The competition of assessment rules for indirect reciprocity

Satoshi Uchida [*] [†] **and Karl Sigmund** [†] [‡]

[*]Research Division, RINRI Institute, 101-0061 Tokyo, Japan,[†] Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria, and [‡]International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria

Submitted to Proceedings of the National Academy of Sciences of the United States of America

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 others, 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 by analytic methods, and show that the sterner rule has a slight advantage. Stable polymorphisms can subsist, but lead to moral consensus: all players' images are the same in each observer's eyes.

replicator dynamics | Prisoner's Dilemma game | indirect reciprocity | leading eight | second-order assessment

In indirect reciprocity, helpful acts are returned, not by the recipient, but by third parties [1, 2, 3]. 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 [4, 5, 6, 7, 8]. 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 [9, 10, 11, 12]. 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 [2], [8], [9] and [11]). 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 [13]. 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 [13, 14].

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* [14, 15]. 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. Stable polymorphisms 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.

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 dif-

Reserved for Publication Footnotes

ferent partners. (This can be replaced by the assumption that the probability w for another round satisfies $w > c/b$.)

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 [13]). 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: (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 [11], [12], [14] and [15], we assume that an intended refusal is always carried out (see also [16], [17] and [18]). 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 [19] 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 [14]. 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 [13] and [20]. 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

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

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

the payoffs P_i are expressed by

$$P_1 = -\bar{\epsilon}c + \bar{\epsilon}(x_1 + r_1 x_3 + s_1 x_4)b, \quad [3]$$

$$P_2 = \bar{\epsilon}(x_1 + r_2 x_3 + s_2 x_4)b, \quad [4]$$

$$P_3 = -\bar{\epsilon}\textstyle\sum_i x_i r_i c + \bar{\epsilon}(x_1 + r_3 x_3 + s_3 x_4)b, \quad [5]$$

$$P_4 = -\bar{\epsilon}\textstyle\sum_i x_i s_i c + \bar{\epsilon}(x_1 + r_4 x_3 + s_4 x_4)b, \quad [6]$$

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} is determined by the assessment dynamics $dr^i_{mn}/d\tau = -r^i_{mn} + F^i_{mn}$ describing the relaxation to an equilibrium $r^i_{mn} = F^i_{mn}$. Here τ represents the time measured by the time scale of the assessment dynamics and F^i_{mn} is given by

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

Due to the linear dependence of F^i_{mn} on $\{r^i_{mn}\}$ (see below), if the time scale of assessment dynamics is much faster than that of replicator dynamics (i.e., if x_i is treated as a constant in the assessment dynamics), the assessment dynamics converges to a fixed point. For this reason, we assume that the assessment dynamics is always at an equilibrium: $r^i_{mn} = F^i_{mn}$.

The above probabilities are expressed by the proportions of (m,n)-players in the whole population, namely P, Q, R and S:

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

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

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

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

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} (or F^i_{mn}) and P, Q, R (and $S = 1 - P - Q - R$):

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

How these equations are obtained is presented in the supporting information.

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

$$c_{11}P + c_{12}Q + c_{13}R = d_1, \quad [\mathbf{13}]$$
$$c_{21}P + c_{22}Q + c_{23}R = d_2, \quad [\mathbf{14}]$$
$$c_{31}P + c_{32}Q + c_{33}R = d_3, \quad [\mathbf{15}]$$

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

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

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

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 [21].

If the unconditional altruists are absent, i.e. 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,

$$P = \frac{1 - \bar{\epsilon}(x_3 + x_4)}{2 - \bar{\epsilon}(x_3 + x_4)}, \quad [\mathbf{17}]$$
$$S = \frac{1}{2 - \bar{\epsilon}(x_3 + x_4)}. \quad [\mathbf{18}]$$

On the face $x_1 = 0$, $P_3 = P_4$ and hence x_3/x_4 is constant (Fig.1-(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.

