International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Moral assessment moral hazard in indirect reciprocity

Sigmund, K.

IIASA Interim Report
2012

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-12-070**

Moral assessment and moral hazard in indirect reciprocity

Karl Sigmund (ksigmund@iiasa.ac.at)

Approved by

Ulf Dieckmann
Director, Evolution and Ecology Program

February 2015

Moral assessment and moral hazard in indirect reciprocity

Karl Sigmund [1,2]

[1] Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria
[2] International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria

February 16, 2011

Keywords: evolutionary game theory; indirect reciprocity; cooperation; reputation;

Corresponding author: Karl Sigmund
Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria
e-mail: karl.sigmund@univie.ac.at, phone: +43 01 4277 50612, fax: +43 01 4277 9506

Abstract: xxx

1 Introduction

In *The Descent of Man* (Darwin 1872), Darwin wrote that in contrast to other social animals such as bees or ants, mans 'motive to give aid no longer consists solely of a blind instinctive impulse, but is largely influenced by the *praise and blame* of his fellow men' (our italics). Why should we attach weight to purely symbolic incentives such as praise and blame? Probably because they are often associated with more material incentives. It would make little sense to strive for a good image if all were treated equally. What others know about us is likely to affect the way we are treated.

In many modern approaches to the evolution of human cooperation, the quest to obtain a good image in the eyes of others is relatively neglected. Both in theoretical investigations and experimental tests, it is often assumed that players are anonymous. In real-life interactions, anonymity is less frequent. Usually, we have some information about the individuals we interact with, and are concerned about our own image.

In this paper, the role of reputation in indirect reciprocity will be reviewed. Indirect reciprocity is one of the Five mechanisms of cooperation (Nowak 2006), and arguably the one that is most special to humans. But it should be stressed right away that (a) reputation plays an important role in other forms of cooperation too (not just in indirect reciprocity), and that (b) conversely, there exist forms of indirect reciprocity which are not based on reputation assessment. This will be taken up in more detail in the discussion.

The canonical approach towards explaining altruistic acts (which, by definition, imply a cost to agents who confer benefits to others) is based on a long philosophical tradition. It aims to show that the costs can be recouped in the long run, so that they are self-interested after all. In other words, it means to take the altruism out of altruism (Trivers 2002).

The simplest scenario in this context is that of reciprocal altruism, usually modeled as a repeated Prisoners Dilemma game (Trivers 1971). The recipient of a helpful action returns help at some later occasion. This is the basis of direct reciprocation. 'You scratch my back, and Ill scratch yours'. With indirect reciprocity, the helpful action is returned, not by the recipient, but by a third party. 'You scratch my back, and someone will scratch yours.' This promise seems even more suspect than the previous one. Why should anyone shoulder my debt, and

pay vicariously, in my stead?

Among the several variants of indirect reciprocity, the best known is based on reputation (Sugden 1986, Alexander 1987, Nowak and Sigmund 2005). Help is channeled toward those who have acquired the reputation to be helpful. In this way, exploiters are repressed.

2 Reputation assessment

The simplest model is based on a large, well-mixed population of players randomly meeting each other (Nowak and Sigmund 1998a,b). The probability that the same two players meet more than once is negligible, in such a scenario. Whenever two players meet, chance decides who is the (potential) Donor and who is Recipient. Donors decide whether or not to confer a benefit b to the Recipient, at a cost c to themselves. As usual, it is assumed that $c < b$. Donors providing help acquire the image G (for good), and Donors refusing help the image B (for Bad). Thus players have binary images, entirely determined by what they decided when last in the position of Donor. We can then consider three strategies: (1) the unconditional helpers $AllC$ who always provide help, (2) the unconditional defectors $AllD$ who always refuse to help, and (3) the conditional co-operators $CondC$, who help Recipients if and only if these have a G-image. This strategy is the obvious analogue of TFT (Tit For Tat). It refuses help to those players who, in their previous round, refused to help. We denote by x, y and z the frequencies of the three strategies ($x + y + z = 1$).

