Journal of Theoretical Biology 299 (2012) 25–30

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Moral assessment in indirect reciprocity

Karl Sigmund [a,b,*]

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

ARTICLE INFO

Available online 5 April 2011

Keywords:
Evolutionary game theory
Indirect reciprocity
Cooperation
Reputation

ABSTRACT

Indirect reciprocity is one of the mechanisms for cooperation, and seems to be of particular interest for the evolution of human societies. A large part is based on assessing reputations and acting accordingly. This paper gives a brief overview of different assessment rules for indirect reciprocity, and studies them by using evolutionary game dynamics. Even the simplest binary assessment rules lead to complex outcomes and require considerable cognitive abilities.

1. Introduction

In *The Descent of Man*, Darwin (1872) 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'. 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, 2006).

The simplest scenario in this context is that of reciprocal altruism, usually modeled as a repeated Prisoner's 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 I'll 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; Brandt and Sigmund, 2006). 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 in the role of 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

* Correspondence address: Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria. Tel.: +43 01 4277 50612; fax: +43 01 4277 9506.
E-mail address: karl.sigmund@univie.ac.at

0022-5193
doi:10.1016/j.jtbi.2011.03.024

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 analog 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 e.g., 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, a sequence of arbitrarily small endogenous perturbations could 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., $wb(1-\varepsilon)$. 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. A conditional co-operator who refuses help to a player with image *B* 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.

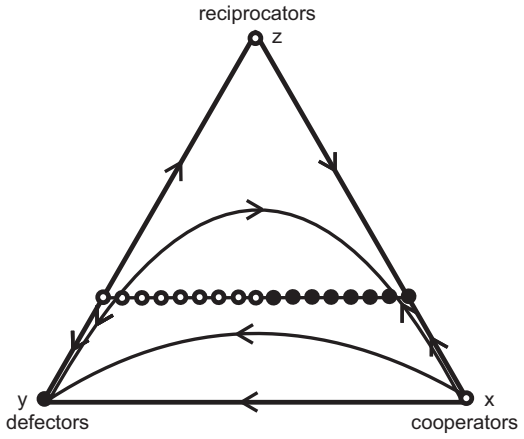

Fig. 1. The replicator dynamics of indirect reciprocity, with the unconditional strategies *AllC*, *AllD* and the reciprocating strategy *CondC*. Full circles correspond to stable fixed points, and empty circles to unstable fixed points (stability being understood in the sense of Lyapunov: all close-by states remain close-by).

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 *Contrite TFT* 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 Sugden's 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 $2^8=256$ 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*, which do not depend on the assessment), and hence $4096=256\times 16$ different strategies conceivable in this set-up (Brandt and Sigmund, 2004; Ohtsuki and Iwasa, 2004). Not surprisingly, most are nonsensical, such as, for instance: 'view everyone as *G* who fails to give to a *G* player, and help if and only if your own image is different from that of your recipient'.

Ohtsuki and Iwasa (2004, 2006) have shown that there exist, among the 256 assessment rules, only eight which can lead to cooperation, if 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 can do better, and invade. (In particular, *AllC* or *AllD* cannot invade.) None of these 'leading eight' is of first order. Each distinguishes between justified and unjustified defection. The eight rules 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 (with image *G* or *B*) 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*. (The former has been studied by Sugden, 1986, the latter by Kandori, 1992). 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*.

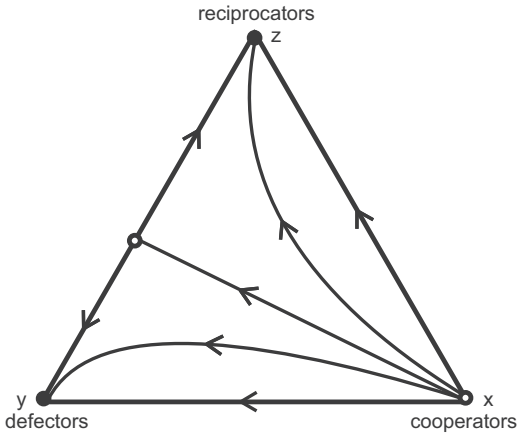

Fig. 2. The replicator dynamics of indirect reciprocity, with the unconditional strategies *AllC*, *AllD* and the reciprocating strategy *MILD* (or *STERN*).

