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

Evolution of Indirect Reciprocity by Image Scoring / The Dynamics of Indirect Reciprocity

Nowak, M.A. and Sigmund, K.

IIASA Interim Report
May 1998

Nowak, M.A. and Sigmund, K. (1998) Evolution of Indirect Reciprocity by Image Scoring / The Dynamics of Indirect Reciprocity. IIASA Interim Report. IIASA, Laxenburg, Austria, IR-98-040 Copyright © 1998 by the author(s). http://pure.iiasa.ac.at/5608/

Interim Reports on work of the International Institute for Applied Systems Analysis receive only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work. All rights reserved. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage. All copies must bear this notice and the full citation on the first page. For other purposes, to republish, to post on servers or to redistribute to lists, permission must be sought by contacting repository@iiasa.ac.at

IIASA

International Institute for Applied Systems Analysis • A-2361 Laxenburg • Austria
Tel: +43 2236 807 • Fax: +43 2236 71313 • E-mail: info@iiasa.ac.at • Web: www.iiasa.ac.at

INTERIM REPORT IR-98-040 / May

Evolution of Indirect Reciprocity by Image Scoring / The Dynamics of Indirect Reciprocity

Martin A. Nowak (martin.nowak@zoo.ox.ac.uk)
Karl Sigmund (ksigmund@esi.ac.at)

Approved by
Ulf Dieckmann (dieckman@iiasa.ac.at)
Project Coordinator, *Adaptive Dynamics Network*

Interim Reports on work of the International Institute for Applied Systems Analysis receive only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.

IIASA Studies in Adaptive Dynamics No. 27

The Adaptive Dynamics Network at IIASA fosters the development of new mathematical and conceptual techniques for understanding the evolution of complex adaptive systems.

Focusing on these long-term implications of adaptive processes in systems of limited growth, the Adaptive Dynamics Network brings together scientists and institutions from around the world with IIASA acting as the central node.

Scientific progress within the network is reported in the IIASA Studies in Adaptive Dynamics series.

The Adaptive Dynamics Network

The pivotal role of evolutionary theory in life sciences derives from its capability to provide causal explanations for phenomena that are highly improbable in the physicochemical sense. Yet, until recently, many facts in biology could not be accounted for in the light of evolution. Just as physicists for a long time ignored the presence of chaos, these phenomena were basically not perceived by biologists.

Two examples illustrate this assertion. Although Darwin's publication of "The Origin of Species" sparked off the whole evolutionary revolution, oddly enough, the population genetic framework underlying the modern synthesis holds no clues to speciation events. A second illustration is the more recently appreciated issue of jump increases in biological complexity that result from the aggregation of individuals into mutualistic wholes.

These and many more problems possess a common source: the interactions of individuals are bound to change the environments these individuals live in. By closing the feedback loop in the evolutionary explanation, a new mathematical theory of the evolution of complex adaptive systems arises. It is this general theoretical option that lies at the core of the emerging field of adaptive dynamics. In consequence a major promise of adaptive dynamics studies is to elucidate the long-term effects of the interactions between ecological and evolutionary processes.

A commitment to interfacing the theory with empirical applications is necessary both for validation and for management problems. For example, empirical evidence indicates that to control pests and diseases or to achieve sustainable harvesting of renewable resources evolutionary deliberation is already crucial on the time scale of two decades.

The Adaptive Dynamics Network has as its primary objective the development of mathematical tools for the analysis of adaptive systems inside and outside the biological realm.

IIASA Studies in Adaptive Dynamics

No. 1 Metz JAJ, Geritz SAH, Meszéna G, Jacobs FJA, van Heerwaarden JS:
Adaptive Dynamics: A Geometrical Study of the Consequences of Nearly Faithful Reproduction.
IIASA Working Paper WP-95-099.
van Strien SJ, Verduyn Lunel SM (eds.): Stochastic and Spatial Structures of Dynamical Systems, Proceedings of the Royal Dutch Academy of Science (KNAW Verhandelingen), North Holland, Amsterdam, pp. 183-231 (1996).

No. 2 Dieckmann U, Law R:
The Dynamical Theory of Coevolution: A Derivation from Stochastic Ecological Processes.
IIASA Working Paper WP-96-001.
Journal of Mathematical Biology (1996) 34, 579–612.

No. 3 Dieckmann U, Marrow P, Law R:
Evolutionary Cycling of Predator-Prey Interactions: Population Dynamics and the Red Queen.
Journal of Theoretical Biology (1995) 176, 91–102.

