REVIEWS

Evolution of indirect reciprocity

Martin A. Nowak[1] & Karl Sigmund[2,3]

Natural selection is conventionally assumed to favour the strong and selfish who maximize their own resources at the expense of others. But many biological systems, and especially human societies, are organized around altruistic, cooperative interactions. How can natural selection promote unselfish behaviour? Various mechanisms have been proposed, and a rich analysis of indirect reciprocity has recently emerged: I help you and somebody else helps me. The evolution of cooperation by indirect reciprocity leads to reputation building, morality judgement and complex social interactions with ever-increasing cognitive demands.

Humans are the champions of reciprocity. Experiments and everyday experience alike show that what Adam Smith called 'our instinct to trade, barter and truck' relies to a considerable extent on the widespread tendency to return helpful and harmful acts in kind. We do so even if these acts have been directed not to us but to others. This has been analysed under the headings of 'third party altruism'[1] or 'indirect reciprocity'[2], and has led to a considerable amount of experimental and theoretical investigation over the past few years.

Direct reciprocity is captured in the principle: 'You scratch my back, and I'll scratch yours'. But it is harder to make sense of the principle 'You scratch my back and I'll scratch someone else's'[3] or 'I scratch your back and someone else will scratch mine' (Fig. 1). Why should this work? Presumably, I will not get my back scratched if it becomes known that I never scratch anybody else's. Indirect reciprocity, in this view, is based on reputation. But why should anyone care about what I did to a third party?

There are two approaches converging on this issue. One is rooted in social science, the other in evolutionary biology.

The main reason why economists and social scientists are interested in indirect reciprocity is that one-shot interactions between anonymous partners in a global market become increasingly frequent and tend to replace the traditional long-lasting associations and exchanges based on repeated give and take between relatives, neighbours, or members of the same village. A substantial part of our life is spent in the company of strangers[4], and many transactions are no longer face-to-face. The growth of web-based auctions and other forms of e-commerce is built, to a considerable degree, on reputation and trust[5–10]. The possibility to exploit such trust raises what economists call moral hazards. How effective is reputation, especially if information is only partial?

In contrast, evolutionary biologists are interested in the emergence of human societies, which constitutes the last (up to now) of the major transitions in evolution[11]. Unlike other eusocial species, such as bees, ants or termites, humans display a large amount of cooperation between non-relatives[12–14]. A considerable part of human cooperation is based on moralistic emotions—for instance, anger directed towards cheats, or the proverbial 'warm inner glow' felt after performing an altruistic action. Neuro-economic experiments relate these emotions to physiological processes[15–17]. Intriguingly, humans not only feel strongly about interactions that involve them directly, they also judge the actions between third parties, as demonstrated by the contents of gossip[18]. Indirect reciprocity is therefore likely to be connected with the origins of moral norms. Such norms are evidently to a large extent culture-specific, but the capacity for moral norms seems to be a human universal for which there is little evidence in other species[19].

Because the recent rapid advance of experimental investigations of indirect reciprocity was in large part driven by theory, we shall discuss the modelling approaches before reviewing the experiments. But first we note, in a wider context, that indirect reciprocity seems to require a 'theory of mind'[20]. Whereas altruism directed towards kin works because similar genomes reside in different organisms, reciprocal altruism recognizes that similar minds emerge from different brains. It is easy to conceive that an organism experiences as 'good' or 'bad' anything that affects the organism's own reproductive fitness in a positive or negative sense. The step from there to judging, as 'good' or 'bad', actions between third parties, is not obvious. The same terms 'good' and 'bad' that are applied to pleasure and pain are also used for moral judgements: this linguistic quirk reveals an astonishing degree of empathy, and reflects highly developed faculties for cognition and abstraction.

This review of theoretical and empirical studies of indirect reciprocity stresses the importance of monitoring not only partners in continuing interactions but also all individuals within the social network. Indirect reciprocity requires information storage and transfer as well as strategic thinking and has a pivotal role in the evolution of collaboration and communication. The possibilities for games of manipulation, coalition-building and betrayal are limitless. Indirect reciprocity may have provided the selective challenge driving the cerebral expansion in human evolution.

Direct versus indirect reciprocity

In the terminology based on Hamilton, Trivers and Wilson[12–14], an act is said to be altruistic if it is costly to perform but confers a benefit on another individual. In evolutionary biology, costs and benefits are measured in darwinian fitness, which means reproductive success. In other contexts, other utility scales such as monetary rewards may be more appropriate. Reciprocal altruism in its original, 'direct' sense is defined as an exchange of altruistic acts between the same two individuals so that, in total, both obtain a net benefit[1]. In the simplest model, the altruistic act consists in conferring a benefit b on the recipient at a cost c to the donor. We shall always assume that the cost is smaller than the benefit, so that if the act is returned, both

[1]Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Faculty for Mathematics, University of Vienna, A-1090 Vienna, Austria. [3]IIASA, A-2631, Laxenburg, Austria.

individuals experience a gain. The payoff structure yields an instance of the familiar Prisoner's Dilemma game[21]. If both players cooperate, each receives $b - c$, which is better than what they would obtain by both defecting, namely 0. But a unilateral defector would earn b, which is the highest payoff, and the exploited cooperator would pay the cost c without receiving any benefit. The payoff-maximizing move is defecting.

This changes if the game is repeated for several rounds. For simplicity we shall assume that in each round both players decide simultaneously. We could also assume that they alternate, which leads to a slightly different game[22–24]. The so-called folk theorem on repeated games implies that if the probability for future rounds is sufficiently high, cooperation can be sustained by so-called trigger strategies, which switch to relentless defection as soon as the co-player defects once[25,26,87]. A rational player must weigh the benefit of exploiting the co-player in one round against the cost of forfeiting collaboration in all future rounds, and would therefore abstain from defection.

