A survey on indirect reciprocity

Hannelore Brandt[1,2], Hisashi Ohtsuki[3], Yoh Iwasa[3] and Karl Sigmund[1,2]

January 14, 2005

[1]: Fakultät für Mathematik, Nordbergstrasse 15, 1090 Wien, Austria
[2]: Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, 2361 Laxenburg, Austria
[3]: Department of Biology, Faculty of Sciences, Kyushu University, Japan

Abstract

This survey deals with indirect reciprocity, i.e. with the possibility that altruistic acts are returned, not by the recipient, but by a third party. After briefly sketching how this question is dealt with in classical game theory, we turn to models from evolutionary game theory. We describe recent work on the assessment of interactions, and the evolutionary stability of strategies for indirect reciprocation. All stable strategies (the 'leading eight') distinguish between justified and non-justified defections, and therefore are based on non-costly punishment. Next we consider the replicator dynamics of populations consisting of defectors, discriminators and undiscriminating altruists. We stress that errors can destabilise cooperation for strategies not distinguishing justified from unjustified defections, but that a fixed number of rounds, or the assumption of an individual's social network growing with age, can lead to cooperation based on a stable mixture of undiscriminating altruists and of discriminators who do not distinguish between justified and unjustified defection. We describe previous work using agent-based simulations for 'binary score' and 'full score' models. Finally, we survey the recent results on experiments with the indirect reciprocation game.

1 Introduction

In evolutionary biology, the two major approaches to the emergence of cooperation are kin-selection, on one hand, and reciprocation, on the other. The

latter, which is essential for understanding cooperation between non-related individuals and very prominent in human societies, can be subdivided into two parts of unequal size. In direct reciprocity, it is the recipient of a helpful action who eventually returns the aid. In indirect reciprocity, the return is provided by a third party. This possibility has originally been named 'third-party altruism' or 'generalised reciprocity' by Trivers (1971). Later, Alexander (1987) explored it under the (now common) heading of 'indirect reciprocity', see also Ferrière (1998) and Wedekind (1998). Indirect reciprocity is much less well studied than direct reciprocity, and offers interesting theoretical challenges.

Several mechanisms for indirect reciprocity are conceivable. It could be, for instance, that a person having been helped is enclined to help a third party in turn. In cyclical networks, this provides a plausible feedback loop. But studies by Boyd and Richerson (1989) and van der Heijden (1996) suggest that such networks have to be rather small and rigid.

Alexander suggested, in contrast, that indirect reciprocation is based on reputation and status. By giving help to others, individuals acquire a high reputation. If help is directed preferentially towards recipients with a high reputation, defectors will be penalised. Such indirect reciprocation based on reputation and status is the topic of this paper.

The two main reasons why reputation mechanisms are interesting show up at two stages in human evolution which could not be further apart. On the one hand, status and reputation may well have played a major role in the evolution of moral systems since the dawn of prehistory, boosting cooperation between non-relatives (a major cause for the evolutionary success of hominids) and possibly providing a major selective impetus for the emergence of language, as a means of transmitting information about group members through gossip (Alexander, 1987, Nowak and Sigmund, 1998a, Panchanathan and Boyd, 2003). On the other hand, the very recent advent of e-commerce makes the efficient assessment of reputations and moral hazard in trust-based transactions a burning issue. Anonymous one-shot interactions in global markets, rather than long-lasting repeated interactions through direct reciprocation, seem to play an ever-increasing role in today's economy (Bolton et al, 2002, Keser, 2002, Dellarocas, 2003).

The aim of this paper is to provide a survey of the model-based theoretical investigations of the concept of indirect reciprocation, and of the remarkable results on experimental economic games inspired by them.

2 Indirect reciprocation for rational players

Before approaching the subject in the spirit of evolutionary game dynamics, we should stress that the same topic can also be addressed within classical game theory. At a first glance, it may almost look like a non-issue in this context. Indeed, it is easy to see that the main classical results on repeated games survive unharmed if the single co-player with whom one interacts in direct reciprocation is replaced by the wider cast of co-players showing up in indirect reciprocation. This holds, in particular, for the folk theorem on repeated games. It states, essentially, that every feasible payoff larger than the maximin level which players can guarantee for themselves is obtainable by strategies in Nash equilibrium, provided that the probability for another round is sufficiently large (Fudenberg and Maskin, 1986, Binmore, 1992). This can be achieved, in particular, by 'trigger strategies' that switch to defection after the first defection of the co-player: for in that case, it makes no sense to exploit the co-player in one round, thereby forfeiting all chances for mutual cooperation in further rounds. Exactly the same argument holds for indirect reciprocation in a population where players are randomly matched between rounds, if they know the case-history of every co-player which they encounter, and refuse help to any individual who ever refused to help someone (Rosenthal, 1979, Okuno-Fujiwara and Postlewait, 1989, Kandori, 1992). The difference between personal enforcement, in the former case, and community enforcement, in the latter, is irrelevant to the sequence of payoffs encountered by an individual player.

It must be noted, however, that with such trigger strategies, the defection of a single player A results in the eventual punishment of all players, and the breakdown of cooperation in the whole population. Indeed, if A defects in a given round, then the next player B who is asked to help A will refuse, and so will C when asked to help B, etc, so that defection spreads rapidly through the population. If the population consists of rational agents, player A will not defect. But if even one player fails to be rational, the whole community is under threat.

As Sugden (1986) suggested, this can be remedied by another trigger strategy, which distinguishes between justified and unjustified defections. Such a strategy is based on the notion of *standing*. Each individual has originally a good standing, and loses this only by refusing help to an individual in good standing. Individuals refusing help to someone in bad standing do not lose their good standing. In this way, cooperation can be channelled towards those who cooperate.

So far, so obvious. The situation becomes more interesting if one assumes that players have only a limited knowledge of their co-players past, or

must cope with unintended defections caused, for instance, by an error, or by the lack of adequate ressources to provide the required help. Kandori (1992) seems to have been the first to study the effects of limited observability in this context. In the extreme case, players know only their own history. Kandori has shown that under certain conditions a so-called 'contagious' equilibrium can still ensure cooperation among rational players: the strategy consists in switching to defection after having encountered the first defection. A single defection by one player is 'signalled', in this sense, to the whole community: but the retaliation may reach the wrong-doer only after many rounds, creating havoc among innocents. Moreover, Kandori has shown that with random matching and no information processing, cooperation cannot be sustained if the population is sufficiently large. Ellison (1994) has shown that cooperation can even be resumed, eventually, if such 'contagious' punishments stop after a signal defined by a public random variable. He notes, however, that such cooperative equilibria are very dependent on the assumption that all players are rational. On the other hand, decentralised mechanisms of local information processing based on a label carried by each agent may allow simple equilibrium strategies leading to cooperation even if occasionally errors occur. After a unilateral defection, players must 'repent' by cooperating, while meekly accepting the defection of their co-players for a certain number of rounds.

3 Indirect reciprocation for evolutionary games

In evolutionary games, it is no longer possible to postulate that players settle on an equilibrium which is sustained by their anticipation of the payoff obtained when they deviate unilaterally. Players are not assumed to be rational, or able to think ahead, deliberate, or coordinate. Strategies are simple behavioral programs; they are supposed to spread within the population if they are successful in the sense of yielding a high payoff (see e.g. Hofbauer and Sigmund, 1998)). Typically, one assumes that such strategies arise randomly within a small minority of the population, by mutation or some other process. The question then becomes whether simple trial-and-error mechanisms resembling natural selection are able to lead, in the long run, to the emergence of cooperative behaviour.

The first papers in this field, by Nowak and Sigmund (1998a,b), led to a number of theoretical and experimental investigations. Roughly speaking, by now the fact that cooperative behaviour based on indirect reciprocity can emerge through evolutionary mechanisms is no longer in doubt, but there is debate on which strategy it is most likely to be based.

In the evolutionary version of the indirect reciprocity game, one considers populations of players which are endowed with some simple strategies. Whenever two players meet in one round of that game, one of them is randomly assigned the role of the donor and the other the role of the recipient. The donor can give help to the recipient: in this case, the recipient's payoff increases by a benefit b whereas the donor's payoff decreases by $-c$, the cost of giving (with $c < b$). The donor can, alternatively, refuse to help, in which case the payoffs of both players are not affected. A player's strategy specifies under which conditions the player should give help, when in the role of the donor.

From time to time, players leave the population and are replaced by new players. The probability that a new player inherits a given strategy occuring within the population is proportional to its frequency, and to the average payoff achieved by players using this strategy. This mimicks selection, but it can just as well be interpreted as a learning process: in that case, players switch their strategies without actually having to die. Some models of evolutionary games also incorporate mutations, which introduce small numbers of players using strategies which were not present in the resident population.