No. 4 Marrow P, Dieckmann U, Law R:
Evolutionary Dynamics of Predator-Prey Systems: An Ecological Perspective.
IIASA Working Paper WP-96-002.
Journal of Mathematical Biology (1996) 34, 556–578.

No. 5 Law R, Marrow P, Dieckmann U:
On Evolution under Asymmetric Competition.
IIASA Working Paper WP-96-003.
Evolutionary Ecology (1997) 11, 485–501.

No. 6 Metz JAJ, Mylius SD, Diekmann O:
When Does Evolution Optimise? On the Relation between Types of Density Dependence and Evolutionarily Stable Life History Parameters.
IIASA Working Paper WP-96-004.

No. 7 Ferrière R, Gatto M:
Lyapunov Exponents and the Mathematics of Invasion in Oscillatory or Chaotic Populations.
Theoretical Population Biology (1995) 48, 126–171.

No. 8 Ferrière R, Fox GA:
Chaos and Evolution.
Trends in Ecology and Evolution (1995) 10, 480–485.

No. 9 Ferrière R, Michod RE:
The Evolution of Cooperation in Spatially Heterogeneous Populations.
IIASA Working Paper WP-96-029.
American Naturalist (1996) 147, 692–717.

No. 10 Van Dooren TJM, Metz JAJ:
Delayed Maturation in Temporally Structured Populations with Non-Equilibrium Dynamics.
IIASA Working Paper WP-96-070.
Journal of Evolutionary Biology (1998) 11, 41–62.

No. 11 Geritz SAH, Metz JAJ, Kisdi E, Meszéna G:
The Dynamics of Adaptation and Evolutionary Branching.
IIASA Working Paper WP-96-077.
Physical Review Letters (1997) 78, 2024–2027.

No. 12 Geritz SAH, Kisdi E, Meszéna G, Metz JAJ:
Evolutionarily Singular Strategies and the Adaptive Growth and Branching of the Evolutionary Tree.
IIASA Working Paper WP-96-114.
Evolutionary Ecology (1998) 12, 35–57.

No. 13 Heino M, Metz JAJ, Kaitala V:
Evolution of Mixed Maturation Strategies in Semelparous Life-Histories: the Crucial Role of Dimensionality of Feedback Environment.
IIASA Working Paper WP-96-126.
Philosophical Transactions of the Royal Society of London Series B (1997) 352, 1647–1655.

No. 14 Dieckmann U:
Can Adaptive Dynamics Invade?
IIASA Working Paper WP-96-152.
Trends in Ecology and Evolution (1997) 12, 128–131.

No. 15 Meszéna G, Czibula I, Geritz SAH:
Adaptive Dynamics in a Two-Patch Environment: a Simple Model for Allopatric and Parapatric Speciation.
IIASA Interim Report IR-97-001.
Journal of Biological Systems (1997) 5, 265–284.

No. 16 Heino M, Metz JAJ, Kaitala V:
The Enigma of Frequency-Dependent Selection.
IIASA Interim Report IR-97-061.
Trends in Ecology and Evolution (1998) in press.

No. 17 Heino M:
Management of Evolving Fish Stocks.
IIASA Interim Report IR-97-062.
Canadian Journal of Fisheries and Aquatic Sciences (1998) in press.

No. 18 Heino M:
Evolution of Mixed Reproductive Strategies in Simple Life-History Models.
IIASA Interim Report IR-97-063.

No. 19 Geritz SAH, van der Meijden E, Metz JAJ:
Evolutionary Dynamics of Seed Size and Seedling Competitive Ability.
IIASA Interim Report IR-97-071.

No. 20 Galis F, Metz JAJ:
Why are there so many Cichlid Species? On the Interplay of Speciation and Adaptive Radiation.
IIASA Interim Report IR-97-072.
Trends in Ecology and Evolution (1998) 13, 1–2.

No. 21 Boerlijst MC, Nowak MA, Sigmund K:
Equal Pay for all Prisoners. / The Logic of Contrition.
IIASA Interim Report IR-97-073.
AMS Monthly (1997) 104, 303–307.
Journal of Theoretical Biology (1997) 185, 281–294.

No. 22 Law R, Dieckmann U:
Symbiosis without Mutualism and the Merger of Lineages in Evolution.
IIASA Interim Report IR-97-074.
Proceedings of the Royal Society of London Series B (1998) 265, 1245–1253.

