International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

Interim Report **IR-10-054**

The competition of assessment rules for indirect reciprocity

Satoshi Uchida (satoshi.uchida@univie.ac.at)
Karl Sigmund (ksigmund@iiasa.ac.at)

Approved by

Ulf Dieckmann
Program Leader, EEP

June 2011

The competition of assessment rules for indirect reciprocity

Satoshi Uchida[a,b,1] and Karl Sigmund[b,c,2]

[a] *Research Division, RINRI Institute,*

Chiyoda-ku misaki-cho 3-1-10, 101-0061 Tokyo, Japan

[b] *Faculty of Mathematics, University of Vienna,*

Nordbergstrasse 15, 1090 Vienna, Austria

[c] *International Institute for Applied Systems Analysis,*

Schlossplatz 1, 2361 Laxenburg, Austria

[1] Corresponding author. E-mail: satoshi.uchida@univie.ac.at

[2] E-mail:karl.sigmund@univie.ac.at

Abstract

Indirect reciprocity is one of the basic mechanisms to sustain mutual cooperation. Beneficial acts are returned, not by the recipient, but by third parties. Indirect reciprocity is based on reputation and status: it pays to provide help because this makes one more likely to receive help in turn. The mechanism depends on knowing the past behavior of other players, and assessing that behavior. There are many different systems of assessing other individuals, which can be interpreted as rudimentary moral systems (i.e. views on what is 'good' or 'bad'). In this paper, we describe the competition of some of the leading assessment rules called *SUGDEN* and *KANDORI* by analytic methods. We show that the sterner rule *KANDORI* has a slight advantage in the sense that *KANDORI*-players have more chance to earn higher payoff than *SUGDEN*-players in the presence of unconditional altruists. On the other hand, we see that the unconditional altruists are eliminated in the long run and that stable polymorphisms of *KANDORI* and *SUGDEN* can subsist, but that a moral consensus is realized even in those polymorphic states: all players' images are the same in each observer's eyes.

Keywords: Replicator Dynamics; Prisoner's Dilemma Game; Leading Eight; Second-order Assessment

1 Introduction

In indirect reciprocity, helpful acts are returned, not by the recipient, but by third parties (Alexander, 1987; Sugden, 1986; Trivers, 1971). If Alice helps Betty, then Alice is helped in turn, not by Betty, as in direct reciprocation, but by some Conny or Claire. Indirect reciprocity has been amply documented in human populations (Camerer and Fehr, 2006; Bolton et al., 2005; Seinen and Schram, 2001; Wedekind and Milinski, 2000; Wedekind and Braithwaite, 2002). In order not to be subverted by exploiters (for instance by defectors who never help others), the help must be channelled away from them, and directed preferentially towards the helpers. For this, two requirements are needed: (a) information about previous interactions, even those in which one has not been involved; and (b) an assessment of these interactions. Thus indirect reciprocity is based on constant monitoring of the other members of the population, and on judging whether they deserve to be helped or not, or in other words whether they have a good image or not (Leimar and Hammerstein, 2001; Nowak and Sigmund, 1998a, b; Panchanathan and Boyd, 2003). This can be viewed as an elementary form of moral judgment. Individuals assess other players' actions as good or bad even if they are not directly affected by them.

The most elementary way for C to assess A simply reflects whether A gave help to B or not. In the first case, A is viewed as good and in the second case as bad. But this leads to an interesting inconsistency: if C refuses to help A, then C is perceived by third parties as bad irrespective of whether the potential recipient A is good or bad. As a result, C is less likely to be helped. Acting on a moral judgment can thus be costly. This suggests that a better assessment rule should also take into account whether a refusal to help was justified or not (see Camerer and Fehr (2006), Leimar and Hammerstein (2001), Nowak and Sigmund (1998a) and Sugden (1986)). However, there exist several ways for doing this, and it is not

clear which assessment should evolve in the long term. To give an example: should the act of helping a bad individual be considered as good or as bad?

There are many possible moral systems. How do they compare? In a first approach, we may consider three different classes of assessment rules (Brandt and Sigmund, 2004). A first-order assessment rule only takes into account whether A helps B or not. A second-order assessment rule takes also into account the image of the recipient B. A third-order assessment rule takes moreover into account the image of the donor A. A strategy in the indirect reciprocity interaction consists of an *assessment rule* together with an *action rule* telling the player which decision to take, as a donor, depending on the image of the recipient and the own image (Brandt and Sigmund, 2004; Ohtsuki and Iwasa, 2004).

Ohtsuki and Iwasa have shown that among the 4096 resulting strategies, only 8 lead to a stable regime of mutual cooperation, if adopted by all members of the population. These are said to be the *leading eight* (Ohtsuki and Iwasa, 2004, 2006). Two of these strategies are based on second-order assessment, none on first-order assessment. In this context, 'stable' means that the corresponding population cannot be invaded by other action rules. However, this does not settle the issue whether other *assessment* rules can invade. In the set-up considered by Ohtsuki and Iwasa, the image of an individual is the same in the eyes of all members of the population. Clearly, this does not allow to compare different assessment rules.

If one wants to analyze the evolution of even the simplest system of morals, one has to consider the competition of several assessment rules in the population. This is what we propose to do in the present paper: we consider the two second-order assessment rules belonging to the 'leading eight', as well as the first-order assessment rule which only registers whether help is given or not. We find that this first-order assessment rule is eliminated (not surprisingly), and that among the second-order assessment rules, the sterner rule has a slight

advantage; if it is as frequent as the other rule (or more frequent), its payoff is at least as high. Stable polymorphisms of the two second-order assessment rules exist, but interestingly, the population always converges to a state where both assessments coincide: evolution leads to moral consensus.

