Evolution of direct and indirect reciprocity

Gilbert Roberts*

*Centre for Behaviour and Evolution, Newcastle University, Henry Wellcome Building, Framlington Place,
Newcastle upon Tyne NE2 4HH, UK*

Indirect reciprocity (IR) occurs when individuals help those who help others. It is important as a potential explanation for why people might develop cooperative reputations. However, previous models of IR are based on the assumption that individuals never meet again. Yet humans and other animals often interact repeatedly within groups, thereby violating the fundamental basis of these models. Whenever re-meeting can occur, discriminating reciprocators can decide whether to help those who helped others (IR) or those who helped them (direct reciprocity, DR). Here I used simulation models to investigate the conditions in which we can expect the different forms of reciprocity to predominate. I show that IR through image scoring becomes unstable with respect to DR by experience scoring as the probability of re-meeting increases. However, using the standing strategy, which takes into account the context of observed defections, IR can be stable with respect to DR even when individuals interact with few partners many times. The findings are important in showing that IR cannot explain a concern for reputation in typical societies unless reputations provide as reliable a guide to cooperative behaviour as does experience.

Keywords: cooperation; reciprocal altruism; indirect reciprocity; reputations

1. INTRODUCTION

It is well established that individuals will tend to help those who help them (Trivers 1971), but it has also been suggested that individuals will tend to help those who help others (Alexander 1987). This process of indirect reciprocity (IR) is important in offering a reason why it might pay to develop a reputation for being cooperative. As such, it could explain helping behaviour that occurs outside of the restrictive conditions required for direct reciprocity (DR; Axelrod & Hamilton 1981). Furthermore, it has been argued that IR is fundamental to human moral systems (Alexander 1987). Simulation models have demonstrated how systems of IR can lead to cooperation (Nowak & Sigmund 1998), and this has spurred a considerable amount of both theoretical and empirical interest (Nowak & Sigmund 2005).

Although a number of authors have proposed ways in which IR may operate (Boyd & Richerson 1989; Pollock & Dugatkin 1992; Roberts 1998), the model by Nowak & Sigmund (1998) has been most influential. Their model was based on a simple reputation measure called an image score that increased when individuals donated help and decreased when they declined to do so. Strategies were determined by thresholds: the higher the threshold, the greater a partner's image score must be before they were helped. Evolutionary simulations showed that cooperation could be established through discriminatory strategies which helped those with higher image scores.

However, there were a number of issues concerning the model that have meant the conditions in which we can expect to find IR remain a subject of intense theoretical interest (Nowak & Sigmund 2005). In particular, genetic drift played a major role in Nowak & Sigmund's

simulations (Leimar & Hammerstein 2001). It is also questionable whether it is in an individual's strategic interests to take account of a recipient's score: by discriminating against low scorers, one lowers one's own score and makes it less likely that one will receive help. Thus, it may be costly to refuse to help someone with a low image score. Furthermore, observing an act of cooperation or defection is not a reliable guide to the strategy of an individual (Pollock & Dugatkin 1992). Image scoring punishes an individual for defecting even if they were playing a tit-for-tat-like strategy and defecting on a defector. This failure to distinguish 'justified' and 'unjustified' defections is a key weakness of image scoring (Panchanathan & Boyd 2003). A way around this is the standing strategy (Sugden 1986). According to this concept, everyone starts in good standing; they lose good standing by failing to help another individual with good standing and regain it by helping. However, failing to help an individual that is not in good standing does not result in a loss of good standing. Thus, the standing strategy escapes the dilemma of justified defections by removing their cost. This standing strategy has been shown to out-compete image scoring (Leimar & Hammerstein 2001; Panchanathan & Boyd 2003).

However, all of this work has been directed towards elucidating which strategy of IR dominates in an environment of other IR strategies and non-cooperation. No models of IR have included the strategy of DR. There is a good reason for this: because DR is known to work, modellers have ensured individuals never have the chance to receive help from someone they have helped: this means that any cooperation found cannot be attributed to DR. Clearly, the question of whether cooperation can ever be sustained under one-shot interactions is an interesting theoretical one. However, the majority of social interactions in humans and other animals occur in groups within which re-meeting is possible, if not frequent.