The first model by Nowak and Sigmund (1998a) was based on the concept of a score, a numerical value for reputation. A player's score, at any given time, is defined as difference between the number of decisions to give help, and the number of decisions to refuse help, up to that time. The score of a player entering the game is zero: it then increases or decreases by one point in each round in which the player is in the position of a donor. The range of the score is the set of all integers. This is called the 'full score' model. In a second, 'binary' model, discussed in Nowak and Sigmund (1998b), the range is reduced to two numbers only, 0 (*bad*) or 1 (*good*). This reflects only the players' behaviour in their previous round as a donor. One can, of course, conceive many other ways for keeping score: for instance, by considering neither all the previous actions of the players, nor their last action only, but their last five or ten actions, etc. The decision whether to give help or not should then be based on the scores of the players involved. In particular, a recipient with a high score should be more likely to receive help.

4 Assessment and reprobation

So far, the length of the memory is an aspect which has not attracted much attention. Most of the debate has concentrated on another issue: how should the score be updated? The basic issue is the same as in the framework of games between rational players. Cooperation cannot be sustained without

discriminating against defectors. Players who discriminate must, on occasion, refuse help. If this lowers their score, they will be discriminated against, in future encounters, and obtain a lower payoff. How can such strategies be selected?

One solution is almost obvious. It is to use the same distinction between justified and non-justified defections as Sugden, and hence to rely on the notion of standing. As Nowak and Sigmund (1998b) described it, 'a player is born with good standing, and keeps it as long as he helps 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. We believe that Sugden's strategy is a good approximation to how indirect reciprocation actually works in human societies.' And to the question of Fehr and Fischbacher (2003): 'Should an individual who does not help a person with a bad reputation lose his good reputation?', the answer is, clearly no.

However, two aspects make it worthwhile to investigate image-scoring more closely: one is the argument that standing is a rather complex notion, and seems to require a constant monitoring of the whole population which may overtax the players. Suppose your recipient A has refused help to a recipient B in a previous round. Was this refusal justified? Certainly so, if B has proved to be a helper. But what if B has refused help to some C? Then you would have to know whether B's defection towards C was justified, etc. With direct reciprocation, you have only to keep track of your previous interactions with B. Even here, an error in perception can lead to a deadlock: it may happen that both players believe that they are in good standing and keep punishing each other in good faith (see Boerlijst et al 1997). With indirect reciprocation the problem becomes much more severe: you have to keep track, not only of the antecedents of your current recipient, but of the past actions of the recipient's former recipients etc.

The second interesting aspect of scoring is related to the concept of costly punishment. It is easy to see that the threat of punishment can keep players on the path of cooperation, and thus can solve the social dilemma, which is resumed in the question: why do players contribute to a public good, instead of just exploiting it? They may simply do it to avoid punishment. But if punishment is costly to the punisher, a 'second order social dilemma' arises: why should players shoulder the burden of punishing others? The doctrine of strong reciprocation asserts that many humans are willing to do it, even if they know that they will not meet the punished (and possibly reformed) wrong-doer ever again. Strong reciprocators contribute to the public good, and punish those who don't. There exist several attempts to explain this trait (e.g. Gintis 2000, Fehr and Fischbacher 2003) of which at least one,

incidentally, is based on reputation (Sigmund et al 2001). In the context of indirect reciprocation, we can view discrimination as a form of punishment: low-scorers are deprived of help. If players assess each other according to their standing, the punishment is not costly for the punisher. But if they register only whether the other defected or not, without distinguishing between justified and non-justified defections, then punishment is costly. In view of the fact that many humans are ready to engage in costly punishment in a great variety of contexts (see e.g. Fehr and Gächter 2002), it cannot altogether be excluded that this factor also plays a role in indirect reciprocation. As we shall see in the last section, experiments support this view (Milinski et al, 2001).

On theoretical grounds, it is therefore not obvious how individuals update the scores of their co-players. In fact, this standard of moral judgement, which eventually leads to a social norm, can also be subject to evolution.

In the following investigation we shall assume that individuals engaged in the indirect reciprocation game keep track of the scores in their community, and then decide, when in the role of the donor, whether to give help or not, depending on the recipient's score, and possibly on their own. Needless to say, one can envisage many other strategies, taking into account the accumulated payoffs for donor and recipient, the prevalence of cooperation within the community, the outcome of the last round as a recipient etc. We shall not consider these possibilities in the following models, but start by describing the recent results obtained in two papers, one by Ohtsuki and Iwasa (2004), the other by Brandt and Sigmund (2004), which both, independently, adress the issue of the evolution of updating mechanisms for the indirect reciprocity game. This can be viewed as investigating simple mechanisms for local information processing. But it has farther-ranging implications for the evolution of social norms, and hence of moral judgements. When is a defection justified, or not? When is a player good, or bad? Let us first consider this question in a very limited context, when the score can only take two values.

5 Binary models

We shall assume that every strategy consists of two modules, an assessment module and an action module. The assessment module comes into play when individuals observe interactions between two players. The image of the player acting as potential donor is possibly changed. The image of the recipient, who is the passive part in the interaction, remains unchanged. The action module prescribes whether a player in the position of a potential donor provides help or not, based on the information obtained through the player's assessment

module.

Starting with the assessment module, we shall for simplicity assume that individual A's score of individual B depends only on how B behaved when last observed by A as a potential donor, i.e. whether B gave or refused help to some third party C. Thus A has a very limited memory, and the score of B can only take two values, *good* and *bad*. (In this context, we note that Dellarocas (2003) found that binary feedback mechanisms publishing only the single most recent rating obtained by an online seller are just as efficient as mechanisms publishing the sellers total feedback history). We shall assume that all players are born *good*. In every interaction observed by A, there are two possible outcomes (B can give help or not), two possible score values for B and two for C. Thus there are eight possible types of interaction, and hence, depending on whether they find A's approval or not, $2^8 = 256$ different value systems.

As intuitively appealing examples of such assessment modules, let us consider three of these value systems, or 'morals'. We shall say that they are based on SCORING, STANDING and JUDGING, respectively (these terms are not completely felicitous, but the names of the first two, at least, are fixed by common use). These morals differ on which of the observed interactions incur reprobation, i.e. count as *bad*. Someone using the SCORING assessment system will always frown upon any potential donor who refuses to help a potential recipient, irrespective of the latter's image. Someone using the STANDING assessment system will condemn those who refuse to help a recipient with a *good* score, but will condone those who refuse to help a recipient with a *bad* score. Those using the JUDGING assessment system will, in addition, extend their reprobation to players who help a co-player with a *bad* score.

Thus these three value systems are of different strictness towards wrongdoers. Roughly speaking, someone who refuses to help is always bad in the eyes of a SCORING assessor. Only those who fail to give to a *good* player are bad in the eyes of a STANDING assessor. Someone who fails to give to a *good* player, but also someone who gives help to a *bad* player is bad in the eyes of a JUDGING assessor (see Table 1).

Turning to the action module, we shall assume that a player's decision on whether to help or not is based entirely on the scores of the two players involved. Since there are four situations (donor and recipient can each be *good* or *bad*), there are $2^4 = 16$ possible decision rules. Four intuitively appealing examples would be CO, SELF, AND and OR. CO is uniquely affected by the score of the potential recipient, and gives if and only if that score is *good*. SELF worries exclusively about the own score, and gives if and only if this score is *bad*. AND gives aid if the recipient's score is *good* and the

own score *bad*, and OR gives aid if the recipient's score is *good* or the own score *bad*. Of course the 16 decision rules also include the two unconditional rules, always to give, and never to give, ALLC and ALLD, which do not rely on scores at all. (see Table 2).

A strategy in this model for indirect reciprocity is determined by a specific combination of action and assessment strategy. This yields altogether $2^4 \times 2^8 = 2^{12} = 4096$ different strategies.

6 The leading eight

Ohtsuki and Iwasa (2004) have investigated the evolutionary stability of these strategies. Thus they looked for strategies with the property that a population whose members all use this strategy cannot not be invaded by a small minority using another strategy. Ohtsuki and Iwasa assumed that players were subject to errors, by implementing an unintended move (with a probability μ) or by assigning an incorrect score to a player (with probability ν). Depending on the values of μ, ν, b and c, they found various evolutionarily stable strategies (ESS), including of course ALLD. Most remarkably, they singled out eight strategies (called 'the leading eight') which are robust against errors and lead to cooperation even if b is only slightly larger than c (the ratio must exceed 1 by a factor proportional to the error probabilities).

Only the CO and the OR action module occur among the leading eight. Such players always give help to a *good* player, and defect (when *good*) against a *bad* player. The assessment module of the leading eight is consistent with this prescription: they all assess players as *good* or *bad* if they give (resp. withhold) help to a *good* player, irrespective of their own score, and they all allow *good* players to refuse help to *bad* players without losing their reputation. Interestingly, all other actions towards a *bad* player are possible, i.e. whether a *good* player gives help to a *bad* player, or a *bad* player gives (or refuses) help to a *bad* player. These are just the eight alternatives making up the leading eight. If the assessment module requires a *bad* player to give to a *bad* player, the corresponding action module is OR; in all other cases it is CO. We note that strategies with the STANDING and the JUDGING assessment module can belong to the leading eight, but not those with the SCORING module.

