

The evolution of judgement bias in indirect reciprocity

Daniel J. Rankin^{1,2,*} and Franziska Eggmann¹

¹*Division of Behavioural Ecology, Institute of Zoology, University of Bern, Wohlenstrasse 50a, 3032 Hinterkappelen, Switzerland*

²*Center for Systems Biology, Bauer Laboratory, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA*

Indirect reciprocity is a form of reciprocity where help is given to individuals based on their reputation. In indirect reciprocity, bad acts (such as not helping) reduce an individual's reputation while good acts (such as helping) increase an individual's reputation. Studies of indirect reciprocity assume that good acts and bad acts are weighted equally when assessing the reputation of an individual. As different information can be processed in different ways, this is not likely to be the case, and it is possible that an individual could bias an actor's reputation by putting more weight to acts of defection (not helping) than acts of co-operation (helping) or *vice versa*. We term this difference 'judgement bias', and build an individual-based model of image scoring to investigate the conditions under which it may evolve. We find that, if the benefits of co-operation are small, judgement bias is weighted towards acts perceived to be bad; if the benefits are high, the reverse is true. Our result is consistent under both scoring and standing strategies, and we find that allowing judgement bias to evolve increases the level of co-operation in the population.

Keywords: information processing; social information; co-operation; game theory; reputation; tragedy of the commons

1. INTRODUCTION

Indirect reciprocity, where individuals help others they have previously observed being helpful, has often been invoked to explain the evolution of co-operation in humans (Boyd & Richerson 1983; Nowak & Sigmund 1998*a,b*, 2005; Wedekind & Milinski 2000; Leimar & Hammerstein 2001; Ohtsuki 2004). Indirect reciprocity (Nowak & Sigmund 2005) differs from direct reciprocity (Trivers 1971) in that the repeated encounters between the same individuals are not necessary, and a donor instead receives a payback not from the beneficiary itself but from another individual in the population. For indirect reciprocity to work, the interacting individuals must acquire information regarding which individuals they should co-operate with and which they should not. This is done by using social information, either through eavesdropping (Johnstone 2001) or through other methods, such as gossip (Sommerfeld *et al.* 2007). Help should then be directed towards individuals with better reputations (Nowak & Sigmund 1998*a*, 2005; Leimar & Hammerstein 2001). Indirect reciprocity therefore leads to reputation building and morality judgement (Nowak & Sigmund 2005; Rockenbach & Milinski 2006).

Indirect reciprocity works by helping individuals with a good reputation, and not helping individuals with a bad reputation. In models of indirect reciprocity, reputation is modelled as a score, which is determined by an actor's previous actions. There are two mainstream strategies for determining an individual's reputation: standing and

scoring (Nowak & Sigmund 2005). In scoring, the reputation of an individual increases with every act of help and decreases whenever help is refused (Nowak & Sigmund 1998*a*). In other words, scoring is a simple binary system where helping is seen as 'good' and not helping is seen as 'bad'. Standing, by contrast, uses not only what the observed actor does, but also the reputation of the recipient, in order to determine the actor's reputation (Sugden 1986). In standing, refusing to help bad individuals does not reduce one's reputation as it would under scoring, and standing has been shown to be evolutionarily stable over scoring (Leimar & Hammerstein 2001; Panchanathan & Boyd 2003; Ohtsuki & Iwasa 2004, 2007). In good standing, the reputation of an actor not only depends on their actions (whether or not they offer help), but can also depend on the reputation of both the actor and the recipient. Taking both reputation-forming rules (i.e. when to see a given act as either good or bad) and helping rules (i.e. whom to offer help to) into account yields a total of 4096 possible discrete strategies (Ohtsuki & Iwasa 2004, 2007; Nowak & Sigmund 2005). Of these strategies, only eight (termed the 'leading eight') are evolutionarily stable (Ohtsuki & Iwasa 2004). These strategies have two characteristics in common: (i) co-operating with good individuals (whose reputation is perceived to be good) is regarded as a good act, while defecting against them is regarded as a bad act; and (ii) defecting against bad individuals (individuals who are perceived to be bad) is seen as a good, sanctioning, act (Ohtsuki & Iwasa 2004). This stands in contrast to the simpler, binary good/bad rule used in image scoring (Nowak & Sigmund 2005), which will always be prone to invasion from standing strategies (Leimar & Hammerstein 2001; Panchanathan & Boyd 2003; Ohtsuki & Iwasa 2004, 2007). The studies of Ohtsuki & Iwasa (2004, 2007) have highlighted the need

* Author and address for correspondence: Department of Biochemistry, University of Zurich, Building Y27, Winterthurststrasse 190, Zurich 8057, Switzerland (d.rankin@bioc.uzh.ch).