In the context of indirect reciprocity, any two players are supposed to interact at most once with each other. Each player can experience many rounds, but never with the same partner twice. Thus it is not possible that a cheat is held to account by the victim. (In a variant of this model, two players could interact on several occasions, one always as the donor, the other as recipient, so that a return is again excluded.) Clearly, trigger strategies can still ensure a cooperative Nash equilibrium, such that if all players use them, no player would have an incentive to deviate. In strategic thinking, only the payoffs matter, not by whom they are provided. In this sense, the step from direct to indirect reciprocity corresponds simply to the step from personal enforcement to community enforcement[27–30]. However, a trigger strategy prescribing each person to cooperate until the first defection is personally experienced, and thenceforth to defect, hurts the original wrong-doer only after many rounds. A strategy triggered by the first defection in the population leaves cooperation at the mercy of the first wrong move. In both cases many innocents would be punished, and errors would cause havoc. Obviously, retaliation should be directed towards the cheat rather than towards the whole community. This requires more detailed information. Game theory shows that even if information is transmitted only locally and errors occur occasionally, cooperation can be sustained: there exist strategies such that no rational player has an interest in deviating unilaterally[28].

In evolutionary game theory it is not assumed that players are rational but only that successful strategies spread—by being inherited, for instance, or copied through imitation or learning[31]. For direct reciprocity, game theoretical analysis and individual-based simulations have shown that a population of defectors can be invaded by a small cluster of retaliators[32] or even by a single retaliator[33]. Typically, one considers a well-mixed population in which individuals meet randomly and play a series of Prisoner's Dilemma games with each other. What counts is the total payoff. Retaliators compensate for the loss of being exploited by a defector in the first round with long sequences of altruistic exchange with other retaliators. Once cooperation is established, a complex evolution takes place, which depends on the size of the population, the cost-to-benefit ratio, the average number of rounds and the probability of errors[32,34,35].

A similar model of indirect reciprocity assumes that, within a well-mixed population, individuals meet randomly, one in the role of the potential donor and the other as a potential recipient (Fig. 2). Each individual experiences several rounds of this interaction in both roles, but never with the same partner twice. Again, all that counts is the total payoff. A player can follow either an unconditional strategy, such as always to cooperate or always to defect, or else a conditional strategy, which discriminates between the potential recipients on the basis of past interactions. In a simple example, a discriminating player can help the co-player if that co-player's score exceeds a certain threshold. A player's score is 0 at birth, increases whenever that player helps and decreases whenever the player withholds help. Individual-based simulations show that if the cost-to-benefit ratio is sufficiently low, and the amount of information about the co-player's past sufficiently high, cooperation based on discrimination can emerge.

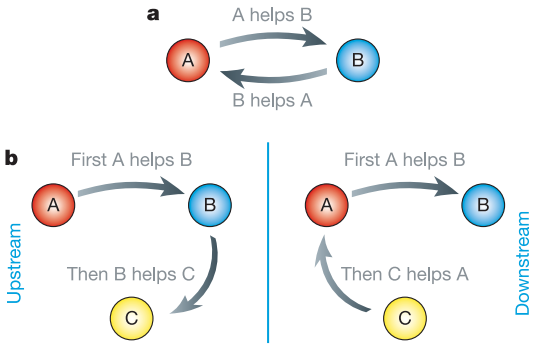


Figure 1 | Direct and indirect reciprocity. **a**, Direct reciprocity means that A helps B and B helps A. **b**, Indirect reciprocity comes in two flavours. 'Upstream reciprocity' (left) is based on a recent positive experience. A person who has been at the receiving end of a donation may feel motivated to donate in turn. Individual B, who has just received help from A, goes on to help C. 'Downstream reciprocity' (right) is built on reputation. Individual A has helped B and therefore receives help from C. Mathematical investigations of indirect reciprocity have shown that natural selection can favour strategies that help others based on their reputation. Upstream reciprocity is harder to understand[2,56,77,78] but is observed in economic experiments. In both cases, the decision to help can be interpreted as a misdirected act of gratitude. In one case recipients are thanked for what another did; in the other case they are thanked by someone who did not profit by what they did.

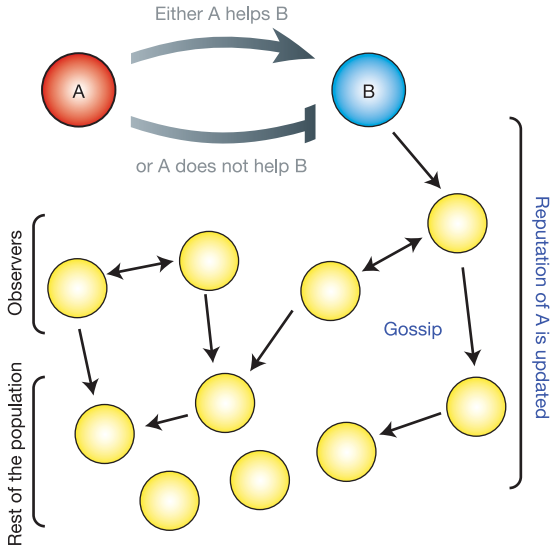


Figure 2 | Building a reputation. In a natural extension of the basic model of indirect reciprocity, an action between donor A and recipient B is observed by a subset of the population[36]. The observers, the donor and the recipient can inform others. People could pass on what has happened (the action) or their assessment of the action. There are many possibilities of error: the action or the intention of the donor can be interpreted differently by different people; some individuals may receive conflicting information from different sources; some individuals may not receive any information at all; people can have different assessment modules. The reputation of a person is therefore not simply a label that is visible to all others, but instead each person has a private list of the reputation of others. Although language could help to synchronize these lists[42], ultimately reputation is in the eyes of the beholder.

In the resulting population, help is channelled towards those who have helped[36–38].

Two features of this model were immediately apparent[36,39,40]. One is the paradoxical nature of the discriminating strategy. In terms of rational game theory, why should players care about the scores of others rather than just about their own payoff, and why should they decrease their own score (and thus their likelihood of receiving help on later occasions) by withholding help from low-scorers? Lower scores can cause lower payoffs. The second issue concerns the lack of stability of the cooperative outcome. The simulations display occasional bursts of defection, which are based on a previous build-up of undiscriminating altruists. In a population of discriminators, unconditional cooperators can increase by random drift and eventually invite the invasion of defectors (Fig. 3).