No. 23 Klinkhamer PGL, de Jong TJ, Metz JAJ:
Sex and Size in Cosexual Plants.
IIASA Interim Report IR-97-078.
Trends in Ecology and Evolution (1997) 12, 260–265.

No. 24 Fontana W, Schuster P:
Shaping Space: The Possible and the Attainable in RNA Genotype-Phenotype Mapping.
IIASA Interim Report IR-98-004.

No. 25 Kisdi E, Geritz SAH:
Adaptive Dynamics in Allele Space: Evolution of Genetic Polymorphism by Small Mutations in a Heterogeneous Environment.
IIASA Interim Report IR-98-038.

No. 26 Fontana W, Schuster P:
Continuity in Evolution: On the Nature of Transitions.
IIASA Interim Report IR-98-039.
Science (1998) 280, 1451–1455.

No. 27 Nowak MA, Sigmund K:
Evolution of Indirect Reciprocity by Image Scoring. / The Dynamics of Indirect Reciprocity.
IIASA Interim Report IR-98-040.
Nature (1998) 393, 573–577.

No. 28 Kisdi E:
Evolutionary Branching Under Asymmetric Competition.
IIASA Interim Report IR-98-045.

Issues of the IIASA Studies in Adaptive Dynamics series can be obtained free of charge. Please contact:

Adaptive Dynamics Network
International Institute for Applied Systems Analysis
Schloßplatz 1
A–2361 Laxenburg
Austria

Telephone +43 2236 807, Telefax +43 2236 71313, E-Mail adn@iiasa.ac.at,
Internet http://www.iiasa.ac.at/Research/ADN

Contents

About this Report

In addition to reciprocation based on repeated interactions within a pair, there exists another, indirect reciprocity, where the donor does not obtain a return from the recipient, but from a third party. Donors provide help if the recipient has helped others in the past. This works if the cost of an altruistic act is offset by a raised 'score', or status, which increases the chance to subsequently become the recipient of an altruistic act. Cooperation is channelled towards the 'valuable' members of the community. For Richard Alexander, 'indirect reciprocity involves reputation and status, and results in everyone in the group continually being assessed an re-assessed'.

In the first part of the report this is modelled by a population of individuals having the options of helping another or not. In each generation, a number of potential donor-recipient pairs are chosen randomly: if the help is actually provided, this implies a cost c to the donor, a benefit b to the recipient, and it increases the donor's score by one. The score of a player refusing to help is decreased by one. Initially all scores are zero. We consider strategies given by integers k; a player with such a strategy helps if and only if the score of the potential recipient is at least k. We can follow the frequencies of the strategies from generation to generation, allowing for occasional mutations.

In the second part, models which are even more simplified help to explain analytically cycling behaviour, with its long bouts of cooperation interspersed by short periods of defection, which is reminiscent of the lack of stability near a critical state. Somewhat surprisingly, cooperation is more robust if the society is challenged more frequently by invasion attempts of defectors. One can compute the minimal amount of discriminators, the minimal number of rounds per generation and the maximal size of the society, for indirect reciprocity to work. This yields as necessary condition for cooperation that the *degree of acquaintanceship* (the probability that a player knows the score of the co-player) is larger than the cost-to-benefit ratio c/b. This result is analogous to Hamilton's rule which states that the *degree of relatedness* (the probability that an allele in the player's genome is also present in the co-player) must exceed c/b.

About the Authors

Martin Nowak
Department of Zoology
University of Oxford
South Parks Road
Oxford OX1 3PS, UK

Karl Sigmund
Institut für Mathematik
Universität Wien
Strudlhofgasse 4
A-1090 Vienna, Austria
and
Adaptive Dynamics Network
International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria

Acknowledgments

Thanks to Marian Dawkins, Axel Kacelnik, John Krebs and Robert May for discussion. Support from the Wellcome Trust is gratefully acknowledged. Part of the work was done at IIASA (Laxenburg).
Correspondence should be addressed to Martin Nowak (e-mail: martin.nowak@zoo.ox.ac.uk). Later this year, Martin Nowak will move from Oxford to Princeton. The new address will be Institute for Advanced Study, Olden Lane, Princeton, New Jersey 08540, U.S.A.