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 together in the population? This is not obvious. It is important to note that the stability of the leading eight merely means that no other action rule can invade. This does not imply that no other assessment rule can invade.

Ohtsuki and Iwasa (2007) and Panchanathan and Boyd (2004) have assumed that all members of the population agree in their assessment. This means that every player has either the *G*- or the *B*-image in the eyes of all players. These authors would accept the view 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. Assuming that assessment rules are private, c.f. Brandt and Sigmund (2005), Pacheco et al. (2006), and Takahashi and Mashima (2006), this raises the question: 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*- and *AllD*-players to the population, we observe a bi-stable 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 *AllC*-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*,

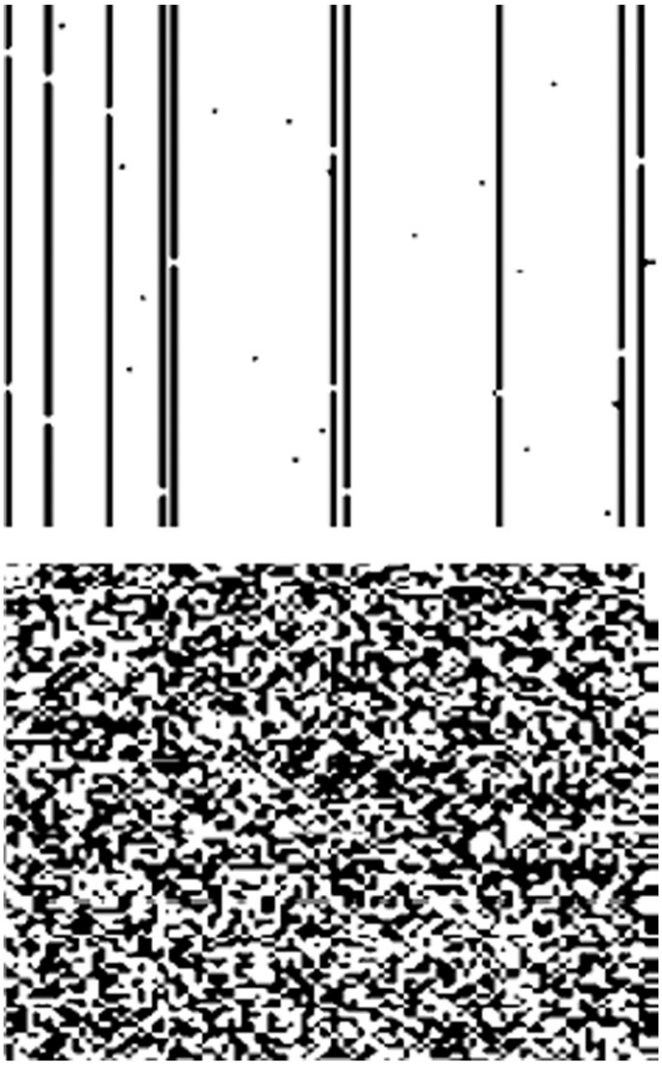

Fig. 3. The image matrix (β_{ij}) in a population of 100 *MILD* players (top) resp. 100 *STERN* players (right). The white and black pixels correspond to good (or nice) resp. bad (or nasty) images. It is assumed that q is 99%. The state, after 20 000 interactions, corresponds to a stationary distribution. The probability ε to mis-implement an intended donation is 0.1. (From Uchida, 2010).

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} and β_{jk}(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.

If observers are chosen at random, this updating process corresponds to a Markov chain on the space of image matrices. A rigorous analysis seems to offer considerable challenges. Uchida (2010) has investigated the stochastic process by means of extensive computer simulations. The outcome is striking (See Fig. 3). 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 mis-implementing 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 *AllC* can keep cooperating: meeting with an *AllC*-player provides the conditional co-operators with an opportunity to redress their image. Clearly, this works also in the case of perfect information.