The two issues are closely related. Considered intuitively, the intentions of players who selfishly care about their own income only, and who accordingly give help just to keep their own score high, are vastly different from the intentions of altruists who have only the interests of their co-players in mind and help them on every occasion. But the effect is the same in both cases: support will be given regardless of the co-players' contributions.

Binary assessment, or the world in black and white

To analyse these questions, an even simpler model was proposed, based on a binary score, taking only the values 'good' or 'bad', depending on what the co-player did when last observed[40–42] (Box 1). This can be viewed as a basic system of moral assessment. In its simplest form (first order), the assessment depends only on the action of the observed player, which means it depends on whether that player gave help or not. A discriminating donor using this assessment rule refuses help to a 'bad' recipient, and therefore becomes 'bad', which reduces the chance of being helped in turn. (With a wider range of score values, a single refusal to help has less impact on the reputation.) Effectively, discriminating players pay a cost for punishing bad co-players. Such a form of altruistic punishment can promote cooperation in the community, but at a cost to the punisher, and thus can be viewed as a social dilemma[43]. The fact that costly punishment has an undisputed role in other contexts, such as public goods games, ultimatum games and trust games[44–47], shows that this form of discrimination is plausible.

A more sophisticated assessment rule should distinguish between justified and unjustified defection and should therefore take into account the score of the receiver: someone withholding help from a

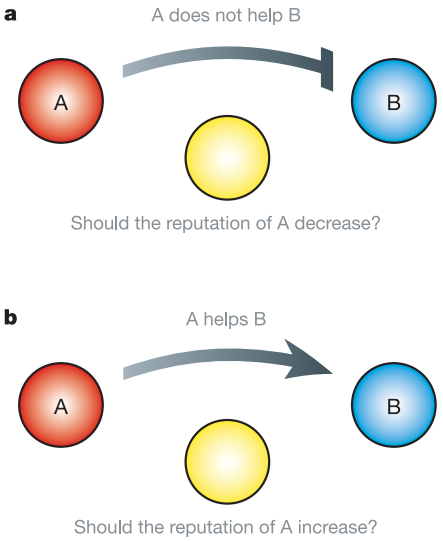


Figure 3 | Two problems with indirect reciprocity. B has defected in previous rounds and therefore has a low reputation. **a**, If A does not help B, so as to punish B for previous defections, then why should the reputation of A be reduced? **b**, If A does help B, although B is a defector, then why should the reputation of A increase? Helping defectors destabilizes cooperation. Strategies (assessment rules) of indirect reciprocity that try to fix these problems are cognitively demanding and vulnerable to deception. Before defecting, A could try to signal that B has a low reputation. We argue that the intricate complexity of indirect reciprocity provided the selective mould for human language and human intelligence.

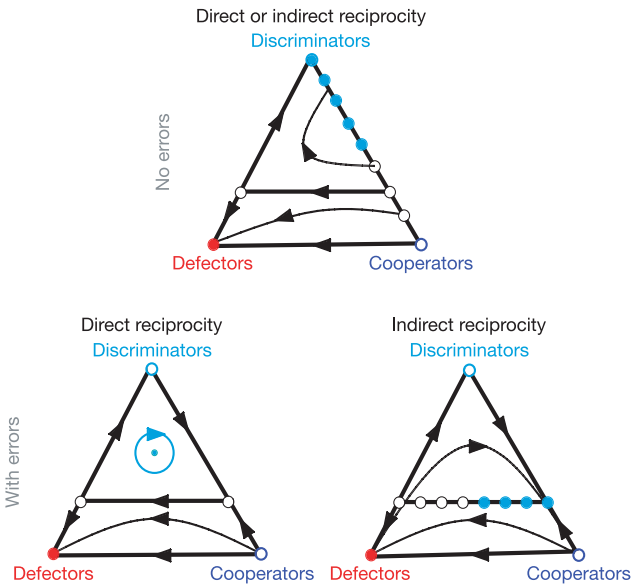
Box 1 | The good, the bad and the discriminating

Similarities and differences between direct and indirect reciprocity become apparent when studying the replicator dynamics of three strategies: always cooperate, always defect and the simplest discriminating strategy. For direct reciprocity, this is Tit For Tat, which helps in the first round and then does whatever the opponent did in the previous round. For indirect reciprocity, this is the strategy that prescribes helping unless the recipient is known to have refused to help in the previous round. In each case, there are many other discriminating strategies, which are likely to take over eventually. This analysis is just a first step.

In the absence of discriminators, defectors win against cooperators. In the absence of cooperators, defectors and discriminators form a bistable system: depending on the initial condition, either one or the other strategy wins. In the absence of defectors, discriminators and cooperators are in equilibrium. Random fluctuations, however, make their frequencies drift up and down. The equilibria can be invaded by defectors if the frequency of discriminators is below a certain threshold. With all three strategies present, the dynamics lead either to defectors only or to a mixture of the two kinds of altruist.

If errors occur, or if an intended donation cannot be implemented through a lack of resources, discriminating and undiscriminating altruists reach, in the absence of defectors, a stable coexistence with a well-defined frequency of discriminators. If all three types are present in the population, the system displays two types of behaviour. If the frequency of discriminators is too low, defectors win. If the discriminators are sufficiently frequent, all three strategies coexist. However, in direct reciprocity the frequencies oscillate periodically, whereas in indirect reciprocity they converge to an equilibrium. In each case, a long series of random fluctuations may eventually destroy the coexistence of the three strategies.

In the deterministic model, only the emergence of other conditional strategies can save cooperation in the long run[79]. For stochastic population dynamics, the time average of the evolutionary oscillations can be centred on discriminators[80].



Box 1 Figure 1 | Basic evolutionary dynamics of direct and indirect reciprocity.

'bad' player should not pay with a reduced score[36,39,40,48]. There are many assessment rules of second order, which also depend on the score of the receiver, and of third order, which depend additionally on the score of the donor (Box 2). Only eight of them lead to cooperation and are at the same time evolutionarily stable: a homogeneous population of players using such a strategy cannot be invaded by players using other strategies[42]. All eight of these strategies distinguish justified from unjustified defections. This property must hold for any strategy that maintains cooperation, eliminates cheats and can overcome errors[49]. However, the problem with the concept of justified defection is that it requires information not only about the past of the co-player but also about the past of the co-player's co-players, and their co-players, and so on[36,40].