If a population contains only two of these strategies, the outcome is the same with direct as with indirect reciprocity (Brandt and Sigmund 2006). $AllD$ players dominate $AllC$ players. The competition of $AllD$ with the conditional strategy is bi-stable, as long as the cost-to-benefit ratio c/b is smaller than the probability w for another round (with the same partner, in direct reciprocity, and with some other partner, in indirect reciprocity). In a mixture of unconditional and conditional co-operators, both do equally well. In order to avoid this dynamic degeneracy, and also to add a realistic feature, we assume that with a probability ϵ, an intended help is not implemented (see also Fishman 2003, Fishman et al 2001, Lotem et al 1999). In this case, there exists a stable coexistence between $AllC$ and $CondC$. In the interior of the simplex Δ_3 which corresponds to the state space of the population (x, y, z), the replicator dynamics (see (Hofbauer and Sigmund 1998) admits a line of rest points, which joins the $AllD + CondC$ equilibrium with the $AllC + CondC$ equilibrium and is given by a constant value of z. In the

vicinity of the $AllC + CondC$ equilibrium, these rest points are stable (but not asymptotically stable, of course). These stable rest points correspond to highly cooperative populations. In the long run, however, random shocks will eventually push the population into the homogeneous state $y = 1$ corresponding to the fixation of $AllD$ (Fig 1). Hence cooperation can prevail for some time, in this model, but will ultimately break down. Although the details of the dynamics differ, the same conclusion holds with direct reciprocity too, if $CondC$ is replaced by TFT. (We assume, in both cases, that the cost c is smaller than the discounted benefit that can be expected in the following round, i.e., $wb(1-\epsilon)$. If this does not hold, the triumph of $AllD$ is immediate.)

One of the reasons for the failure of $CondC$ lies in its paradoxical nature. If a conditional co-operator refuses help to a player with image B, it acquires that image too. The $CondC$-player can, by helping a G-recipient on the next opportunity, redress that image. But during some time, the player is branded, and less likely to receive help. In this sense, the act of punishing a B-player is costly. The strategy can help to uphold cooperation in the population (for a while), but this comes at a price.

There is an obvious way to repair this weakness. It consists in discriminating between justified and unjustified defection. The same problem had already been treated in the context of direct reciprocation. It is well known that a pure TFT-population is greatly plagued by errors in implementation. Each such error provokes a chain of backbiting. A variant of TFT called $ContriteTFT$ can overcome this problem. It is based on the notion of standing (Sugden 1986). In a similar vein, Sugden suggested that assessments, in indirect reciprocity, should take into account whether the Recipient of a refusal to help had a B- or a G-image. Only the latter refusal should be considered as bad, and entail a B image to the non-helping Donor. 'A player can keep his good standing even as he defects, as long as the defection is directed at a player with bad standing. We believe that Sugdens strategy is a good approximation to how indirect reciprocation actually works.'(Nowak and Sigmund 1998a) This point was taken up by a number of authors (Panchanathan and Boyd 2003, Leimar and Hammerstein 2001).

This opens up a vast range of ways of assessing actions, (i.e., attributing a G- or a B-image), even if the actions are not directed at the observer. A first-order assessment rule simply depends on whether the Donor helps the Recipient or not. A second-order assessment rule takes into account, additionally, whether the Recipient has a G-image or a B-image. A third order assessment rule can depend, additionally, on the image of the Donor. It may make a difference whether a B-player or a G-player provides help to a B-player. Altogether, there are 256

third-order assessment rules.

A strategy, in this indirect reciprocity game, depends not only on the assessment rule (i.e., how the player judges actions between two other players), but also how such an assessment is used to reach a decision on whether to help or not. A player could, for instance, decide to give help only to G-players. But the player could also take into account the own image, and help, for instance, whenever the own image is B, so as to remove the blemish as quickly as possible. There are 16 such action rules (including the two unconditional rules $AllC$ and $AllD$), and hence 4096 different strategies conceivable in this set-up (Brandt and Sigmund 2004, Ohtsuki and Iwasa 2004). Not surprisingly, most are nonsensical.

Ohtsuki and Iwasa (2004, 2006) have shown that there exist, among the 256 assessment rules, only eight which can lead to cooperation, when the whole population embraces them. Each of these 'leading eight' is stable in the following sense: there exists a specific action rule such that no dissident minority using another action rule (such as $AllC$ or $AllD$) can do better, and invade. None of these 'leading eight' is of first order. Each distinguishes between justified and unjustified defection. They agree on several points. It is always good to give help to a G-player, and always bad to withhold help from a G-player. Moreover, a good player refusing help to a B-player does not loose the G-image. There remain three situations: namely when someone (good or bad) helps a B-player, or when a B-player refuses help to a B-player. This yields the $2^3 = 8$ assessment systems belonging to the leading eight. Two of them are of second order, and in the following we shall only deal with them. They both agree in viewing (rather oddly) that a B-player refusing to help a B-player obtains a G-image. They disagree on whether it is good to help a B-player or not. The assessment that views it as good will be termed $MILD$, the other $STERN$. For both $MILD$ and $STERN$, the corresponding action rule is: give help if and only if the Recipient has image G. (In particular, the own image will not influence the decision). The corresponding strategy will again be denote by $MILD$ resp. $STERN$.