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rspb.2008.1715> or via <http://journals.royalsociety.org>.

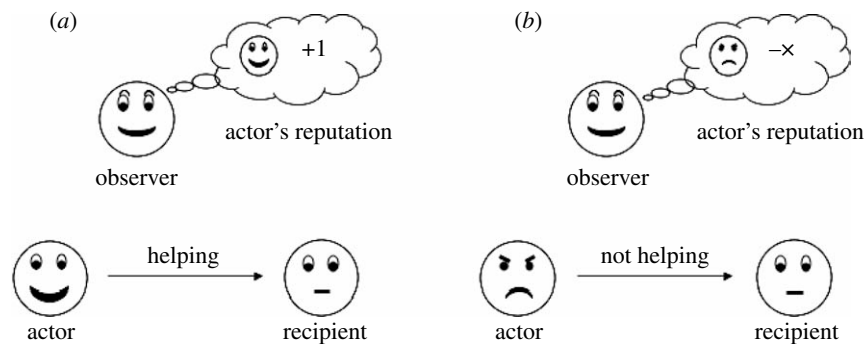


Figure 1. Illustration of judgement bias. (a) An individual (the actor) helps a recipient. This act is observed by an observer, who then adds 1 to the reputation of the actor. (b) The actor does not help the recipient and the observer subtracts x (according to the observer's own judgement bias) from the reputation of the actor. The definition of good and bad acts change in our study, depending on whether we are looking at a scoring strategy (which ignores the reputation of the recipient) or a standing strategy (for which the definition of a good or bad act depends on the reputation of the recipient, as perceived by the observer).

to fully examine the adaptive significance of information processing and information use in social interactions.

The majority of models on indirect reciprocity assign equal weight to good and bad acts (Nowak & Sigmund 1998b, 2005; Leimar & Hammerstein 2001; Panchanathan & Boyd 2003; Brandt & Sigmund 2004, 2005, 2006; Ohtsuki 2004; Ohtsuki & Iwasa 2004; Roberts 2008). In these models, the reputation of an actor is often calculated either by a simple binary system, based on an individual's previous action (i.e. helped or did not help), or, more often, by adding one to their score (in the case of a good act) or subtracting one from their score (in the case of a bad act). Any behaviour involving information use may involve individuals processing information differently according to which behaviours they have observed, and there is therefore no *a priori* reason to expect good and bad acts to have the same influence on an individual's reputation. We term the act of weighting good and bad acts differently, while assessing one's reputation, judgement bias (see figure 1 for illustration). This may come about, for example, by remembering more bad acts that individuals commit than good acts (e.g. Mealey *et al.* 1996; Oda 1997), or by making it harder for an individual to regain a bad reputation than to have a good reputation ruined. In order to be specific about what we mean by judgement bias, we defined 'positive judgement bias' as the case where good acts are weighted more than bad acts (i.e. $x < 1$ in figure 1) and 'negative judgement bias' as the case where bad acts are weighted more than good acts (i.e. $x > 1$ in figure 1). We use the terms 'positive' and 'negative' not to refer to the numerical value of x , but rather to denote acts perceived to be good (positive, i.e. co-operating) and those perceived to be bad (negative, i.e. not co-operating), respectively. The absence of judgement bias corresponds to the case where $x = 1$.

Weighting good and bad acts in the same way could influence the evolution of indirect reciprocity and co-operation. Reciprocity can be seen to be a similar mechanism to punishment, where the refusal of help is a way of sanctioning unco-operative individuals (Rockenbach & Milinski 2006). Indirect reciprocity, however, only sanctions by refraining from helping, while punishment sanctions by actively imposing a cost on a social partner. To some extent, negative judgement bias may potentially work in a similar way to costly punishment, where bad acts are sanctioned through the lowering of an unco-operative actors' reputation (rather than actively imposing

a cost on one's partner, as in the cost of punishment). In this case, if an observer puts more weight on bad acts committed by an actor than on good acts that actor may have done, then the observer will be less likely to offer help to them in future rounds. As such, negative judgement bias results in individuals defecting against others that have a propensity to defect, while rewarding pure co-operators with acts of helping. This should, in turn, promote co-operation. In this study, we investigate the conditions under which judgement bias may evolve. We use a simple individual-based simulation allowing indirect reciprocity and judgement bias to evolve, and compare our results for both scoring and standing strategies.