In the appendix, it is shown that in the interior of the state space,

$$Q < R, \quad Q < P < S \quad [\mathbf{19}]$$

and

$$r_2 < s_1 < s_3 < r_4 < r_1 = r_3 < s_4 \quad [\mathbf{20}]$$

are always valid. The proportion of nice *AllD*-players s_2 is somewhere between r_2 and $r_1 = r_3$. This implies that P_3 is greater than P_1 if $x_1 > 0$. Indeed, using $\bar{P}_i := P_i/\bar{\epsilon}$, we see that

$$\bar{P}_3 - \bar{P}_1 = \big((r_3 - r_1)x_3 + (s_3 - s_1)x_4\big)b + (1 - \sum_i x_i r_i)c > 0. \quad [\mathbf{21}]$$

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.

Moreover, condition $x_3 \leq x_4$ implies $P_3 < P_4$ (see Appendix). 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 $Q < R$ means 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 a 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 proportion increases.

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. 2-(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 supporting information. The fixed point *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.2-(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.2-(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.3). 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.

Discussion

There are several other papers highlighting the merits of *KANDORI*. We mention, in particular, [20] and [22], which apply numerical simulations to a group selection scenario.

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 [12, 14, 20]. 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 [23].

Just as in [21], we have not considered third-order assessment rules. We have made another departure from the model by Ohtsuki and Iwasa, which concerns a technical point. In that 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. The assumption that the number of rounds is infinite is mostly made for notational convenience. Sufficiently large probabilities of a further round lead to the same outcomes. We only have to replace b by wb (see [24]).

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 [10, 25]. 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. 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 [26] on the role of errors in perception). Exchange of information and opinions via gossip and other forms of communication is important, but not faultless [26, 23].

If we admit that players can mis-perceive whether an act of 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 [27]. 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 [28, 29, 30, 31, 32, 33]. As mentioned in [34], 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 [35], [36], [37], [38],[39], [40], [41], [42] and [43]). Recently, evidence for indirect reciprocation in other species has also been uncovered [44]. 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 [45] and [46]) 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 [20] and [22]), but evolution converges to a state where both rules can coexist and always agree.

Appendix

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.[**20**] 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

$$\bar{P}_4 - \bar{P}_3 = \hat{C}c + \hat{B}b, \qquad [\mathbf{22}]$$

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

$$\begin{aligned} \hat{C} &= R - Q > 0, \\ \hat{B} &= -\bar{\epsilon}Rx_3 + \bar{\epsilon}(Q + R)x_4. \end{aligned} \qquad [\mathbf{23}]$$

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

$$\begin{aligned} \hat{C} &= \bar{\epsilon}(1 + c_{23})R = \bar{\epsilon}(x_1 + x_3)R, \\ \hat{B} &= -\bar{\epsilon}x_3R + \bar{\epsilon}x_4(1 - c_{23})R \\ &= \bar{\epsilon}\big(-x_3 + x_4(2 - \bar{\epsilon}x_1 - \bar{\epsilon}x_3)\big)R. \end{aligned} \qquad [\mathbf{24}]$$

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

$$c(x_1 + x_3) - x_3b + (2 - \bar{\epsilon}x_3 - \bar{\epsilon}x_1)x_4b > 0, \qquad [\mathbf{25}]$$

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 $x_3 < c/b(1 - x_4) + (2 - \bar{\epsilon} + \bar{\epsilon}x_4)x_4$.

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

1. Trivers R (1971) The evolution of reciprocal altruism. Quart Rev Biol 46: 35-57.
2. Sugden R (1986) The Economics of Rights, Cooperation and Welfare (Basil Blackwell, Oxford).
3. Alexander RD (1987) The Biology of Moral Systems (Aldine de Gruyter, New York).
4. Wedekind C, Milinski M (2000) Cooperation through image scoring in humans. Science 288: 850-852.
5. Seinen I, Schram A (2001) Social status and group norms: indirect reciprocity in a Repeated helping experiment. European Economic Review 50: 581-602.
6. Wedekind C, Braithwaite VA (2002) The long-term benefits of human generosity in indirect reciprocity. Curr Biol. 12: 1012-1015.
7. Bolton G, Katok E, Ockenfels A (2005) Cooperation among strangers with limited information about reputation. Journal of Public Economics 89: 1457-1468.
8. Camerer C, Fehr E (2006) When does "economic man" dominate social behaviour? Science 311: 47–52.
9. Nowak MA, Sigmund K (1998a) Evolution of indirect reciprocity by image scoring. Nature 282: 462-466.