Again, the stability of cooperation is threatened by unconditional altruists who merely wish to keep their own good score. In a population consisting entirely of discriminators, undiscriminating cooperators fare just as well and thus spread by neutral drift. If their frequency exceeds a certain threshold, defectors can invade and take over. The situation is altered in a significant way if players occasionally fail to cooperate even though their strategy prescribes cooperation[40,50–53]. This can be due to errors in implementation or lack of resources. It is plausible that when recipients need help, donors are also short of means[53]. In this situation, a population consisting only of conditional and unconditional altruists is not subject to random drift, but selection leads to a well-defined, stable mix of the two types. This mixture can be vulnerable to the invasion of defectors[40,41].

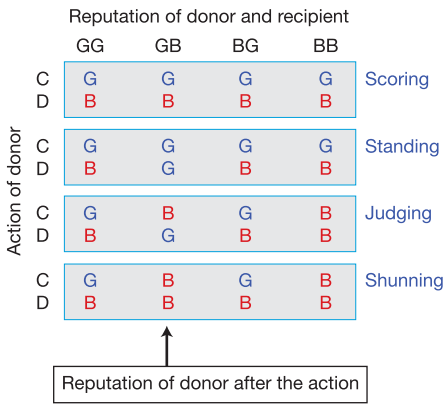
There are several ways out of this impasse. Various assumptions on the distribution of the number of rounds lead to a bistable system in which, depending on the initial state, the population converges either to the fixation of defectors or to a stable mix of altruists that cannot be invaded by defectors[53,54]. In this case, the very fact that individuals are not perfect and sometimes defect involuntarily promotes the stability of cooperation[50].

Punishing a player with a bad score creates another player with a bad score; but this 'passing the buck along' can stop in two ways, by encountering either a discriminator who is uninformed or an undiscriminating altruist. Therefore, both the lack of information and the prevalence of unconditional altruists may, surprisingly, stabilize cooperation. A stable mix of discriminating and undiscriminating altruists that cannot be invaded by defectors is also obtained by assuming that, as players grow older, their social network grows and so does their information about their co-players' past[55,56]. Alternatively, if the discriminating strategy distinguishes between justified and unjustified defection, the population can converge to discriminators only, which cannot be invaded by unconditional strategies[40].

Can discrimination based on the concept of 'justified defection' be destabilized by errors in perception? Not if players have the same reputation in the eyes of all members of their population. If such a

Box 2 | Let a hundred morals bloom

In a world of binary moral judgements there are four ways of assessing donors in terms of 'first-order assessment': always consider them as good, always consider them as bad, consider them as good if they refuse to give, or consider them as good if they give. Only this last option makes sense. Second-order assessment also depends on the score of the receiver; for example, it can be deemed good to refuse help to a bad person. There are 16 second-order rules. Third-order assessment also depends on the score of the donor; for example, a good person refusing to help a bad person may remain good, but a bad person refusing to help a bad person remains bad. There are 256 third-order assessment rules. We display four of them in Box 2 Fig. 1. With Scoring, cooperation, C, always leads to a good reputation, G, whereas defection, D, always leads to a bad reputation, B. Standing is like Scoring, but it is not bad if a good donor defects against a bad recipient. With Judging, in addition, it is bad to cooperate with a bad recipient. For another assessment rule, Shunning, all donors who meet a bad recipient become bad, regardless of what action they choose. Shunning strikes us as grossly unfair, but it emerges as the winner in a computer tournament if errors in perception are included and if there are only a few rounds in the game[57].



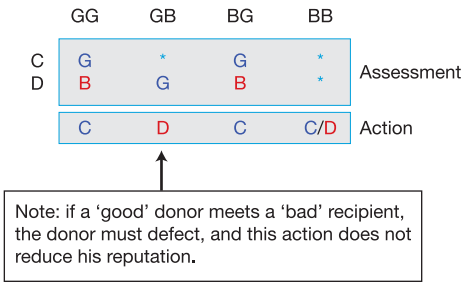
Box 2 Figure 1 | Four assessment rules.

An action rule for indirect reciprocity prescribes giving or not giving, depending on the scores of both donor and recipient. For example, you may decide to help if the recipient's score is good or your own score is bad. Such an action might increase your own score and therefore increase the chance of receiving help in the future. There are 16 action rules.

If we view a strategy as the combination of an action rule and an assessment rule, we obtain 4,096 strategies. In a remarkable calculation, Ohtsuki & Iwasa[42,49] analysed all 4,096 strategies and proved that only eight of them (the 'leading eight'; Box 2 Fig. 2) are evolutionarily stable under certain conditions and lead to cooperation.

The three asterisks in the assessment module of the 'leading eight' indicate a free choice between G and B. There are therefore $2^3 = 8$ different assessment rules. The action module is built as follows: if the column in the assessment module is G and B, then the corresponding action is C, otherwise the action is D.

Both Standing and Judging belong to the leading eight, but neither Scoring nor Shunning do. However, we expect that Scoring has a similar role in indirect reciprocity to that of Tit For Tat in direct reciprocity. Neither strategy is evolutionarily stable, but their ability to catalyse cooperation in adverse situations and their simplicity constitute their strength. In extended versions of indirect reciprocity, in which donors can sometimes deceive others about the reputation of the recipient, Scoring is the 'foolproof' concept of 'I believe what I see'. Scoring judges the action and ignores the stories.



Box 2 Figure 2 | Ohtsuki & Iwasa's 'leading eight'.

consensual assessment can be achieved, the corresponding strategy is robust[39,40,42]. But if players have different views about the reputation of others, then errors in perception can undermine cooperation[57]. Such private lists of the scores of co-players are very plausible if individual interactions are observed by only a fraction of the population[36]. Gossip might be a way of achieving consensus, but it can also be used for spreading unfounded rumours and manipulating co-players. The co-evolution of human language and cooperation by indirect reciprocity is a fascinating and as yet unexplored topic.