It seems obvious that in an ESS leading to cooperation, assessment rules and action rules should correspond. This requirement does not hold for CO-SCORING, for instance, where *good* players have to refrain from helping *bad* players although this makes them lose their *good* score. Interestingly, there is one exception to this requirement, among the leading eight: for the

last two strategies displayed on Table 3, *bad* players meeting *bad* co-players cannot redress their score one way or the other. However, in a homogenous population playing this strategy, encounters between two *bad* players are exceedingly rare.

Ohtsuki and Iwasa obtained their analytical results under the assumption that players experience infinitely many interactions during their life-time (an approximation which implies that the population is very large). Furthermore, they demand from their ESS strategies only that they are able to repel invasions by strategies with the same assessment module. They also assume that a player's score is the same in the eyes of all co-players. This last assumption is justifed by the so-called 'indirect observation model', which postulates that an interaction between A and B, say, is observed by one player only, for instance C, and that all other members of the population adopt C's assessment. A similar model is used in Panchanathan and Boyd (2003). Other authors, for instance Nowak and Sigmund (1998), Lotem and Fishman (1999) or Leimar and Hammerstein (2001), adopt a 'direct observation model' where all players keep their own, private score of their co-players. Ultimately, it would seem that the evolution of assessment modules will have to be addressed in this context. It is argued that thanks to language, all members of a population should agree on their scores, and it may well be indeed that gossip is powerful enough to furnish all individuals with information about all past interactions. But it is common-day experience that even if two people witness the same interaction directly, they can differ in their assessment of that interaction. This strongly argues for private scores, and has strong implications: as Ohtsuki and Iwasa stressed, CO-STANDING is not an ESS in the direct observation model, but can be invaded, if errors in perception occur, by the undiscriminating ALLC.

7 Replicator dynamics

Another way to approach analytically the evolution of indirect reciprocity is via replicator dynamics. For this, one clearly has to drastically reduce the number of strategies involved. Typically, one considers only three: ALLC, ALLD and a discriminating strategy. Indeed, the main problem for the emergence of discriminating cooperation is that it is threatened by strategies which do not punish defection, and eventually undermine the stability of the helping behavior.

The discriminating strategy usually investigated in this context is CO-SCORING. Let us assume that each player has two interactions per round, one as a donor and one as a recipient, against two different, randomly chosen

co-players. (Assuming one interaction only, with equal probability as donor or recipient, changes the expressions but not the conclusions). We denote the frequency of the indiscriminate altruists, i.e. the ALLC-players, with x, that of defectors, i.e. the ALLD-players, with y, and the frequency of the discriminate altruists, i.e. the CO-SCORING players, with $z = 1 - x - y$. To begin with, we assume that in the first round, discriminators consider their co-players as *good*. With $P_x(n)$, $P_y(n)$ and $P_z(n)$ we denote the expected payoff in the n-th round for ALLC, ALLD and CO-SCORING, respectively. It is easy to see that

$$P_x(1) = -c + b(x + z),$$

$$P_y(1) = b(x + z),$$

and

$$P_z(1) = -c + b(x + z).$$

In the n-th round (with $n > 1$) it is

$$P_x(n) = -c + b(x + z),$$

$$P_y(n) = bx$$

and

$$P_z(n) = -cg_n + b(x + zg_{n-1})$$

where g_n denotes the frequency of *good* players at the start of round n (with $g_1 = 1$) and g_{n-1}, therefore, is the probability that the discriminator has met a *good* player in the previous round and has a *good* score at the start of round n. Clearly $g_n = x + zg_{n-1}$ for $n = 2, 3, ...$ (the *good* players consist of the ALLC players and those discriminators who have met players with a *good* score in the previous round). Hence

$$P_z(n) = (b - c)g_n$$

and by induction

$$g_n = \frac{x}{x + y} + z^{n-1}\frac{y}{x + y}.$$

In the limiting case $n \to +\infty$ this yields

$$P_z = (b - c)\frac{x}{1 - z}.$$

If there is only one round per generation, then defectors win, obviously. This need no longer the case if there are N rounds, with $N > 1$. The total payoffs $\hat{P}_i := P_i(1) + ... + P_i(N)$ are given by

$$\hat{P}_x = N[-c + b(x + z)],$$

$$\hat{P}_y = Nbx + bz,$$

and

$$\hat{P}_z = N(b-c) + y[-b + \frac{b-c}{1-z}(1 + z + ... + z^{N-1} - N)].$$

Let us now assume that the frequencies of the three strategies evolve under the action of selection, with growth rates given by the difference between their payoff $\hat{P}_i$ and the average $\hat{P} = x\hat{P}_x + y\hat{P}_y + z\hat{P}_z$. This yields the replicator equation $\dot{x} = x(\hat{P}_x - \hat{P})$, $\dot{y} = y(\hat{P}_y - \hat{P})$ and $\dot{z} = z(\hat{P}_z - \hat{P})$ on the unit simplex S_3 spanned by the three unit vectors $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ of the standard base.

In there are exactly N rounds in the game, this equation has no fixed point with $x > 0, y > 0$ and $z > 0$, hence the three types cannot co-exist in the long run. The fixed points are: the defectors corner $\mathbf{e}_y$ with $y = 1$; the point $\mathbf{F}_{yz}$ with $x = 0$ and $z + ... + z^{N-1} = c/(b-c)$; and all the points on the edge $\mathbf{e}_x\mathbf{e}_z$. Hence in the absence of defectors, all mixtures of discriminating and indiscriminating altruists are fixed points.

The overall dynamics can be most easily described in the case $N = 2$ (see Fig.1). The parallel to the edge $\mathbf{e}_x\mathbf{e}_y$ through $\mathbf{F}_{yz}$ is invariant. It consists of an orbit with ω-limit $\mathbf{F}_{yz}$ and α-limit $\mathbf{F}_{xz}$. This orbit l acts as a separatrix. All orbits on one side of l converge to $\mathbf{e}_y$. This means that if there are too few discriminating altruists, i.e. if $z < c/(b-c)$, then defectors take over. On the other side of l, all orbits converge to the edge $\mathbf{e}_x\mathbf{e}_y$. In this case, the defectors are eliminated, and a mixture of altruists gets established.

This leads to an interesting behaviour. Suppose that the society consists entirely of altruists. Depending on the frequency z of discriminators, the state is given by a point on the fixed point edge $\mathbf{e}_x\mathbf{e}_z$. We may expect that random drift makes the state fluctuate along this edge and that from time to time, mutation introduces a small quantity y of defectors. What happens then? If the state is between $\mathbf{F}_{xz}$ and $\mathbf{e}_x$, the defectors will take over. If the state is between $\mathbf{e}_z$ and $\mathbf{F}$, the state with $z = 2c/b$, they will immediately be selected against, and promptly vanish. But if a minority of defectors invades while the state is between $\mathbf{F}$ and $\mathbf{F}_{xz}$, then defectors thrive at first on the indiscriminating altruists and increase in frequency. But thereby, they deplete their resource, the indiscriminating altruists. After some time, the discriminating altruists take over and eliminate the defectors. The population returns to the edge $\mathbf{e}_x\mathbf{e}_z$, but now somewhere between $\mathbf{e}_x$ and $\mathbf{F}$, where the ratio of discriminating to indiscriminating altruists is so large that defectors can no longer invade. The defectors have experienced a Pyrrhic victory. They can only take over if their invasion attempt starts when the state is between $\mathbf{F}_{xz}$ and $\mathbf{e}_x$. For this, the fluctuations have to cross the gap between $\mathbf{F}$ and

$\mathbf{F}_{xz}$. This takes some time. If defectors try too often to invade, they will never succeed.

In the limiting case that the number of rounds N is infinite, we obtain for the average payoffs P_i per round, that

$$P_x - P_y = bz - c$$

and

$$P_z - P_y = \frac{x}{1-z}(P_x - P_y).$$

In the interior of S_3, the fixed points form a line $z = c/b$ parallel to the edge $\mathbf{e}_x\mathbf{e}_y$. We denote this line by l (it is just the limit of the separatrix l in the previous paragraph, for $N \to +\infty$). The edges with $x = 0$ and $z = 0$ consist of fixed points. In the interior of S_3 all orbits are parallel to l. Those with $z < c/b$ converge to the left (the discriminating altruists vanish), those with $z > c/b$ to the right (the undiscriminate altruists vanish) (see Fig.2).