2. THE MODEL

We built an individual-based simulation to investigate how bias in judging good acts relative to bad acts can be favoured in an indirect reciprocity setting (figure 1 for illustration of judgement bias). We started with a population of 250 individuals. At the start of each simulation, all individuals were assigned different degrees of judgement bias x , which were drawn randomly from an exponential distribution with mean 1. A good act increased the score of the actor by one. A bad act on the other hand reduced the player's score by the judgement bias x (figure 1 for details). As such, each individual had a different perception of the reputations of other members of the population, according to how they bias judgement of good versus bad acts (i.e. the reputation of a given individual was based on the value of x a particular observer has). All individuals in the population had a value of x drawn from an exponential distribution to ensure that the average value of x in the population was 1, in order to fairly evaluate whether positive (where $x < 1$) or negative (where $x > 1$) judgement bias would be favoured. This meant that the judgement bias had the potential to evolve in both directions. Using an exponential distribution meant that the starting median of x was not necessarily 1, although the starting mean was $x = 1$, although we expected that this would not affect our results. We therefore tested our results, where the values of x corresponding to positive and negative judgement bias were reversed (i.e. $x > 1$ and $x < 1$, respectively—fig. S3 in the electronic supplementary material), as well as testing our results with a mutation-limited model, where mutations were drawn from a normal distribution (fig. S8 in the electronic supplementary

material). Both analyses produced qualitatively similar results, corresponding to the main model described here, suggesting that the distribution of x used did not affect the conclusions drawn from our results.

Each player was further assigned two other traits at the start of the simulation: the propensity to discriminate, y , which represented how much an individual will use the score of their partner to make the decision to co-operate; and z , which represented the propensity of an individual to co-operate in the absence of discrimination. If $y=0$, then an individual co-operated purely based on its value of z , while if $y=1$, an individual co-operated based on the perceived image score of their partner. Both y and z could take values between zero and one, and at the start of each simulation, every individual was given a random value of y and z drawn from a uniform distribution.

(a) Interaction

In each round, every individual acted as the focal individual (the actor) in an interaction with a randomly chosen partner (the recipient). As we consider a population size of 250 individuals, the probability of meeting the same individual in two consecutive rounds was negligible (0.004). Each individual had a reputation score based on past actions. At the beginning of the model, every player's reputation score was set to zero, and after each round the image score of all individuals was updated. Discriminating individuals always co-operated (helped) if their partner had a reputation score of zero or above, and defected (did not help) if their partner had a reputation score of less than zero (by default, this means that an individual who discriminates will help in the first round, since all reputations are zero at the beginning of each generation). For each interaction, a uniform random number was drawn between zero and one. If the chosen actor had a discrimination value y greater than this value, then the player used the reputation of the actor to decide whether or not to help the recipient (see §2*b* for details). If the reputation of the recipient, from the viewpoint of the actor, was positive then the actor would help, while if the recipient's reputation was negative, then the actor would not help.

If y was less than the randomly chosen number, then another uniform random number was chosen to decide whether the actor will choose to help the recipient. In this latter case, if z (the propensity for indiscriminate co-operation) is greater than the random number, then the individual helps; otherwise it does not help. In all cases, if the focal individual chose to help, it conferred a benefit b on its partner and paid a cost c . If the focal individual did not help, then neither individual gained, nor paid, anything.

(b) Reputation

We assumed that each act was observed by every member of the population (although we later relaxed this assumption to take into account limitations in the number of acts a given individual could observe—see figs S5 and S6 in the electronic supplementary material). If an observer witnessed an actor committing a good act, then the reputation of the actor increased by 1 point. If an observer witnessed a bad act, then the reputation of the actor decreased by x (where x is the individual judgement bias of the observer). As such, all reputations in our model

were personal and with respect to the observer. After every round, the reputations, as well as the pay-offs, of all individuals were updated.

We decided to compare two different reputation systems. The first was the classical 'scoring' system (Nowak & Sigmund 1998*b*), where defection is always seen as bad, and therefore decreases one's score by x (the individual judgement bias of the observer), while co-operation is always seen as good and increases one's score by 1. In the second reputation system, we used 'standing' from one of the leading eight criteria for reputations (Ohtsuki & Iwasa 2004). This is based not only on what the focal individual does, but also on the reputation of their partner. This avoids the dilemma posed by good individuals in image scoring: if a good individual defects against a known defector, they will lower their image score, even though it would be the best not to co-operate with a known defector (Nowak & Sigmund 2005). We compared the standing and scoring in order to evaluate how judgement bias would evolve under these two reputation-forming mechanisms. In our analysis, we defined standing as the case where co-operation was always seen as good, regardless of the reputation of the recipient, and therefore increased one's score. In line with keeping within the reputation-forming rules defined by the leading eight criteria (Ohtsuki & Iwasa 2004), we assumed that a good individual defecting against an individual who had a bad (i.e. negative) image score was seen as good, while a bad individual defecting against another bad individual was ignored, and had no effect on the actor's score. As judgement bias x differed between individuals, image scores were always based on the reputation of the focal individual as judged by any given observer.

