We consider the same situation as for Fig. 2, but this time there is not a fixed number of rounds, but a probability, w, of a next round. The separatrix becomes a line of fixed points. The edge $\mathbf{e}_1\mathbf{e}_3$ is also a line of fixed points. Again there are two regions in phase space. If there are sufficiently many discriminators then defectors become eliminated, if the frequency of discriminators drops below a critical level then defectors take over.

quite distinct from the repeated Prisoner's Dilemma game, it yields a remarkably similar dynamics. Indeed, let us consider the Prisoner's Dilemma (PD) game, where each of the two players has, in each round, two options: to play **C** (to cooperate) or **D** (to defect). The payoff matrix is given by

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} \tag{39}$$

where $T > R > P > S$, i.e. the reward R for mutual cooperation is larger than the punishment P for joint defection, but a unilateral defector receives the highest payoff T (the temptation) and a unilateral cooperator the lowest payoff S (the sucker's payoff). Let us assume that in each generation, each player is matched with one randomly chosen co-player for a variable number of rounds. Again, we assume that the probability for a further round is constant and given by some $w < 1$. Let us assume that the population contains only three types of players, the unconditional cooperators, the unconditional defectors, and the Tit For Tat players. Let x_1, x_2 and x_3 be their respective frequencies. The expected payoffs are (as is well known, see for instance Nowak & Sigmund, 1987)

$$\hat{P}_1 = \frac{1}{1-w}\left[R(x_1 + x_3) + Sx_2\right] \tag{40}$$

$$\hat{P}_2 = \frac{Px_2 + Tx_1}{1-w} + \left(T + \frac{wP}{1-w}\right)x_3 \tag{41}$$

$$\hat{P}_3 = \frac{R(x_1 + x_3)}{1-w} + \left(S + \frac{wP}{1-w}\right)x_2. \tag{42}$$

If we normalise these payoff values, such that $P_2 = 0$, and if we set, as is natural, for the temptation by unilateral defection $T = b$, for the reward by mutual cooperation $R = b - c$, for the punishment of bilateral defection $P = 0$ and for the cost of being suckered $S = -c$, then the payoffs in the PD model become

$$P_1 = \frac{bwx_3 - c}{1-w} \tag{43}$$

and

$$P_3 = \frac{bwx_3 + cwx_2 - c}{1-w} = P_1 + \frac{cwx_2}{1-w}, \tag{44}$$

which behaves like the dynamical system with $N = 2$. In fact, for $w = 1/2$ it is exactly the same system. (If however $w = (N-1)/N$ for $N > 2$, then the equations do not agree with the dynamics given by (10)–(11); we also note that the system (29)–(32) with a random number of rounds is different, and in particular contains higher order terms.)

8. A Model with Incomplete Information

Even in small groups, where everyone knows everyone else, it is unlikely that all group members witness all interactions. Therefore each player has a specific perception of the image score of the other players. The same player can have different image scores in the eyes of different individuals. Furthermore, it is unrealistic to assume that episodes as donor and recipient alternate in a well synchronised way. Some individuals will be more often in a position to give help than others.

We shall therefore assume from now on that in each round, a given individual is with probability 1/2 either a donor or a recipient. If there are only few rounds, it is quite possible that

a given individual is never a donor. This is more in line with the stochastic simulations in Nowak & Sigmund (1998). We extend the previous two-score model by assuming that with probability q a given individual knows the score of a randomly chosen opponent. A discriminator who does not know the score of the co-player will assume with probability 1 that this score is G. If g_n denotes, as before, the frequency of G-scorers in the population, and $x_{1G}(n)$, $x_{2G}(n)$ and $x_{3G}(n)$ are the frequencies of indiscriminate altruists, unconditional defectors resp. discriminators in round n, then clearly $x_{1G}(n) = x_1$ and $x_{2G}(n) = (1/2)x_{2G}(n-1)$, since a defector is with probability 1/2 in the role of a donor and then unmasks himself. Therefore

$$x_{2G(n)} = \frac{x_2}{2^{n-1}}. \tag{45}$$

The score of a discriminator remains unchanged if he is a recipient. If he is a potential donor, he will either know the co-player (with probability q) and help if the co-player has score G (as happens with probability g_n), or else he will not know the co-player's score, and help (this happens with probability $1-q$). Altogether, this yields

$$x_{3G}(n) = (1/2)x_{3G}(n-1) + (1/2)x_3(1-q+qg_n). \tag{46}$$

Since $g_n = x_{1G}(n) + x_{2G}(n) + x_{3G}(n)$, it follows that

$$g_n = sg_{n-1} + (x_1 + (1-q)x_3) \tag{47}$$

with $s = (1+qx_3)/2$. This recurrence relation implies (together with $g_1 = 1$) that

$$g_n = \left(\frac{1+qx_3}{2}\right)^{n-1} \frac{x_2}{1-qx_3} + \frac{x_1 + (1-g)x_3}{1-qx_3}. \tag{48}$$