10. Nowak M A, Sigmund K (1998b) The dynamics of indirect reciprocity. J Theor Biol 194: 561-574.
11. Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocation. Proc R Soc Lond B 268: 745-753.
12. 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
13. Brandt H, Sigmund K (2004) The logic of reprobation: action and assessment rules in indirect reciprocity. J Theor Biol 231: 475-486.
14. Ohtsuki H, Iwasa Y (2004) How should we define goodness? – Reputation dynamics in indirect reciprocity. J Theor Biol 231: 107-120.
15. Ohtsuki H, wasa Y (2006) The leading eight: social norms that can maintain cooperation by indirect reciprocity. J Theor Biol 239: 435-444.
16. Lotem A, Fishman M A, Stone L (1999) Evolution of cooperation between individuals. Nature 400: 226-227.
17. Fishman MA, Lotem A, Stone L (2001) Heterogeneity stabilises reciprocal altruism interaction. J Theor Biol 209: 87-95.
18. Fishman MA (2003) Indirect reciprocity among imperfect individuals. J Theor Biol 225: 285-292.
19. Hofbauer J, Sigmund K (1998) Evolutionary Games and Population Dynamics (Cambridge UP), p 87.
20. Pacheco J, Santos F, Chalub F (2006) Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLOS Computational Biology 2: e178.
21. 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.
22. Chalub F, Santos FC, Pacheco JM (2006) The evolution of norms. J Theor Biol 241: 233-240.
23. Sommerfeld R, Krambeck HJ, 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.
24. Brandt H, Sigmund K (2006) The good, the bad and the discriminator – errors in direct and indirect reciprocity. J Theor Biol 239: 183-194.
25. 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.
26. Takahashi N, Mashima R (2006) The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. J Theor Biol 243: 418-436.
27. Milinski M, Semmann D, Bakker TCM, Krambeck H J (2001) Cooperation through indirect reciprocity: image scoring or standing strategy? Proc Roy Soc London B 268: 2495-2501.
28. Rosenthal R W (1979) Sequences of games with varying opponents. Econometrica 47: 1353-1366.
29. Kandori, M (1992) Social norms and community enforcement. Review of Economic Studies 59: 63-80.
30. Pollock G B, Dugatkin L A (1992) Reciprocity and the evolution of reputation. J Theor Biol 159: 25-37.
31. Ellison G (1994) Cooperation in the Prisoner's Dilemma with anonymous random matching. Review of Economic Studies 61: 567-588.
32. Okuno-Fujiwara M, Postlewaite A (1995) Social norms in matching games. Games and Economic Behavior 9: 79-109.
33. Yamagishi T, Jin N, Kiyonari T (1999) Bounded generalized reciprocity: ingroup boasting and ingroup favouritism. Advances in Group Processes 16: 161-197.
34. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437: 1292-1298.
35. Dufwenberg M, Gneezy, U, Gueth, W, van Damme E (2001) Direct vs indirect reciprocation – an experiment. Homo Oeconomicus 18: 19-30.
36. Masuda N, Ohtsuki H (2007) Tag-based indirect reciprocity by incomplete social information. Proc Roy Soc London B 274: 689-695.
37. Milinski M, Semmann D, Krambeck HJ (2002a) Donors in charity gain in both indirect reciprocity and political reputation. Proc Roy Soc London B 269: 881-883.
38. Milinski M, Semmann, D, Krambeck HJ (2002b) Reputation helps solve the 'Tragedy of the Commons'. Nature 415: 424-426.
39. Panchanathan K, Boyd R (2004) Indirect reciprocity can stabilise cooperation without the second-order free-rider problem. Nature 432: 499-502.
40. Roberts G (2008) Evolution of direct and indirect reciprocity. Proc Roy Soc London B 275: 173-179.
41. Semmann D, Krambeck HJ, Milinski M (2004) Strategic investment in reputation. Journal of Behavioral Ecology and Sociobiology 56: 248-252.
42. 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.
43. Suzuki S, Akiyama E (2007b) Three-person game facilitates indirect reciprocity under image scoring. J Theor Biol 249: 93-100.
44. Bshary R, Grutter AS (2006) Image scoring causes cooperation in a cleaning mutualism. Nature 441: 975-978.
45. Bolton G, Katok E, Ockenfels A (2004) How effective are on-line reputation mechanisms? An experimental investigation. Management Science 50: 1587-1602.
46. Keser C (2002) Experimental games for the design of reputation management systems. IBM Systems Journal 43: 498-503.