It is straightforward to analyze the replicator dynamics for a population consisting of the two unconditional strategies $AllC$ and $AllD$ and either the $MILD$ or the $STERN$ strategy (Ohtsuki and Iwasa 2007, Sigmund 2010). In each case, we obtain a bi-stable situation. (Fig.2) But what happens if both the $MILD$ and the $STERN$ strategy occur in the population? This is not obvious. It is important to note that the stability of the leading eight means: no other action rule can invade. This does not imply that no other assessment rule can invade.

Ohtsuki and Iwasa have assumed, like several other authors (Panchanathan and Boyd 2004, etcXXX), that all members of the population agree in their as-

sessment. This means that every player has either the G- or the B-image in the eyes of all players. These authors would agree that it is unlikely that all players observe all interactions, but they assume that every interaction is observed by one player, whose assessment is then shared by all. No matter whether this is a likely scenario or not, it has clearly to be abandoned as soon as one is interested in the competition of several assessment rules. Which moral norm is likely to become established in the population?

Thus G and B mean different things in the eyes of a $MILD$ or a $STERN$ observer. To distinguish them, we may say that a player can be good or bad when assessed according to the $MILD$ rules, and nice or nasty when assessed by the $STERN$ rules. A priori, then, a player can be good and nice, good and nasty, bad and nice or bad and nasty.

The replicator dynamics of a population consisting only of players adopting the $MILD$ or the $STERN$ strategy is disappointing. There is no selective advantage one way or the other, the segment representing all possible mixtures of $MILD$ and $STERN$ consists of rest points. If we add unconditional $AllC$- or $AllD$-players to the population, we observe a bistable outcome. Depending on the initial condition, either a homogeneous $AllD$ population will emerge, or a stable mixture of $MILD$ and $STERN$. The best that can be said is that $STERN$ has a slight advantage, in the sense that whenever there are equally many $STERN$ and $MILD$ players (together with unconditional players), the ratio of $STERN$ to $MILD$ will increase (Uchida and Sigmund, 2010).

This analysis, so far, has relied on the assumption of perfect information. Every players knows about every interaction, either by direct observation or through gossip. This is clearly an unrealistic assumption. If we want to give it up, we must assume that every player has a private list of the images of all other players. Thus the image matrix (β_{ij}) consists of entries G or B, depending on whether player j has image G or B in the eyes of player i. Whenever player j is Donor to some Recipient player k, then those players i who observe the interaction will have an occasion for updating their image of j. The new entries will depend on β_{ik} (since we assume only second-order assessments, the image of the Donor plays no role). But if player i does not observe the interaction between j and k, the value β_{ij} remains unchanged.

This updating process corresponds to a Markov chain on the space of image matrices. A rigorous analysis seems to offer considerable challenges. Uchida has investigated the stochastic process by means of extensive computer simulations (Uchida 2011). The outcome is striking. The smallest deviation from the perfect-information condition has disastrous consequences for a homogeneous population

of $STERN$ players. In the long run, every entry of the image matrix is G or B with equal probability. The entries are uncorrelated. Thus effectively, a $STERN$ player is not doing any better than a player letting a coin-toss decide between helping or not. Compared with this, a homogeneous population of $MILD$ players does much better. A large majority of them will keep agreeing on the images of their co-players. (The percentage depends only on the probability ϵ of misimplementing an intended donation, and on the probability q to observe a given interaction.) A $CondC$ population, on the other hand, ends up with a bad image for everyone. But a mixture of $CondC$ and $AllD$ can keep cooperating: meeting with an $AllC$-player provides the conditional co-operators with an opportunity to redress their image.