*gilbert.roberts@ncl.ac.uk.

Previous authors have implicitly assumed that the conclusions from models generated with no re-meeting can be applied to groups with re-meeting. For example, IR has been advanced as fundamental to human sociality, morality and language (Nowak & Sigmund 2005) despite the fact that human societies (in all but special cases such as internet transactions) typically violate a basic assumption of the models. Similarly, IR has been cited as an explanation for cooperation in other animals such as babblers (Ferriere 1998) even though the small groups in which these birds live make the model inappropriate.

This is an important point because once re-meeting occurs, individuals can choose between indirect and DR strategies. All reciprocity depends upon discrimination whereby altruists restrict their altruistic acts to other altruists. However, the question becomes how such discrimination is best effected, whether by deciding to cooperate based on how a partner has behaved with others, or based on how that partner has behaved with you. This comes down to the relative merits of reputations and experience. Only when individuals are using reputations to inform their decisions can it be worth being seen to be altruistic. Thus, in order to test whether IR can explain a concern for reputation and consequently cooperation in groups with re-meeting, we need to test its stability with respect not just to non-altruism or other strategies of IR but to DR.

Here I investigate the conditions in which we can expect to find IR using a simulation model in which indirect strategies compete with their DR analogues. In order to construct a framework in which strategies of both direct and IR could play against each other, I used a simplified system with features of image scoring, standing and classical DR strategies. IR is based on reputational information. Assuming this is public, an image score will then be a property of an individual, reflecting its history of interactions, regardless of the identity of its partners. However, in DR, it is a partner's previous interactions with the focal individual only that is of importance, so an experience score needs to be calculated for each individual's interactions with each other individual. The aim was to determine when individuals should use reputation or experience in scoring partners, and thereby to determine the extent to which we can expect to see IR and the associated reputation-based behaviour when DR is possible.

2. MATERIAL AND METHODS

I played strategies of IR against their DR analogues in evolutionary simulations. In order to investigate both forms of reciprocity within the same framework, I used a modified version of Nowak & Sigmund's (1998) image scoring model. A simple index of cooperation history was provided by an integer scale running from -1 to 1 . Scores were initialized at neutral (0); if an individual defected on its last move then its score became -1 ; and if it cooperated it became 1 . Note that this differs from Nowak & Sigmund in that their scale went from -5 to 5 ; each interaction either added or subtracted a point, rather than just taking the last interaction; and they limited the number of interactions each individual had so that these bounds were not exceeded. This was not practical when looking at repeated interactions. It also differs from the standing (Leimar & Hammerstein 2001) concept of good and bad by having a neutral rather than positive starting score.

Indirect and DR were implemented by calculating the score using reputational- and experience-based information, respectively. When an individual was using reputational information, the score assigned to a partner reflected that partner's last interaction with any other individual. Such strategies of IR are denoted $IR[k]$. Conversely, when an individual was using experience, the score assigned reflected the partner's last interaction with the focal individual itself. The latter strategies are denoted $DR[k]$. Note that where partners re-met, the last move may have been with the same individual and in these cases the image score will be equivalent to the experience score. While it would be possible to play a strategy that relied purely on second-hand information, it is not a very plausible strategy that discounts information about a partner simply because it is based on their own experience.

The standing strategy of IR was implemented by assigning an image score of -1 only when a defection was against an individual of image score 0 or 1 . To retain a level playing field, a DR version of this was introduced by assigning an experience score of -1 only when a partner had defected on an individual of experience score 0 or 1 . As experience scoring is analogous to tit-for-tat, so the standing version of experience scoring is analogous to contrite tit-for-tat (Boyd & Richerson 1989).

Strategies were defined by their thresholds k , such that a focal individual cooperated only when the score of its partner was greater than or equal to its threshold. For comparability, an individual's threshold applied to both direct and indirect strategies. Retaining Nowak & Sigmund's threshold concept meant that the potential strategy set was simplified by excluding unrealistic strategies of cooperating only with low image score individuals. Where $k = -1$, individuals cooperated even when their partner defected; $k = 0$ strategists cooperated provided the partner had never played or had cooperated; $k = 1$ strategists cooperated provided the partner had cooperated and $k = 2$ strategists never cooperated.