In order to obtain an intuitive feeling for these results, we may look at the updating process for β_{ij}. For ease of notation, we replace the entries G resp. B by 1 resp. 0. With probability $(1-q)$, the entry 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 (b) j refuses to help k, and i approves. The probability that j helps k is $(1-\varepsilon)\beta_{jk}$, and the probability that i approves is 1 if i follows the *MILD* or *CondC* assessment rule, and it is β_{ik} in the case of *STERN*. The probability that j refuses to help k is $1-(1-\varepsilon)\beta_{jk}$, and the probability that i approves is $(1-\beta_{ik})$ if i follows the *MILD* or *STERN* assessment rule, and it is 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 the *CondC*, *MILD* and *STERN* assessment rules, respectively:

$$(1-\varepsilon)h = h,$$

$$(1-\varepsilon)h+(1-(1-\varepsilon)h)(1-h) = h,$$

$$(1-\varepsilon)h^2+(1-(1-\varepsilon)h)(1-h) = h.$$

The corresponding solutions are $h=0$, $h=(1+\sqrt{\varepsilon})^{-1}$ and $h=\frac{1}{2}$, respectively. Of course the independence assumption is false, but in the case of small q it is a justifiable assumption, since different players are unlikely to base their assessments on the same observations.

This handful of results is a striking illustration of the fact that information conditions are of the utmost importance, for reputation-based indirect reciprocity (cf. Masuda and Ohtsuki, 2007). This was stressed already in the first papers on this topic. In Nowak and Sigmund (1998a,b), 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 recipient 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 $c<qwb(1-\varepsilon)$). In Uchida (2010) 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 Alice's eyes, player Bob is a stranger. It could also be that Alice knows Bob, but has missed Bob's 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 their recipients, or to have observed their latest interactions. In Fishman et al. (2001), Mohtashemi and Mui (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, but suddenly that player defects in one interaction, we will not necessarily lose our good opinion of that player (but rather assume that the recipient deserved no better). In particular, Berger (forthcoming) has shown that a tolerant first-order assessment rule (*Tolerant Scoring*) 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). This provides stability to cooperation: a few isolated defections will not destroy the good reputation that a player has accumulated, but only slightly reduce it. In another vein, Suzuki and Akiyama (2007a,b) have analyzed indirect reciprocity for interactions in larger groups, and found a wealth of interesting dynamic behavior. The evolution of norms in multi-level selection models has been studied by Chalub et al. (2006) and Pacheco et al. (2006).

3. Discussion

Historically, studies of indirect reciprocity were based on direct reciprocity (see, e.g., Rosenthal, 1979; Ellison, 1994; Okuno-Fujiwara and Postlewaite, 1995). 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-player's action. In any case, direct and indirect reciprocation are likely to interact (see Dufwenberg et al., 2001). Thus, players who start a repeated Prisoner's Dilemma interaction with some co-player are likely to be guided by that co-player's past behavior towards others, and to defect in their first move if they have seen their co-player defect on others. 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 Prisoner's 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 (2008) 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). 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 and Barclay (in press), 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., 2005; Seinen

and Schram, 2005; Engelmann and Fischbacher, 2009). Interestingly, many players seem to content themselves with first-order assessment, possibly because higher-order assessment is cognitively taxing (Milinski et al., 2001). Obviously, information transfer is of utmost importance (see, e.g., Sommerfeld et al., 2007). 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 Triver's 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 (see e.g., Boyd and Richerson, 1989). 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 (Yamagishi et al., 1999; Wedekind and Milinski, 2000; Engelmann and Fischbacher, 2009; Rutte and Taborsky, 2007; Rutte and Pfeiffer, 2009; Barta et al., 2011). 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). Switching from positive to negative incentives, we note that an individual with a reputation for punishing cheaters is less likely to be exploited. Hence, acquiring such a reputation can be beneficial (Hauert et al., 2001; Semmann et al., 2004; Hilbe and Sigmund, 2010). So far, there seems only one experimental paper supporting this view, see Barclay (to appear). All these mechanisms (indirect reciprocity, positive and negative incentives) can be viewed as instances of generalized reciprocity, and the corresponding strategies as offspring of *TFT*.

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. This agrees well with Darwin's view that 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.