In order to obtain an intuitive feeling for these results, we may look at the updating process for β_{ij}. With probability $(1-q)$, it remains unchanged. With probability q, it will be replaced by the new image of j in the eyes of player i. This is 1 if either (a) j gives to k, and i approves, or j refuses to help k, and i approves. The probability that j helps k is $(1-\epsilon)\beta_{jk}$, and the probability that i approves is 1 if i follows the $MILD$ or $CondC$ assessment rule, and β_{ik} in the case of $STERN$. The probability that j refuses to help k is $1-(1-\epsilon)\beta_{jk}$, and the probability that i approves is $(1-\beta_{ik}$ if i follows the $MILD$ or $STERN$ assessment rule, and 0 if i plays $CondC$. If we assume (wrongly) that the images of k in the eyes of i and j, i.e., β_{ik} and β_{jk}, are independent, and if we denote by h_{ij} the expected value of β_{ij} etc, then in the stationary equilibrium, where $h_{ij} = h_{jk} = h$ by symmetry, we obtain for $CondC$, $MILD$ and $STERN$, respectively

$$(1-\epsilon)h = h$$
$$(1-\epsilon)h + (1-(1-\epsilon)h)(1-h) = h$$
$$(1-\epsilon)h^2 + (1-(1-\epsilon)h)(1-h) = h$$

. The corresponding solutions are $h = 0$, $h = (1+\sqrt{\epsilon})^{-1}$ and $h = 1/2$, respectively. Of course the independence assumption is false, but in the case of small q it is almost satisfied.

This handful of results is a striking illustration of the fact that information conditions are of the utmost importance, for reputation-based indirect reciprocity. This was stressed already in the first papers on this topic. In (Nowak and Sigmund 1998b), q denotes the probability that a player knows about the reputation of another player, i.e., has some information about the behavior of that player. With probability $1-q$, the co-player is unknown. In this case, it is assumed that the co-player receives the benefit of doubt, i.e., is held to be a G-player. $CondC$-players

could resist invasion by $AllD$ players if $q > c/b$ (or, in a more elaborate model, if $c < qwb(1-\epsilon)$. In (Uchida 2011) q is the probability that a given player observes the last action of a co-player. If not, then the co-players former image will remain unaltered. Eventually, models will have to encompass both types of uncertainty. It could be that in Alices eyes, player Bob is a stranger. It could also be that Alice knows Bob, but has missed Bobs last action as a Donor.

Whatever the interpretation of q, it seems likely that it is not a constant. In particular, it is reasonable to assume that the social network of a player grows with time. In this case, the player will be more and more likely to know the reputation of a recipient. In (Fishman et al 2001), (Mohtashemi and Lui 2003) and (Brandt and Sigmund 2005), it is shown that appropriate assumptions can turn the $CondC + AllC$ equilibrium into a stable attractor, able to repel invasion attempts by $AllD$-minorities.

It is an obvious weakness of all models considered so far that they are based on a very short memory only. Assessments are updated according to the action last observed. In real life, reputations are not always based on one action only. If we know that a player has cooperated for a long time, and we suddenly see him defecting in one interaction, we will not necessarily lose our good opinion of him (but rather assume that the recipient deserved no better). In particular, (Berger 2011) has shown that a tolerant first-order assessment rule ($TolerantScoring$) can stably sustain cooperation. Such an assessment with built-in tolerance against single defections can be based on sampling two actions in the recipients past.

Several models consider a more sophisticated evaluation system, for instance with a score that is not binary (see e.g. Nowak and Sigmund 1998a, or Leimar and Hammerstein 2001). It provides stability to cooperation: a few isolated defections will not destroy the good reputation that a player has accumulated, but only slightly reduce it.

3 Discussion

Historically, studies of indirect reciprocity were based on direct reciprocity. In a certain sense, however, indirect reciprocity can be viewed as the primary phenomenon, and direct reciprocity as a special case, based on direct experience (as a recipient) of the co-players action. In any case, direct and indirect reciprocation are likely to interact. Thus, players who start a repeated Prisoners Dilemma interaction with some co-player are likely to be guided by that co-players past behavior towards others, and to defect in the first move. The corresponding strategy

$ObserverTFT$ (Pollock and Dugatkin 1992) is an interesting link between TFT and $CondC$. (Whereas the usual TFT-player, on engaging with a new partner in a repeated Prisoners Dilemma game, always provides help, an $ObserverTFT$ also takes into account how that new partner behaved in interactions with others, and in particular defects in the first round if and only if this new partner was last seen defecting.)

Roberts (Roberts 2005) has pointed out that in small populations, the assumption that players interact at most once is implausible. If the probability of re-meeting is sufficiently large, $CondC$ will be superseded by strategies based on direct experience. But a second-order assessment based on three images (good, bad and neutral) exploits advantageously the supplementary information conveyed by reputation and proves superior to strategies based on direct experience only.