If there is a fixed probability $w < 1$ for a further round (see Nowak and Sigmund, 1998b), we obtain for the total payoff values:

$$P_x - P_y = \frac{wbz - c}{1-w}$$

and

$$P_z - P_y = \frac{1 - w + wx}{1 - wz}(P_x - P_y)$$

and the fixed points form the line l defined by $z = c/wb$, as well as the $\mathbf{e}_x\mathbf{e}_z$-edge. In the interior of S_3 the orbits are on the lines with $z = ax^{1-w}$ (see Fig. 3). Above l the orbits converge to the fixed point edge with $y = 0$, below l to the vertex with $y = 1$. The state will drift along the fixed point lines until a mutation sends it to the region below l, where the defectors win.

It is clear that such a degenerate behaviour is rather sensible to perturbations. Let us assume that errors in implementation can occur. For simplicity, we consider only errors turning an intended cooperation into a defection with a certain probability $1 - r$. Equivalently we may assume, following Lotem et al (1999), that $1 - r$ is the probability that an individual is actually unable to perform the intended act of giving help (this incapacity may be due, for instance, to a lack of resources or an injury). Such an incapacity is highly likely: as Fishman (2004) wrote, 'individuals who are always able to help do not need help from others... In practice, one donates help when the costs are small, in order to secure reciprocity in the hour of need.' The defectors' payoff in the first round is $P_y(1) = rb(x + z)$, and in all further rounds it is $P_y(n) = rbx$. In the n-th round ($n > 1$) we obtain $P_x(n) = -rc + br^2z + P_y(n)$,

and $P_z(n) = -rcg_n + rbzrg_{n-1} = r(b-c)g_n - br^2x + P_y(n)$, where g_n, the frequency of players with a *good* image at the start of the n-th round, satisfies $g_n = r(x + zg_{n-1})$ and is given by

$$g_n = \frac{rx}{1-rz} + (rz)^{n-1}\frac{1-rx-rz}{1-rz}$$

(clearly $g_1 = 1$ and $P_x(1) = P_z(1) = -rc + P_y(1)$). These expressions have been obtained by Panchanathan and Boyd (2003) and by Fishman (2004). In the limiting case of infinitely many rounds,

$$P_z - P_y = \frac{rx}{1-rz}(P_x - P_y).$$

Once more, we obtain a line l of equilibria in the interior of S_3, given by $z = c/br$. This line intersects the edge $\mathbf{e}_x\mathbf{e}_z$ (where all players are altruists) at the point $\mathbf{F}_{xz}$. This time, the edge does not consist of fixed points: in fact $\mathbf{F}_{xz}$ is the only equilibrium mixture of discriminating and indiscriminating altruists, and it is stable within the edge $\mathbf{e}_x\mathbf{e}_z$ of altruists. Indeed, if almost all altruists are indiscriminating, then the unintended defections which cause discriminating altruists to refuse help in the next round will allow them to obtain a higher payoff than ALLC-players without being taken to account too frequently; whereas if most altruists are discriminating, most refusals to help will be severely punished. If we consider an arbitrary mixture of defectors and altruists, the orbit will either converge to $\mathbf{e}_y$ or it will first converge to the line l, drift along this line and then, if a random shock introduces some more defectors while the frequency of undiscriminating altruists is sufficiently low, to$\mathbf{e}_y$. In any case the evolution will ultimately lead to the fixation of defectors (see Fig.4).

Panchanathan and Boyd (2003) have noticed that the same happens if the number of rounds is not infinite, but a random variable with a geometric distribution given by a parameter w (a constant probability for a further round). In contrast, they found that if the discriminating strategy is OR-STANDING or CO-STANDING, then the monomorphic state with all players discriminating is stable. They also found that OR-STANDING is slightly superior to CO-STANDING (which does not belong to the 'leading eight', incidentally). Panchanathan and Boyd concluded that 'when errors are added, indirect reciprocity cannot be based on an image-scoring strategy'. And indeed they have pointed out an important vulnerability of the CO-SCORING strategy. Nevertheless, the verdict seems to depend on the modelling assumptions.

Indeed, as shown by Fishman (2004), if one assumes that the number N of rounds is constant, then the equilibrium $\mathbf{F}_{xz}$ is transversally stable, i.e. it cannot be invaded by defectors if $c/b < 1 - 1/N$, and if r is sufficiently

large: for this, one has only to check that the payoff values $P_x = P_z$ at $\mathbf{F}_{xz}$ exceed P_y. Hence cooperation can be stably sustained. Brandt and Sigmund (2004) showed that the same holds if the number of rounds is a random variable with a Poisson distribution with parameter λ, provided $b > 2c$ and λ is sufficiently large. In both cases, the model leads to a bistable dynamics (see Fig.5). Depending on the initial condition, either defectors take over, or the population converges to a stable mixture of discriminating and undiscriminating altruists, and hence to a cooperative regime.

Fishman stressed, therefore, that involuntary defection (caused by errors, or by incapacitation) stabilises indirect reciprocity. He states: 'Indirect reciprocity, at least in the current case, is stable only among imperfect individuals.' In Lotem et al (1999), Lotem et al (2002), and Sherratt and Roberts (2001), this inability of giving help, due to lack of quality, is further analysed: helping behaviour is used as a way of signalling high quality (see also Zahavi, 1995).

Ohtsuki (2004) studied adaptive dynamics for stochastic strategies of the CO-SCORING type. His strategies are given by triples (p_0, p_1, p_2) where p_0 is the probability to help a player in the absence of information about his score (an event whose likelihood is q), whereas p_1 and p_2 are the probabilities to help an individual with *good* resp. *bad* score. In his analysis of monomorphic populations, Ohtsuki finds that there exist two regions, one in which all p_i-values increase and one in which all decrease. As in the case of direct reciprocation (see Nowak and Sigmund, 1990), the discriminating strategies (with $p_1 = 1$ and $p_2 = 0$) act not as end-points but rather as pivots of the evolution: in their neighborhood all p_i-values increase but the degree of discrimination $p_1 - p_2$ decreases so that eventually a continuum of equilibria is approached. Once there, mutations can send the population towards defection. This instability is even more pronounced if errors in perception or implementation are included.

8 Asynchronous entry

So far, we have assumed that the whole population lives according to the same schedule: all players engage together in the first round (once as donor and once as recipient), then all in the second round etc... This can indeed model what happens with a group of persons volunteering for an experimental game. But for real-life interactions, it may seem more appropriate to model a population with generations blending into each other. Occasionally, a new player is born, and will experience exactly n rounds of the game with probability $e(n)$ (for $n = 0, 1, 2, ...$). In contrast to the previous model, differ-

ent players will usually experience a different number of rounds: these rounds are no longer synchronised. Let us denote by g the frequency of individuals with *good* reputation in the population. If the population is sufficiently large, and stationary, then g will not be affected by the birth of a new individual, or its age. Let us assume, to begin with, that newcomers are considered as *good*. After the first round, an ALLC player will have a *good* reputation with probability r, an $ALLD$ player with probability 0 and a discriminator with probability rg. Hence $g = rx + rzg$ and thus

$$g = \frac{rx}{1 - rz}.$$

The payoff for an ALLC player is $-cr + br(x + z)$ in the first round and $-cr + br(x + zr)$ in all following rounds. For an ALLD player it is $br(x + z)$ in the first round and brx in all following rounds. For a discriminator the payoff is$-cgr + br(x + z)$ in the first round and $-cgr + br(x + gzr)$ in all following rounds. Thus $P_z - P_y = g(P_x - P_y)$.

In the limiting case of infinitely many rounds we see that

$$P_z - P_x = r(1 - g)(c - brz)$$

which yields again a line l of fixed points satisfying $z = c/br$. The phase portrait looks like that of Fig 4 (if $r > c/b$), and defectors will always win in the end. This holds also if the number of rounds is any random variable with expectation value E, except that the z-value of l has to be multiplied by $E/(E - 1 - w(0))$. Indeed, since $P_x = P_z$ holds if and only if $P_x = P_y$, it follows that $P_x = P_z$ always defines a line of fixed points.

A similar result is obtained in a model where discriminators know their co-player's reputation (i.e. their behaviour in the last round) only with a certain probability q, and assume that it is *good* if they have no information We note in this context that Panchanathan and Boyd (2003) have shown that for sufficiently large q and b/c, it is selectively advantageous to be trustful in this sense. For ALLC players, the payoff is $-cr + br(x + z)$ in the first round and $-cr + br[x + (1 - q)z + qzr]$ in all subsequent rounds. For ALLD players, it is $br(x+z)$ in the first and $br[x+(1-q)z]$ in all subsequent rounds. For discriminators, it is $-cr(1 - q) - crqg + br(x + z)$ in the first round and $-cr(1 - q) - crqg + brx + br(1 - q)z + brqzr[(1 - q) + qg]$ in all subsequent rounds. ALLC players and discriminators have the same payoff iff $z = c/br$, in the limiting case of infinitely many rounds; Since

$$P_x - P_z = rq(1 - g)(P_x - P_y)$$

holds in every round, there exists, for sufficiently large r, an equilibrium mixture of discriminating and undiscriminating altruists, but this equilibrium

can always be invaded by defectors. The same holds also for other scoring strategies, as for instance for OR-SCORING; it also holds if we assume that a discriminator who does not know the recipient's score defects. Thus we see that the argument of Panchanathan and Boyd (2003) is even more robust for the asynchronous entry case than for the case of synchronised rounds: it holds whenever the probability that a discriminator gives help is the same from one round to the next.