Another debated issue concerns the underlying population structure[39]. Analytical models are often based on the idealization of a very large, well-mixed population. Individual-based simulations typically assume population sizes of 50–100 individuals, on the basis of estimates for the group size of hunter–gatherers. In such small populations, random drift can strongly affect evolution. Populations consisting of many separate groups with a modicum of exchange between them have also been modelled[39,54]. However, such a population structure facilitates the evolution of altruism through group selection[57]. It could be that cooperators fare less well than defectors within each group but that groups of cooperators fare better than groups of defectors[58]. In extreme cases this leads to cooperation even in the absence of indirect reciprocity. Although it is most interesting to study the interaction between group selection and indirect reciprocity, it is equally important not to confuse the two effects.

Expensive games and economic experiments

The basic experimental set-up for testing indirect reciprocity studies a group of players equipped with an initial monetary endowment. Each player is repeatedly given the opportunity of donating money to a specific co-player, thus increasing the account of this person by an amount b. Players know that if they choose to do so, an amount c will be deducted from their own account—the cost for providing the gift. To eliminate confounding effects, players usually do not interact face-to-face but are given some information about the past actions of their potential recipients. They know that their recipients will never be their donors on future occasions and therefore that there is no scope for direct reciprocity. The interactions both with the co-players and with the experimenters are kept as anonymous as possible, usually under double-blind conditions. Many parameters can be varied within this basic set-up, for instance the cost-to-benefit ratio, the size of the starting account, the number of interactions, the size of the group, the degree of information about the co-players' behaviour, the length of the game or the social backgrounds of the players (Box 3).

From the first experiments onwards, it was clear that a substantial proportion of the players frequently decide to donate. The propensity for indirect reciprocation is apparently widespread. As expected, donations occur more frequently if the cost-to-benefit ratio is lower or the starting account is higher. Reputation has a considerable influence on the decisions. In particular, the image score of potential recipients correlates well with their expectation to actually receive money[59]. Players who donate less often display a higher degree of discrimination. Players who are more open-handed care less about the recipient's score. Conversely, if players know that their own score is passed on, they are much more likely to donate than otherwise[60]. Many players donate even when they are assured of complete anonymity, possibly because they are not fully convinced. Recent experiments suggest that a nagging suspicion remains: 'What if someone is watching?' Intriguingly, even stylized eyespots suffice to influence the giving behaviour[61].

The hypothesis that more information leads to more cooperation has been confirmed in experiments, which compare three information conditions[62]. In one condition, players have no information about their co-players; in the second they are told about what their co-players have decided when last in the role of a donor; and in the third they also know about the score of the recipient of the co-player. We note that this is not always enough to decide whether a previous defection was justified or not. However, the additional knowledge did enhance cooperation[62].

In this series of experiments there is a significant positive correlation between the number of gifts given and received, but a slightly negative correlation between the number of gifts given and the total payoff obtained. In another experiment[63], however, those who give

Box 3 | Games of cooperation

In the Trust Game there are two players, one in the role of the donor, the other in the role of the responder. The donor can transfer some money to the responder. Upon arrival, the amount is multiplied by three. The responder, then, has the possibility of sending some of it back to the donor. A responder with an income-maximizing strategy should send nothing back. Any donor expecting this should therefore transfer nothing. In real experiments, many donors transfer substantial amounts, and some obtain large returns, so that both players win.

In the Public Goods Game, each of N players can independently decide to transfer some money to a common pool, where it is multiplied by some factor r (smaller than N) and then divided equally between all players irrespective of whether they have contributed or not. Because each player receives, in return for his or her own contribution, only the fraction r/N, the income-maximizing strategy is to contribute nothing. However, in real experiments many players contribute. If all do, they multiply their endowment by the factor r.

The Public Goods Game for $N = 2$ players has the structure of a Prisoner's Dilemma. Two players who cooperate earn more than two players who defect; but a defector cheating on a cooperator earns the highest payoff, and the exploited cooperator earns the lowest. If two players, in a trust game, are simultaneously in the role of the donor, and then simultaneously in the role of the responder, they play two rounds of a Prisoner's Dilemma.

Experimental economists and experimental psychologists have studied these games, and diverse variations, intensively[81,82].

Box 4 | Bidding for trust

Trust, 'a lubricant of social life'[83], is essential in many types of economic transaction and is also linked to physiological processes[84]. In the Trust Game, donors who trust their responder will expect to gain from transferring money. In contrast, donors in the indirect reciprocity game know that they can expect no direct return, even if their recipient is trustworthy. All they can gain from the transfer is an increased reputation for altruism and trustworthiness.

Game theory shows that cooperation can be sustained in the indirect reciprocity game if each player carries a label[28]. The strategy prescribes that players who deviate from it have to be punished (by not being helped) for a number of rounds T. A player's label specifies for how many rounds that player has to be punished. If a donor and a receiver meet, the action prescribed by the strategy depends only on their labels, and their labels will be updated depending on the donor's action. No player has an incentive to deviate if all other players adopt this strategy, and the effect of an error will be overcome after T rounds. However, settling on this strategy, for instance on the specific number T, seems to require an institution able to guarantee honest labelling.

Subscribers to eBay auctions are asked to state, after every transaction, whether they were satisfied with their partner or not. Their partner's score can accordingly increase or decrease by one point. The ratings of eBay members, accumulated over 12 months, are public knowledge. This very crude form of assessment seems to suffice for the purpose of reputation-building and seems to be reasonably proof against manipulation. Social history knows many other instances of public scorekeeping: 'Societies have sewn scarlet letters to people's garments, shaved heads, cut off fingers and given medals to signal to strangers some aspect of an individual's past deeds or misdeeds'[30]. Reputation mechanisms were also important in the emergence of medieval trade[85].

often end up with the highest payoff, so that there is a strategic advantage to generosity. The discrepancy could be due to the larger number of rounds—the advantage showed up only after a dozen rounds—or to the fact that players were informed not about the recipients' last move only but about their whole history of giving.