Evolution of Indirect Reciprocity by Image Scoring

Martin A. Nowak
Karl Sigmund

Abstract

The question of cooperation is crucial for understanding Darwinian evolution. Theories of cooperation have been based on kin selection[1,2], group selection[3–5], and reciprocal altruism[6–9]. The idea of reciprocal altruism usually involves direct reciprocity: repeated encounters between the same individuals allow for the return of an altruistic act by the recipient[10–16]. Here we present a new theoretical framework, which is based on indirect reciprocity[17] and does not require the same two individuals ever to meet again. Individual selection can nevertheless favour cooperative strategies directed towards recipients that have helped others in the past. Cooperation pays because it confers the image of a valuable community member. We present computer simulations and analytic models to specify the conditions for evolutionary stability[18] of indirect reciprocity. In particular, we show that the probability of knowing the image of the recipient must exceed the cost-to-benefit ratio of the altruistic act. We argue that the emergence of indirect reciprocity was a decisive step for the evolution of human societies.

Humans have achieved one of the pinnacles of sociality, and the complexity of their cooperative actions is without parallels. In constrast to other examples of ultrasociality[19–22] (e.g. clones, or bee hives, or termite colonies), human cooperation is due less to kin selection based on genetic similarity, than to cultural forces rooted in pervasive moral systems. From hunter tribes and village communities to nation states and global enterprises, the economic effects of nepotism, while certainly present, are minor compared with those of reciprocity. The latter is usually understood as direct reciprocity: help someone who may later help you. But there exists another, indirect reciprocity prevailing in human communities. In this case, one does not expect a return from the recipient, but from someone else, according to the pious advice of ‘give, and you shall be given’. Cooperation is channeled towards the ‘valuable’ members of the community. This has been called the ‘I won’t scratch your back if you won’t scratch their backs’-principle[23]. A donor provides help if the recipient is likely to help others (which often means, if the recipient has helped others in the past). In this case, it pays to advertise cooperation, since the cost of an altruistic act is offset by a greater chance to subsequently become the recipient of an altruistic act. Animal and human behaviour may be influenced by the attempt to increase the image (or status) in the group[24–25].

According to Richard Alexander[17], indirect reciprocity, which ‘involves reputation and status, and results in everyone in the group continually being assessed and reassessed’ plays a large role in human societies (and possibly in some primates, social canines, etc). Alexander interprets moral systems as systems of indirect reciprocity. Clearly, indirect reciprocity presupposes rather sophisticated players, and therefore is likely to be affected by anticipation, planning, deception, and manipulation. The politicking needed to continually assess the status of all members of our community and to bolster our own has probably been a major force for shaping our intelligence. But if we want to understand the basic mechanisms of indirect reciprocity, we have to analyse drastically simplified models.

Imagine a population of individuals having the options to help one another or not. Random pairs of players are chosen, of which one is the potential donor of some altruistic act and the other is the recipient. The donor can cooperate and help the recipient at a cost c to himself, in which case the recipient receives a benefit of value b (with $b > c$). If the donor decides not to help, both individuals receive zero payoff. Each player has an image score, s, which is known to every other player. If a player is chosen as donor and decides to cooperate then his (or her) image score increases by one unit; if the donor does not cooperate then it decreases by one unit. The image score of a recipient does not change. At first, we consider strategies where donors decide to help according to the image score of the recipient. A strategy is given by a number k: a player with this strategy provides help if and only if the image score of the potential recipient is at least k.

Figure 1 shows computer simulations of a population consisting of n players. The strategies are given by k_i and the image levels by s_i. In the beginning of each generation, the image levels of all players are zero (assuming that children do not inherit the image of their parents). In succession, m donor-recipient pairs are chosen. A donor, i, cooperates with a recipient, j, if $k_i \leq s_j$. The fitness of a player is given by the total number of points received during the m interactions. Some players may never be chosen, in which case their payoff from the game will be zero. On average, a player will be chosen $2m/n$ times, either as donor or as recipient. At the end of each generation, players leave offspring in proportion to their fitness. We find that if the game is played for a large number of generations, then eventually all players will adopt the same strategy. If the k-value of this strategy is 0 or less then cooperation is established; if the value is 1 or more then defection has won. Cooperation is more likely to win the greater the number m of interactions per generation. (A totally different model of indirect reciprocity has been studied by Boyd and Richerson[26], who assumed that individuals interact in loops such that a cooperative action can be returned, after several steps, to the original donor. According to Boyd and Richerson their model is unlikely to lead to a cooperative outcome, as it requires the loops to be relatively small, closed, and long-lasting. We think that this is because their model does not include image scores.)