But assume now that q_n, the probability that a discriminator engaged in round n knows the score of the co-player, is increasing in n. This assumption is plausible: with time, a player's social network grows, and therefore also the player's probability to have information about the recipient. Of course, if the population has reached a steady state, then the average probability that a randomly chosen player knows a co-player's score is just the mean value of the q_n, i.e. some constant q. If we assume, as before, that discriminators are trustful, in the sense that they provide help if they do not know the co-player's score, then we obtain as payoffs in the n-th round:

$$P_x(n) = -cr + brx + br(1-q)z + br^2qz$$
$$P_y(n) = brx + br(1-q)z$$

and

$$P_z(n) = -cr[(1-q_n) + q_n g] + brx + br(1-q)z + brqzr[(1-q_{n-1}) + q_{n-1}g].$$

Thus

$$P_x(n) - P_y(n) = -cr + br^2qz$$

and

$$P_z(n) - P_y(n) = P_x(n) - P_y(n) + r(1-g)[cq_n - brzqq_{n-1}].$$

Clearly

$$P_x(n) - P_z(n) = r(1-g)(-cq_n + zbrqq_{n-1}).$$

Let $w(n)$ be the probability that a randomly chosen donor is in round n, then

$$q = \sum w(n)q_n > \hat{q} := \sum w(n)q_{n-1}.$$

We have $P_x(n)(\hat{z}) - P_y(n)(\hat{z}) = 0$ and

$$P_z(n)(\hat{z}) - P_y(n)(\hat{z}) = cr(1-g)(q_n - q_{n-1}) > 0.$$

In Brandt and Sigmund (2005) it is shown that with $z_{cr} = \frac{c}{br\hat{q}}$, there exists a mixture of discriminating and indiscriminating altruists $\mathbf{F}_{xz} = (1 -$

$z_{cr}, 0, z_{cr}$) which is a fixed point. For sufficiently small $w(1)$ (i.e. a sufficiently large likelihood of having more than one round)

$$P_x(z_{cr}) > P_y(z_{cr}).$$

Hence $\mathbf{F}_{xz}$ cannot be invaded by the defectors. The resulting replicator equation is bistable: one attractor consists of defectors only, the other of a mixture of discriminating and undiscriminating altruists.

If $q_n < q_{n-1}$ this would not be valid: except if we assume that discriminators who do not know the recipient's score, instead of helping, i.e. according the benefit of doubt, prefer to refuse help, i.e. to act distrustfully. To resume, we see that if either players are trusting and have a growing social net, or if they are distrustful and have a shrinking net of acquaintances, a stable mixture of discriminating and indiscriminating altruists can be supported by the SCORING assessment module.

9 Numerical simulations

It seems hard to derive analytical expressions for the payoff values if several discriminating strategies are present, and errors in perception and implementation, limited observability etc are taken into account. Thus while it is easy to compute the payoff expressions for mixtures of CO-SCORING with ALLC and ALLD, merely adding OR-SCORING or CO-STANDING to the cast greatly complicates things. Often, pairs of discriminating strategies perform equally well against each other, so that their frequencies drift randomly around: but the success of other strategies at invading them depends on their frequencies, etc. One is often reduced to numerical simulations to investigate such polymorphic states.

In Nowak and Sigmund (1998a,b), well-mixed populations are considered, consisting of some 100 individuals each engaged in some five or ten interactions, sometimes as as a donor, and sometimes as a recipient. But in order to avoid spurious effects of random drift, it is convenient to adopt, following Leimar and Hammerstein (2001), a population structure conveying a more realistic image of prehistoric mankind, and consider some 100 tribes, for instance, with 100 players each, with some modest gene flow between the tribes. We shall start by describing the extensive statistical investigations of Brandt and Sigmund (2004), based on such a population structure, and the assumption of a binary score.

Let us consider the case of separate generations. During one generation, there will be 1000 games within each tribe, so that on average each player is engaged in 10 rounds (a larger number does not significantly change the

outcome). Each individual keeps a private score of all tribe-members. We normalise payoffs by setting $c = 1$, so that b is now the cost-to-benefit ratio. At the end of each generation, each tribe forms a new generation of 100 individuals: with probability p the new individual will be 'locally derived' and inherit a strategy from a member of the tribe, and with probability $1-p$, the new individual will inherit a strategy from some member at large, in each case with a probability which is proportional to that member's total payoff. In order to avoid transitional effects, we present averages over 1000 generations, after an initial phase of 9000 generations. (Usually, a stable composition is reached within 100 generations). In Brandt (2004) one can find an online approach to such numerical simulations which allows the visitors of that site a great deal of experimentation.

Let us first ask which strategies are best at invading a population of defectors, when introduced as a minority of, for instance, 10 percent. It turns out that in the absence of errors, STANDING and JUDGING, together with the CO and the OR module, do best and lead to cooperation whenever $b \geq 4.5$, whereas SCORING requires considerably higher b-values. In the presence of errors, this is attenuated: if, for instance, ALLC, ALLD and a single discriminating strategy are initially equally frequent, then CO-STANDING and OR-STANDING eliminate defectors whenever $b > 3.5$, whereas CO-JUDGING and CO-SCORING require $b > 4.5$, and OR-JUDGING and OR-SCORING even $b > 6.5$.

If a given assessment module is held fixed and several action-modules start at similar frequencies, then cooperation dominates for STANDING and for SCORING as soon as $b > 4$, usually with the CO or the OR module (together with a substantial ALLC population). Less cooperative action modules, as for instance SELF or AND, are rapidly eliminated.

There is a strong propensity for cooperation based on polymorphisms. Let us, for instance, start with a population where the three assessment modules SCORING, STANDING and JUDGING as well as the action modules AND, OR, CO and SELF, together with the indiscriminate strategies ALL C and ALL D are present in equal frequencies. Even if only every second interaction is observed, a cooperative outcome is usually achieved as soon as $b > 2.5$, and CO-SCORING, OR-SCORING, CO-STANDING and OR-STANDING prevail at nearly equal frequencies. JUDGING is grealy penalised by the lack of reliable information. On the other hand, if all interactions are observed and only errors in implementation occur, then CO-JUDGING and OR-JUDGING dominate, eliminating ALLC players and establishing a very stable cooperative regime. If errors in perception occur, then JUDGING is completely eliminated, and SCORING and STANDING perform on a similar level. This also holds if errors in implementation or limited observability are

taken into account.

In a recent and as yet unpublished paper, Takahashi and Mashima (2004) have shown that STANDING is highly vulnerable to errors in perception, if one does not consider a subdivided population linked by migration, as in Leimar and Hammerstein, but a single well-mixed tribe. On the other hand, they emphasised the success of a strategy which had not been considered before, and in particular is not a member of the 'leading eight'. Its action module is CO, and its assessment module ascribes a *bad* score, not only to those refusing help to a *good* player, but to all those who interacted with a *bad* player (irrespective of whether they provided help or not). Players who have met a *bad* player are *bad* and remain so until they are able to redeem themselves by giving to a *good* player. According to Takahashi and Mashima, it remains still to be checked whether such intriguing strategies can get established in more polymorphic populations.

10 Spatial Indirect Reciprocation

In a variant of evolutionary games, spatially distibuted populations are considered, with each individual interacting only with the closest neighbors and updating by switching to the strategy of a random neighbor with a probability proportional to the payoff difference. Let us assume, for instance, that the players sit on an $N \times N$-lattice, with the usual identification of opposite borders, and that the neighborhood of site (i, j) consists of the 8 sites whose coordinates differ by at most one unit (a Moore neighborhood). Since the score depends on how many games have already been played, it is important to introduce no systematic bias in the ordering of the games. A simple approach is to arrange all individuals in a random sequence and let the interactions take place in that order, with this individual as recipient, and one of the neighbors (randomly chosen) as potential donor. Individuals cannot receive help more than once per round, but they may be asked more than once to help a co-player. Not surprisingly, the spatial games lead to the evolution of cooperation for even smaller b/c-values than in the well-mixed case (see Fig. 6, Table 4 and, for interactive experimentation, Brandt 2004). In these simulations, a small mutation probability and a probability of not being able to cooperate, due to lack of resources for example, is included. Moreover, discriminators are tempted to defect instead of helping with a small temptation rate. Every generation consists of 5 rounds played as described. Then, in the spatial case, sites are updated by comparing their payoff with that of a randomly chosen neighboring site (a randomly chosen site of the full lattice, in another variant) and switching to the strategy at that site with a

probability proportional to the payoff difference, if the own payoff is lower. In the spatial case, when updating occurs only between neighbors, cooperation dominates for $b/c \geq 2$, whereas if the population is well-mixed, it takes $b/c \geq 3.5$ to suppress defectors to a small minority.