It seems plausible that humans do not have separate modules for playing direct reciprocity or indirect reciprocity. Similarly, behavior in direct or indirect reciprocity affects, and is affected, by behavior in public good games (Milinski et al 2002a,b, Panchanathan and Boyd 2004 xxx). A good reputation for cooperating in dyadic interactions is likely to promote the reputation for cooperating in larger groups, and vice versa. (In this context, it may be noted that non-punishers will, in general, not be punished, see Kiyonari et al 2004, just as rewarders will often be rewarded in turn. The former issue is an Achilles heel for cooperation based on negative incentives. The latter is an advantage for cooperation based on positive incentives.)

Both direct and indirect reciprocity rely on the implicit assumption that players act consistently, and that past behavior allows to infer future actions.

An impressive number of experiments have shown that indirect reciprocity works (Wedekind and Milinski 2000, Wedekind and Braithwaite 2002, Bolton et al 2004, 2005, Seinen and Schram 2005, Engelmann and Fischbacher 2009). Interestingly, many players seem to content themselves with first-order assessment, possibly because higher-oder assessment is cognitively taxing (Milinski et al 2001). Of particular interest are the large-scale experiments unwittingly provided by e-trading (Keser 2002, Bolton et al 2004). In e-Bay, for instance, the remarkably high level of honesty is supported by a very simple assessment system based on the satisfaction of customers with their partners. This measure (amalgamated over six months) does not take into account the reputation of the customers themselves who evaluate their partner, and hence is of first-order.

Ever since Trivers seminal paper on reciprocal altruism (Trivers 1971), it is known that reciprocation need not be based on repeated interactions between the same two players only. There exist different notions of generalized reciprocation.

What we have described is reciprocation based on reputation: players known for being helpful are more likely to be helped not necessarily by their recipients, but possibly by others who return the help vicariously, so to speak. Vicarious reciprocation is also known as up-stream reciprocity. We may say that help is caused by a feeling of admiration (Shalizi 2011). Down-stream reciprocity occurs when a player who has been helped returns the help, not to the donor, but to a third party. This can be viewed as misguided reciprocation, caused by a feeling of gratitude. Such misguided reciprocation is well documented by experiments, not only on humans (Wedekind and Milinski 2000, Engelmann and Fischbacher 2009, Rutte and Taborsky 2007, Rutte and Pfeiffer 2009, Barta et al 2010). So far, the only theoretical models that support it seem to require some structured population, and localized interactions (Pfeiffer et al 2005).

The promise of a reward (i.e., a positive incentive) can be used to promote cooperation, if individuals are opportunistically motivated to help whenever they can expect a reward (Hauert et al 2001). The mechanisms are not quite the same. In indirect reciprocity, players reward a co-player because they know that this co-player has performed a helpful action. In the context of positive incentives, players perform a helpful action because they know that they will receive a reward. Switching from positive to negative incentives, we note that an individual with a reputation for punishing cheaters is more likely not to be exploited. In several papers, it has been argued that a player with a reputation as a punisher is less likely to encounter exploiters. Hence, acquiring such a reputation can be beneficial (Hauert, Sigmund,etc XXX). (So far, there seems only one experimental paper supporting this view, see Barclay 2011). All these mechanisms (indirect reciprocity, positive and negative incentives) can be viewed as instances of generalized reciprocity, and the corresponding strategies as offspring of Tit For Tat.

In a larger context, explanations of cooperation based on the handicap principle, such as competitive altruism, also rely on reputation (Zahavi 1995, Roberts 1998, Bshary and Grutter 2006, Sylwester and Roberts 2010). An individual who is known as a good co-operator is more likely to be chosen as partner than an individual known for free-riding. The resulting partner-market may well be the most important aspect of reputation-based cooperation. Our reputation can greatly affect our economic opportunities. As Darwin said, praise and blame can have an important influence on our willingness to help others.

Acknowledgment: Part of this work was supported by TECT I 104-G15.

4 References

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

barclay xxx

Barta Z., McNamara J.M., Huszr D.B. and Taborsky M. (2010): Cooperation among non-relatives evolves by state-dependent generalized reciprocity. Proceedings of the Royal Society London, B, doi:10.1098/rspb.2010.1634.