The payoff for the indiscriminate altruists in round n is

$$\hat{A}_1(n) = -(c/2) + (b/2)(x_1 + x_3). \tag{49}$$

The payoff P_2 for the unconditional defectors depends on their score. Those with score B receive $b(x_1 + (1-q)x_3)/2$ and those with score G in addition $qbx_3/2$, so that

$$\hat{A}_2(n) = (b/2)[x_1 + (1-q)x_3 + x_3q(x_{2G}(n)/x_2)]. \tag{50}$$

Finally, a discriminator receives $[-c(qg_n + 1 - q) + bx_1 + (1-q)bx_3)]/2$ if he has score B, and in addition $bqx_3/2$ if he has score G, so that we obtain

$$\hat{A}_3(n) = -(c/2)(qg_n + 1 - q) + (b/2)(x_1 + x_3) - (b/2)qx_3[1 - (x_{3G}(n)/x_3)]. \tag{51}$$

Normalising by subtracting $\hat{A}_2(n)$, this yields

$$A_1(n) = -(c/2) + (b/2)qx_3(1 - 2^{-(n-1)}) \tag{52}$$

and

$$A_3(n) = -(c/2)(1-q) + (q/2)(b-c)g_n - (b/2)qx_1 - (b/2)q(x_2 + x_3)2^{-(n-1)}. \tag{53}$$

If we assume that $w < 1$ is the probability for a further round, then the total payoff for unconditional defectors is $P_2 = 0$, that for indiscriminate altruists is

$$P_1 = \frac{1}{2(1-w)}\left[-c + \frac{bwqx_3}{2-w}\right] \tag{54}$$

and that for discriminators is

$$P_3 = \frac{(bqx_3 - c)(1 - q + qx_1)}{2(1-w)(1-qx_3)} - \frac{bq(x_2 + x_3)}{2-w} + \frac{q(b-c)x_2}{(1-qx_3)(2-w-wqx_3)}, \tag{55}$$

and hence

$$P_3 = P_1 + \frac{qx_2}{1-qx_3}\left[\frac{c - bqx_3}{2(1-w)} + \frac{b-c}{2-w-wqx_3}\right]. \tag{56}$$

It is obvious that $P_1 = 0$ holds iff

$$x_3 = \frac{c(2-w)}{bwq}. \tag{57}$$

A straightforward computation shows that for this x_3-value, $P_3 = 0$. Hence the fixed points of the corresponding replicator equation are (apart from the vertices of the simplex S_3) the edge $\mathbf{e}_1\mathbf{e}_3$

and the line l given by $bwqx_3 = c(2 - w)$. This line divides the interior of S_3 into two regions: in one region, all orbits converge to $\mathbf{e}_2$, in the other region, towards a point on the $\mathbf{e}_1\mathbf{e}_3$-edge which depends on the initial value. This is exactly as in Section 6 (see Fig. 4).

Of course this holds only if the value of x_3 is less than 1, i.e. if $w(c + bq) > 2c$, in other words if the expected number of rounds, i.e. $(1 - w)^{-1}$, satisfies

$$1/(1 - w) > (bq + c)/(bq - c). \qquad (58)$$

If we consider only the two strategies defector and discriminator, then discriminator can be evolutionarily stable only if

$$q > c/b. \qquad (59)$$

This looks exactly like Hamilton's rule for altruism through kin selection, except that the coefficient of relatedness, k, is replaced by the probability to know the co-player's score, q.

9. Discussion

Several authors, starting with Trivers himself, have stressed that reciprocal altruism need not be restricted to dyads of interacting individuals (see Trivers, 1971; Boyd, 1988; Dugatkin *et al*., 1992; May, 1987; Axelrod & Dion, 1988; Binmore, 1992 and Chap. 7 of Sugden, 1986, for instance.)