Models were based on a metapopulation or 'island' structure (Leimar & Hammerstein 2001) in which a population of p individuals was distributed over i islands with $n = 100$ on each. In all simulations presented here, $p = 1000$ and $i = 10$. Within each island, individuals were distributed in groups of size $1 \leq g \leq 100$ within which they interacted.

Simulations started with equal proportions of all strategies. Note that previous implementations of the standing strategy have been based upon all individuals starting in good standing (Brandt & Sigmund 2004). However, this begs the question of the origins of such good reputations if what we are trying to explain is the origins of cooperation. Individuals therefore began with neutral image and experience scores of 0 .

The instability of cooperation is a well-known feature of many models. It arises from the tendency of a population of discriminating cooperators to be invaded by indiscriminating cooperators who are then vulnerable to defectors. However, introducing a realistic assumption that not all individuals will be able to cooperate, even if this would be in their own interests, means that the discrimination essential for stability is retained in the population (Lotem *et al.* 1999; Sherratt & Roberts 2001). I therefore included a proportion (10% unless otherwise stated) of phenotypic defectors in each generation. These individuals were constrained from cooperation and did not contribute to strategy evolution through reproduction.

Figure 1. Evolution of indirect versus direct reciprocity (DR) in large groups with few meetings. Strategies of indirect reciprocity (IR) through image scoring played against DR through experience scoring in groups of size $g=100$ and number of meetings (opportunities for unilateral altruism) $m=1$. Data plotted are from the first of 10 replicates (summarized in figure 3). (a) The percentage of moves that were cooperative and (b) the percentage of each strategy, broken down by reciprocity type and threshold value.

Interactions proceeded by randomly choosing a focal individual which then had an opportunity to help a randomly chosen partner. If it helped, then it incurred a cost $c=-1$, and the partner received a benefit $b=2$. This process was repeated such that every group member had on average m meetings with every other, either as donor or recipient. Simulations were evolutionary in that those strategies accumulating the most points produced most offspring. Reproduction was based on the success of a strategy both within and between islands. This was implemented by summing pay-offs within each island to produce the within-island expected relative reproductive success of each strategy and summing across the entire population to give global expected relative reproductive success (Leimar & Hammerstein 2001). An individual for the next generation was derived locally with probability 0.9 and globally with probability 0.1.

Figure 2. Evolution of indirect versus DR in small groups with many meetings. (a) Strategies of IR by image scoring versus DR by experience scoring and (b) IR as a form of contrite tit-for-tat. In both cases, parameters were $g=10$ and $m=40$ and the values plotted are the percentages of each strategy, as defined by reciprocity type and threshold value.

This reduced the potential for genetic drift and allowed migration of successful strategies to other islands. Reproduction was accompanied by mutation: with probability $\mu=0.01$ an individual's strategy was replaced at random.

3. RESULTS

I began by playing image scoring against experience scoring in conditions where we would expect to find IR, namely where individuals have a large number of partners ($g=100$) but only meet each other once, either as donor or recipient. Here, a high level of cooperation was quickly established and remained stable (mean cooperation = $65.99 \pm 0.17\%$ s.e.; all summary statistics are calculated

across 10 replicates for generations 4000–5000; see figure 1a for an example and figure 3a for an overview). Note that the below maximal level of cooperation reflects the presence in the model of phenotypic defectors (Lotem *et al.* 1999; Sherratt & Roberts 2001). This cooperation was almost entirely through IR (mean IR = $99.75 \pm 0.21\%$; figures 1b and 3b). A mixture of IR strategies was found (see also Brandt & Sigmund 2005), with the discriminating strategy of IR[0] (cooperate if the partner has not played or has cooperated with its last partner) somewhat more frequent than the unconditionally cooperative strategy of IR[−1] which exhibited frequency dependence on the former (means 56.49 ± 0.09 and $43.19 \pm 0.30\%$, respectively).