In the following sections, we describe the model, derive the results, and discuss both outcomes and methods.

2 The model

We consider a large, well-mixed population. From time to time, two individuals are randomly matched in a one-shot interaction, a so-called donation game. A coin toss decides who is the potential donor and the potential recipient (we suppress the 'potential' from now on). The donor can, at a personal cost c, provide a benefit b to the recipient, with $b > c$. We shall actually assume (as is usually done) that both players are simultaneously donor and recipient: this does not affect the outcome of the model. The interaction is an example of a Prisoner's Dilemma game. We assume that each individual experiences an infinity of such interactions, always with different partners.

Furthermore, we assume that the players can observe each other. (If not, cooperation cannot evolve.) Each player A has an assessment rule by which to judge others according to their behavior as donor in their previous interaction. Player A's judgment is binary: it assigns either γ (for 'good') or β (for 'bad') to all other players. The action rules of all players are the same: they give help if they assess the recipient as γ, and they refuse help otherwise. (In technical terms, all action rules are of Co-type, see Brandt and Sigmund (2004)). The assessment rules, however, can be different. The corresponding strategies, therefore, depend

entirely on the assessment rule. We shall consider only the following assessment rules (see also Table1 and Fig.1): (1) *AllC* (view everyone as γ); (2) *AllD* (view everyone as β); (3) *SUGDEN*, also known as Simple Standing (view everyone as γ except those who, in their previous round, refused help to a γ-recipient); (4) *KANDORI* (view exactly those as γ who, in their previous round, gave help to a γ-recipient or refused help to a β-recipient); and finally (5) *SCORING* (the first-order assessment that views exactly those as γ who, in their previous round, gave help, no matter to whom). We see that the second-order assessment rules *SUGDEN* and *KANDORI* differ in their view of those who give help to a β-player: *KANDORI*, the sterner assessment, condemns this.

We shall moreover assume that players sometimes commit an error. With a certain probability ϵ, they fail to implement an intended help. Following Leimar and Hammerstein (2001), Ohtsuki and Iwasa (2004, 2006) and Panchanathan and Boyd (2003), we assume that an intended refusal is always carried out (see also Fishman et al. (2001), Fishman (2003) and Lotem et al. (1999)). Finally, we assume that from time to time, a randomly chosen individual switches strategy by adopting the strategy i of a model chosen with a probability proportional to that model's fitness $F_i = (1-s)F + sP_i$. Here, F is a baseline fitness (the same for all), P_i is the average payoff for an individual of type i, and $s \in]0,1]$ is a parameter measuring the importance of the game for overall success. The resulting dynamics is given (up to a change in velocity) by the replicator equation $\dot{x}_i = x_i(P_i - \bar{P})$, where x_i is the frequency of strategy i in the population and $\bar{P} = \sum_k x_k P_k$ is the average payoff in the population (see Hofbauer and Sigmund (1998) p. 87).

Ohtsuki and Iwasa showed that *SUGDEN* and *KANDORI* belong to the leading eight: if everyone in the population shares the corresponding assessment rule, it is best to follow the corresponding action module of giving help exactly to the γ-recipients (Ohtsuki and Iwasa,

2004). No other action module (such as, for instance: 'always refuse help', or 'help only if, in addition, the own image is β') can invade. But this does not settle the issue of the assessment rule itself. Is there a selective advantage in choosing one rule rather than another? For this, we have to assume that any given player A can have different images in the eyes of different observers. All individuals form their own opinion on the interactions they observe. This approach is not used by Ohtsuki and Iwasa, who assume that the image is public (decided, for instance, by one observer who acts as a referee). Private images are used in the individual-based simulations in Brandt and Sigmund (2004) and Pacheco et al. (2006). Here, we present an analytical approach to deal with the competition of several assessment rules.

Let us first consider the competition of *SUGDEN* and *KANDORI* only. We allow for *AllC* and *AllD* players in the population, but not for *SCORING*. Thus we consider only the strategies (1) to (4). We denote γ as 'good' respectively 'nice' in the eyes of an *SUGDEN*- resp. *KANDORI*-player, and β as 'bad' resp. 'nasty'. We denote the proportions of players of type i who are evaluated as (a) both bad and nasty by r^i_{00}, (b) bad and nice by r^i_{01}, (c) good and nasty by r^i_{10} and (d) good and nice by $r^i_{11} (= 1 - r^i_{00} - r^i_{01} - r^i_{10})$.

These quantities determine the payoffs. In fact, if we define

$$r_i = r^i_{10} + r^i_{11} \text{ (prop. of good players of type } i), \tag{1}$$

$$s_i = r^i_{01} + r^i_{11} \text{ (prop. of nice players of type } i), \tag{2}$$

the payoffs P_i are expressed by

$$P_1 = -\bar{\epsilon}c + \bar{\epsilon}(x_1 + r_1x_3 + s_1x_4)b, \tag{3}$$

$$P_2 = \bar{\epsilon}(x_1 + r_2x_3 + s_2x_4)b, \tag{4}$$

$$P_3 = -\bar{\epsilon}\textstyle\sum_i x_i r_i c + \bar{\epsilon}(x_1 + r_3x_3 + s_3x_4)b, \tag{5}$$

$$P_4 = -\bar{\epsilon}\textstyle\sum_i x_i s_i c + \bar{\epsilon}(x_1 + r_4x_3 + s_4x_4)b, \tag{6}$$

where $\bar{\epsilon} := 1 - \epsilon$ is the probability that an intended help is actually given. For example, $\bar{\epsilon}\sum_i x_i s_i$ in Eq.(6) is the probability that a player of type 4 gives a help to another player, and thus incurs cost c. The term $\bar{\epsilon}(x_1 + r_4 x_3 + s_4 x_4)$ is the probability that a player of type 4 is helped by a randomly chosen donor, and thus provided with a benefit b.