References

Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter, New York.
Barclay, P. Reputational benefits of altruistic behavior. In: Columbus, F. (Ed.), Advances in Psychology, Nova Science Publishers, Hauppauge, NY, to appear.
Barta, Z., McNamara, J.M., Huszár, D.B., Taborsky, M., 2011. Cooperation among non-relatives evolves by state-dependent generalized reciprocity. Proc. R. Soc. London B 278, 843–848.
Berger, U. Learning to cooperate via indirect reciprocity. Games and Econ. Behav., forthcoming.
Bolton, G., Katok, E., Ockenfels, A., 2004. How effective are on-line reputation mechanisms? An experimental investigation. Manage. Sci. 50, 1587–1602.
Bolton, G., Katok, E., Ockenfels, A., 2005. Cooperation among strangers with limited information about reputation. J. Public Econ. 89, 1457–1468.
Boyd, R., Richerson, P.J., 1989. The evolution of indirect reciprocity. Soc. Networks 11, 213–236.
Brandt, H., Sigmund, K., 2004. The logic of reprobation: action and assessment rules in indirect reciprocity. J. Theor. Biol. 231, 475–486.
Brandt, H., Sigmund, K., 2005. Indirect reciprocity, image scoring, and moral hazard. Proc. Natl. Acad. Sci. 102, 2666–2670.
Brandt, H., Sigmund, K., 2006. The good, the bad and the discriminator: errors in direct and indirect reciprocity. J. Theor. Biol. 239, 183–194.
Bshary, R., Grutter, A.S., 2006. Image scoring causes cooperation in a cleaning mutualism. Nature 441, 975–978.
Chalub, F., Santos, F.C., Pacheco, J.M., 2006. The evolution of norms. J. Theor. Biol. 241, 233–240.
Darwin, C., 1872. The Descent of Man, and Selection in relation to Sex, (reprinted Princeton University Press, 1981).
Dufwenberg, M., Gneezy, U., Gueth, W., van Damme, E., 2001. Direct vs indirect reciprocation—an experiment. Homo Oeconomicus 18, 19–30.
Ellison, G., 1994. Cooperation in the Prisoner's Dilemma with anonymous random matching. Rev. Econ. Stud. 61, 567–588.
Engelmann, D., Fischbacher, U., 2009. Indirect reciprocity and strategic reputation-building in an experimental helping game. Games and Econ. Behav. 67, 399–407.
Fishman, M.A., Lotem, A., Stone, L., 2001. Heterogeneity stabilizes reciprocal altruism interaction. J. Theor. Biol. 209, 87–95.
Fishman, M.A., 2003. Indirect reciprocity among imperfect individuals. J. Theor. Biol. 225, 285–292.
Hauert, C., Nowak, M.A., Sigmund, K., 2001. Reward and punishment. Proc. Natl. Acad. Sci. 98, 10757–10762.
Hilbe, C., Sigmund, K., 2010. Incentives and opportunism: from the carrot to the stick. Proc. R. Soc. London B 277, 2427–2433.
Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge University Press.
Kandori, M., 1992. Social norms and community enforcement. Rev. Econ. Stud. 59, 63–80.
Keser, C., 2002. Experimental games for the design of reputation management systems. IBM Syst. J. 43, 498–503.
Kiyonari, T., Barclay, P. Cooperation in social dilemmas: free-riding may be thwarted by second-order reward rather than punishment. J. Pers. Soc. Psychol., in press.
Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocation. Proc. R. Soc. London B 268, 745–753.
Lotem, A., Fishman, M.A., Stone, L., 1999. Evolution of cooperation between individuals. Nature 400, 226–227.
Masuda, N., Ohtsuki, H., 2007. Tag-based indirect reciprocity by incomplete social information. Proc. R. Soc. London B 274, 689–695.
Milinski, M., Semmann, D., Krambeck, H.J., 2002a. Donors in charity gain in both indirect reciprocity and political reputation. Proc. R. Soc. London B 269, 881–883.
Milinski, M., Semmann, D., Krambeck, H.J., 2002b. Reputation helps solve the 'Tragedy of the Commons'. Nature 415, 424–426.
Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.J., 2001. Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. London B 268, 2495–2501.
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.
Nowak, M.A., Sigmund, K., 1998a. Evolution of indirect reciprocity by image scoring. Nature 282, 462–466.
Nowak, M.A., Sigmund, K., 1998b. The dynamics of indirect reciprocity. J. Theor. Biol. 194, 561–574.
Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1292–1298.
Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.
Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness? Reputation dynamics in indirect reciprocity. J. Theor. Biol. 231, 107–120.
Ohtsuki, H., Iwasa, Y., 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theor. Biol. 239, 435–444.
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.
Okuno-Fujiwara, M., Postlewaite, A., 1995. Social norms in matching games. Games and Econ. Behav. 9, 79–109.
Pacheco, J., Santos, F., Chalub, F., 2006. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLOS Comput. Biol. 2, e178.
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.