Evidence for strategic reputation building is found in many experiments. Donations are more frequent in earlier rounds, when the player's own reputation has a higher impact on the future income. But several experiments show that even players who know that their score will not be communicated show generous behaviour[62,64]. Such donors cannot be motivated by selfish interests. For many players, however, the propensity to donate more than doubles if they know that their action will be communicated and will therefore affect their own score. The influence of the recipient's score decreases accordingly. Often there is evidence for a dual motivation—players give donations if the recipient's score is high or their own score is low[36].

Experiments investigating whether a player's justified defections lowers his or her chance to receive subsequent donations indicate that cognitive problems challenge the donor[65]. Players faced with a full history of all previous rounds take in general a longer time to reach their decision, suggesting that they attempt to take into account not only their recipient's last move but also that of their recipient's recipient. Nonetheless, the statistics of such games with full information look surprisingly similar to those obtained when players know only the score of the recipient. Moreover, players who justifiably refuse to donate to a defector show an increased tendency to provide donations in the following round, as if to make up for that refusal. This indicates that they expect their refusal to lower their score in the co-players' eyes and that they do not rely on the community's understanding.

Many experiments have shown that players who have just received a donation are more likely to give a donation in turn. There is evidence for 'upstream' indirect reciprocity in cyclical networks: as expected, short loops and a high benefit-to-cost ratio favour cooperation[66]. A variant of the Trust Game has two donor–responder pairs, but such that the transfers are criss-cross: the responder of one donor can return money only to the other donor (and does not know the amount transferred by that donor). The return rates turn out to be no lower than if they were addressed to the player's own donor[67,68]. In experimental situations that are not based on rigid networks, the decisions of donors also tend to mirror their own recent experience[59,64]. People are nicer to others if third parties have been nicer to them. More generally, it seems that decisions often depend on both the donor's payoff and the recipient's score, but such strategies have not been analysed so far.

Box 5 | Social viscosity

Altruism towards genetic relatives can evolve by kin selection provided that Hamilton's rule[14] holds: the coefficient of relatedness, R, between the donor and the recipient has to exceed the cost-to-benefit ratio of the altruistic act:

$$R > c/b.$$

As Haldane has said, 'I will jump into the river to save two brothers or eight cousins.' The probability that brothers share a 'selfish gene' is 1/2; the same probability for cousins is 1/8. Kin selection works in 'viscous' populations in which chances are high that neighbours are genetic relatives.

For indirect reciprocity a similar rule holds[36,86]: the probability, q, of knowing the social score of another person must exceed the cost-to-benefit ratio:

$$q > c/b.$$

The role of genetic relatedness that is crucial for kin selection is replaced by social acquaintanceship. In a fluid population, in which most interactions are anonymous and people have no possibility of monitoring the social score of others, indirect reciprocity has no chance. In a socially viscous population, in which people know each other's reputation, cooperation by indirect reciprocity can thrive.

The strategic links of indirect reciprocity

Indirect reciprocity is situated somewhere between direct reciprocity and public goods. On the one hand it is a game between two players only, the donor and the recipient, but on the other hand it has to be played within a larger group.

Richard Alexander claimed that indirect reciprocity originates from direct reciprocity in the presence of interested audiences[2]. A good strategy for the latter is Observer Tit For Tat[36,69]. Players using this strategy for the repeated Prisoner's Dilemma are following Tit For Tat, except that in the first round they defect if they know that their co-player, in a previous repeated game against another player, has defected. Observer Tit For Tat relies on reputation in the first round and on personal experience in all further rounds against the same co-player. Conversely, experiments show that if several rounds of a Prisoner's Dilemma are appended to an indirect reciprocity game, the display of the previous score increases a player's probability of cooperating with generous players in the first few rounds. After a couple of rounds the personal experience obtained with the given co-player becomes more decisive[63].

The widespread tendency to judge actions between third parties, and the readiness for cooperation combined with altruistic punishment (also known as strong reciprocity[70]) has been neatly captured in an experiment involving three players[71,72]. First, players A and B engage in one round of a Prisoner's Dilemma game. Then player C has the possibility to mete out costly punishment on A and B. Defectors are often punished, although this reduces the endowment of the punisher. It would be interesting to see whether observers of a repeated Prisoner's Dilemma game are using a similar level of punishment, or whether they reduce it because wronged players have the opportunity of avenging themselves.

Indirect reciprocity and public goods games are also closely connected. For example, donors are more generous if they learn that the recipient has recently made a donation to a charitable institution[73]. An even more remarkable effect was found in an experiment alternating rounds of the Public Goods Game with rounds of the indirect reciprocity game[74]. It is known that many players show an initial willingness to contribute to the public good a substantial amount of their endowment, but this willingness often vanishes within a few rounds. This is not the case if indirect reciprocity games are sandwiched between the rounds of the Public Goods Game. If players are informed about their recipient's action in the Public Goods Game, they tend to be more generous towards recipients who contributed much. Conversely, players are more willing to contribute to the public good if they know that this will be communicated before the start of the indirect reciprocity game. The contributions to the public good do not deteriorate from one round to the next. The donations in the indirect reciprocity game, which are channelled towards those who contributed much to the public good, can be viewed as rewards.

Whereas most experiments, in indirect reciprocity, were motivated by models, this last experiment led to a model[75,76]. A discriminating strategy, which defects in all rounds of the indirect reciprocity game if the recipient is known to have defected in the Public Goods Game, can guarantee stable cooperation. Because this discriminating strategy distinguishes between justified and unjustified defection, it is effectively a non-costly form of punishing free riders in the Public Goods Game.

Future directions

Thus indirect reciprocity based on reputation serves as a link between diverse forms of cooperative interaction. The moralistic assessment of the other members in the population, even if they are observed only at a distance, provides a powerful tool for channelling support towards those who collaborate, and an incentive to join group efforts.

The exploration of the links between indirect reciprocity and other game theoretical models for cooperative interaction promises to offer further opportunities for a better understanding of human traits (Boxes 4 and 5). Future theoretical and experimental work on indirect reciprocity is likely to go beyond the context of economic interactions in the narrow sense and to address such issues as the physiological correlate of trust and decision-making, the emergence of language capabilities and moral norms, subliminal effects framing our beliefs, and the pervasive role of individual reputations and social prejudice.