11 Full score

The original numerical simulations of Nowak and Sigmund (1998a) considered the case, not of a binary score, but of a full score ranging through all integer values. This means that if, on average, individuals experience only five rounds as a donor, their score cannot exceed ± 5. This score range seems much more natural than the restriction to a binary score. In fact, binary scores were only introduced as a crude simplification to allow for analytical results.

With a full score, one can again consider the same assessment modules as before, and in particular SCORING or STANDING. One can also consider different action modules, but their number vastly incrases. For instance, in the OR-family, we would find all strategies of the type $(k \vee h)$, meaning 'help if the recipient's score exceeds k or if your own score is below h'. It seems intuitively clear that the main disadvantage of SCORING, namely that punishing is costly, is greatly reduced. Indeed, players with a high reputation for helping will be able to refrain occasionally from helping a low-scorer without threatening their own score, which will be reduced by one unit but remain in the high range. The numerical simulations of Nowak and Sigmund (1998a) were confirmed by Leimar and Hammerstein (2001), who found, however, that a modest gene-flow between groups reduced the success of SCORING. Thus while, for $b/c = 4$, AND-SCORING produces on average 40 percent of cooperation in isolated groups without migration, it does much less well if mixing occurs between the groups. We note that the poor showing of the AND-module is also reflected in the simulations with a binary module. Such strategies are not cooperative in the sense that they do not always lead to help-giving in a monomorphic population.

Leimar and Hammerstein also reported an interesting robustness of the STANDING module against errors of perception, adding that the issue was not fully resolved yet. Indeed, a systematic investigation of the different assessment modules for the full-score case is lacking so far, due in part to analytical difficulties, and in part to the fact that the proliferation of strategies for each assessment module often leads to neutral polymorphisms which are dominated by random drift rather than a clear-cut selective force. It seems safe to predict that the costs of complexity, the prevalence of phenotypic defectors (i.e. players unable to give help even if they want to) and the issue of

public vs private scores will become essential topics for these investigations.

In an interesting approach, Mohtashemi and Mui (2003) have performed agent-based simulations based on the SCORING module, using players with growing networks of acquaintances (in every round, the donors and their acquaintance are added to the acquaintance of the recipients). They found that this greatly promotes the emergence of cooperation, a result which is well in tune with our analysis of the replicator dynamics in the asynchronous entry case.

12 Experimental Games

Wedekind and Milinski (2000) set up experiments with 79 undergraduate students, who were divided into eight groups. All players were provided with a starting account, and were repeatedly offered the possibility to give 4 Swiss Francs to another person of the same group, at a cost of 1 (or, in some groups, 2) Swiss Franc to themselves. Players knew that they would never meet the same person in the reciprocal role. The interactions were anonymous, but the potential donors were shown the history of giving or not giving of the potential recipient before they were asked for their decision. There were six rounds in each group, and each player was once per round a potential donor, and twice per round a receiver, although this was not announced beforehand. The frequency of giving ranged from 48 to 87 percent, depending on the group. As expected, those groups with a lower cost of giving (or with a higher starting account) donated more often. The image score of potential recipients correlated well with their expectation to actually receive money. The amount of discrimination was higher among those players who donated less often: apparently, those who were more generous cared less about the recipient's image score.

In a similar experiment, Seinen and Schram (2001) found corresponding results. In particular, they concluded that subjects are much more likely to help if they know that their score is passed on. They also found that groups develop different norms, i.e. minimal thresholds for the score. 'Finding a norm that is consistent with the own social status... is important in synchronizing norms within a group'. Seinen and Schram also found clear evidence that the own score becomes an important factor in the decision to help, when players know that it is communicated to future donors.

In an interesting variant of the indirect reciprocation game, Milinski et al (2002a) showed that if players were given the opportunity, between rounds, to make a public donation to a charity, the amount of their donation correlated positively with the likelihood that they would receive money, in subsequent

rounds, from their co-players.

Bolton, Katok and Ockenfels (2001) performed a variety of experiments with high or low costs ($b/c = 5$ or $= 5/3$) and with three different information conditions: (a) no information, (b) first order information (whether the recipient gave help when last in the role of the donor) and (c) second order information (the recipient's decision when last in the role of the donor, and the previous move of the recipient of that game). We note that (b) and (c) both allow SCORING to be implemented, but that (c) does not provide all the information needed to implement a STANDING strategy. The hypothesis that more information leads to more giving is confirmed in the experiments of Bolton et al (2001). Also, giving is higher in earlier rounds, when reputation has a higher impact on the future income. However, it appears that even if there is no information, some players are prone to cooperate. Furthermore, the decisions of the donors seem also to be affected by how often they were given. This shows that some relevant aspects of the game have not yet been covered by models. Players seem to be affected by what they have received, and tend to give because they received help. Strategies basing the decision to give on the score and, additionally, on the payoff history, i.e. the donor's past income, seem plausible, but apparently have not yet been investigated. But let us stress that Bolton et al (2001) found a significant positive correlation between the number of gifts given by players and the number they receive. They also found that there is a slight, negative correlation between the number of gifts given and the total payoff obtained by a player.

This last result stands in contradiction to the findings obtained by Wedekind and Braithwaite (2002): in their experiment, those who gave much ended up with the highest payoff. Donors knew only the score of the recipient, calculated according to the SCORING rule, on a scale of integers ranging from -6 to +6. Wedekind and Braithwaite found evidence for the OR module. From the twelfth round onward, there was a positive correlation between image score and total payoff, statistically significant in most rounds. Thus generosity pays in this kind of game, which argues for its selective advantage. The correlation within a population increases with the mean generosity of the group. In a subsequent game of direct reciprocity (six rounds of the Prisoner's Dilemma game between the same two players) the display of the previous end score tended to boost cooperation towards generous players in the first three rounds, and then was superseded, reasonably enough, by the personal experience obtained with the given co-player.

A similar interconnection between direct and indirect reciprocation can be found in Milinski et al (2002b). They combine rounds of an indirect reciprocity game with rounds of a public good game. The donors in the indirect reciprocity game are also informed about the recipients' actions in

the public goods game. If the two games alternate, contributions to the public good game remain high, while they quickly deteriorate in an unbroken succession of public good games. This experiment provides evidence that indirect reciprocity has a similar impact as direct reciprocity. Moreover, the version with alternating rounds can be viewed as a sequence of public good games with the possibility, after each round, of rewarding contributors. It thus offers an intriguing complement to the literature on public goods with punishment (see, e.g., Fehr and Gächter 2000).

Engelmann and Fischbacher (2002) ran experiments designed to find out whether donors were more motivated with keeping up their own score or with reacting to the recipients' score. At any time, only half of the players had a public score (assessed according to SCORING), which was displayed when they were recipients. There was clear evidence that donors without score react to the recipient's score; such donors cannot be guided by selfish motives. On the other hand, the propensity to give more than doubled, for many players, if they were told that their action would affect their own score. Such subjects also seem to be less influenced by the recipient's score. This provides strong evidence for selfish reputation-building. Further evidence for such 'strategic' use of reputation has been obtained by Semmann et al (2004).

In another series of experiments, Milinski et al (2001) addressed the question of STANDING versus SCORING. Each group included a bogus player who always refused to help. Discriminating player should always refuse to give aid to such a player. The question was: would these players, in turn, be penalised by their co-players or not? The former outcome would speak for the prevalence of a SCORING strategy, the latter for STANDING. Players were again anonymous, and were given, not only the history of the receiver, but also that of the receivers' previous receivers, so that they could judge whether a defection by the receiver was justified or not. ($b/c = 4/5$). It was found that the potential donors of the sham defector (whose refusals were justified) experienced significantly more defections than STANDING would predict, but less than SCORING would predict. Interestingly, the donors of the sham defector tended to be more generous in their other interactions, as if they expected to be punished and wanted to redress their score. This suggests that players do not expect that other players follow a STANDING strategy.

The same result held, surprisingly, when the experimenters provided only the history of the receiver (so that a STANDING strategy was actually impossible to implement). Indeed, the statistics of the games with full information (where donors were provided with the complete histories of all co-players) and with restricted information (where players were only provided with the list of previous actions of the potential recipient) look remarkably similar.

With full information, the players took a longer time to reach their decision. This suggests that they tried to interpret the complex histories. But after three or four rounds, it becomes rather complicated to work back through the histories of the recipient's recipients etc., so that players most likely were overburdened, cognitively, and simply stopped to care about details, possibly falling back to some mixture of SCORING and STANDING.