(c) Selection

Every individual started each generation with a base-line pay-off of 100, to avoid negative pay-offs. After every generation, the total pay off of each individual was calculated. Selection took place at the end of each generation. Pay-offs were converted into probabilities (by dividing by 10 000), and these probabilities were then used to select individuals at random, but proportional to their pay-offs, to form a pool of potential offspring. Offspring for the next generation were selected from this pool. Every simulation ran for 5000 generations, which was long enough for selection to reduce the variance of all three traits in the population to a very small value (less than 10^{-10}). The simulation was run 100 times and for a range of different benefits. The costs were always chosen to be smaller than the benefits ($b > c$, which is usual in co-operation games), and we set the costs to 1. We ran the simulation for 25 rounds in each generation. To test the robustness of our results, we compared the simulation for different parameters. We ran the simulation for different numbers of rounds (both 5 and 50) in each generation, and found them to be qualitatively the same (fig. S3 and S4, respectively, in the electronic supplementary material). Testing our model for the case where individuals in the population can only remember 10 per cent of all encounters (fig. S5 in the electronic supplementary material) or when the population size was 25 (fig. S6 in the electronic supplementary material), we found the results to be qualitatively the same. We additionally compared our results with a model that started from

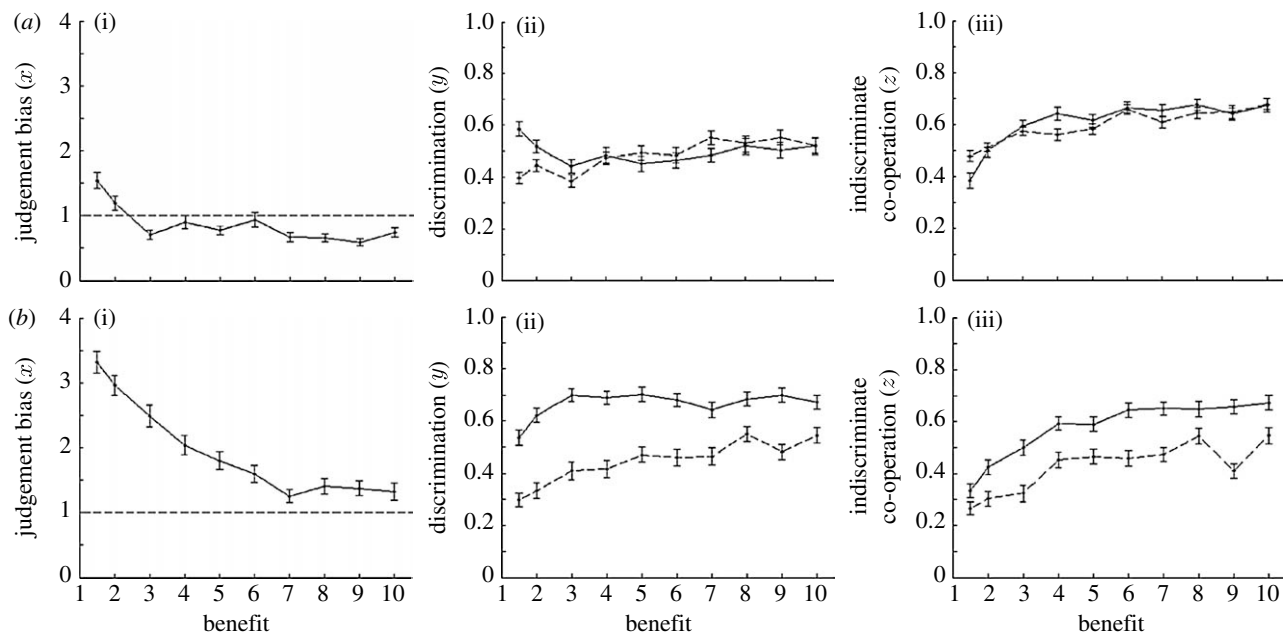


Figure 2. The evolution of judgement bias x , in combination with discrimination y and co-operation x coevolving. (a) The results from the image-scoring simulation and (b) the results from the standing simulation. (a(i), b(i)) show judgement-bias, (a(ii), b(ii)) show propensity to discriminate and (a(iii), b(iii)) show propensity to co-operate in the absence of discrimination. Solid lines indicate results where judgement bias could evolve and dashed lines indicate results where judgement bias did not evolve (classical indirect reciprocity, where $x=1$ for all members of the population). The horizontal line in (a(i)) and (b(i)) shows the threshold where there is no judgement bias (i.e. $x=1$). The simulation was run for 25 rounds and costs were set to $c=1$.

a homogeneous population and randomly introduced mutations in all three traits (see fig. S7 in the electronic supplementary material for details), and found that the results were overall qualitatively similar.