We can also include mutation, by assuming that there is a small probability that a strategy does not reproduce accurately but gives rise to an offspring adopting a different strategy (Fig 2). In this case, several strategies can persist. We study the frequency distribution of various strategies and analyse how often a cooperative regime is achieved. A minimum number of rounds per generation is needed for cooperation to prevail. Interestingly this number can be very small: it suffices that each player is chosen only for about 2 interactions per life-time. (In this case there is only a probability of 1/4 that a defector can be punished; namely if he is first chosen as a donor and subsequently as a recipient.) Below we present an analytic model for evaluating the minimum number of interactions compatible with cooperation.

Long term simulations, including mutation, usually do not converge to a simple equilibrium distribution of strategies, but show endless cycles. In very simple terms, what happens is that defectors are invaded by discriminators, who only help players whose score exceeds some threshold. Next, discriminators are undermined by unconditional cooperators. The prevalence of these indiscriminate altruists subsequently allows the return of defectors. In a population consisting only of discriminators and unconditional cooperators,

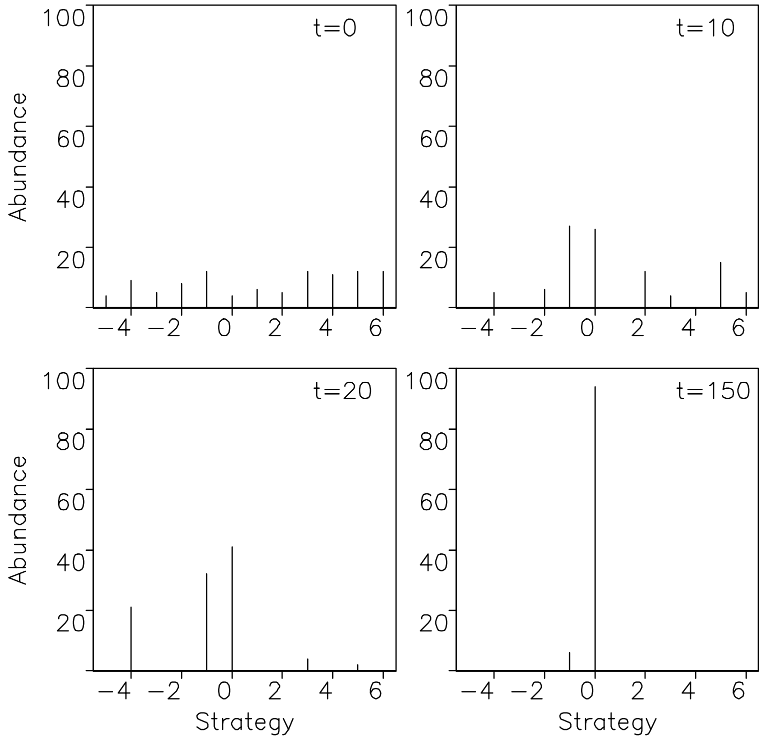

Figure 1. Cooperation wins in a computer simulation of indirect reciprocity. The population consists of $n = 100$ individuals. The image scores range from -5 to $+5$, the k-values from -5 to $+6$. The strategy, $k = -5$, represents unconditional cooperators, while the strategy, $k = +6$, represents defectors. In each round of the game, two individuals are chosen at random; one as donor, the other as recipient. The donor cooperates if the image score of the recipient is greater than or equal to the donor's k. Cooperation means the donor pays a cost, c, and the recipient obtains a benefit, b. There is no payoff in the absence of cooperation. In the beginning of each generation all players have image score 0. Hence, strategies with $k \leq 0$ are termed "cooperative", because they cooperate with individuals that have not had an interaction. In each generation $m = 125$ pairs are chosen; each player has, on average, 2.5 interactions. The chance that a given player meets the same co-player again, or that a chain of possible altruistic acts ever leads back to the original donor, is negligibly small. Therefore, direct reciprocity cannot work here. At the end of each generations, players produce offspring proportional to their payoff. At generation, $t = 0$, we start with a random distribution of strategies. After $t = 10$ generations, the strategies $k = -1, 0, +2, +5$ have increased in abundance. After $t = 20$ generations, the strategies $k = -4, -1, 0$ dominate the population. After $t = 150$ generations, the population consists almost entirely of the strategy $k = 0$, which is the most discriminating among all cooperative strategies. It cooperates with everyone who has image score 0 or greater. After $t = 166$ generations, all other strategies have become extinct and $k = 0$ is fixed in the population. Parameter values: $b = 1$, $c = 0.1$ (to avoid negative payoffs we add 0.1 in each interaction).