1. Trivers, R. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46,** 35–57 (1971).
2. Alexander, R. D. *The Biology of Moral Systems* (Aldine de Gruyter, New York, 1987).
3. Binmore, K. G. *Game Theory and the Social Contract* (MIT Press, Cambridge, Massachusetts, 1994).
4. Seabright, P. *The Company of Strangers: A Natural History of Economic Life* (Princeton Univ. Press, Princeton, 2004).
5. Bolton, G. E., Katok, E. & Ockenfels, A. How effective are electronic reputation mechanisms? An experimental investigation. *Manage. Sci.* **50,** 1587–1602 (2004).
6. Keser, C. Trust and reputation building in e-commerce. IBM Watson Research Center, CIRANO working paper no. 2002s-75 ⟨http://www.cirano.qc.ca/en/publication_detail.php?id = 2002s-75⟩ (2002).
7. Dellarocas, C. Sanctioning reputation mechanisms in online trading environments with moral hazard. MIT Sloan School of Management working paper no. 4297–03 ⟨http://ssrn.com/abstract = 393043⟩ (2003).
8. Bolton, G., Katok, E. & Ockenfels, A. in *Applications of Supply Chain Management and e-Commerce Research* (eds Geunes, J., Akcali, E., Pardalos, P. M., Romeijn, H. E. & Shen, Z.-J.) 195–216 (Springer, Dordrecht, 2005).
9. Resnick, P., Zeckhauser, R., Friedman, E. & Kuwabara, K. Reputation systems. *Commun. ACM* **43,** 45–48 (2000).
10. Lucking-Reiley, D., Bryan, D., Prasad, N. & Reeves, D. *Pennies from eBay: The Determinants of Price in Online Auctions*. Vanderbilt Univ., Univ. Arizona working paper (1999).
11. Maynard Smith, J. & Szathmary, E. *The Major Transitions in Evolution* (Oxford Univ. Press, Oxford, 1997).
12. Wilson, E. O. *Sociobiology* (Harvard Univ. Press, Cambridge, Massachusetts, 1975).
13. Trivers, R. *Social Evolution* (Benjamin Cummings, Menlo Park, 1985).
14. Hamilton, W. D. *Narrow Roads of Gene Land* Vol. 1 (Freeman, New York, 1996).
15. Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. The neural basis of economic decision-making in the ultimatum game. *Science* **300,** 1755–1758 (2003).
16. Rilling, J. K. *et al.* A neural basis for social cooperation. *Neuron* **35,** 395–405 (2002).
17. Fehr, E. The neural basis of altruistic punishment. *Science* **305,** 1254–1258 (2004).
18. Dunbar, R. *Grooming, Gossip and the Evolution of Language* (Harvard Univ. Press, Cambridge, Massachusetts, 1996).
19. Brown, D. E. *Human Universals* (McGraw-Hill, New York, 1991).
20. Whiten, A. & Byrne, R. W. (eds) *Machiavellian Intelligence II: Extensions and Evaluations* (Cambridge Univ. Press, Cambridge, UK, 1997).
21. Axelrod, R. *The Evolution of Cooperation* (Reprinted by Penguin, Harmondsworth, 1989.) (Basic Books, New York, 1984).
22. Nowak, M. A. & Sigmund, K. The alternating prisoner's dilemma. *J. Theor. Biol.* **168,** 219–226 (1994).
23. Frean, M. R. The prisoner's dilemma without synchrony. *Proc. R. Soc. Lond. B* **257,** 75–79 (1994).
24. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity and social history. *Games Econ. Behav.* **10,** 122–142 (1995).
25. Fudenberg, D. & Maskin, E. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **50,** 533–554 (1986).
26. Binmore, K. *Fun and Games: A Text on Game Theory* (Heath, Lexington, Massachusetts, 1992).
27. Rosenthal, R. W. Sequences of games with varying opponents. *Econometrica* **47,** 1353–1366 (1979).
28. Kandori, M. Social norms and community enforcement. *Rev. Econ. Stud.* **59,** 63–80 (1992).
29. Ellison, G. Cooperation in the prisoner's dilemma with anonymous random matching. *Rev. Econ. Stud.* **61,** 567–588 (1994).
30. Okuno-Fujiwara, M. & Postlewaite, A. Social norms in matching games. *Games Econ. Behav.* **9,** 79–109 (1995).
31. Nowak, M. A. & Sigmund, K. Evolutionary dynamics of biological games. *Science* **303,** 793–799 (2004).
32. Nowak, M. A. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355,** 250–253 (1992).
33. Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428,** 646–650 (2004).
34. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211,** 1390–1396 (1981).
35. Nowak, M. A. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364,** 56–58 (1993).
36. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573–577 (1998).
37. Wedekind, C. Give and ye shall be recognized. *Science* **280,** 2070–2071 (1998).
38. Ferrière, R. Help and you shall be helped. *Nature* **393,** 517–519 (1998).
39. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocation. *Proc. R. Soc. Lond. B* **268,** 745–753 (2001).
40. Panchanathan, K. & Boyd, R. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224,** 115–126 (2003).
41. Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194,** 561–574 (1998).
42. Ohtsuki, H. & Iwasa, Y. How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231,** 107–120 (2004).
43. Dawes, R. M. Social dilemmas. *Annu. Rev. Psychol.* **31,** 169–193 (1980).
44. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90,** 980–994 (2000).
45. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415,** 137–140 (2002).
46. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425,** 785–791 (2003).
47. Sigmund, K., Hauert, C. & Nowak, M. A. Reward and punishment. *Proc. Natl Acad. Sci. USA* **98,** 10757–10762 (2001).
48. Sugden, R. *The Economics of Rights, Cooperation and Welfare* (Blackwell, Oxford, 1986).
49. Ohtsuki, H. & Iwasa, Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.*(in the press).
50. Lotem, A., Fishman, M. A. & Stone, L. Evolution of cooperation between individuals. *Nature* **400,** 226–227 (1999).
51. Lotem, A., Fishman, M. A. & Stone, L. Evolution of unconditional altruism through signalling benefits. *Proc. R. Soc. Lond. B* **270,** 199–205 (2002).
52. Fishman, M. A., Lotem, A. & Stone, L. Heterogeneity stabilises reciprocal altruism interaction. *J. Theor. Biol.* **209,** 87–95 (2001).
53. Fishman, M. A. Indirect reciprocity among imperfect individuals. *J. Theor. Biol.* **225,** 285–292 (2003).
54. Brandt, H. & Sigmund, K. The logic of reprobation: Assessment and action rules for indirect reciprocity. *J. Theor. Biol.* **231,** 475–486 (2004).
55. Mohtashemi, M. & Mui, L. Evolution of indirect reciprocity by social information: The role of trust and reputation in evolution of altruism. *J. Theor. Biol.* **223,** 523–531 (2003).
56. Brandt, H. & Sigmund, K. Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl Acad. Sci. USA* **102,** 2666–2670 (2005).
57. Takahashi, N. & Mashima, R. The emergence of indirect reciprocity: Is the standing strategy the answer? Hokkaido Univ. working paper no. 29 ⟨http://lynx.let.hokudai.ac.jp/COE21/pdf/029.zip⟩ (2003).
58. Wilson, D. S. A theory of group selection. *Proc. Natl Acad. Sci. USA* **72,** 143–146 (1975).
59. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288,** 850–852 (2000).
60. Seinen, I. & Schram, A. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *Eur. Econ. Rev.* (in the press).
61. Haley, K. J. & Kessler, D. M. T. Nobody is watching?: Subtle cues affect generosity in an anonymous economic game. *Evol. Hum. Behav.* **26,** 245–256 (2005).
62. Bolton, G. E., Katok, E. & Ockenfels, A. Cooperation among strangers with limited information about reputation. *J. Public Econ.* **89,** 1457–1468 (2005).
63. Wedekind, C. & Braithwaite, V. A. The long-term benefits of human generosity in indirect reciprocity. *Curr. Biol.* **12,** 1012–1015 (2002).
64. Engelmann, D. & Fischbacher, U. Indirect reciprocity and strategic reputation building in an experimental helping game. Univ. Zürich working paper no. 132 ⟨http://www.iew.unizh.ch/wp/iewwp132.pdf⟩ (2002).
65. Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H. J. Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proc. R. Soc. Lond. B* **268,** 2495–2501 (2001).
66. Greiner, B. & Levati, M. V. Indirect reciprocity in cyclical networks: An experimental study. Max Planck Institute, Jena, working paper no. 2003-15 ⟨ftp://papers.mpiew-jena.mpg.de/esi/discussionpapers/2003-15.pdf⟩ (2003).
67. Dufwenberg, M., Gneezy, U., Güth, W. & van Damme, E. Direct vs indirect reciprocity: An experiment. *Homo Oecon.* **18,** 19–30 (2001).
68. Güth, W., Königstein, M., Marchand, N. & Nehring, K. Trust and reciprocity in the investment game with indirect reward. *Homo Oecon.* **18,** 241–262 (2001).
69. Pollock, G. B. & Dugatkin, L. A. Reciprocity and the evolution of reputation. *J. Theor. Biol.* **159,** 25–37 (1992).
70. Gintis, H. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Stategic Interaction* (Princeton Univ. Press, Princeton, 2000).
71. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends Cogn. Sci.* **8,** 185–190 (2004).