In general, r^i_{mn} changes in time, but such changes must be much faster than the changes of the frequencies of strategies x_i. Therefore we ignore the transient of the evolution of r^i_{mn} and only consider the equilibrium after many rounds of games with x_i being fixed. At the equilibrium, r^i_{mn} satisfies the following equality:

$$\begin{aligned} r^i_{mn} &= (\text{the probability that } i \text{ actually helps} \\ &\quad \text{and the action is evaluated as } m \text{ resp. } n) \\ &+ (\text{the probability that } i \text{ defects erroneously} \\ &\quad \text{and the action is evaluated as } m \text{ resp. } n) \\ &+ (\text{the probability that } i \text{ defects intentionally} \\ &\quad \text{and the action is evaluated as } m \text{ resp. } n). \end{aligned} \tag{7}$$

These probabilities can be expressed by the proportions of (m, n)-players in the whole population, namely P, Q, R and S:

$$P = \sum_i x_i r^i_{00} \text{ (prop. of bad-nasty players)}, \tag{8}$$

$$Q = \sum_i x_i r^i_{01} \text{ (prop. of bad-nice players)}, \tag{9}$$

$$R = \sum_i x_i r^i_{10} \text{ (prop. of good-nasty players)}, \tag{10}$$

$$S = \sum_i x_i r^i_{11} \text{ (prop. of good-nice players)}. \tag{11}$$

We note that $R + S$ is the proportion of good players, $Q + S$ that of nice players, $Q + P$ that of bad players and $R + P$ the proportion of nasty players.

This yields the following relations between r^i_{mn} and P, Q, R (and $S = 1 - P - Q - R$):

$$\left[\begin{array}{cc} r^1_{11} = \bar{\epsilon}(Q+S) + \epsilon P & r^1_{10} = \bar{\epsilon}(R+P) + \epsilon Q \\ r^1_{01} = \epsilon R & r^1_{00} = \epsilon S \\ r^2_{11} = P & r^2_{10} = Q \\ r^2_{01} = R & r^2_{00} = S \\ r^3_{11} = \bar{\epsilon}S + P & r^3_{10} = \bar{\epsilon}R + Q \\ r^3_{01} = \epsilon R & r^3_{00} = \epsilon S \\ r^4_{11} = \bar{\epsilon}(Q+S) + P & r^4_{10} = \epsilon Q \\ r^4_{01} = R & r^4_{00} = \epsilon S \end{array}\right]. \tag{12}$$

How these equations are obtained is presented in the supplementary material.

If we substitute these relations into Eqs.(8), (9) and (10), we obtain a linear system for the unknowns P, Q, R:

$$c_{11}P + c_{12}Q + c_{13}R = d_1, \tag{13}$$

$$c_{21}P + c_{22}Q + c_{23}R = d_2, \tag{14}$$

$$c_{31}P + c_{32}Q + c_{33}R = d_3, \tag{15}$$

with $d_1 = \epsilon(x_1 + x_3 + x_4) + x_2, d_2 = d_3 = 0$ and

$$\left[\begin{array}{ccc} c_{11} = d_1 + 1 & c_{12} = d_1 & c_{13} = d_1 \\ c_{21} = 0 & c_{22} = 1 & c_{23} = -d_1 - \bar{\epsilon}x_4 \\ c_{31} = -\bar{\epsilon}x_1 & c_{32} = -d_1 - \bar{\epsilon}x_3 & c_{33} = 1 - \bar{\epsilon}(x_1 + x_3) \end{array}\right]. \tag{16}$$

By solving, we obtain the payoff values as functions of the frequencies (x_1, x_2, x_3, x_4) of the strategies.

3 Results

The determinant of the matrix (c_{ij}) is zero only on the edge between *AllD* and *SUGDEN* (i.e., if $x_1 = x_4 = 0$). The dynamics on that edge is bistable, with the unstable fixed point determined by $x_3 = c/\bar{\epsilon}b$, see also Ohtsuki and Iwasa (2007).

From Eqs.(13)-(16), it follows that, in the presence of the unconditional altruists (i.e., if $x_1 > 0$),

$$Q < R, \quad Q < P < S \tag{17}$$

and

$$r_2 < s_1 < s_3 < r_4 < r_1 = r_3 < s_4 \tag{18}$$

are always valid (see Appendix A for the detailed calculation). The proportion of nice *AllD*-players s_2 is somewhere between r_2 and $r_1 = r_3$. From these inequalities, we see that the condition $x_3 \leq x_4$ implies $P_3 < P_4$ (see Appendix A). Thus if *KANDORI* and *SUGDEN* are equally frequent, the former wins whenever unconditional altruists are present.

The advantage of *KANDORI* can be understood by the following argument: in order that a cooperative player A obtains a nice image from *KANDORI*, A's recipient must also be nice, whereas A always obtains a good evaluation from *SUGDEN*. Therefore a cooperative player who is nice is always good, whereas the inverse is not necessarily true; thus it is more difficult to obtain nice images than good ones. The inequality $Q < R$ implies that *KANDORI*-players incur less cost than *SUGDEN*-players on average if *AllC* is present. At the same time, the inequality $s_3 < r_4$ implies that the probability that *KANDORI*-players evaluate *SUGDEN*-players as nice is less than that *SUGDEN*-players evaluate *KANDORI*-players as good. Therefore, *KANDORI*-players are more likely to obtain a cooperative offer

from *SUGDEN*-players than vice-versa. Moreover, we find from $s_3 < r_4 < r_3 < s_4$ that *KANDORI*-players are more likely to give help to *KANDORI*-players and less likely to give help to *SUGDEN*-players. If the two types of discriminators are equally frequent, *KANDORI* obtains a higher payoff than *SUGDEN* and its relative frequency increases.