3. RESULTS

We first consider the case where both y and z are fixed (and cannot evolve), and look at the influence of these two parameters on the evolution of judgement bias x (fig. S1 in the electronic supplementary material plots the effect of changing the propensity to discriminate y and indiscriminate co-operation z on the evolution of judgement bias for the scoring simulation; fig. S2 in the electronic supplementary material shows corresponding results under good standing). If $y=0$ for all members of the population, judgement bias never evolves (i.e. $x=1$), as no individuals discriminate in the population. If $y=1$, or $z=1$, all individuals always co-operate in the first round and so judgement bias also never evolves as there is no variation in the reputation of individuals. It can be seen that increasing the level of discrimination (y) or decreasing the level of indiscriminate co-operation (decreasing z) both have the overall effect of favouring negative judgement bias (i.e. weighting bad acts more strongly than good acts—where x evolves to be less than 1).

In both the scoring and standing simulations, some judgement bias evolved (figure 2). In both cases, smaller values of b (i.e. lower benefit-to-cost ratios) favoured negative judgement bias, where more weight was attached to remembering bad acts than to good acts. In the scoring simulation, up to the point where the benefits were approximately three times greater than the cost (i.e. $b=3$, $c=1$), positive judgement bias evolved, where judgement bias was skewed towards weighting bad acts more strongly than good acts. In the standing simulations,

for the parameter range analysed, only negative judgement bias evolved—bad acts were always weighted more heavily than good acts—while after this point the reverse became true. We tested the standing simulation for extreme values and found that positive judgement bias was still favoured when the benefits were $b=100$ (i.e. 100 times greater than the costs; average values at the end of 100 simulations: $x^{\text{end}}=0.651\pm 0.076$, $y^{\text{end}}=0.634\pm 0.030$ and $z^{\text{end}}=0.776\pm 0.023$). However, there was still no judgement bias for standing when $b=100$ (average values at the end of 100 simulations: $x^{\text{end}}=1.019\pm 0.095$, $y^{\text{end}}=0.753\pm 0.023$ and $z^{\text{end}}=0.737\pm 0.025$). We further ran the simulation with no cost to help, to investigate what happened at the extreme benefit-to-cost ratios, in the absence of any costs at all. In this case (where $c=0$, $b=1$), judgement bias was skewed towards judging good acts better than bad acts in the scoring simulations (average values at the end of 100 simulations: $x^{\text{end}}=0.698\pm 0.068$, $y^{\text{end}}=0.574\pm 0.032$ and $z^{\text{end}}=0.698\pm 0.024$) but not in the standing simulations (average values at the end of 100 simulations: $x^{\text{end}}=1.143\pm 0.112$, $y^{\text{end}}=0.685\pm 0.025$ and $z^{\text{end}}=0.625\pm 0.029$). We also compared the simulations where judgement bias was allowed to evolve to those where it was not (i.e. where we fixed $x=1$ for all individuals). We found that, in fact, allowing judgement bias to evolve had the result of favouring discrimination (middle graphs in figure 2), and also favoured increased co-operation among non-discriminators (lower graphs in figure 2). This was observed in both the standing and the scoring simulations, although the effect was much stronger under standing. We additionally ran the simulation where judgement bias was inverted, such that a value of $x>1$ corresponded to positive judgement bias, where more weight was placed on good acts rather than bad acts (i.e. the reverse of above: a judgement bias greater than 1 put

more weight on good acts, while a judgement bias less than 1 puts more weight on bad acts). We found that the results were qualitatively the same (fig. S3 in the electronic supplementary material).

4. DISCUSSION

Our results show that judgement bias evolves under indirect reciprocity. While judging good and bad acts equally allows for a very simple method of reputation, our results suggest that the information used to assess the reputation of conspecifics should not carry the same weight. In our model, judgement bias (x) was with respect to the weight placed on bad acts relative to good acts. This means that, for a judgement bias of three, the negative effect of a single bad act on one's reputation is only rectified after subsequently committing three good acts. The intensity of judgement bias depends very much on the ratio of costs and benefits: for lower costs, the weight that a badly perceived act carries will be greater than that of acts perceived to be good. As the benefits of helping increase, judgement bias becomes skewed towards good acts (fig. S1 in the electronic supplementary material); however, this result varies depending on the reputation rules employed (i.e. scoring versus standing; figure 2).