The next step was to consider image and experience scoring in conditions where both direct and IR should be possible. Figure 2a gives a representative example of the evolutionary dynamics where individuals meet many times ($m=40$) in small groups ($g=10$). Once again, cooperation was readily established (mean = $66.43 \pm 0.39\%$; figure 3a) but this time DR in the form of the tit-for-tat-like (Axelrod & Hamilton 1981) discriminating strategy DR[0] dominated IR through image scoring (mean IR = $10.98 \pm 2.33\%$; figure 3b; mean DR[0] = $88.99 \pm 2.33\%$).

I went on to test whether IR is any more robust to DR when IR used a standing-like strategy (Sugden 1986) rather than image scoring and DR used the contrite version of experience scoring. Using the same conditions as above of many meetings in small groups ($m=40$, $g=10$), I once again found a high level of cooperation was quickly established (mean cooperation $66.43 \pm 0.39\%$; figure 3a). This time though, IR was robust against DR (mean IR = $99.97 \pm 0.00\%$; example in figure 2b and summary statistics in figure 3b). Unlike the case when image scoring was dominant, standing was represented almost entirely by the discriminating standing strategy IR[0] (mean IR[0] = $97.21 \pm 0.20\%$).

Further simulations for a range of meeting numbers and group sizes confirmed the pattern suggested by the above examples. Whether IR was represented by image scoring or standing had little effect on overall cooperation levels, apart from in a few cases of low meeting numbers and group sizes when image scoring sometimes failed to establish cooperation (figure 3a). As the number of meetings increased, the percentage of IR decreased, but only where this was represented by image scoring (figure 3b). This trend was clearer where group size was larger.

I investigated whether the percentage of phenotypic defectors affected the competition between direct and IR. Reducing phenotypic defectors from 10 to 1% had negligible effect on the relative percentages (IR was 99.76 ± 0.01 and 99.48 ± 0.05 , respectively), while the percentage of cooperation increased from 66.09 ± 0.05 to 79.60 ± 0.01 as expected due to there being fewer defectors in the population. However, removing phenotypic defectors reduced the percentage of IR to 63.44 ± 1.70 . Furthermore, it had a dramatic effect on the evolutionary dynamics which became highly unstable with no strategy being most frequent for more than a few hundred generations at a time.

Errors in perception of a partner's score were introduced by replacing that score with a random score with probability ε . (Note that the actual rate at which this will lead to an

Figure 3. Comparisons of IR (represented by image scoring and standing and plotted as circles) with DR (represented by experience scoring and plotted as crosses) for a range of meeting numbers and group sizes. (a) The percentage of cooperative moves summed across all strategies in simulations with image scoring compared to those with standing strategies. (b) The percentage of IR (as opposed to DR) strategies in simulations with image scoring and with standing. In both (a,b), points plotted are means with standard errors calculated across 10 replications and including only data for generations 4000–5000. Data are presented for an illustrative range of numbers of meetings m and group sizes g .

individual cooperating when it should have defected or vice versa will be lower than ε and will depend on the actor's strategy and the frequency of cooperation in the population.) I assumed that IR will be subject to sources of error on top of those which apply to DR. This is because information must be gleaned from reputation or gossip rather than by direct experience. Errors were therefore introduced only for those using IR in order to investigate whether such errors would lead to reduced usage of indirect as opposed to DR. I focused on conditions where IR through standing would normally be stable, namely where $g=10$ and $m=20$. As shown in figure 4, the proportion of IR decreased with increasing errors in image scoring such that DR dominated above around $\varepsilon=0.3$.

