Indirect reciprocity, image scoring, and moral hazard

Hannelore Brandt and Karl Sigmund*

Fakultät für Mathematik, University of Vienna, Nordbergstrasse 15, Vienna, A-1090 Austria; and IIASA, A-2361 Laxenburg, Austria

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved December 1, 2004 (received for review October 7, 2004)

Whether one-shot interactions can stably sustain mutual cooperation if they are based on a minimal form of reputation building has been the subject of considerable debate. We show by mathematical modeling that the answer is positive if we assume an individual's social network evolves in time. In this case, a stable mixture of discriminating and undiscriminating altruists can be proof against invasion by defectors. This sheds light on current discussions about the merits of different types of moral assessment, an issue where theoretical arguments and experimental findings seem at odds.Unexpectedly, our approach also relates to the proverbial observation that people tend to become more tightfisted with age.

cooperation | reputation | evolutionary dynamics

"Give and it shall be given unto you." But by whom? Luke's account (New Testament, Luke 6:38) was not specific on that point. A helpful action, or a gift, can be returned by the recipient, in which case one speaks of direct reciprocation. But it can also be returned by a third party. Alexander (1) called this "indirect reciprocity," emphasizing its reliance on status and reputation.

In a simple model, Nowak and Sigmund (2) attach a binary score ("Good" or "Bad") to each individual in the population. From time to time, two individuals meet randomly, one as donor, the other as recipient. At some cost c to one's own payoff, the donor can help the recipient, i.e., increase the recipient's payoff by a benefit $b>c$. In that case, the donor's score will be Good in the eyes of all observers, whereas the score of a donor refusing to confer the benefit will be Bad. A discriminating strategy of helping only those with a Good score would channel benefits toward those who help and discourage defectors.

The question is whether such a strategy can evolve in the population, assuming, as usual in evolutionary game dynamics, that strategies yielding a total payoff above average increase in frequency. The issue has attracted considerable attention for two major reasons. One lies in the potential of indirect reciprocity for explaining the emergence, among humans, of cooperation among nonrelatives. Alexander (1) viewed this as the biological basis of morality; others (3, 4) saw in it a major motivation for language, gossip being a way of spreading reputations. The recent advent of e-commerce provides the other reason why understanding the assessment of reputations matters: the prevalence of anonymous one-shot interactions in global markets raises the issues of trust building and moral hazard (5–7).

Although economic experiments have strongly bolstered the concept of indirect reciprocity (8–12), the radically simplified model of Nowak and Sigmund (2) has raised the skepticism of theoreticians (4, 13, 14). A discriminator who refuses to help recipients with a Bad score receives a Bad score and risks getting no help in the next round. In this sense, punishing defectors by withholding help is costly. Can such a trait evolve? Would it not be advantageous to distinguish justifiable defections (against a Bad recipient) from nonjustifiable defections (against a Good recipient) and attach a Bad score only to the latter? This would constitute a noncostly form of punishment and would greatly alleviate the discriminator's task. But such a distinction requires considerable cognitive capacities. Not only the recipient's past but also that of the recipient's recipients, etc., must be taken into account. If information spreads through rumor, rather than direct observation, the task may be alleviated, but the likelihood of misperception and deception grows. Conceivably, noncostly punishment cannot be realized, and many experiments show, anyway, that humans do not shrink from using costly punishment (15).

Materials and Methods

To return to theory, Ohtsuki and Iwasa (14), as well as Brandt and Sigmund (16), analyzed all conceivable strategies based on a binary score (some 4,096, at first count). Indeed, each such strategy can be viewed as a combination of two modules. The action module prescribes to a donor whether to give, depending on the recipient′s score and one's own (there are four possible combinations of the two scores and hence $2^4 = 16$ action modules). The assessment module prescribes how to assess the players' scores as a result of their action as donors in the previous round. Because there are two possible scores for a donor, two for a recipient, and two possible actions (to give or not), an assessment module has to state, for each of the eight combinations, how to assess the resulting action (for instance, is it Good if a donor with a Good score refuses help to a recipient with a Bad score, etc.). Thus, there are $2^8 = 256$ assessment modules and hence 16×256 strategies. Ohtsuki and Iwasa (14) found that eight of them are evolutionarily stable and lead to cooperation even if the benefit b is only slightly larger than cost c. All these "leading eight" strategies differentiate between justifiable and nonjustifiable defection. Nevertheless, the less-sophisticated discrimination mechanism suggested in ref. 2 can promote cooperation if it leads to a stable mixture of discriminators and undiscriminating altruists. After all, a population need not be homogenous, although this is required for evolutionary stability. But Panchanathan and Boyd (4) showed that, in the presence of errors (or other causes for unintended defections, for instance, lack of resources), such a mixture can be invaded by defectors. This blow was softened by Fishman (17), who found that if the game extends over a constant number of rounds, the mixture of discriminating and undiscriminating altruists can repel defectors. But what is more likely, a constant number of rounds per lifetime, as assumed in ref. 17, or a constant probability for a further round, as assumed in ref. 4?