The second result obtained from the inequalities is that P_3 is greater than P_1 if $x_1 > 0$. Indeed, using $\bar{P}_i := P_i/\bar{\epsilon}$, we see that

$$\bar{P}_3 - \bar{P}_1 = \big((r_3 - r_1)x_3 + (s_3 - s_1)x_4\big)b + (1 - \sum_i x_i r_i)c > 0. \tag{19}$$

Hence x_1/x_3 converges to 0, so that all orbits in the interior of the state simplex converge to the face $x_1 = 0$, i.e., *AllC* is eliminated in the long run.

If $x_1 = 0$ then $c_{31} = 0$, hence Eqs. (14) and (15) imply $Q = R = 0$. This means that in the absence of *AllC*-players, *SUGDEN* and *KANDORI* always agree in their assessment and hence do not differ in their behavior. In this case,

$$P = \frac{1 - \bar{\epsilon}(x_3 + x_4)}{2 - \bar{\epsilon}(x_3 + x_4)}, \tag{20}$$

$$S = \frac{1}{2 - \bar{\epsilon}(x_3 + x_4)}. \tag{21}$$

On the face $x_1 = 0$, $P_3 = P_4$ and hence x_3/x_4 is constant (Fig.2-(c)). Each solution remains on a half ray through $x_2 = 1$; it is easy to see that the segment with $x_3 + x_4 = c/\bar{\epsilon}b$ consists of fixed points. Depending on which side of that segment they start, orbits converge either to $x_2 = 1$ or $x_2 = 0$. Hence the evolution, in the absence of *AllC*, leads either to *AllD* or else to a stable mixture of *KANDORI* and *SUGDEN*. These states are the only Nash equilibria.

To describe the competition of *SCORING* with one of the second-order assessment rules (for instance, *KANDORI*), we can use equations up to Eq.(11), replacing the other assessment rule with *SCORING*. Fig. 3-(a) shows the vector field of the replicator dynamics if *SCORING*

and *KANDORI* are present in the population. The equations used in this simulation are given in the supplementary material. A homogeneous population of *SCORING* is unstable, as *AllC* and *KANDORI* can invade. The edge *AllD*-*SCORING* consists of fixed points. But the stable ones are only those with $x_3 \leq c/\bar{\epsilon}b$. At these stable fixed points, all players defect and their payoff is zero. These fixed points cannot be invaded by *KANDORI* or *AllC* and hence are Nash equilibria (Fig.3-(c)).

The segment given by $x_3 = c/\bar{\epsilon}b$ and $x_4 = 0$ also consists of fixed points. However, these are unstable since these states can be invaded by *KANDORI*, see Fig.3-(b).

The same holds for the competition of *SCORING* with *SUGDEN*.

If all 5 types of strategies are present, *AllC* is again eliminated in the long run. If $x_1 = 0$, the replicator dynamics leads either to a mixture of *AllD* and *SCORING* (with the frequency of defectors at least $1 - c/\bar{\epsilon}b$), or to a mixture of *KANDORI* and *SUGDEN* (see Fig.4). The two types of players agree in their assessment (in the former case, all are evaluated as β, thus all defect, in the latter case, the assessment of *SUGDEN* and *KANDORI* are equivalent as mentioned above), and moral consensus is achieved.

4 Discussion

There are several other papers highlighting the merits of *KANDORI*. We mention, in particular, Chalub et al. (2006) and Pacheco et al. (2006), which apply numerical simulations to a group selection scenario. We also refer to Brandt and Sigmund (2004), where two third-order rules called *STANDING* and *JUDGING* are compared (which are closely related to *SUGDEN* and *KANDORI* respectively). It is shown that the sterner rule *JUDGING* has advantages compared to the milder rule *STANDING*, based on individual-based simulations in group

structured populations.

Our paper relies entirely on analytic methods and uses an individual selection scenario. We extend the investigations of Ohtsuki and Iwasa in one direction, by allowing different players to judge their co-players by different assessment rules. This is an important issue, as it allows to investigate the competition of different 'moral systems'. In particular, this approach no longer makes use of the assumption that one player acts as a referee whose public assessment is adopted by all other players (Ohtsuki and Iwasa, 2004; Pacheco et al., 2006; Panchanathan and Boyd, 2003). It is common-day experience that different people can assess one and the same action in different ways. While gossip can greatly help to spread information, it need not lead to consensual assessment (Sommerfeld et al., 2007).

Just as in Ohtsuki and Iwasa (2007), we have not considered third-order assessment rules. The complications of an analytical treatment of such rules seem too arduous at present. We have made another departure from the model by Ohtsuki and Iwasa, which concerns a technical point. In their model, generations are separate: all players are born at the same time and their rounds are synchronized. We assume that the strategies spread by imitation, rather than by inheritance. Instead of producing offspring, players switch their strategy. This does not affect the mathematical model, but makes the interpretation somewhat more natural. Moreover, we assume asynchronous updating: players update their strategy one at a time, and their rounds are not synchronized. This modeling assumption, however, has hardly any effect on the outcome. One interesting issue is to investigate the effects of a random number of rounds and analyze whether the probability of a further round smaller than one qualitatively changes the results obtained here (see Brandt and Sigmund (2006)).