Previous work has shown that image scoring is generally favoured, if the probability q of knowing the reputation of one's social partner exceeds the benefit-to-cost ratio (i.e. $q > b/c$; Nowak & Sigmund 1998a; Taylor & Nowak 2007). This means that image scoring is increasingly favoured as the benefit of helping increases. If the costs to help are relatively high, choosing to co-operate with an individual who defects can be costly. An individual should therefore avoid any risks and stay on the safer side, i.e. avoiding helping defecting individuals. In our model, weighting bad acts more strongly than good acts (i.e. with a higher value of x) when determining an actor's reputation can help to avoid the cost of helping defectors. When the costs of helping are high, helping is expensive and should therefore only be directed to individuals with good reputations (i.e. to more co-operative individuals). As such, under lower cost-to-benefit ratios, stronger negative judgement bias (which puts greater weight on bad acts) evolves.

As the benefits of helping increase relative to the costs, co-operating with a bad individual becomes less costly. An individual should now care more about its own reputation in order to earn the high benefits from being helped. Under these circumstances, it becomes more beneficial to try to foster as good a reputation as possible, and this can be achieved through positive judgement bias by weighting good acts more strongly than bad acts. This can be seen by comparing the results of the standing simulation to those of the scoring simulation. Under image scoring, defecting against a defector is unconditionally seen as a bad act, and will therefore reduce one's reputation and the chance of being helped in the future. If the cost of helping is cheap (e.g. when the benefit-to-cost ratio is high), then individuals who have a positive judgement bias (and put more weight on good acts) will co-operate more, and will therefore gain a better reputation. This, in turn, will lead to them experiencing help from others, and so positive judgement bias will evolve in the population. Under standing, a good individual who defects against bad individuals is seen as good (as opposed to scoring, where

they are regarded as bad acts), while a bad individual defecting against another bad individual is ignored. As such, defecting against a bad individual will do nothing to harm one's reputation (and will even further raise the reputation of an individual with a good reputation). This means that, even under relatively high benefits, judgement bias towards bad acts (negative judgement bias) is favoured, as one can still defect with bad individuals without ruining one's reputation.

An interesting result of our model is that, under standing, there is a much larger range of benefits over which judgement bias towards bad acts is favoured, than under scoring (figure 2). In other words, standing promotes judgement bias towards bad acts. This is likely to be due to a feedback between the coevolution of discrimination and judgement bias under scoring, in that an individual whose judgement bias is weighted towards bad acts will not help as much (as they discriminate against a larger number of individuals) and will lower their reputation. This is caused by a limitation in scoring, where the reputation of the recipient is ignored: under image scoring, defecting against a bad individual is simply seen as a bad act (Nowak & Sigmund 1998a), while under good standing this limitation is removed, and discriminating individuals take into account not only whether an individual co-operated, but also the reputation of who they interacted with (Leimar & Hammerstein 2001). Dramatically, biasing the reputation of an observed individual through judgement bias can almost be seen as a punishing act, as it reduces the chance that the observer will help them in a future encounter. As such, we note that strong judgement bias works in a similar way to punishment or strong reciprocity (Fehr & Gächter 2002; Fehr & Fischbacher 2004; Gardner & West 2004; Hauert *et al.* 2007), where a bad act is sanctioned by a social partner. In the case of judgement bias, this sanctioning comes in the form of a dramatically lowered reputation (e.g. if $x = 4$, it will take four good acts to regain a good reputation). The difference with punishment is that, with reciprocity, sanctions occur through not co-operating with unco-operative individuals (and hence, despite the potential reduction in one's reputation, a focal individual will benefit from not paying the cost of co-operation), while punishment involves active and costly sanctioning of unco-operative players (Rockenbach & Milinski 2006).

A consequence of the evolution of judgement bias is that it favours both indirect reciprocity and indiscriminate co-operation (compare the solid and dotted lines in figure 2, under both standing and scoring, although the effect is stronger under standing). This is because, when judgement is biased towards weighting bad acts more strongly than good acts, committing a bad act becomes much more costly for one's reputation. This implies that not only is it evolutionarily favourable to process information in different ways, but also this will in turn promote the evolution of co-operation by favouring discriminative strategies. An interesting question to examine further would be whether standing still remains evolutionarily stable over scoring when judgement bias is allowed to evolve. However, based on previous studies, we strongly expect it to remain so (Leimar & Hammerstein 2001; Ohtsuki & Iwasa 2004, 2007).