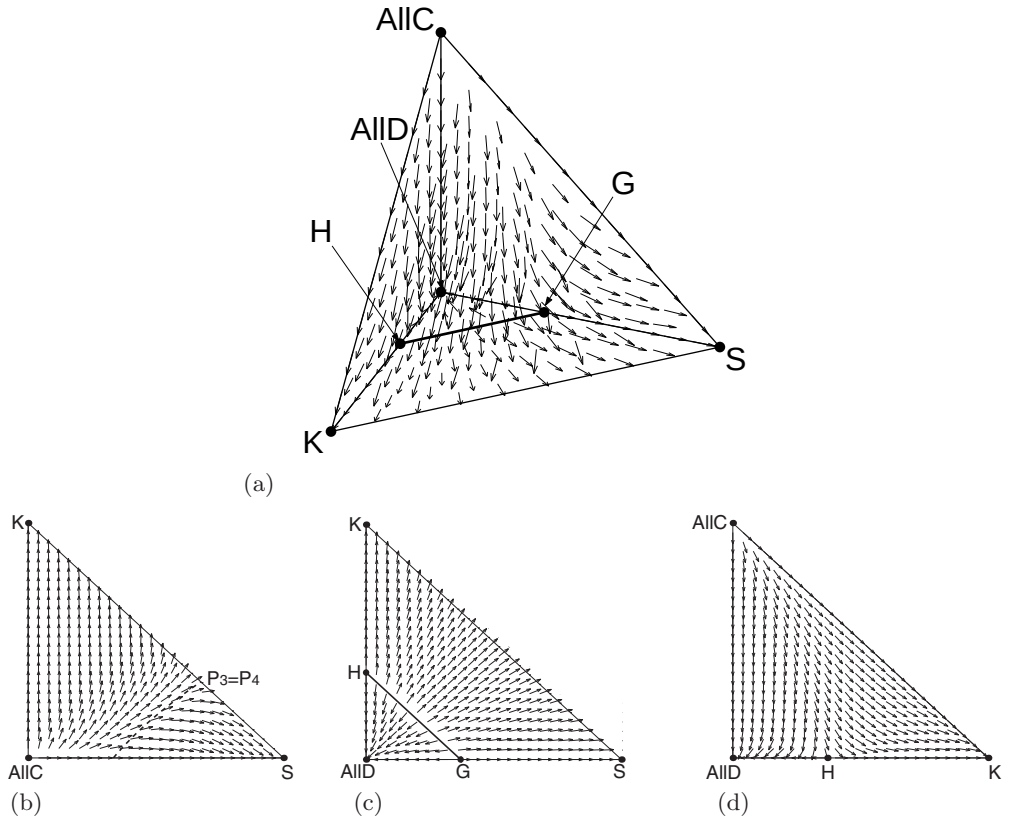

Fig. 1. The vector field generated by replicator dynamics in the whole state space (a) and on each face (b-d). The vector field on the face $x_4 = 0$ is similar to (d) (see also [21]). The abbreviation S corresponds to $SUGDEN$ and K to $KANDORI$. To produce the figure, we normalized the vector at each point (except for the case where the vector vanishes) so that the direction is easily recognized. Parameters: $c = 1, b = 3$ and $\epsilon = 0.1$.