Both our model and that of Ohtsuki and Iwasa suffer from two limitations which are more serious. One concerns the assumption that players are assessed according to their last

interaction only: their actions in previous rounds are not taken into account. In reality, reputations are often based on a longer data-base. Moreover, they are not 'binary': the moral world is not just black or white. The second limitation is due to the assumption that players have perfect information. Again, this is unrealistic. Usually, players often have only limited information, and sometimes none at all (Mohtashemi and Mui, 2003; Nowak and Sigmund, 1998b). If they do not know the antecedents of their co-player, they need a 'default' rule. Since this rule describes whether the individual is trustful or suspicious, this clearly introduces an important distinction. We were not able yet to overcome the complications raised by this possibility. Moreover, an assessment can be erroneous. Again, this is a possibility which we encounter every day. Misunderstandings and mis-perceptions have possibly a more devastating effect than mis-implementations (see Takahashi and Mashima (2006) on the role of errors in perception). Exchange of information and opinions via gossip and other forms of communication is important, but not faultless (Sommerfeld et al., 2007; Takahashi and Mashima, 2006).

If we admit that players can mis-perceive whether help has been given or refused, or that they can be confused about the reputation of the recipient, we introduce a source of errors which is extremely complicated to analyze. It seems not unlikely that these errors affect the more complex second-order assessment rules, such as *SUGDEN* and *KANDORI*, to a greater degree than the more simple-minded first-order *SCORING*. In fact, there is experimental evidence to support the view that second-order assessment can overtax human cognitive abilities (Milinski et al., 2001). We know no empirical work permitting to conclude whether *SUGDEN* or *KANDORI* is more frequent.

Indirect reciprocity based on reputation systems has a long history (Ellison, 1994; Kandori, 1992; Okuno-Fujiwara and Postlewaite, 1995; Pollock and Dugatkin, 1992; Rosenthal,

1979; Yamagishi et al., 1999). As mentioned in Nowak and Sigmund (2005), there are two main motivations to pursue its investigation. One concerns the evolution of human communities: how does cooperation work in villages and small-scale societies? (See Dufwenberg et al. (2001), Masuda and Ohtsuki (2007), Milinski et al. (2002a, b), Panchanathan and Boyd (2004), Roberts (2008), Semmann et al. (2004) and Suzuki and Akiyama (2007a, b)). Recently, evidence for indirect reciprocation in other species has also been uncovered (Bshary and Grutter, 2006). The other motivation concerns the rapid growth of anonymous interactions on a global scale, made possible by the spread of communication networks: how can cheating be avoided in on-line trading? (see Bolton et al. (2004) and Keser (2002)) In both cases, simple, robust methods for assessing others are essential.

The present investigation can clearly be no more than a first step in analyzing the competition of different rudimentary forms of moral systems. Within the context of second-order assessment rules belonging to the leading eight, the sterner rule has an advantage (see also Chalub et al. (2006) and Pacheco et al. (2006)), but evolution converges to a state where both rules can coexist and always agree.

Acknowledgements

We wish to thank Ulrich Berger for his useful comments. Part of this work is funded by EUROCORES TECT I-104 G15.

Appendix A Inequalities

We mention some inequalities that help us understand the system better. Let us assume $x_1 > 0, x_3 > 0$ and $x_4 > 0$. From Eq.(13) together with $P + Q + R = 1 - S$, we immediately

find $P = d_1 S < S$. From Eq.(14) we have $Q = -c_{23}R < R$. This relation between R and Q together with Eq.(15) yields a relation $c_{31}P + c_{32}Q - c_{33}/c_{23}Q = 0$ that is simplified to $Q = c_{31}c_{23}/(c_{33} - c_{32}c_{23})P$. Here $c_{31}c_{23} = \bar{\epsilon}\big(1 - \bar{\epsilon}(x_1 + x_3)\big)x_1 \geq 0$ and $c_{33} - c_{32}c_{23} = \bar{\epsilon}\big(1 - \bar{\epsilon}(x_1 + x_3)\big)(x_1 + x_4) = c_{31}c_{23} + \bar{\epsilon}\big(1 - \bar{\epsilon}(x_1 + x_3)\big)x_4 > c_{31}c_{23}$. Hence $Q < P$.

From these inequalities, Eq.(18) is derived. In fact from Eq.(12), $s_4 - r_1 = \epsilon(R - Q) > 0$, $r_1 - r_3 = 0$ and $r_1 - r_4 = \bar{\epsilon}R > 0$. The difference between r_4 and s_3 is calculated as $r_4 - s_3 = Q - \epsilon R$. Substituting $Q = -c_{23}R$ from Eq.(14), we have $r_4 - s_3 = (-c_{23} - \epsilon)R$, where $-c_{23} - \epsilon = \epsilon(x_1 + x_3) + (x_2 + x_4) - \epsilon(x_1 + x_2 + x_3 + x_4) = \bar{\epsilon}(x_2 + x_4) > 0$. Further, $s_3 - s_1 = \bar{\epsilon}(P - Q) > 0$ and $s_1 - r_2 = \bar{\epsilon}(S - P) + \epsilon(R - Q) > 0$.

We can also find an explicit expression for the region $P_3 = P_4$, using

$$\bar{P}_4 - \bar{P}_3 = \hat{C}c + \hat{B}b, \tag{A.1}$$

where $\hat{C}$ is the cost term $\sum_i (r_i - s_i)x_i$ and $\hat{B}$ the benefit term $(r_4 - r_3)x_3 + (s_4 - s_3)x_4$. By the definitions of r_i and s_i, we have

$$\begin{aligned} \hat{C} &= R - Q > 0, \\ \hat{B} &= -\bar{\epsilon}Rx_3 + \bar{\epsilon}(Q + R)x_4. \end{aligned} \tag{A.2}$$

Taking the relation between R and Q into account, we obtain

$$\begin{aligned} \hat{C} &= \bar{\epsilon}(1 + c_{23})R = \bar{\epsilon}(x_1 + x_3)R, \\ \hat{B} &= -\bar{\epsilon}x_3 R + \bar{\epsilon}x_4(1 - c_{23})R \\ &= \bar{\epsilon}\big(-x_3 + x_4(2 - \bar{\epsilon}x_1 - \bar{\epsilon}x_3)\big)R. \end{aligned} \tag{A.3}$$

Therefore if $-x_3 + x_4(2 - \bar{\epsilon}x_1 - \bar{\epsilon}x_3) \leq 0$ or equivalently $x_4 \leq x_3/(2 - \bar{\epsilon}x_1 - \bar{\epsilon}x_3)$, $P_4 > P_3$ holds regardless of the values of b and c. This region completely includes the plane $x_3 = x_4$ in the state space.