There are several ways to model generalised or indirect reciprocity. Alexander, who elaborated on the importance of this notion, did not fully specify the mechanisms involved, but mentioned several possibilities. One conceivable form of reward (see e.g. Alexander, 1987, p. 94) consists in having the success of the group contribute to the success of his own descendants, which is simply group selection in the modern sense, see Wilson & Sober (1994). One other form has been investigated by Boyd & Richerson (1989): individual A helps B, who helps C, who helps D, who finally returns the help to A. Thus individuals are arranged in closed, oriented loops, reminiscent of the hypercycles in the theory of Eigen & Schuster (1979) on catalytic loops of selfreplicating molecules. Boyd & Richerson investigate two strategies: upstream Tit For Tat (A keeps helping B if D keeps helping A) and downstream TFT (A keeps helping B if A observes that B keeps helping C). They find that the second type is much more efficient than the first, but that it is also more difficult to perform. (It should be noted that for two-member loops, both strategies reduce to Tit For Tat.) Boyd & Richerson conclude that this type of indirect reciprocity is less likely to evolve than pairwise reciprocity, and is only effective for relatively small, closed, long-lasting loops.

In a sense, this indirect reciprocity is still quite direct, and the social networks in human groups (or pirmates, for that matter—see de Waals, 1996) are much more fluid than the "long-lasting loops" indicate. Alexander (1987) envisions a more diffuse mechanism when he stresses (p. 85) that "the return [of the beneficence] may come from essentially any individual or collection of individuals in the group", and emphasised the importance of assessment and status. We have tried to model this in Nowak & Sigmund (1998) by means of "scores" assigned to each group member. If the model is reduced to the minimum (two scores only), we obtain the discriminator strategy.

The same strategy has been reached, through a different approach, in Pollock & Dugatkin (1992), who termed it Observer Tit For Tat. They studied it in the context of the repeated Prisoner's Dilemma, which is the usual framework for analysing direct reciprocity. Pollock & Dugatkin allowed the players to occasionally observe a co-player before starting the repeated interaction. If the future co-player was seen defecting in his last interaction, then Observer Tit For Tat prescribes to defect in the first round. Pollock & Dugatkin were mostly interesting in comparing this strategy with the usual Tit For Tat, but they also found that it could hold its own against defectors when no degree of future interaction with the current partner was presumed. They also obtained a condition similar to (53), but without modelling the different rounds in an individual's lifetime, and in particular without (52). The approach by Pollock & Dugatkin is truly remarkable. They did not aim at a model of indirect reciprocity, but actually investigated what Alexander would view as its prerequisite, namely "direct reciprocity occurring in the presence of interested audiences"

(Alexander, 1987, p. 93), and came out with what we believe is the simplest strategy under which indirect reciprocity can be implemented—an unintended support for the correctness of Alexander's intuition.

The success of a discriminating player is somewhat hampered by the fact that whenever he refuses to help a *B*-scorer, he loses his *G*-score. A more sophisticated strategy has been studied by Sugden (1986) in a context which is only slightly different. In Sugden's model, in each round a randomly chosen player needs help, and each of the other players can provide some help (thus the needy player can get as payoff $(m-1)b$, where m is the group size). Sugden's T_1 strategy is based on the notion of standing: a player is born with good standing, and keeps it as long as he helps needy players who are in good standing. Such a player can therefore keep his good standing even when he defects, as long as the defection is directed at a player with bad standing (this is in contrast to the discriminator strategy). We believe that Sugden's strategy is a good approximation to how indirect reciprocity actually works in human communities: it offers, as Sugden remarks, a workable insurance principle. But as stressed in Boerlijst *et al*. (1997) in connection with Contrite Tit For Tat, strategies based on standing are prone to be affected by errors in perception. If information is incomplete, then a player observed while withholding his help may be misunderstood; he may have defected on a player with good standing, or punished someone with bad standing. An eventual error can spread. The discriminator rule is less demanding on the player's capabilities, and still works. We expect that in actual human communities, indirect reciprocity is based on more complex reckonings, and believe that this should be amenable to experimental tests.

Finally, we mention that according to Zahavi (1995), Arabian babblers "compete with each other to invest in the interests of the group, and often interfere with the helping of others". This jostling for the position of the helper cannot be explained in terms of group selection, kin selection or direct reciprocation. However, if helping raises one's score and therefore one's fitness, this type of competition can easily be understood: indirect reciprocity based on image scoring provides a simple explanation.

We wish to thank Martin Posch, Immanuel Bomze, Josef Hofbauer, Robert May and Alex Kacelnik for helpful discussions. The FWF grant P11144 and The Wellcome Trust of Great Britain are gratefully acknowledged.

REFERENCES

Alexander, R. D. (1979). *Darwinism and Human Affairs*. Seattle: University of Washington Press.

Alexander, R. D. (1987). *The Biology of Moral Systems*. New York: Aldine de Gruyter.

Axelrod, R. (1984). *The Evolution of Cooperation*, reprinted 1989. Harmondsworth: Penguin.