Berger U (2011) Learning to cooperate via indirect reciprocity. Games and Economic Behavior, forthcoming

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

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

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

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

Brandt, H. and Sigmund K (2005) Indirect reciprocity, image scoring, and moral hazard, PNAS 102 2666-2670

Brandt, H., and Sigmund, K. 2006. 'The good, the bad and the discriminator: errors in direct and indirect reciprocity', *Journal of Theoretical Biology* 239, 183-194.

Bshary, R., and Grutter, A.S. 2006. 'Image scoring causes cooperation in a cleaning mutualism', *Nature* 441, 975-978.

Chalub, F., Santos, F.C., and Pacheco, J.M. 2006. 'The evolution of norms', *Journal of Theoretical Biology* 241, 233-240.

Darwin, C. 1872 The Descent of Man, and Selection in relation to Sex, (reprinted Princeton UP, 1981)

Dufwenberg M, Gneezy, U, Gueth, W and E. van Damme, (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

Engelmann, D. and U. Fischbacher (2009) Indirect reciprocity and strategic reputation-building in an experimental helping game, Games and Economic Behavior, 67(2), 399-407

Fishman, M.A., Lotem, A., and Stone. L. 2001. 'Heterogeneity stabilizes reciprocal altruism interaction', *Journal of Theoretical Biology* 209, 87-95.

Fishman, M.A. 2003. 'Indirect reciprocity among imperfect individuals', *Journal of Theoretical Biology* 225, 285-292.

Hauert et al 2001

Hilbe

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

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.

Kiyonari, T, Barclay, P, Wilson, M and Daly, M (2004) Second-order punishment in the one-shot social dilemma, Int J Psychology 39, 329-xx

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

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

Lotem, A., Fishman, M.A., and Stone, L. 2002. 'Evolution of unconditional altruism through signaling benefits', *Proceedings of the Royal Society B*, 270, 199-205.

Masuda, N., and Ohtsuki, H. 2007. 'Tag-based indirect reciprocity by incomplete social information', *Proceedings of the Royal Society B* 274, 689-695.

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

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

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

Mohtashemi, M., and Mui, L. 2003. 'Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism', *Journal of Theoretical Biology* 223, 523-531.

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

Nowak, M A and Sigmund, K (1998b) The dynamics of indirect reciprocity, JTB 194, 561-574

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

Nowak MA (2006). Five rules for the evolution of cooperation, *Science* 314, 1560-1563.

Ohtsuki H., and Iwasa, Y. 2004. 'How should we define goodness? – Reputation dynamics in indirect reciprocity', *Journal of Theoretical Biology* 231, 107-120.

Ohtsuki, H., and Iwasa, Y. 2006. 'The leading eight: social norms that can maintain cooperation by indirect reciprocity', *Journal of Theoretical Biology* 239, 435-444.

Ohtsuki, H. and Iwasa, Y. 2007. 'Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation', *Journal of Theoretical Biology* 244, 518-531.

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

Pacheco, J., Santos, F., and Chalub, F. 2006. 'Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity', *PLOS Computational Biology* 2, e178.

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

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

Pfeiffer, T., Rutte, C., Killingback, T., Taborsky, M., and Bonhoeffer, S. 2005. 'Evolution of cooperation by generalized reciprocity', *Proceedings of the Royal Society B* 272, 1115-1120.

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

G. Roberts. Competitive altruism: from reciprocity to the handicap principle. Proceedings of the Royal Society of London Series B-Biological Sciences 1998, 265(1394), 427-431.

Roberts, G. 2008. 'Evolution of direct and indirect reciprocity', *Proceedings of the Royal Society B* 275, 173-179.

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

Rutte C and Pfeiffer T (2009): Evolution of reciprocal altruism by copying observed behaviour. Current Science 97(11): 1-6

Rutte C. and Taborsky M. (2007): Generalized reciprocity in rats. PLoS Biology 5, 1421-1425

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

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

Sigmund, K (2010) The Calculus of Selfishness, Princeton UP, Princeton

Shalizi, C. 2011, Honor among thieves, American Scientist ??? 87-88

Sommerfeld, R., Krambeck, H.J., Semmann, D., and 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., and Akiyama, E. 2007a. 'Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity', *Journal of Theoretical Biology* 245, 539-552.

Suzuki, S., and Akiyama, E. 2007b. 'Three-person game facilitates indirect reciprocity under image scoring', *Journal of Theoretical Biology* 249, 93-100.