P_4 is larger than P_3 if

$$c(x_1 + x_3) - x_3 b + (2 - \bar{\epsilon} x_3 - \bar{\epsilon} x_1) x_4 b > 0, \tag{A.4}$$

which depends on the benefit-cost ratio c/b. In particular, for $x_2 = 0$, i.e., if $AllD$ is absent, this region is given by $x_3 < c/b(1 - x_4) + (2 - \bar{\epsilon} + \bar{\epsilon} x_4) x_4$.

References

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

Bolton, G., Katok, E., Ockenfels, A., 2004. How effective are on-line reputation mechanisms? An experimental investigation. Management Science 50, 1587-1602.

Bolton, G., Katok, E., Ockenfels, A., 2005. Cooperation among strangers with limited information about reputation. Journal of Public Economics 89, 1457-1468.

Brandt, H., Sigmund, K., 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. J. Theor. Biol. 231, 475-486., doi:10.1016/j.jtbi.2004.06.032.

Brandt, H., Sigmund, K., 2006. The good, the bad and the discriminator – errors in direct and indirect reciprocity. J. Theor. Biol. 239, 183-194., doi:10.1016/j.jtbi.2005.08.045.

Bshary, R., Grutter, A.S., 2006. Image scoring and cooperation in a cleaner fish mutualism. Nature 441, 975-978.

Camerer, C., Fehr, E., 2006. When does "economic man" dominate social behaviour? Science 311, 47-52.

Chalub, F., Santos, F.C., Pacheco, J.M., 2006 The evolution of norms. J. Theor. Biol 241, 233-240., doi:10.1016/j.jtbi.2005.11.028.

Dufwenberg, M., Gneezy, U., Gueth, W., van Damme, E., 2001. Direct vs indirect reciprocation – an experiment. Homo Oeconomicus 18, 19-30.

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

Fishman, M.A., 2003. Indirect reciprocity among imperfect individuals. J. Theor. Biol. 225, 285-292., doi:10.1016/S0022-5193(03)00246-7.

Fishman, M.A., Lotem, A., Stone, L., 2001. Heterogeneity stabilises reciprocal altruism interaction. J. Theor. Biol. 209, 87-95., doi:10.1006/jtbi.2000.2248.

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge UP, Cambridge, p 87.

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

Keser, C., 2002. Experimental games for the design of reputation management systems. IBM Systems Journal 43, 498-503.

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

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

Masuda, N., Ohtsuki, H., 2007. Tag-based indirect reciprocity by incomplete social information. Proc. Roy. Soc. Lond. B 274, 689-695.

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

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

Mohtashemi, M., Mui, L., 2003. Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism. J. Theor. Biol. 223, 523-531., doi:10.1016/S0022-5193(03)00143-7.

Nowak, M.A., Sigmund, K., 1998a. Evolution of indirect reciprocity by image scoring. Nature 282, 462-466.

Nowak, M. A., Sigmund, K., 1998b. The dynamics of indirect reciprocity. J. Theor. Biol. 194, 561-574., doi:10.1006/jtbi.1998.0775.

Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1292-1298.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness? – Reputation dynamics in indirect reciprocity. J. Theor. Biol. 231, 107-120., doi:10.1016/j.jtbi.2004.06.005.

Ohtsuki, H., Iwasa, Y., 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theor. Biol. 239, 435-444., doi:10.1016/j.jtbi.2005.08.008.

Ohtsuki, H., Iwasa, Y., 2007. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. J. Theor. Biol. 244, 518-531., doi:10.1016/j.jtbi.2006.08.018.

Okuno-Fujiwara, M., Postlewaite, A., 1995. Social norms and random matching games. Games and Economic Behavior 9, 79-109.

Panchanathan, K., Boyd, R., 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J. Theor. Biol. 224, 115-126., doi:10.1016/S0022-5193(03)00154-1.

Panchanathan, K., Boyd, R., 2004. Indirect reciprocity can stabilise cooperation without the second-order free-rider problem. Nature 432, 499-502.

Pacheco, J., Santos, F., Chalub, F., 2006. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLOS Computational Biology 2, 178.

Pollock, G.B., Dugatkin, L.A., 1992. Reciprocity and the evolution of reputation. J. Theor. Biol. 159, 25-37., doi:10.1016/S0022-5193(05)80765-9.

Roberts, G., 2008. Evolution of direct and indirect reciprocity. Proc. Roy. Soc. Lond. B 275, 173-179.

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

Semmann, D., Krambeck, H.J., Milinski, M., 2004. Strategic investment in reputation. Journal of Behavioral Ecology and Sociobiology 56, 248-252.

Seinen, I., Schram, A., 2001. Social status and group norms: indirect reciprocity in a repeated helping experiment. European Economic Review 50, 581-602.

Sommerfeld, R., Krambeck, H.J., Semmann, D., Milinski, M., 2007. Gossip as an alternative for direct observation in games of indirect reciprocity. Proceedings of the National Academy of Sciences 104, 17435-17440.