Panchanathan, K., Boyd, R., 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., Bonhoeffer, S., 2005. Evolution of cooperation by generalized reciprocity. Proc. R. Soc. London B 272, 1115–1120.
Pollock, G.B., Dugatkin, L.A., 1992. Reciprocity and the evolution of reputation. J. Theor. Biol. 159, 25–37.
Roberts, G., 1998. Competitive altruism: from reciprocity to the handicap principle. Proc. R. Soc. London B 265, 427–431.
Roberts, G., 2008. Evolution of direct and indirect reciprocity. Proc. R. Soc. London B 275, 173–179.
Rosenthal, R.W., 1979. Sequences of games with varying opponents. Econometrica 47, 1353–1366.
Rutte, C., Pfeiffer, T., 2009. Evolution of reciprocal altruism by copying observed behavior. Curr. Sci. 97 (11), 1–6.
Rutte, C., Taborsky, M., 2007. Generalized reciprocity in rats. PLoS Biol. 5, 1421–1425.
Seinen, I., Schram, A., 2005. Social status and group norms: indirect reciprocity in a repeated helping experiment. Eur. Econ. Rev. 50, 581–602.
Semmann, D., Krambeck, H.J., Milinski, M., 2004. Strategic investment in reputation. J. Behav. Ecol. Sociobiol. 56, 248–252.
Sigmund, K., 2010. The Calculus of Selfishness. Princeton University Press, Princeton.
Shalizi, C., 2011. Honor among thieves. Am. Sci. 99 (1), 87–88.
Sommerfeld, R., Krambeck, H.J., Semmann, D., Milinski, M., 2007. Gossip as an alternative for direct observation in games of indirect reciprocity. Proc. Natl. Acad. Sci. 104, 17435–17440.
Sugden, R., 1986. The Economics of Rights, Cooperation and Welfare. Basil Blackwell, Oxford.
Suzuki, S., Akiyama, E., 2007a. Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. J. Theor. Biol. 245, 539–552.
Suzuki, S., Akiyama, E., 2007b. Three-person game facilitates indirect reciprocity under image scoring. J. Theor. Biol. 249, 93–100.
Sylwester, K., Roberts, G., 2010. Cooperators benefit through reputation-based partner choice in economic games. Biol. Lett. 6, 659–662.
Takahashi, N., Mashima, R., 2006. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. J. Theor. Biol. 243, 418–436.
Trivers, R., 1971. The evolution of reciprocal altruism. Quart. Rev. Biol. 46, 35–57.
Trivers, R., 2006. Reciprocal altruism: 30 years later. In: Kappeller, P.M., van Schaik, C.P. (Eds.), Cooperation in Primates and Humans. Springer, Berlin, pp. 67–83.
Uchida, S., 2010. Effect of private information on indirect reciprocity. Phys. Rev. E 82, 036111(8).
Uchida, S., Sigmund, K., 2010. The competition of assessment rules for indirect reciprocity. J. Theor. Biol. 263, 13–19.
Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.
Wedekind, C., Braithwaite, V.A., 2002. The long-term benefits of human generosity in indirect reciprocity. Curr. Biol. 12, 1012–1015.
Yamagishi, T., Jin, N., Kiyonari, T., 1999. Bounded generalized reciprocity: ingroup boasting and ingroup favoritism. Adv. Group Process. 16, 161–197.
Zahavi, A., 1995. Altruism as a handicap—the limitations of kin selection and reciprocity. Avian Biol. 26, 1–3.