Sylwester K, Roberts G. 2010 Cooperators benefit through reputation-based partner choice in economic games. Biology Letters 6, 659-662.

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

Takahashi, N. and Mashima, R. 2006. 'The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity', *Journal of Theoretical Biology* 243, 418-436.

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

Trivers. R. (2002) Natural Selection and Social Theory: Selected Papers of Robert Trivers. New York: Oxford University Press

Uchida, S. (2011) Effect of private information on indirect reciprocity, Physical Review E 82, 036111(8)

Uchida, S. and Sigmund, K. (2010) The competition of assessment rules for indirect reciprocity. Journ. Theor. Biol. 263, 13-19

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

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

Yamagishi, T., Jin, N., and Kiyonari, T. 1999. 'Bounded generalized reciprocity: ingroup boasting and ingroup favoritism', *Advances in Group Processes* 16, 161-197.

Zahavi, A. (1995): Altruism as a handicap - The limitations of kin selection and reciprocity. Avian Biol. 26: 1-3.

Figure 1.

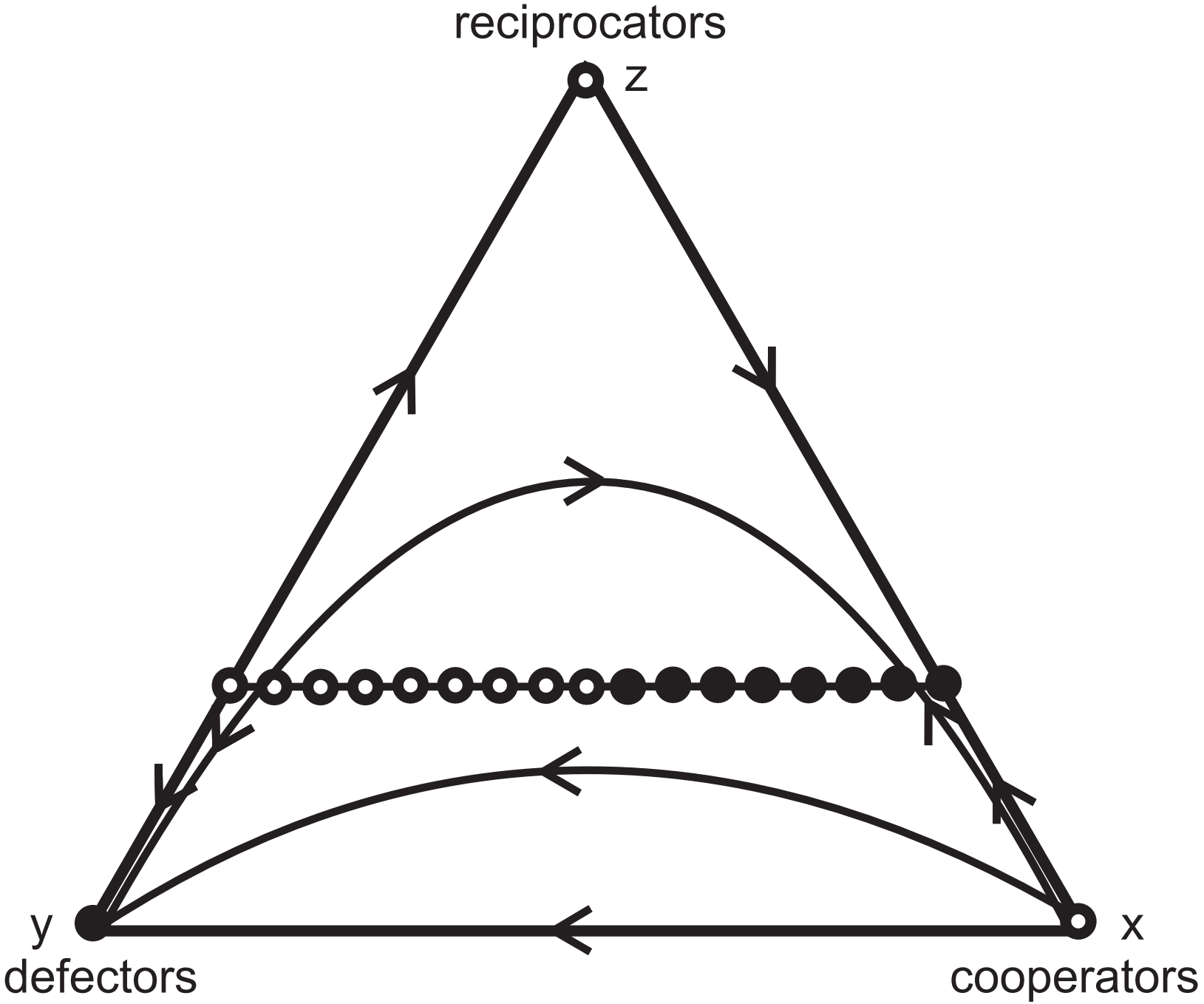

Figure 2.