We have not specified which mechanisms may cause judgement bias, and have simply assumed that one's

reputation will be weighted in a particular way. Judgement bias could work if cheaters were simply remembered better than co-operators. There is some evidence for this in the psychological literature. For example, photographs of people with tags attached to them were more likely to be remembered when the tag read 'cheater' than when it read 'co-operator' or 'trustworthy' (Mealey *et al.* 1996). This suggests that defecting from co-operation elicits a stronger psychological response than co-operating (Mealey *et al.* 1996). In a similar vein, it has been shown experimentally that human subjects devote more attention to unco-operative individuals than to co-operative individuals (Vanneste *et al.* 2007). Memory has already been shown to constrain strategies under direct reciprocity (Milinski & Wedekind 1998). It is likely that additional distractions during co-operative scenarios could mean that, if any acts are remembered in such a situation, they will be bad acts.

Additional evidence for judgement bias comes from a recent study involving customer-added book reviews on online booksellers (Chevalier & Mayzlin 2006). Reputation plays a large part in many areas of human life, particularly now with the rise of the internet as a marketplace (Resnick & Zeckhauser 2002; Resnick *et al.* 2006). Online book reviews can be seen as the collective reputation of a given book, as they are made by the public at large, and a potential buyer may form an opinion online regarding the quality of a given book when he/she views the book. Studies have found that bad reviews had a much greater (negative) impact on book sales than good reviews (Chevalier & Mayzlin 2006). This suggests that customers of online booksellers attach more weight to negative reviews than they do to positive reviews, and thus exhibit negative judgement bias.

We expect judgement bias to evolve in any organism that exhibits image scoring, whether that be cleaner fish (Bshary 2002; Bshary & Grutter 2006) or humans (Wedekind & Milinski 2000; Wedekind & Braithwaite 2002). Repeated interactions between the same individuals are not necessary for indirect reciprocity and judgement bias to evolve. However, reputation-building systems are costly and will require demanding cognitive abilities (Nowak & Sigmund 2005). As such, we expect our results to apply to animals with higher cognition: as the cost of obtaining information increases, reputation gathering will become less favoured. Our results could potentially apply not only to indirect reciprocity but also to direct reciprocity, where individuals score their partners based not just on what they did in the last round, but from all previous rounds. Based on our results, one might expect that, under direct reciprocity, the tendency would be to attach more weight to remembering acts where one's partner refused to help over those where they did help. For example, refusing to help after a partner has abstained from helping twice ('tit-for-two-tats', Axelrod 1984) is a way of weighting the act of helping over not helping (by not helping after one's partner did not return help twice). We could imagine that if players helped according to a cumulative 'reputation' or 'score' based on what their partner did, the acts of helping and not helping should have different influences on the decision of a focal individual to co-operate with its partner.

Reputation involves a complex processes, and there are many types of information available on which an individual can base his or her perception of another's

reputation. Weighting good or bad acts in the same way is not evolutionarily stable according to our model, and judgement bias can additionally help to promote indirect reciprocity and co-operation. Consistent with predictions on how humans remember individuals who commit bad acts (Mealey *et al.* 1996; Oda 1997; Barclay & Lalumière 2006; Vanneste *et al.* 2007), we expect judgement bias to be an important part of reputation building in indirect reciprocity.

D.J.R. was funded by the Swiss National Science Foundation (grant no. 3100A0-105626 to M. Taborsky) and the Cogito Foundation. We are grateful to Claus Wedekind for encouragement and very helpful discussions and comments on the manuscript. We additionally thank Kevin Foster, Andy Gardner, Christoph Hauert, Dik Heg and Michael Taborsky for their discussions and Joao Alpedrinha, Sasha Dall, Andy Gardner and Laurent Lehmann for their helpful comments. The comments of four anonymous reviewers greatly improved the manuscript.

REFERENCES

- Axelrod, R. 1984 *The evolution of cooperation*. New York, NY: Perseus Books Group.
- Barclay, P. & Lalumière, M. L. 2006 Do people differentially remember cheaters? *Hum. Nat. Interdiscip. Biosoc. Perspect.* **17**, 98–113. (doi:10.1007/s12110-006-1022-y)
- Boyd, R. & Richerson, P. J. 1983 The evolution of indirect reciprocity. *Soc. Networks* **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 reciprocity. *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)
- Brandt, H. & Sigmund, K. 2006 The good, the bad and the discriminator—errors in direct and indirect reciprocity. *J. Theor. Biol.* **239**, 183–194. (doi:10.1016/j.jtbi.2005.08.045)
- Bshary, R. 2002 Biting cleaner fish use altruism to deceive image-scoring client reef fish. *Proc. R. Soc. B* **269**, 2087–2093. (doi:10.1098/rspb.2002.2084)
- Bshary, R. & Grutter, A. S. 2006 Image scoring and cooperation in a cleaning mutualism. *Nature* **441**, 975–978. (doi:10.1038/nature04755)
- Chevalier, J. A. & Mayzlin, D. 2006 The effect of word of mouth on sales: online book reviews. *J. Mark. Res.* **43**, 345–354. (doi:10.1509/jmkr.43.3.345)
- Fehr, E. & Fischbacher, U. 2004 Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87. (doi:10.1016/S1090-5138(04)00005-4)
- Fehr, E. & Gächter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
- Gardner, A. & West, S. A. 2004 Cooperation and punishment, especially in humans. *Am. Nat.* **164**, 753–764. (doi:10.1086/425623)
- Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A. & Sigmund, K. 2007 Via freedom to coercion: the emergence of costly punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)
- Johnstone, R. A. 2001 Eavesdropping and animal conflict. *Proc. Natl Acad. Sci. USA* **98**, 9177–9180. (doi:10.1073/pnas.161058798)
- Leimar, O. & Hammerstein, P. 2001 Evolution of co-operation through indirect reciprocity. *Proc. R. Soc. B* **268**, 745–753. (doi:10.1098/rspb.2000.1573)