This cognitive problem is, in part, due to the design of the economic experiments. The players do not have a close acquaintance with each other, and can distinguish their group members just by their pseudo-names, so that they are not really involved with them. It could be argued that in more life-like interactions within a real group, individuals are familiar with each other's personalities, and thus find it easier to update their image scores in real time. It would facilitate the players' task of keeping track of their co-players' standing if they were told, after each round, to update all the image scores within the group, and note them down. The drawback of such an instruction is that it necessarily suggests to the participants that these image scores form a key element of the game. The players would no longer be 'naive' with respect to the experiment, but approach it with a certain bias. On the other hand, given that it can by now be granted that *some* type of image score is involved in this kind of game, it could be worth trying to provide players with an instruction like: 'Write down, between each round, who did the right thing, in your eyes, and who did not.' From the resulting protocols, it should be possible to find out the assessment modules and, comparing this with the decisions taken by the players, the action modules.

Another possible way of clarifying the situation would be to subject players to very short histories only. For instance, one could start by explaining the rules of the game, and then let groups of six or ten players actually play ten or twelve rounds, sitting face to face with each other, so that they thoroughly understand what they are about. Then, one could separate the players, place each into some cubicle, and tell them that they would now play the same game, with a new group of co-players with whom they could interact only via computer. In reality, they would all be confronted, in the third round, with a fictitious co-player who had given in the first round, but refused to give in the second round against a recipient who had refused to give in the first round. This should disentangle the SCORING vs STANDING issue. It is considered bad form, in economic games, to mislead players. But in view of the importance of the question, this may be considered a white lie. Morally it may not be quite right, but it can help us to better understand morals and their evolution.

Acknowledgements

H. Brandt and K. Sigmund acknowledge support from the Wissenschaftskolleg "Differential Equations" of the Austrian Science Fund FWF. We thank Josef Hofbauer, Martin Nowak and Drew Fudenberg for helpful conversations.

References

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

Axelrod, R and Hamilton, W D 1981 The evolution of cooperation, Science 211, 1390-6

Berg, J. Dickhaut, J. and McCabe, K. 1995 Trust, Recipocity and Social History, Games and Economic Behavior 10, 122-142

Binmore, K G (1992) Fun and Games: a text on game theory, Health and Co, Lexington, Mass

Boerlijst, M C, Nowak, M A and K. Sigmund (1997) The logic of contrition, JTB 185, 281-294

Bolton, G., Katok, K. and Ockenfels, A., 2004, Cooperation among strangers with limited information about reputation, Journal of Public Economics, in press

Bolton, G., Katok, E. and Ockenfels, A 2004 How effective are online reputation mechanisms? An experimental investigation, Management Science, in press

Bolton, G.E. and Ockenfels, A. 2000, ERC: a theory of equity, reciprocity and competition, American Economic Review 90, 166-193

Boyd, R and Richerson, P J 1989 The evolution of indirect reciprocity, Social Networks 11, 213-236

Boyd, R and Richerson, P J 1992 Punishment allows the evolution of cooperation (or anything else) in sizeable groups, Ethology and Sociobiology 113, 171-195

Brandt, H. 2004, URL: http://homepage.univie.ac.at/hannelore.brandt/simulations Interactive java applets for online experimentation on presented simulations

Brandt, H. and Sigmund, K (2004) The logic of reprobation: assessment and action rules for indirect reciprocity, JTB, 231, 475-486

Brandt, H. and Sigmund, K. (2005) Indirect reciprocity, image scoring, and moral hazard, to appear in PNAS

Camerer, C E 2003 Behavioral Game Theory, Princeton UP

Colman, A. M. 1995 Game Theory and its Applications in the Social and Biological Sciences, Oxford: Butterworth-Heinemann.

Dawes, R. M. 1980 Social Dilemmas. Ann. Rev. Psychol. 31, 169.

Dellarocas, C 2003 Efficiency and Robustness of binary feedback mechanisms in trading environments with moral hazard (working paper, MIT Sloan School of Management)

Ellison, G. (1994) Cooperation in the Prisoner's Dilemma with anonymous random matching, Review of Economic Studies, 61, 567-588

Engelmann, D. and U. Fischbacher 2002 Indirect reciprocity and strategic reputation-building in an experimental helping game, working paper, Univ of Zürich

Fehr, E. and Gächter, S. 2000 Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90, 980.

Fehr, E. and Gächter, S. 2002 Altruistic punishment in humans. Nature 415, 137.

Fehr, E. and Fischbacher, U. (2003) The nature of human altruism, Nature 425, 785-791

Ferrière, R. 1998, Help and you shall be helped, Nature 393, 517-519

Fishman, M.A. 2003 Indirect reciprocity among imperfect individuals, JTB 225, 285-292

Fishman, M.A., Lotem, A. and Stone. L. 2001 Heterogeneity stabilises reciprocal altruism interaction, JTB 209, 87-95

Fudenberg D and Maskin, E 1986 The folk theorem in repeated games with discounting or with incomplete information, Econometrica 50, 533-554

Gintis, H 2000 Game Theory Evolving, Princeton UP

Harbaugh, W T 1998 The prestige motive for making charitable transfers, American Economic Review 88, 277-282

Hofbauer, J. and Sigmund, K (1998) Evolutionary Games and Population Dynamics, Cambridge UP

Kandori, M 1992 Social norms and community enforcement, The Review of Economic Studies 59, 63-80

Keser, Claudia 2002 Trust and Reputation Building in e-Commerce, Working paper, IBM Watson Research Center

Leimar, O. and Hammerstein, P. 2001, Evolution of cooperation through indirect reciprocation, Proc R Soc Lond B, 268, 745-753

Lotem, A., Fishman, M A and Stone, L 1999 Evolution of cooperation between individuals, Nature 400, 226-227

Lotem, A., Fishman, M A and Stone L 2002 Evolution of unconditional altruism through signalling benefits, Proc R. Soc London B, 270, 199-205

Milinski, M., Semmann, D., Bakker, T.C.M. and Krambeck, H. J. 2001 Cooperation through indirect reciprocity: image scoring or standing strategy? Proc Roy Soc Lond B 268, 2495-2501

Milinski, M., Semmann, D. and Krambeck, H.J. 2002a Donors in charity gain in both indirect reciprocity and political reputation, Proc Lond Soc B 269, 881-883

Milinski, M., Semmann, D. and Krambeck, H.J. 2002b Reputation helps solve the 'Tragedy of the Commons', Nature 415, 424-426

Mohtashemi, M. and L. Mui 2003 Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism,

JTB 223, 523-531

Nowak, M.A. and Sigmund, K (1990) The evolution of stochastic strategies in the prisoner's dilemma, Acta Appl. Math. 20, 247-265.

Nowak, M.A. and Sigmund, K. 1998a Evolution of indirect reciprocity by image scoring, Nature 282, 462-466 Nowak, M.A. and Sigmund, K. 1998b, The dynamics of indirect reciprocity, JTB 194, 561-574

Ohtsuki, H (2004) Reactive strategies in indirect reciprocity, JTB 227, 299-314

Ohtsuki H, and Iwasa, Y (2004) How should we define goodness? – Reputation dynamics in indirect reciprocity, JTB 231, 107-120

Okuno-Fujiwara, M and Postlewaite, A 1995 Social Norms in Matching Games, Games and Economic Behaviour, 9, 79-109

Panchanathan, K. and R. Boyd 2003 A tale of two defectors: the importance of standing for evolution of indirect reciprocity, JTB 224, 115-126

Pollock, G B and L A Dugatkin 1992 Reciprocity and the evolution of reputation, JTB 159, 25-37

Rosenthal, R W 1979 Sequences of games with varying opponents, Econometrica 47, 1353-1366

Seinen, I. and Schram, A. 2001 Social status and group norms: indirect reciprocity in a helping experiment, working paper, CREED, Univ. of Amsterdam

Semmann D, Krambeck, H J and Milinski, M (2004) Strategic investment in reputation, J Behav. Ecol. Sociobiol. 56, 248-252

Sheratt, T N and Roberts, G 2001 The importance of phenotypic defectors in stabilising reciprocal altruism, Behav. Ecol. 12, 313-317

Sigmund, K., Hauert, C., and Nowak, M. A. 2001 Reward and punishment. Proc. Natl. Acad. Sci. 98, 10757-10763

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

Takahashi, N. and R. Mashima 2004, The importance of indirect reciprocity: is the standing strategy the answer? Working Paper Hokkaido Univ

Trivers, R 1971 The evolution of reciprocal altruism, Quart Rev Biol 46, 35-57

Wedekind, C and Milinski, M 2000 Cooperation through image scoring in humans, Science 288, 850-852

Wedekind, C 1998 Give and ye shall be recognised, Science 280,2070-2071

Wedekind, C. and Braithwaite, V.A. 2002 The long-term benefits of human generosity in indirect reciprocity, Curr Biol. 12, 1012-1015

van der Heijden, E C M (1996) Altruism, Fairness and Public Pensions, PhD thesis Amsterdam

Zahavi, A 1995 Altruism as a handicap – the limitations of kin selection and reciprocity, J Avian Biol 26, 1-3

Assessment Modules

situation/strategy	SCORING	STANDING	JUDGING
good $\rightarrow$ good	good	good	good
good $\rightarrow$ bad	good	good	bad
bad $\rightarrow$ good	good	good	good
bad $\rightarrow$ bad	good	good	bad
good $\not\rightarrow$ good	bad	bad	bad
good $\not\rightarrow$ bad	bad	good	good
bad $\not\rightarrow$ good	bad	bad	bad
bad $\not\rightarrow$ bad	bad	bad	bad

Table 1: The assessment module specifies which image to assign to the potential donor of an observed interaction ('good $\rightarrow$ bad' means 'a good player helps a bad player', 'bad $\not\rightarrow$ good' means 'a bad players refuses to help a good player', etc).