4. DISCUSSION

The predominance of IR when individuals meet only once is readily understood since there is no opportunity for

Figure 4. As the rate of error in perception of image scores increases, the proportion of indirect as opposed to DR declines. Data are shown as means with standard errors across 10 simulations for each of a representative range of error rates.

using experience of a partner, yet individuals can use reputational information to decide with whom to cooperate. The result is nevertheless interesting in confirming that Nowak & Sigmund's (1998) essential result that image scoring can support cooperation is robust when criticisms relating to the roles of genetic drift (Leimar & Hammerstein 2001) and errors (Panchanathan & Boyd 2003) are taken into account. In the model presented here, genetic drift is reduced using an island model and a form of error is considered through the inclusion of the phenotypic defectors. These were included on the basis that individuals are likely to differ in their quality as partners (Leimar 1997) and some may lack the resources to cooperate at all. By failing to cooperate even when it would be in their interests to do so, phenotypic defectors maintain the selection pressure for discriminating strategies which are essential for the stability of cooperation (Lotem *et al.* 1999; Sherratt & Roberts 2001).

However, when re-meeting was frequent, DR dominated. Why should experience scoring be better than image scoring in this case? A problem with image scoring is that observing an act of defection is not a reliable guide to the strategy of an individual (Pollock & Dugatkin 1992). An individual who defects suffers a reduction in image score even if it was actually playing a tit-for-tat-like strategy and would have cooperated with a cooperator. The experience-based strategy of testing a partner's response to cooperation therefore provides a more reliable guide to when to cooperate. To the extent that the image scoring strategy of Nowak & Sigmund (1998) mirrors Pollock & Dugatkin's (1992) reputation-based tit-for-tat strategy, so this finding mirrors the latter's result that observer tit-for-tat could out-compete the classical experience-based version only where there were few meetings per partner. Note that the advantage of experience was most apparent in larger groups. This is probably because in small groups there is a higher probability that the last move was with the same partner so image scoring is at less of a disadvantage.

However, the tables are turned again when IR is represented by the standing strategy. According to this system, failing to help only results in a loss of good standing when the partner was itself in good standing. Standing has been shown to be evolutionarily stable where image scoring is not (Leimar & Hammerstein 2001; Panchanathan & Boyd 2003). However, it is important to

appreciate that these authors only considered the stability of standing with respect to other IR strategies and non-cooperation: their results do not tell us how standing would fare against DR. In fact, it fares well: the advantage of the standing strategy here is that it not only gives individuals reputational information about how others have behaved but it puts that information into context by treating observed defections differently according to whether they were justified or not. This gives the standing strategists an advantage over experience scorers since they are able to use reputational information to avoid being exploited on the first move with a partner, but at the same time are able to distinguish those who would cooperate with a cooperator.

A number of factors may affect the balance between direct and IR. Temporal discounting has been shown to be important in determining the profitability of reciprocal altruism (Stephens *et al.* 2002). At first sight, it may appear that DR would be more subject to discounting effects than IR since the return from any one particular individual will tend to come further in the future. However, if one considers the returns at the level of the strategy, both direct and indirect reciprocators will benefit each move in a cooperative population so both will be equally subject to any discounting effect. Another factor is the relative memory requirements for direct and IR. Both direct and IR strategies as implemented here need to remember a minimum of one previous move for every partner. The only difference is that in the case of DR this is the partner's last move with the focal individual whereas in the case of IR it is the partner's last move with another player. The amount of memory required is therefore the same. The standing strategy requires more memory than the scoring strategy because it also takes into account context, i.e. it requires memory for not only how a partner behaved but what this behaviour was in response to. Once again, there should be no differences in memory required between direct and indirect strategies because whenever a standing strategy was employed for IR I applied the analogous 'contrite tit-for-tat' strategy for DR.

These results are crucial in understanding when we can expect to see altruistic acts outside of the confines of DR. It can only be worth being seen to be cooperative when others are using that information in their decisions, that is when they are using reputational as opposed to experience-based information. At first sight, an ability to use reputational information would seem to provide an 'essential advantage' (Nowak & Sigmund 1998). This may be true where interactions are constrained such that individuals never re-meet, e.g. in internet transactions, but in more typical groups this advantage is not so clear. On first meeting, using information on how a partner has previously played with others would seem to give an advantage. Yet unless this information also includes how the partner's partner had played then it is actually better to have no information at all and to probe their behaviour, as does the classical tit-for-tat strategy, by cooperating on first meeting and then using this test result to determine the following interaction. On re-meeting, experience would seem to have the advantage of reliability. Nevertheless, the success of the standing strategy suggests that reputational information, put in context, can perform as well as experience. A strategy of basing cooperative decisions on what a partner last did, regardless of whether

it was with oneself or with others, is remarkably robust against one based on direct experience only. This is akin to the idea of generalized reciprocity (Trivers 1971). Generalized reciprocity has been the subject of a recent model (Pfeiffer *et al.* 2005), though that model is very different in that the decision to cooperate is based on one's own last outcome rather than on a partner's last move.