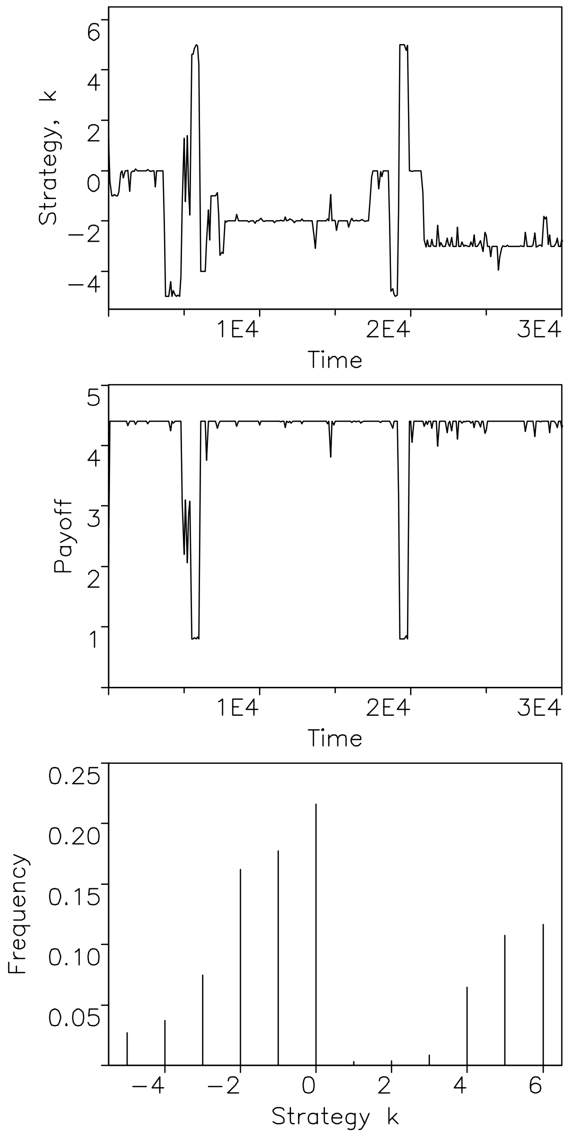

Figure 2. Long-term evolution of indirect reciprocity under mutation and selection. We perform the same computer simulation as in figure 1, but include mutation: there is a probability of 0.001 that an offspring does not act like its parent, but uses another randomly chosen strategy. We observe endless cycles of cooperation and defection. Cooperative populations are relatively stable if they consist of discriminating strategies such as $k = 0$ or -1. But after some time these populations get undermined (through random drift) by strategies such as $k = -4$ or -5 which are too cooperative. Then defectors, $k = 4$ or 5, can invade, which in turn can be overcome by stern discriminators again. In the long run, cooperation is harmed by unconditional cooperators, because they enable defectors to invade. In the absence of unconditional cooperators, cooperative populations persist much longer. (a) The average k-value of the population. (b) The average payoff per individual, per generation. (c) Frequency distribution of strategies sampled over many generations ($t = 10^7$). Parameter values: as for figure 1, but $m = 300$ rounds per generation.

there is no selection against the latter, who can spread by random drift. In simulations without unconditional cooperators, cooperative populations persist much longer.

Cooperation based on indirect reciprocity depends crucially on the ability of a player to estimate the image score of the opponent. In the above model we assume that the image score of each individual is known to every other member of the population. This should only be seen as an idealised scenario. It is more realistic to assume that an interaction between two individuals is observed by a (possibly small) subset of the population. Only these "on-lookers" (and, of course, the recipient) have the possibility to update their perception of the donor's image score. The on-lookers are chosen at random for each particular interaction. Therefore each player has a specific perception of the image score of the other players. The same player can have different image scores in the eyes of different individuals. The information is contained in a matrix whose elements s_{ij} denote the image score of player i as seen by player j. In a donor-recipient interaction between j and i, player j will cooperate if $s_{ij} \geq k_j$. If j has no information on i then $s_{ij} = 0$.

The model now depends on the probability that a given individual observes an interaction between two other individuals. Figure 3 shows computer simulations of this extended model. Again cooperation can easily be established and dominate the population, but a larger number of interactions per generation is needed. There is also an effect of group size. For larger groups, it is more difficult to establish cooperation, because the fraction of individuals that obtain information about any particular interaction will be smaller. Therefore, more interactions are required (relative to group size) in order to discriminate against defectors.