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science* **211,** 1390–1396.

Axelrod, R. & Dion, D. (1988). The further evolution of cooperation. *Science* **242,** 1385–1390.

Binmore, K. G. (1992). *Fun and Games*: *a Text on Game Theory*, Lexington, MA: Heath & Co.

Boerlijst, M., Nowak, M. A. & Sigmund, K. (1997). The logic of contrition. *J. theor. Biol*. **185,** 281–293.

Bomze, I. (1983). Lotka–Volterra eqations and replicator dynamics: a two dimensional classification. *Biol. Cybernetics* **48,** 201–211.

Boyd, R. (1988). Is the repeated Prisoner's Dilemma a good model of reciprocal altruism? *Ethol. Sociobiol*. **9,** 278–305.

Boyd, R. & Lorberbaum, J. P. (1987). No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma. *Nature* **327,** 58–59.

Boyd, R. & Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social Networks* **11,** 213–236.

Crowley, P. H. (1996). Evolving cooperation: strategies as hierarchies of rules. *Biosystems* **37,** 67–80.

de Waals, F. (1996). *Good Natured*: *the Origins of Right and Wrong in Humans and other Animals*. Cambridge, MA: Harvard UP.

Dugatkin, L. A., Mesterton-Gibbons, M. & Houston, A. I. (1992). Beyond the Prisoner's Dilemma: towards models to discriminate among mechanisms of cooperation in nature. *TREE* **7,** 202–205.

Eigen, M. & Schuster, P. (1979). *The Hypercycle*: *a Principle of Natural Selforganization*. Berlin-Heidelberg: Springer.

Hamilton, W. D. (1963). The evolution of altruistic behaviour. *Am. Nat*. **97,** 354–356.

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

Leimar, O. (1997). Repeated games: a state space approach. *J. theor. Biol*. **184,** 471–498.

Lindgren, K. (1991). Evolutionary phenomena in simple dynamics. In: *Artificial Life II* (Langton, C. G. *et al*., ed.), Vol. X, pp. 295–312. Santa Fe Institute for Studies in the Sciences of Complexity.

May, R. M. (1987). More evolution of cooperation. *Nature* **327,** 15–17.

MAYNARD SMITH, J. (1982). *The Theory of Games and Evolution*. Cambridge: Cambridge University Press.

METZ, J. A. J., NISBET, R. M. & GERITZ, S. A. H. (1992). How should we define fitness for general ecological scenarios? *Trends Evol. Ecol.* **7,** 198–202.

MYLIUS, S., DOEBELI, M. & DIEKMANN, O. (1998). Can initial invasion dynamics correctly predict phenotypic substitutions? Preprint.

NOWAK, M. A. & SIGMUND, K. (1987). Oscillations in the evolution of reciprocity. *JTB* **137,** 21–26.

NOWAK, M. A. & SIGMUND, K. (1992). Tit for tat in heterogeneous populations. *Nature* **355,** 250–252.

NOWAK, M. A. & SIGMUND, K. (1993). Win-stay, lose-shift outperforms tit-for-tat. *Nature* **364,** 56–58.

NOWAK, M. A. & SIGMUND, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573–577.

NOWAK, M. A., MAY, R. M. & SIGMUND, K. (1995). The arithmetics of mutual help. *Scient. Am.* **272,** 76–81.

POLLOCK, G. B. & DUGATKIN, L. A. (1992). Reciprocity and the evolution of reputation. *JTB* **159,** 25–37.

SHERRATT, T. N. & ROBERTS, G. (1998). The evolution of generosity and choosiness in cooperative exchanges. *JTB* (in press).

SIGMUND, K. (1995). *Games of Life*. Harmondsworth: Penguin.

SUGDEN, R. (1986). *The Evolution of Rights, Co-operation and Welfare*. Oxford: Blackwell.

TRIVERS, R. (1971). The evolution of reciprocal altruism. *Q. Rev. Biol.* **46,** 35–57.

TRIVERS, R. (1985). *Social Evolution*. Menlo Park, CA: Benjamin Cummings.

WILSON, D. S. & SOBER, E. (1994). Re-introducing group selection to human behavioural sciences. *Behav. Brain Sci.* **17,** 585–654.

ZAHAVI, A. (1995). Altruism as a handicap—the limitations of kin selection and reciprocity. *J. Avian Biol.* **26,** 1–3.

ZEEMAN, E. C. (1980). Population dynamics from game theory. In: *Global Theory of Dynamical Systems*. Lecture Notes in Mathematics 819. New York: Springer.