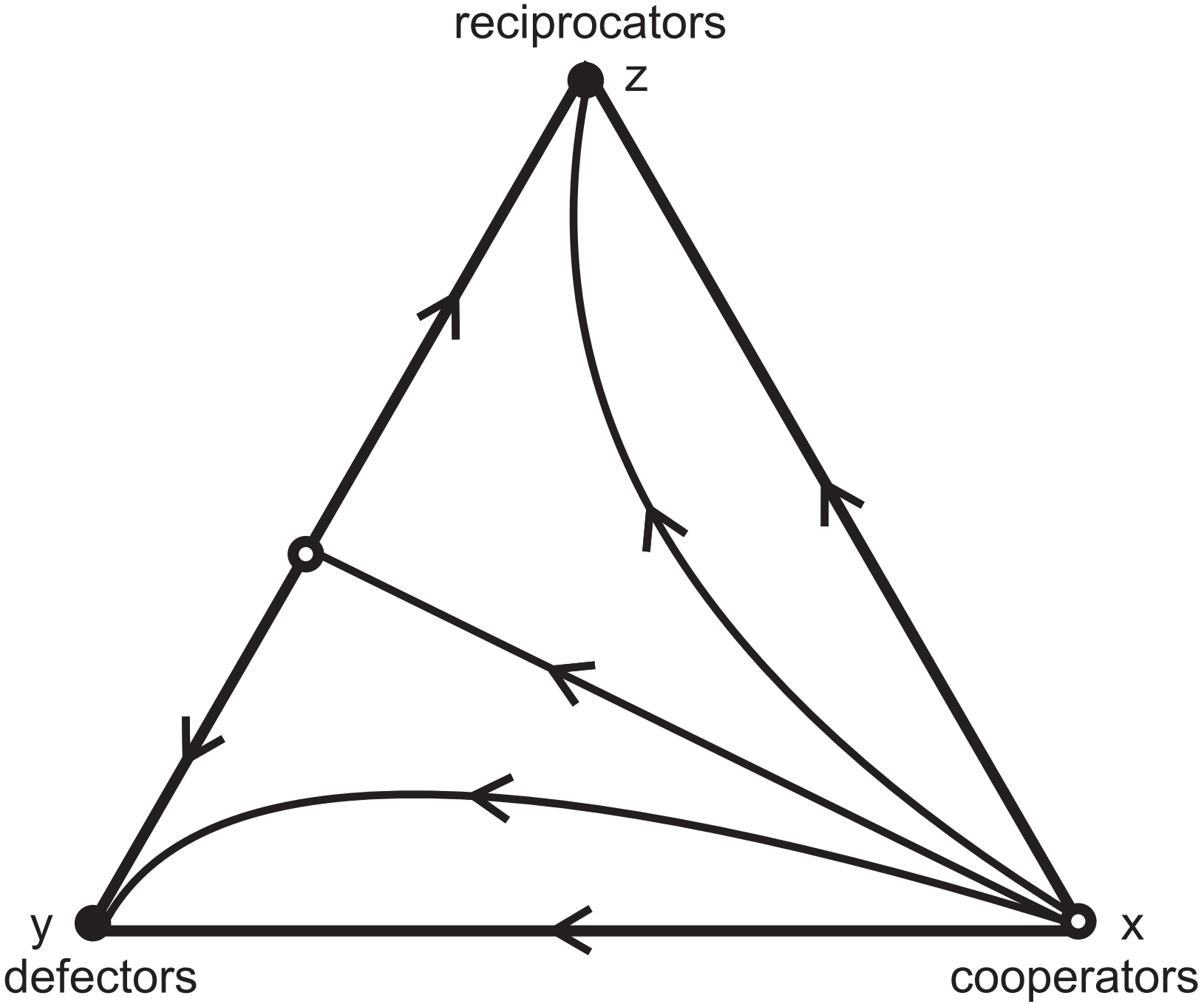

Figure 3.