In fact, both assumptions appear unrealistic. Whereas in an experimental game all players may start at the same time and play their rounds synchronously, it seems plausible to assume that under natural conditions, players enter the population one by one, at random times, and interact asynchronously. Under the assumption of stable age distribution, the analysis of this model becomes even simpler and boosts the conclusion of ref. 4.

Indeed, let us denote by q the probability that a player knows the score of a randomly chosen coplayer (via either direct observation or gossip, through acquaintances), and that discriminators are trustful in the sense that, if they have no information to the contrary, they assume that their recipient's score is Good. In the next section, we will see by a simple calculation that whenever discriminating and undiscriminating altruists do

This paper was submitted directly (Track II) to the PNAS office.

*To whom correspondence should be addressed. E-mail: karl.sigmund@univie.ac.at.

www.pnas.org/cgi/doi/10.1073/pnas.0407370102

equally well, defectors do just as well. This means that, ultimately, they will take over. This is an extremely robust result, independent of the probability distribution of the number of rounds (which could also be constant or infinite) and holding even if different strategies have different error probabilities, if discriminators are suspicious rather than trustful, or if they adopt strategies that also take into account their own score (for instance, by helping whenever the recipient's score is Good or their own score is Bad).

However, there is a way out, which extends an approach due to Mohtashemi and Mui (18), who assumed in their model that whenever a donor provides help, the donor's set of acquaintances is added to the recipient's. We need not be so specific but only assume that a player's network of acquaintances grows, i.e., the probability q_n that a player in round n is informed about the recipient's image grows with n, i.e., $q_n > q_{n-1}$. To keep the model as simple as possible, we assume the number of rounds is proportional to age (later we will consider more general scenarios). It will be shown in the next section that, whenever the average level of information is sufficiently high, there exists a mixture of discriminating and undiscriminating altruists that is an attractor, so defectors cannot invade. Thus let w_n be the probability that a randomly chosen individual is in round n. The average level of information in the population is $q = \Sigma\, w_n q_n$, which by our assumption is larger than $s := \Sigma\, w_n q_{n-1}$. If s is sufficiently large (or, equivalently, if the cost-to-benefit ratio c/b is small enough), and if w_1 is sufficiently small (i.e., if a second round is likely to occur), then a cooperative equilibrium exists, and defectors will be repelled. We emphasize that it is the individual's probability q_n to know the coplayer's score, which increases with age. We do not assume that q, i.e., the overall probability that two randomly chosen individuals know each other's score, increases with time, i.e., that the social network in the group, and thus the average level of acquaintance, grow.

We can, incidentally, also use the opposite condition, $q_n < q_{n-1}$, if we correspondingly suppose the discriminators are distrustful and refuse to help in the absence of information. We do not claim this is a reason why people whose social circle shrinks (the very old, for instance) tend to become suspicious. But both mechanisms, intriguingly, imply people should become more tightfisted with age. In one case, trustful individuals know more and more people and are therefore less and less ready to give the benefit of the doubt. In the other case, suspicious individuals know fewer and fewer people and are willing to support only those of whose Good score they are certain. We note the average frequency of Good persons in the population does not change within an individual's lifetime.

In *Results and Discussion*, we shall return to the current debate about the relative merits of different moral assessment rules giving a binary score (Good or Bad) to coplayers according to how these coplayers act toward third parties. This problem, which seems essential for discussions on the biological evolution of moral norms, offers a wide scope for further experimental and theoretical work.

A Continuous-Entry Model for Indirect Reciprocity

Let us denote by x, y, and z the relative frequencies of indiscriminate altruists, defectors, and discriminators in the population. We assume a continuous model: from time to time, a birth or death occurs, changing the frequencies x, y, and z (with $x + y + z = 1$) according to a differential equation, for instance, the replicator equation (see ref. 19). One could also use some other evolutionary game dynamics describing the transmission of strategies (for instance, through inheritance or imitation) in the population. All we need to assume is that strategies that yield a payoff above the population average will increase in frequency.