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

Suzuki, S., Akiyama, E., 2007a. Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. J. Theor. Biol. 245, 539-552., doi:10.1016/j.jtbi.2006.11.002.

Suzuki, S., Akiyama, E., 2007b. Three-person game facilitates indirect reciprocity under image scoring. J. Theor. Biol. 249, 93-100., doi:10.1016/j.jtbi.2007.07.017.

Takahashi, N., Mashima, R., 2006. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. J. Theor. Biol. 243, 418-436., doi:10.1016/j.jtbi.2006.05.014.

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

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

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

Yamagishi, T., Jin, N., Kiyonari, T., 1999. Bounded generalized reciprocity: ingroup boasting and ingroup favouritism. Advances in Group Processes 16, 161-197.

Table 1: Definitions of the five assessment rules studied in this paper

Situation	*AllC*	*AllD*	*SUGDEN*	*KANDORI*	*SCORING*
Give help to γ-recipient	γ	β	γ	γ	γ
Give help to β-recipient	γ	β	γ	β	γ
Refuse help to γ-recipient	γ	β	β	β	β
Refuse help to β-recipient	γ	β	γ	γ	β

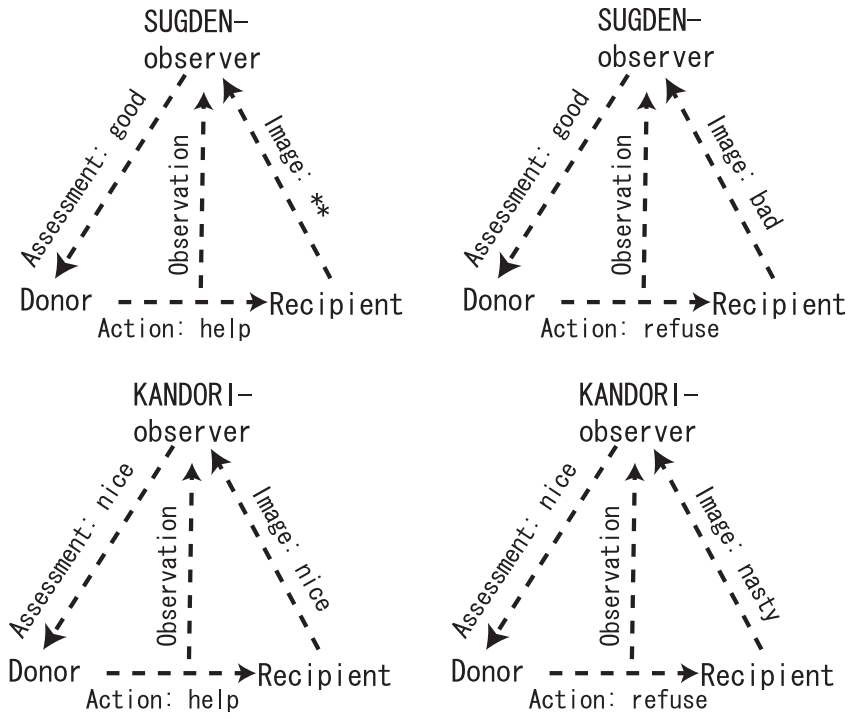

Figure 1: The situations that are positively assessed by the second-order assessment rules in the leading eight (top: *SUGDEN*-observer, bottom: *KANDORI*-observer). All the other situations will be negatively assessed. The positive (or negative) assessment corresponds to γ (or β) in the main text. To distinguish the two rules, the positive (or negative) assessment is denoted by good (or bad) in the eyes of the *SUGDEN*-observer and by nice (or nasty) in the eyes of the *KANDORI*-observer. The symbol "**" in the top left figure is an abbreviation for "good or bad".