Nevertheless, there are other reasons why DR may be superior: the stability of IR is strongly dependent upon q , the probability of knowing the image score of a partner (Nowak & Sigmund 1998). This probability depends upon factors such as how many interactions are public and how reliably they are observed. These introduce sources of error on top of memory failures which are the main source of error in experience scoring. Information about others' past behaviour may be passed on in gossip. However, if talk is cheap, rumours can be as misleading as they are instructive, so only where rumours can effectively be used to detect cheats are they likely to evolve (Nakamaru & Kawata 2004). Therefore, experience scoring will be more reliable than using reputation. As shown by the simulations in which errors were introduced into the perception of partners' standings, inaccurate knowledge of how a partner has behaved towards others reduces the use of indirect as opposed to DR, giving the latter an important advantage.

The relative merits of direct and IR can be understood with reference to the conditions required for each to be an evolutionarily stable strategy. In the case of DR, the key factor is the probability of meeting w again. Cooperation through DR can be stable provided $w > c/b$, where c is the cost and b is the benefit of an altruistic act (Axelrod & Hamilton 1981). Similarly, cooperation through IR can be stable provided $q > c/b$, where q is the probability of knowing a partner's reputation (Nowak & Sigmund 1998). The relative magnitudes of w and q therefore predict the relative success of DR and IR, respectively. Thus, a high number of meetings will increase w with respect to q , thereby favouring DR. Conversely, employing more sophisticated strategies (such as standing as opposed to scoring) will increase q with respect to w and so favour IR.

This paper provides a starting point for bringing recent advances in IR together with DR so we can predict when we might find IR in social groups. Further work will be required using more sophisticated strategies which use fuller interaction histories. These might be expected to have an advantage, yet it is important to be realistic about how much information will be available and how much information individuals can process (Milinski *et al.* 2001). As a consequence, it has been argued that image scoring has merits over standing (Nowak & Sigmund 2005). However, the results here suggest that if that is true, IR would then lose out to direct.

IR has been advanced as an important explanation for human cooperation, yet the theory rests on an assumption of never meeting again (Nowak & Sigmund 2005). This paper has highlighted how we should be cautious about applying models based on systems with no re-meeting to explaining cooperation in human societies and animal groups. Essentially, when re-meeting is possible, we can expect reputation-based altruistic acts to be favoured only where reputation is as reliable as direct experience in deciding whether to cooperate. A number of experiments have shown how people do use reputational information in deciding whether to give to others (Wedekind & Milinski

2000); there is now a need for experimental tests of whether people will use reputations if they also have experience. If birds can combine the two (Peake *et al.* 2002) then it is probable that humans will be able to as well. There is also a need to test to what extent reputation-based behaviour can be explained by IR or may be due to alternatives such as competitive altruism (Roberts 1998; Lotem *et al.* 2003; Barclay 2004; Van Vugt *et al.* 2007).

I thank the JIF/Wellcome Trust and the BBSRC for funding.