Another interesting expansion of the basic model is to include strategies that consider both the recipient's and the donor's image score. We explored two types of strategies. "AND"–strategies cooperate if the image score of the recipient is larger than a certain value *and* the own image score is less than a certain value. The idea is that if an individual has already a high image score, it is not necessary to aim for a still higher image score (by helping others). On the other hand, "OR"–strategies cooperate if the image score of the recipient is larger than a certain value *or* the own image score is less than a certain value. Here the idea is that if an individual has a low image score it may be advantageous to increase the score by helping others regardless of how low their image score is. In both cases we find highly cooperative societies (Fig 4). If, in contrast, we simulate strategies that only consider their own image and do not take into account the image of the recipient, then cooperation does not emerge.

The above models are based on computer simulations, but we can derive analytic insights from a simplified model. Suppose that there are only 2 image levels: 0 (for bad) and 1 (for good). The image of a player depends on his or her last action as a donor: players who defected have score 0, and players who cooperated score 1. Let us only consider two strategies: (i) defectors, who never provide assistance, and (ii) discriminators who help players having image 1, but not players having image 0. A given player knows the score of only a fraction, q, of the population. A discriminator who has no information on potential recipients will assume, with a certain probability, p, that they have image 1. In each round of the game all individuals of the population are chosen, each with the same probability as a donor or a recipient. If $w < 1$ denotes the probability of another round, there are on average $1/(1-w)$ rounds per generation. In the Methods Section we derive the equations that describe how the frequencies of discriminators and defectors of image 0 and 1 change during subsequent rounds of the game. We also calculate the average payoff in each round and analyse how the frequencies of discriminators and defectors change from one generation to the next. It should be stressed that discriminators are not Tit For Tat

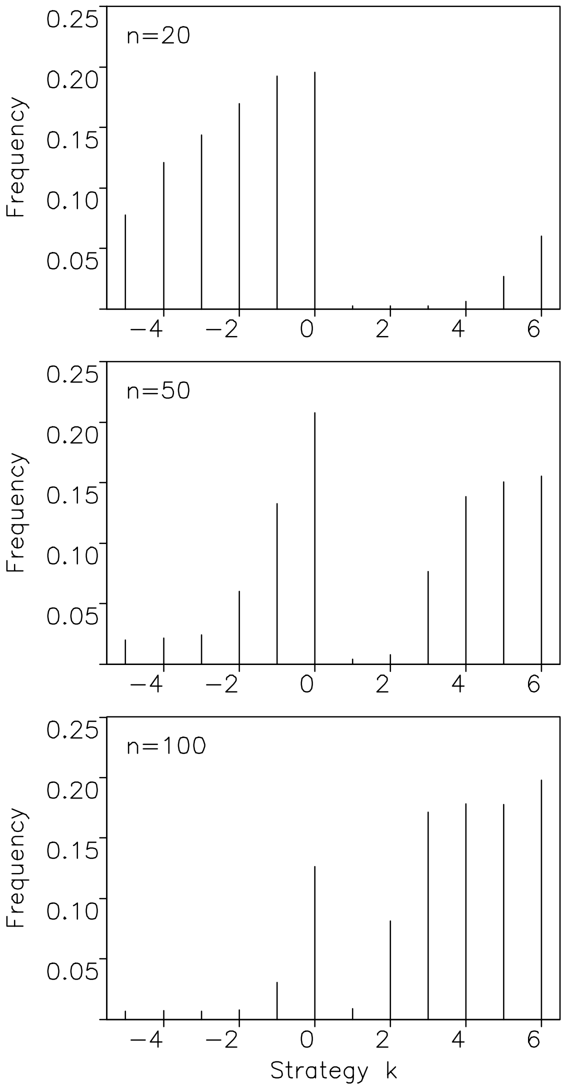

Figure 3. Indirect reciprocity with incomplete information about the image score of other players. We perform the same simulation as in figure 2, but the image score of a donor is updated only for the recipient and the observers of an interaction. Each interaction is observed, on average, by 10 randomly chosen players. The figure shows the frequency distribution of strategies for three different population sizes, $n = 20$, $n = 50$, and $n = 100$, sampled over many generations ($t = 10^7$) in order to obtain representative results. There is a clear effect of group size: cooperation predominates for $n = 20$, but is rare for $n = 100$. For $n = 50$ we find cooperative and defective strategies at roughly equal frequencies. The time averages of the frequency of cooperative strategies (defined by $k \leq 0$) are 90%, 47% and 18% for, respectively, $n = 20$, 50 and 100. Parameter values: as for figure 2, but $m = 10n$ rounds per generation.