- Mealey, L., Daoood, C. & Krage, M. 1996 Enhanced memory for faces of cheaters. *Ethol. Sociobiol.* **17**, 119–128. (doi:10.1016/0162-3095(95)00131-X)
- Milinski, M. & Wedekind, C. 1998 Working memory constrains human cooperation in the Prisoner's Dilemma. *Proc. Natl Acad. Sci. USA* **95**, 13 755–13 758. (doi:10.1073/pnas.95.23.13755)
- Nowak, M. A. & Sigmund, K. 1998a Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577. (doi:10.1038/31225)
- Nowak, M. A. & Sigmund, K. 1998b The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574. (doi:10.1006/jtbi.1998.0775)
- Nowak, M. A. & Sigmund, K. 2005 Evolution of indirect reciprocity. *Nature* **427**, 1291–1298. (doi:10.1038/nature04131)
- Oda, R. 1997 Biased face recognition in the Prisoner's Dilemma game. *Evol. Hum. Behav.* **18**, 309–315. (doi:10.1016/S1090-5138(97)00014-7)
- Ohtsuki, H. 2004 Reactive strategies in indirect reciprocity. *J. Theor. Biol.* **227**, 299–314. (doi:10.1016/j.jtbi.2003.11.008)
- Ohtsuki, H. & Iwasa, Y. 2004 How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120. (doi:10.1016/j.jtbi.2004.06.005)
- Ohtsuki, H. & Iwasa, Y. 2007 Global analysis of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *J. Theor. Biol.* **244**, 518–531. (doi:10.1016/j.jtbi.2006.08.018)
- 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)
- Resnick, P. & Zeckhauser, R. 2002 Trust among strangers in internet transactions: empirical analysis of eBay's reputation system. In *Advances in applied microeconomics: the economics of the internet and e-commerce* (ed. M. R. Baye), pp. 127–157. Amsterdam, The Netherlands: Elsevier Science.
- Resnick, P., Zeckhauser, R., Swanson, J. & Lockwood, K. 2006 The value of reputation on eBay: a controlled experiment. *Exp. Econ.* **9**, 79–101. (doi:10.1007/s10683-006-4309-2)
- Roberts, G. 2008 Evolution of direct and indirect reciprocity. *Proc. R. Soc. B* **275**, 173–179. (doi:10.1098/rspb.2007.1134)
- Rockenbach, B. & Milinski, M. 2006 The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718–723. (doi:10.1038/nature05229)
- Sommerfeld, R. D., Krambeck, H.-J., Semmann, D. & Milinski, M. 2007 Gossip as an alternative for direct observation in games of indirect reciprocity. *Proc. Natl Acad. Sci. USA* **104**, 17 435–17 440. (doi:10.1073/pnas.0704598104) 0704598104
- Sugden, R. 1986 *The economics of rights, co-operation and welfare*. Oxford, UK: Blackwell.
- Taylor, C. & Nowak, M. A. 2007 Transforming the dilemma. *Evolution* **61**, 2281–2292. (doi:10.1111/j.1558-5646.2007.00196.x)
- Trivers, R. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
- Vanneste, S., Verplaetse, J., Van Hiel, A. & Braeckman, J. 2007 Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evol. Hum. Behav.* **28**, 272–276. (doi:10.1016/j.evolhumbehav.2007.02.005)
- Wedekind, C. & Braithwaite, V. A. 2002 The long-term benefits of human generosity in indirect reciprocity. *Curr. Biol.* **12**, 1012–1015. (doi:10.1016/S0960-9822(02)00890-4)
- Wedekind, C. & Milinski, M. 2000 Cooperation through image scoring in humans. *Science* **288**, 850–852. (doi:10.1126/science.288.5467.850)