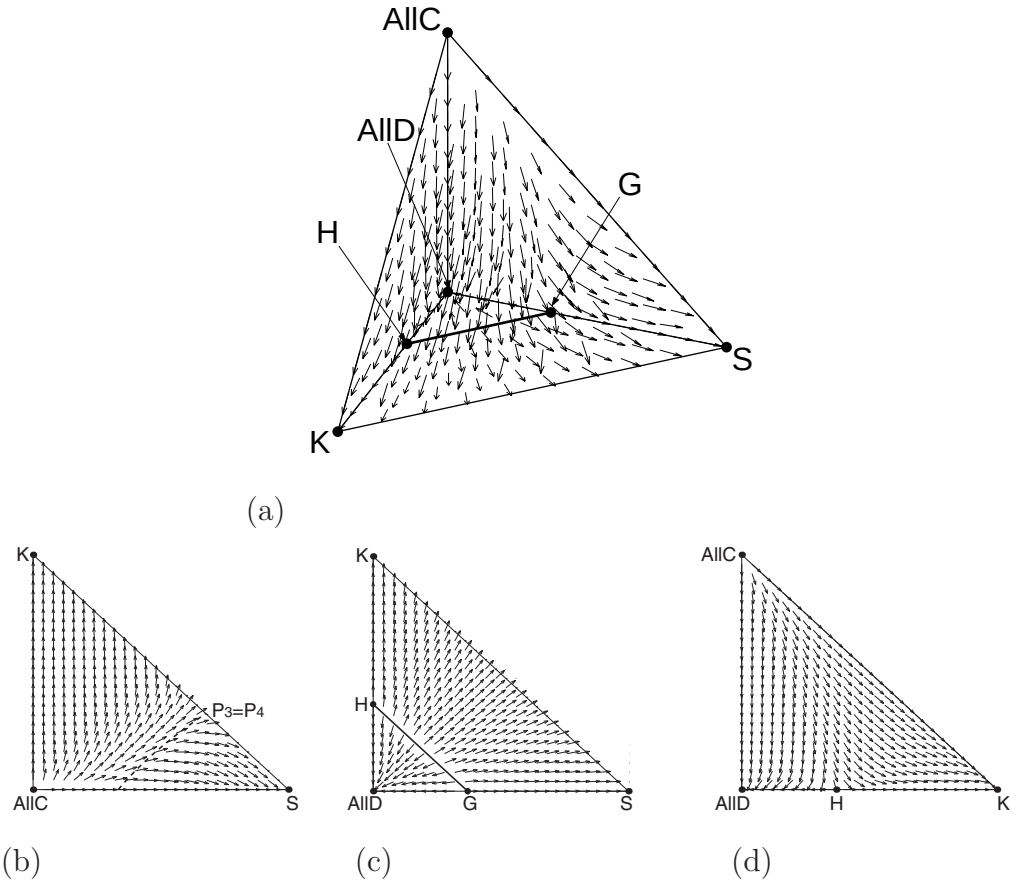

Figure 2: The vector field generated by replicator dynamics in the whole state space (a) and on each face (b-d). The vector field on the face $x_4 = 0$ is similar to (d) (see also (Ohtsuki and Iwasa, 2007)). The abbreviation S corresponds to *SUGDEN* and K to *KANDORI*. To produce the figure, we normalized the vector at each point (except for the case where the vector vanishes) so that the direction is easily recognized. Parameters: $c = 1, b = 3$ and $\epsilon = 0.1$.

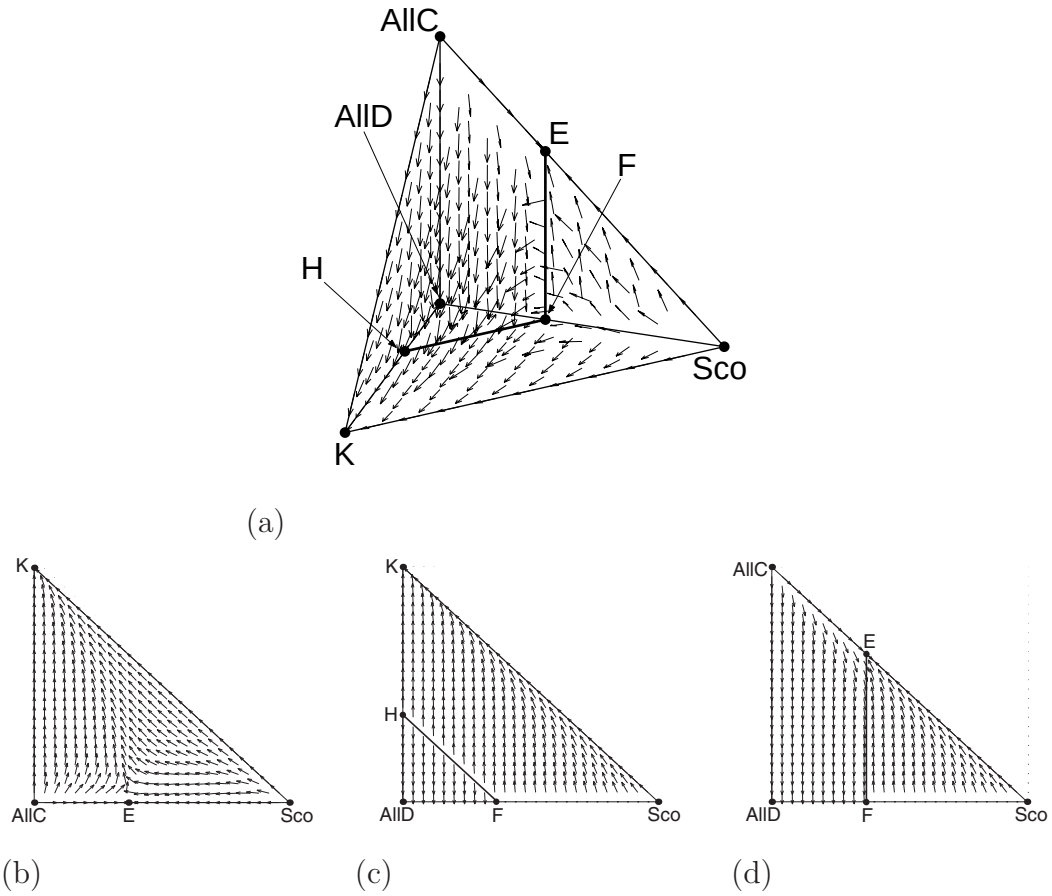

Figure 3: Same as in Fig.2. Here the abbreviation *Sco* corresponds to *SCORING* and *K* to *KANDORI*. The segments *EF* and *HF* consist of fixed points.

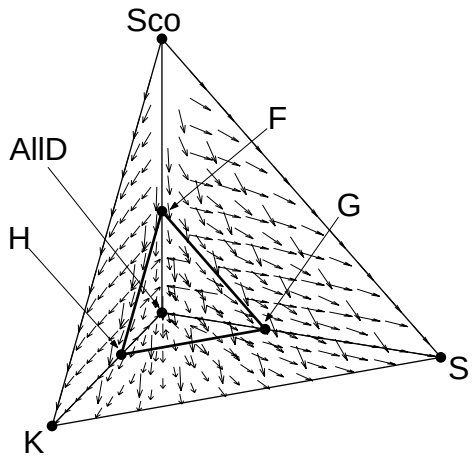

Figure 4: The system with *AllD* and 3 types of discriminators. *AllC* is not involved. The surface *FGH* given by $x_2 = 1 - c/\bar{\epsilon}b$ consists of unstable fixed points, where x_2 is the frequency of *AllD*. Depending on which side of that surface they start from, orbits converge either to the segment *S*-*K* or to the segment *AllD*-*F*.