72. Fehr, E. & Fischbacher, U. Third party punishment and social norms. *Evol. Hum. Behav.* **25,** 63–87 (2004).
73. Milinski, M., Semmann, D. & Krambeck, H. J. Donors in charity gain in both indirect reciprocity and political reputation. *Proc. R. Soc. Lond. B* **269,** 881–883 (2002).
74. Milinski, M., Semmann, D. & Krambeck, H. J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415,** 424–426 (2002).
75. Panchanathen, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free-rider problem. *Nature* **432,** 499–502 (2004).
76. Fehr, E. Don't lose your reputation. *Nature* **432,** 449–450 (2004).
77. Boyd, R. & Richerson, P. J. The evolution of indirect reciprocity. *Soc. Networks* **11,** 213–236 (1989).
78. Pfeiffer, T., Rutte, C., Killingback, T., Taborsky, M. & Bonhoeffer, S. Evolution of cooperation by generalized reciprocity. *Proc. R. Soc. B* **272,** 1115–1120 (2005).
79. Brandt, H. & Sigmund, K. The good, the bad and the discriminator. *J. Theor. Biol.* (in the press).
80. Imhof, L., Fudenberg, D. & Nowak, M. A. Evolutionary cycles of cooperation and defection. *Proc. Natl Acad. Sci. USA* **102,** 10797–10800 (2005).
81. Camerer, C. E. *Behavioral Game Theory* (Princeton Univ. Press, Princeton, 2003).
82. Colman, A. M. *Game Theory and its Applications in the Social and Biological Sciences* (Butterworth-Heinemann, Oxford, 1995).
83. Arrow, K. *The Limits of Organization* (Norton, New York, 1974).
84. Kosfeld, M., Heinrichs, M., Zak, P., Fischbacher, U. & Fehr, E. Oxytocin increases trust in humans. *Nature* **435,** 673–676 (2005).
85. Milgrom, P., North, D. & Weingast, B. The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Econ. Politics* **2,** 1–23 (1990).
86. Suzuki, Y. & Toquenaga, Y. Effects of information and group structure on evolution of altruism: Analysis of two-score model by covariance and contextual analyses. *J. Theor. Biol.* **232,** 191–203 (2005).
87. Fudenberg, D. & Maskin, E. Evolution of cooperation in noisy repeated games. *American Economic Reviewl.* **80,** 274–279 (1990).

Acknowledgements Support from the John Templeton Foundation is gratefully acknowledged. The Program for Evolutionary Dynamics at Harvard University is sponsored by Jeffrey Epstein.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to K.S. (karl.sigmund@univie.ac.at).