Occasionally, a player will be selected to play one round of the indirect reciprocity game. We shall assume for convenience that such a player will actually play two games in one round, one as potential donor and the other as recipient (always with different, randomly chosen coplayers, so there is no scope for direct reciprocity). We could just as well assume that players play only one game per round and are, with equal probability, donors or recipients.

Let us denote by $1 - r$ the probability of an unintended defection, and by q the probability that a player knows the score of a randomly chosen coplayer (via either direct observation or gossip). Let g be the frequency of players with a Good score. Clearly, g depends on x, y, and z. If the population is sufficiently large, g can be taken to be stationary throughout one individual's lifetime, Finally, let us posit that discriminators are trustful in the sense that, if they have no information about the score of the recipient (for instance, in the case of a newborn), they assume their recipient's score is Good.

Thus an undiscriminating altruist will always try to give (but will fail with a probability $1 - r$, possibly through lack of resources). A defector will never give, and a discriminator will try to help, when the recipient's score is Good (or unknown) but will succeed only with probability r. We assume for simplicity there are no mistakes in implementing a defection. This assumption, however, is not necessary for the following.

The payoff in the nth round ($n > 1$) for an undiscriminating altruist is

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

for a defector it is

$$P_y(n) = brx + br(1 - q)z,$$

and for a discriminator it is

$$P_z(n) = -cr(1 - q + qg) + brx + br(1 - q)z + br^2qz(1 - q + qg).$$

The last term in this sum is obtained as follows: the discriminating recipient meets, with probability z, another discriminator, who, with probability q, knows the recipient's score. If that score is Good, the recipient receives payoff b with probability r (because $1 - r$ is the probability that the intended donation fails). The score is Good if the recipient, in the previous round, succeeded in an intended donation (probability r), either not knowing the coplayer's score (probability $1 - q$) or else knowing the coplayer's score (probability q), which was Good (probability g). It is easy to see that $g = (1 - rqz)^{-1}r[z(1 - q) + x]$.

A straightforward computation shows that

$$P_z(n) - P_y(n) = [P_x(n) - P_y(n)](1 - q + qg).$$

The same relation holds for the first round and hence also for total payoff values P_x, P_y, and P_z. These values are obtained as $P_x = \Sigma\, w_n P_x(n)$, etc., because in a stable population, the probability w_n to be in round n is proportional to an individual's probability of reaching round n during his or her lifetime.

The replicator dynamics on the unit simplex S_3 is given by $\dot{x} = x(P_x - \bar{P})$, etc., where $\bar{P} = xP_x + yP_y + zP_z$ is the average payoff in the population. This is an ordinary differential equation on the state space $S_3 = \{(x, y, z): x \geq 0, y \geq 0, z \geq 0, x + y + z = 1\}$, the unit simplex in 3D space. The usual normalization (setting $P_y = 0$ and multiplying the right-hand side by the positive expression $(1 - rqz)/r$) yields $P_x = (1 - rqz)(-c + brqz)$ and $P_z = (1 - q + rqx)(-c + brqz)$. The fixed points are the corners of S_3 (where the population consists of one type only). In addition, if $q > c/br$, all of the points on the segment with $z = c/brq$ are fixed (see Fig. 1).

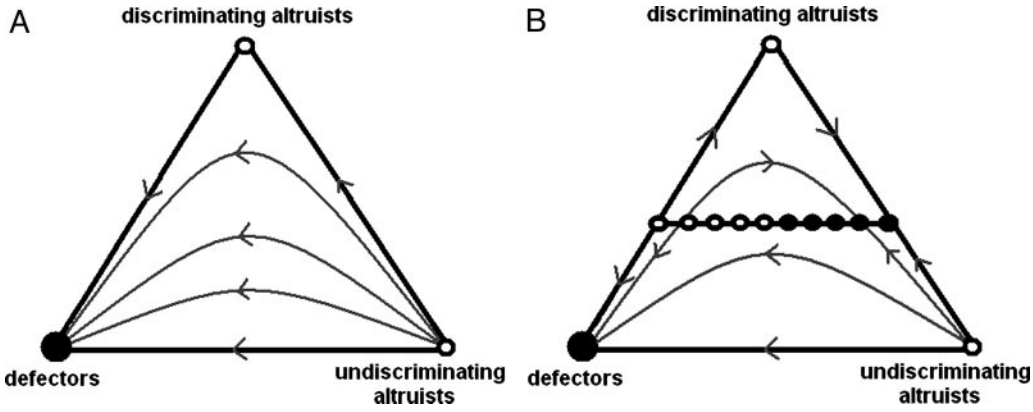

Fig. 1. The dynamics if the probability q of knowing the coplayer's score is constant. (*A*) If $q < c/br$, defectors always win. (*B*) If $q > c/br$, there exists a segment of fixed points in the interior of the state space (filled circles represent stable fixed points, open circles represent unstable fixed points).