Action modules

situation/strategy	SELF	CO	AND	OR	AllC	AllD
good $\stackrel{?}{\rightarrow}$ good	no	yes	no	yes	yes	no
good $\stackrel{?}{\rightarrow}$ bad	no	no	no	no	yes	no
bad $\stackrel{?}{\rightarrow}$ good	yes	yes	yes	yes	yes	no
bad $\stackrel{?}{\rightarrow}$ bad	yes	no	no	yes	yes	no

Table 2: The action module prescribes whether to help or not given the own image, and the image of the potential recipient ('bad $\stackrel{?}{\rightarrow}$ good' prescribes whether a bad player should help when faced with a good co-player, etc.)

The leading eight

situation/strategy	1	2	3	4	5	6	7	8
good $\rightarrow$ good	good	good	good	good	good	good	good	good
good $\rightarrow$ bad	good	bad	good	good	bad	bad	good	bad
bad $\rightarrow$ good	good	good	good	good	good	good	good	good
bad $\rightarrow$ bad	good	good	good	bad	good	bad	bad	bad
good $\not\rightarrow$ good	bad	bad	bad	bad	bad	bad	bad	bad
good $\not\rightarrow$ bad	good	good	good	good	good	good	good	good
bad $\not\rightarrow$ good	bad	bad	bad	bad	bad	bad	bad	bad
bad $\not\rightarrow$ bad	bad	bad	good	good	good	good	bad	bad
good $\stackrel{?}{\rightarrow}$ good	yes	yes	yes	yes	yes	yes	yes	yes
good $\stackrel{?}{\rightarrow}$ bad	no	no	no	no	no	no	no	no
bad $\stackrel{?}{\rightarrow}$ good	yes	yes	yes	yes	yes	yes	yes	yes
bad $\stackrel{?}{\rightarrow}$ bad	yes	yes	no	no	no	no	no	no

Table 3: The leading eight ESS strategies, specified by as assessment module (first 8 rules) and an action module (last 4 rules), obtain highest payoffs among all ESS pairs, and keep their evolutionary stability even for benefit-to-cost ratios close to one. Strategy 1 corresponds to OR-STANDING (Contrite Tit For Tat, or CTFT, in Panchanathan and Boyd, 2003), strategy 8 corresponds to CO-JUDGING. Note that neither CO-STANDING, the RDISC strategy from Panchanathan and Boyd, 2003, nor any SCORING strategy occurs in the list.

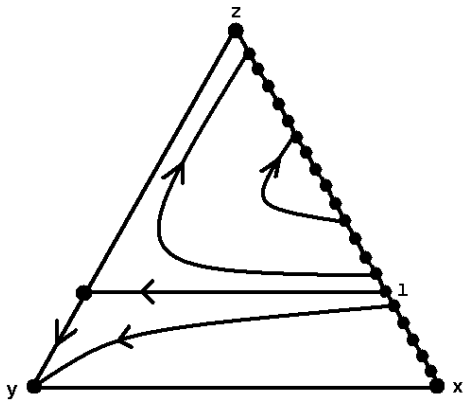

Figure 1: Replicator dynamics when the number of rounds is constant. In the absence of errors, any mixture of AllC and CO-SCORING is a fixed point.

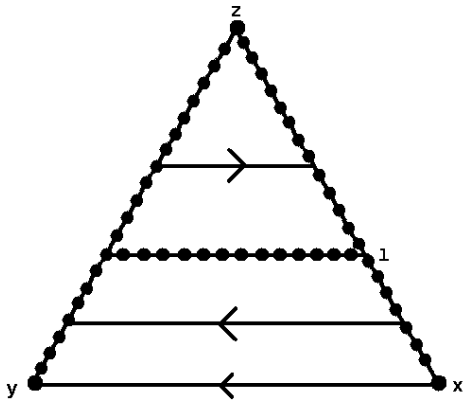

Figure 2: Replicator dynamics in the limiting case of infinitely many rounds, and no errors. In addition to the fixed point edges, we obtain a line of fixed points in the interior of the simplex.

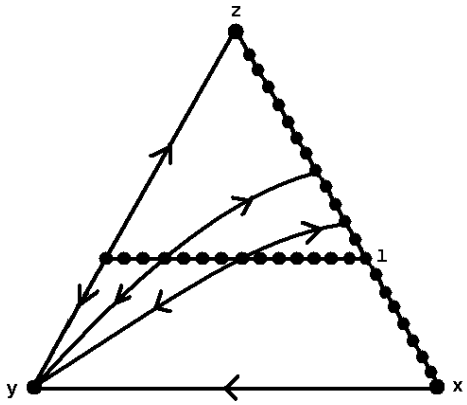

Figure 3: Replicator dynamics when the number of rounds follows a geometric distribution and no errors occur.

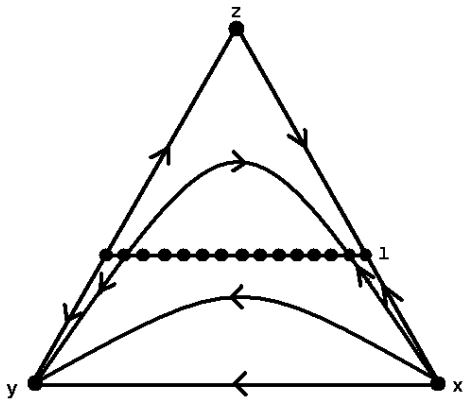

Figure 4: Replicator dynamics if individuals make errors in implementation, and the number of rounds follows a geometric distribution. In the long run, AllD is established. A similar dynamics holds for the asynchronous entry case, for all probability distributions of rounds.

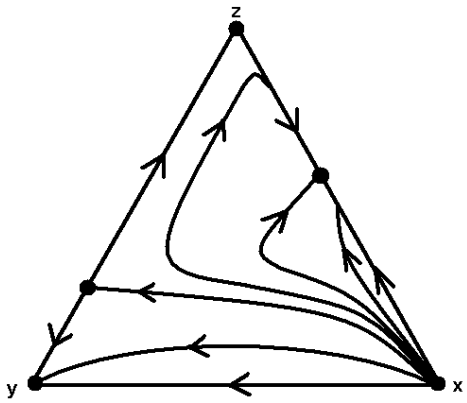

Figure 5: Replicator dynamics when individuals make errors in implementation and the number of rounds is constant. A bistable outcome results. The same holds if the rounds are Poisson distributed, or in the asynchronous entry case when each player's social network grows with time.

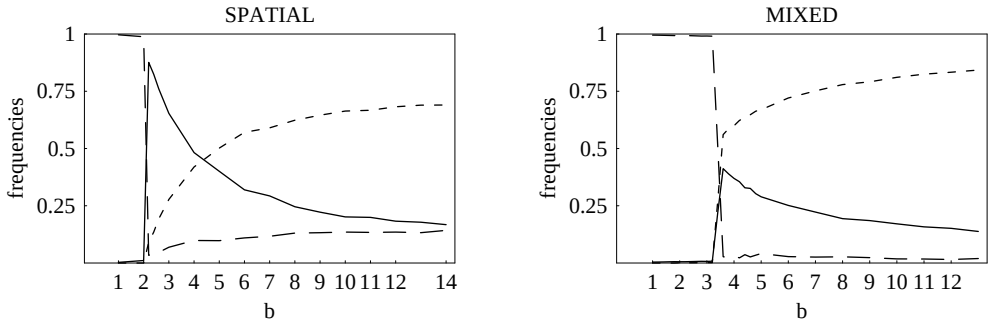

Figure 6: In both graphs, long-term frequencies for a population initialized randomly with strategies AllC (short dashes), AllD (long dashes), and CO-SCORING (solid line) are shown, a mutation rate 0.001, an error rate of 0.05, and a temptation rate for discriminators to defect of 0.05 are included. Five rounds per generation are played. Spatial indirect reciprocity, where individuals are confined to the sites of a square lattice and interact only with their neighbors, promotes cooperation for smaller benefits (with $c = 1$) than in the well-mixed case. In the spatial case, however, defectors can survive more easily within clusters of AllC players, and subsist at frequencies of around 15%.