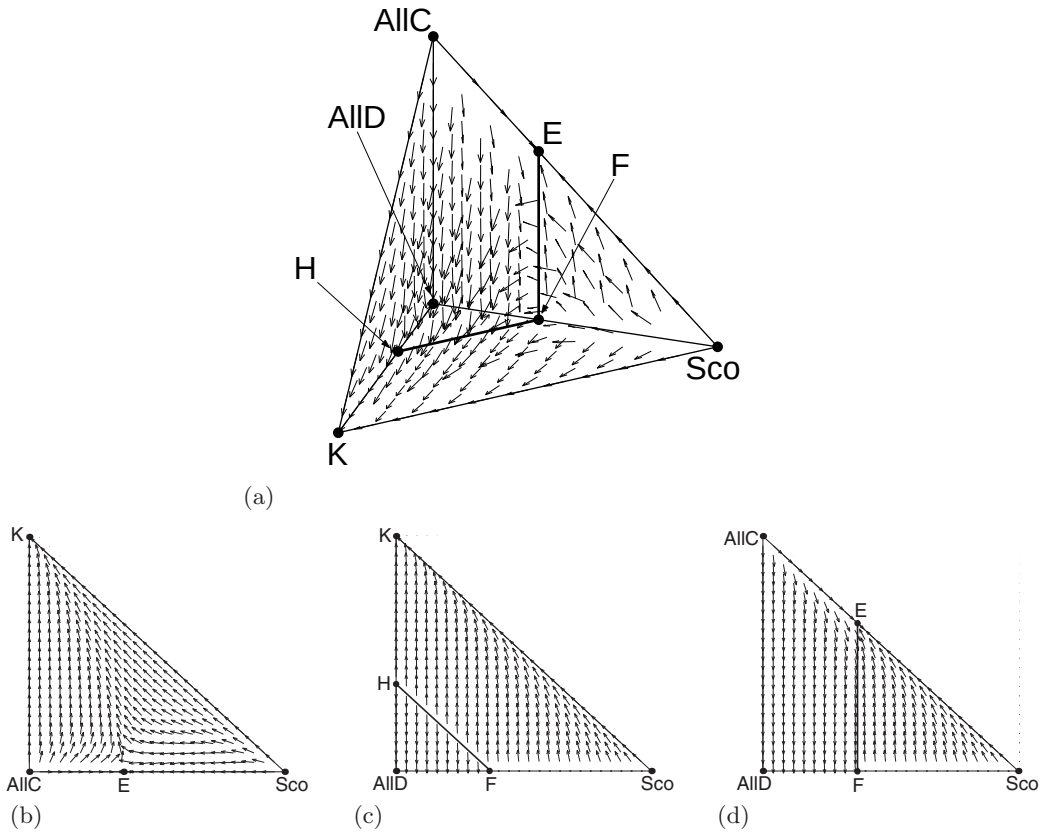

Fig. 2. Same as in Fig.1. Here the abbreviation Sco corresponds to $SCORING$ and K to $KANDORI$. The segments EF and HF consist of fixed points.

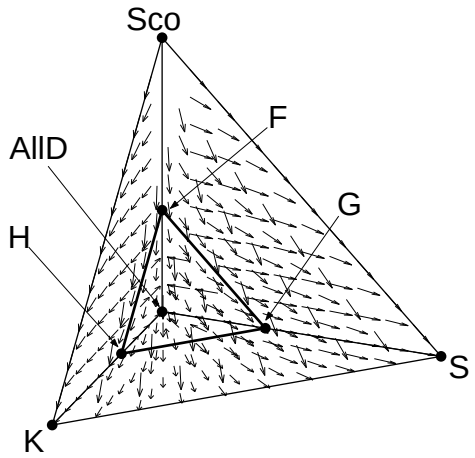

Fig. 3. The system with $AllD$ and 3 types of discriminators. $AllC$ is not involved. The surface FGH given by $x_2 = 1 - c/\bar{e}b$ consists of unstable fixed points, where x_2 is the frequency of $AllD$. Depending on which side of that surface they start from, orbits converge either to the segment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