If $q < c/br$, the undiscriminating altruists are dominated by both the discriminators and the defectors, whereas the discriminators are dominated by the defectors. All orbits in the interior of the simplex lead from $x = 1$ (undiscriminating altruists only) to $y = 1$ (defectors only). This means that if the probability q of knowing the coplayer's type is too small (i.e., if there is not much scope for reputation), cooperation cannot evolve, a well known result from ref. 2 (see Fig. 1*A*).

If $q > c/br$, then line $z = c/brq$ intersects the interior of the simplex and defines a segment of Nash equilibria. The orbits lie on the same curves as before, but the orientation has changed in the region with $z > c/brq$ (see Fig. 1*B*). This means that the mixture of discriminating and undiscriminating altruists given by $z = c/brq$ and $y = 0$ corresponds to a fixed point of the evolutionary dynamics. A cooperative population of two types of altruists can exist, if the average level of information within the population is sufficiently high.

We note that this equilibrium is stable. However, it is not asymptotically stable, because it is contained in a segment of fixed points. The dynamic behavior along the segment of Nash equilibria is interesting. One part of the segment is transversally stable, in the sense that small perturbations away from the segment are opposed by the dynamics. In the other part of the segment, small perturbations will be amplified by the dynamics. A small deviation to higher z values will lead first to an increase and then to a decrease of discriminators and eventually back to the stable part of the segment. A small deviation to lower z values will lead to the fixation of defectors.

In this sense, it must be admitted that, although the mixture of discriminating and undiscriminating altruists is stable, a sufficiently long sequence of random shocks can lead to the eventual fixation of defectors.

However, let us assume now that the probability of knowing a coplayer's score is not a constant but depends on experience and is denoted by q_n in round n. Then

$$P_z(n) = -cr(1 - q_n + q_n g) + brx + br(1 - q)z + br^2 qz(1 - q_{n-1} + q_{n-1} g),$$

where q now is the average of the q_n (i.e., because w_n is the probability of being in round n, we have $q = \Sigma\, w_n q_n$). If $q_n > q_{n-1}$ for all n, then $q > s := \Sigma\, w_n q_{n-1}$. We note that

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

and hence

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

For total payoffs P_x, P_y, and P_z, we obtain

$$P_x(z_{cr}) = P_z(z_{cr})$$

for $z_{cr} := c/brs$. We note that $z_{cr} > c/brq$ and assume in the following that $s > c/br$, i.e., $z_{cr} < 1$. This condition is of the same type as $q > c/br$, i.e., it requires a sufficient amount of information. It is somewhat stronger but leads, as we shall see, to the asymptotic stability of the mixture of discriminating and undiscriminating altruists.

The relation $P_x(n) - P_y(n) = -cr + br^2 qz$ implies that for $z = z_{cr}$, one has $P_x(n) - P_y(n) = cr(q - s)/s$ for $n > 1$ (and $= -cr$ for $n = 1$). It follows that 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 there exists a mixture consisting of discriminating and undiscriminating altruists only, $\mathbf{F}_{xz} = (1 - z_{cr}, 0, z_{cr})$, which cannot be invaded by defectors. The resulting replicator equation is bistable: one attractor consists of defectors only, the other of a mixture of discriminating and undiscriminating altruists (see Fig. 2).

We note that, instead of $q_n > q_{n-1}$ for all n, we need only $\Sigma\, w_n(q_n - q_{n-1}) > 0$, i.e., that, on average, individuals learn about reputations from one round to the next. If rounds are not equally spaced, i.e., more precisely, if the time intervals among rounds within an individual's lifetime do not obey the same probability distribution, then probability w_n that a randomly