REFERENCES

- Alexander, R. D. 1987 *The biology of moral systems*. New York, NY: Aldine de Gruyter.
- Axelrod, R. & Hamilton, W. D. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)
- Barclay, P. 2004 Trustworthiness and competitive altruism can also solve the "tragedy of the commons". *Evol. Hum. Behav.* **25**, 209–220. (doi:10.1016/j.evolhumbehav.2004.04.002)
- Boyd, R. & Richerson, P. J. 1989 The evolution of indirect reciprocity. *Social Netw.* **11**, 213–236. (doi:10.1016/0378-8733(89)90003-8)
- Brandt, H. & Sigmund, K. 2004 The logic of reprobation: assessment and action rules for indirect reciprocation. *J. Theor. Biol.* **231**, 475–486. (doi:10.1016/j.jtbi.2004.06.032)
- Brandt, H. & Sigmund, K. 2005 Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl Acad. Sci. USA* **102**, 2666–2670. (doi:10.1073/pnas.0407370102)
- Ferriere, R. 1998 Evolutionary biology—help and you shall be helped. *Nature* **393**, 517–518. (doi:10.1038/31102)
- Leimar, O. 1997 Reciprocity and communication of partner quality. *Proc. R. Soc. B* **264**, 1209–1215. (doi:10.1098/rspb.1997.0167)
- Leimar, O. & Hammerstein, P. 2001 Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. B* **268**, 745–753. (doi:10.1098/rspb.2000.1573)
- Lotem, A., Fishman, M. A. & Stone, L. 1999 Evolution of cooperation between individuals. *Nature* **400**, 226–227. (doi:10.1038/22247)
- Lotem, A., Fishman, M. A. & Stone, L. 2003 From reciprocity to unconditional altruism through signalling benefits. *Proc. R. Soc. B* **270**, 199–205. (doi:10.1098/rspb.2002.2225)
- Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H. J. 2001 Cooperation through indirect reciprocity: image scoring or standing strategy? *Proc. R. Soc. B* **268**, 2495–2501. (doi:10.1098/rspb.2001.1809)
- Nakamaru, M. & Kawata, M. 2004 Evolution of rumours that discriminate lying defectors. *Evol. Ecol. Res.* **6**, 261–283.
- Nowak, M. A. & Sigmund, K. 1998 Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577. (doi:10.1038/31225)
- Nowak, M. A. & Sigmund, K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291–1297. (doi:10.1038/nature04131)
- Panchanathan, K. & Boyd, R. 2003 A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126. (doi:10.1016/S0022-5193(03)00154-1)
- Peake, T. M., Terry, A. M. R., McGregor, P. K. & Dabelsteen, T. 2002 Do great tits assess rivals by combining direct experience with information gathered by eavesdropping? *Proc. R. Soc. B* **269**, 1925–1929. (doi:10.1098/rspb.2002.2112)
- Pfeiffer, T., Rutte, C., Killingback, T., Taborsky, M. & Bonhoeffer, S. 2005 Evolution of cooperation by

- generalized reciprocity. *Proc. R. Soc. B* **272**, 1115–1120. (doi:10.1098/rspb.2004.2988)
- Pollock, G. B. & Dugatkin, L. A. 1992 Reciprocity and the evolution of reputation. *J. Theor. Biol.* **159**, 25–37. (doi:10.1016/S0022-5193(05)80765-9)
- Roberts, G. 1998 Competitive altruism: from reciprocity to the handicap principle. *Proc. R. Soc. B* **265**, 427–431. (doi:10.1098/rspb.1998.0312)
- Sherratt, T. N. & Roberts, G. 2001 The role of phenotypic defectors in stabilizing reciprocal altruism. *Behav. Ecol.* **12**, 313–317. (doi:10.1093/beheco/12.3.313)
- Stephens, D. W., McLinn, C. M. & Stevens, J. R. 2002 Discounting and reciprocity in an iterated Prisoner's Dilemma. *Science* **298**, 2216–2218. (doi:10.1126/science.1078498)
- Sugden, R. 1986 *The economics of rights, cooperation and welfare*. Oxford, UK: Blackwell.
- Trivers, R. L. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
- Van Vugt, M., Roberts, G. & Hardy, C. 2007 Competitive altruism: a theory of reputation-based cooperation in groups. In *Handbook of evolutionary psychology* (eds R. I. M. Dunbar & L. Barrett), pp. 531–540. Oxford, UK: Oxford University Press.
- Wedekind, C. & Milinski, M. 2000 Cooperation through image scoring in humans. *Science* **288**, 850–852. (doi:10.1126/science.288.5467.850)