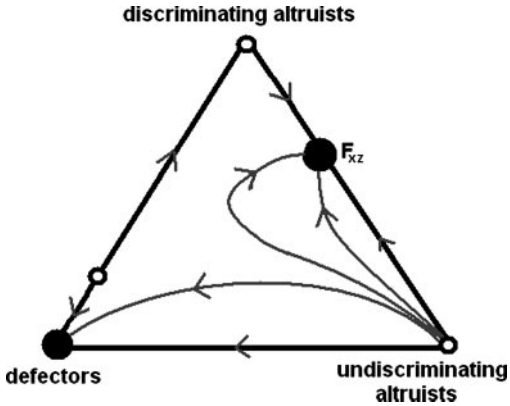

Fig. 2. If the probability of knowing the coplayer's score grows during an individual's lifetime, the dynamics is bistable. Depending on the initial condition, the population in the end consists of defectors only or of a mixture of discriminating and nondiscriminating altruists.

chosen individual is in round n need not be proportional to probability v_n that an individual reaches round n. Instead of $s > c/brq$, we now need the analogous condition $\hat{s} > c\hat{q}/brq$, where $\hat{q} := \Sigma\, v_n q_n$ and $\hat{s} := \Sigma\, v_n q_{n-1}$.

If we assume, in contrast, that discriminators are suspicious, then

$$P_x(n) = -cr + brx + bqr^2 z, \qquad P_y(n) = brx,$$

and

$$P_z(n) = -crq_n g + brx + br^2 zqgq_{n-1}.$$

An argument similar to that above shows that the same bistable dynamics holds if q_n is decreasing.

Results and Discussion

Indirect reciprocity stands somewhere between direct reciprocity and public Goods games. It is based on dyadic interactions, but within a larger group. There are interesting links among the different concepts. Thus, the discriminating strategy considered here is a close relative of Observer Tit for Tat, a strategy for playing the repeated Prisoner's Dilemma in a larger group (see refs. 2 and 20), and Contrite Tit for Tat is analogous to strategies for indirect reciprocity that distinguish between justified and unjustified defections and thus are based on "standing" (21, 22). Both empirical and theoretical findings show that, if public Goods games and indirect reciprocity games alternate, they strongly influence each other; in fact, reputation building through indirect reciprocity may help in solving the "tragedy of the commons" (23, 24).

Many recent experiments have shown that indirect reciprocity can often lead to cooperation (see, e.g., refs. 8–12). However, it is difficult to disentangle the different factors behind the decisions of the players. Thus it appears, for instance, that players who have recently received a donation are more prone to give, in their turn, than those who have experienced a refusal. Furthermore, a sizeable number of players seem motivated not only by the coplayer's score but also by their own. Finally, there is currently a debate between the relative merits of "scoring" and "standing," essentially tackling the issue of whether players distinguish between justified and unjustified defections.

In this paper, we have concentrated on a model that is minimalistic in several points of view. It considers only binary scores, hence it does not count how often a player has given or refused to give, but only what the player did in the last round (or when last observed). This restriction to a binary score may seem artificial, and it is indeed a device to keep the model analytically tractable (for larger score ranges, see refs. 2 and 13).

However, it helps to understand the principles of decentralized mechanisms of local information processing in the context of reputation and moral hazard. For instance, Dellarocas (6) found, in the context of online selling, that binary feedback mechanisms publishing only the single most-recent rating obtained by the seller are just as efficient as mechanisms publishing the seller's total feedback history.

Basically, there are three degrees of sophistication in assessing whether an observed interaction between two coplayers is Good or Bad. In first-order assessment, this judgement depends entirely on whether the potential donor gives or refuses to give. In second-order assessment, the score of the receiver is taken into account; it may make a difference whether help is refused to a Good or a Bad person. In third-order assessment, the score of the donor is also taken into account (see refs. 14 and 16).

The discriminating strategy considered here is based on first-order assessment. It cannot be implemented without cost: a player refusing to help a Bad coplayer will be Bad in the next round and less likely to be helped. In this sense, such discriminators engage in costly punishment, and such punishment, although certainly widespread among humans (15, 25, 26), raises a second-order social dilemma. There is an obvious way out, namely, to use a standing strategy, as suggested by Sugden (21). To the question, "Should an individual who does not help a person with a Bad reputation lose his Good reputation?" posed in ref. 15, the obvious answer is "no."

However, standing strategies are based on second- and even third-order assessment, and hence they are complicated to perform if information is not perfect. If, for instance, the probability that a player's action is known is only 50%, then a second- or third-order assessment module is often useless, because it is unlikely the player will know what the recipient and the recipient's recipient have done in previous rounds. The possibility of error due to misperception or lack of information is large. One may, of course, assume that gossip reduces this uncertainty (4), but gossip increases the possibility of cheating through lies and manipulations (3). In numerical simulations (27) that tested different second-order assessment rules under errors of perception, the standing strategy did much less well than another rather paradoxical strategy that assigns a Bad score to any player meeting a Bad coplayer (irrespective of the player's decision). The reason is essentially the following. Discriminators are threatened because they do less well against invading defectors than undiscriminating altruists; their score is reduced. Standing repairs that defect by not reducing the score of the discriminator; the paradoxical strategy repairs it instead by reducing the score of the undiscriminating altruist as well.

Finally, little experimental evidence of standing strategies has been found so far (28). This makes it important to analyze whether first-order assessment rules, despite their obvious drawbacks, can sustain cooperation. Here we have shown that, under the reasonable assumption that each person's social knowledge increases with experience, a mix of discriminating and undiscriminating altruists can stably resist invasion by defectors.

We thank Josef Hofbauer for helpful advice. H.B. acknowledges support by Austrian Science Funds WK 1009.

1. Alexander, R. D. (1987) *The Biology of Moral Systems* (de Gruyter, New York)
2. Nowak, M. A. & Sigmund, K. (1998) *Nature* **393,** 573–577.
3. Dunbar, R. (1996) *Grooming, Gossip and the Evolution of Language* (Harvard Univ. Press, Cambridge, MA).
4. Panchanathan, K. & Boyd, R. (2003) *J. Theor. Biol.* **224,** 115–126.
5. Bolton, G., Katok, E. & Ockenfels, A. (2004) *Manage. Sci*., in press.
6. Dellarocas, C. (2004) *Sanctioning Reputation Mechanisms in Online Trading Environments with Moral Hazard*, Working Paper (MIT Sloan School of Management, Cambridge, MA).
7. Keser, C. (2002) *Trust and Reputation Building in e-Commerce*, Working Paper (IBM Watson Research Center, Yorktown Heights, NY).
8. Wedekind, C. & Milinski, M. (2000) *Science* **288,** 850–852.
9. Milinski, M., Semmann, D. & Krambeck, H. J. (2002) *Proc. R. Soc. London Ser. B* **269,** 881–883.
10. Wedekind, C. & Braithwaite, V. A. (2002) *Curr. Biol.* **12,** 1012–1015.
11. Engelmann, D. & Fischbacher, U. (2002) *Indirect Reciprocity and Strategic Reputation-Building in an Experimental Helping Game*, Working Paper (Univ. of Zürich, Zurich).
12. Seinen, I. & Schram, A. (2001) *Social Status and Group Norms: Indirect Reciprocity in a Helping Experiment*, Working Paper (CREED, Univ. of Amsterdam, Amsterdam).
13. Leimar, O. & Hammerstein, P. (2001) *Proc. R. Soc. London Ser. B* **268,** 745–753.
14. Ohtsuki, H. & Iwasa, Y. (2004) *J. Theor. Biol.* **231,** 107–120.
15. Fehr, E. & Fischbacher, U. (2003) *Nature* **425,** 785–791.
16. Brandt, H. & Sigmund, K. (2004) *J. Theor. Biol.* **231,** 475–486.
17. Fishman, M. A. (2003) *J. Theor. Biol.* **225,** 285–292.
18. Mohtashemi, M. & Mui, L. (2003) *J. Theor. Biol.* **223,** 523–531.
19. Hofbauer, J. & Sigmund, K. (1998) *Evolutionary Games and Population Dynamics* (Cambridge Univ. Press, Cambridge, U.K.)

20. Pollock, G. B. & Dugatkin, L.A. (1992) *J. Theor. Biol.* **159,** 25–37.
21. Sugden, R. (1986) *The Economics of Rights, Cooperation and Welfare* (Blackwell, Oxford, U.K.).
22. Boerlijst, M. C., Nowak, M. A. & Sigmund, K. (1997) *J. Theor. Biol.* **185,** 281–294.
23. Milinski, M., Semmann, D. & Krambeck, H. J. (2002) *Nature* **415,** 424–426.
24. Panchanathan, K. & Boyd, R. (2004) *Nature* **432,** 499–502.
25. Fehr, E. & Gächter, S. (2002) *Nature* **415,** 137–140.
26. Boyd, R. & Richerson, P. J. (1992) *Ethol. Sociobiol.* **113,** 171–195.
27. Takahashi, N. & Mashima, R. (2004) *The Emergence of Indirect Reciprocity: Is the Standing Strategy the Answer?*, Working Paper (Hokkaido Univ. Press, Hokkaido, Japan).
28. Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H. J. (2001) *Proc. R. Soc. London Ser. B* **268,** 2495–2501.