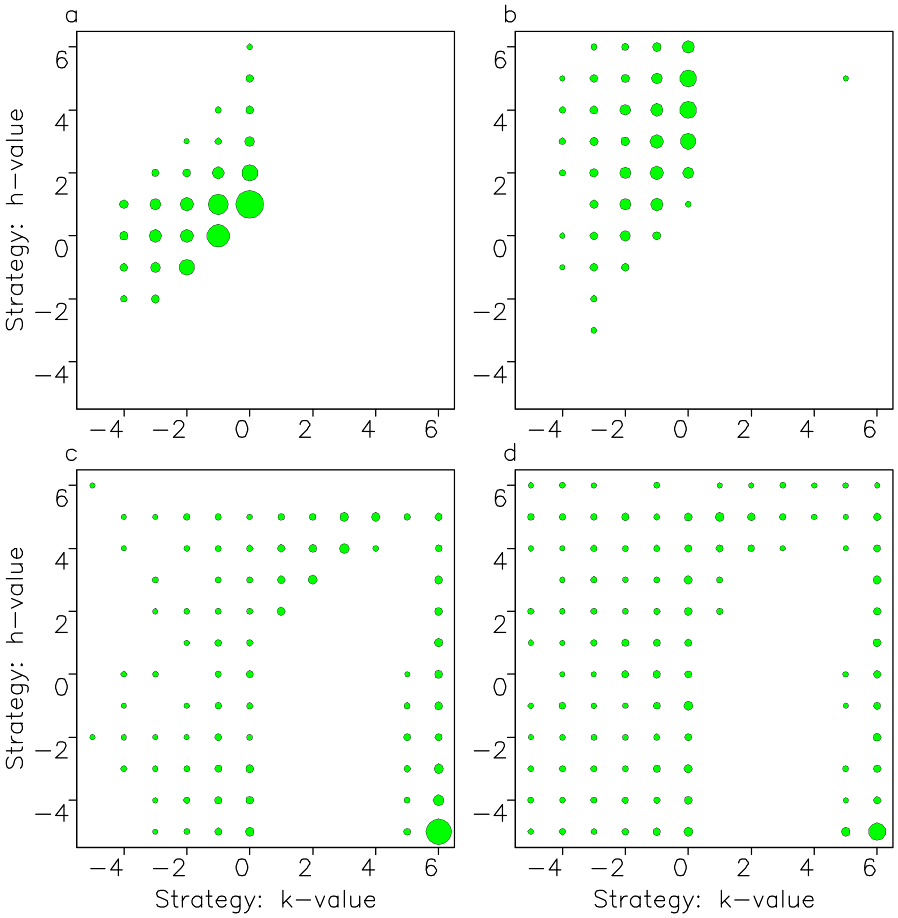

Figure 4. A further dimension is added to the game if donors base their decision to cooperate not only on the image score of the recipient but also on their own score. In (a) and (b), we consider strategies that cooperate if the image score of the opponent is at least k AND the own image score is less than h. The idea is that if the image score of an individual is already high, then it makes no sense to invest in a still higher image. The figures show the frequency distribution of strategies that are defined by their k and h–values sampled over many generations. In (a), we assume perfect information about the image of all players. The most frequent strategy is ($k = 0$,$h = 1$). This strategy cooperates if the image score of the opponent is at least 0 and the own image score is less than 1. If the whole population adopts this strategy then it clearly does not pay to aim for an image exceeding 0. For the same reason, other strategies with $h = k + 1$ are successful in this simulation. Strategies with $k > 0$ are unsuccessful, because they are too uncooperative. In (b), we assume imperfect information about other players' image. Here it pays to invest in a higher image than strictly necessary, because a given altruistic act is only seen by a subset of other players. The most frequent strategy is $k = 0$, $h = 4$. In (c) and (d), strategies are explored that cooperate if the image score of the recipient is at least k OR the own image score is less than h. The idea is that players with low image may want to increase their image by helping others indiscriminately. Such a scenario also leads to cooperative societies (dominated by strategies with $k \leq 0$) but unconditional defectors ($k = 6$,$h = -5$) benefit from the reduced level of discrimination and represent the most frequent single strategy. Again (c) is based on perfect information while (d) assumes imperfect information of the co-players' image score. In (a), (b), (c), and (d), respectively, the frequency of cooperative interactions is 55%, 57%, 70% and 80%. This has to be compared with less than 0.1% cooperation in simulations where strategies only consider their own image score and do not discriminate according to the image score of the recipient. Parameter values: (a,c) as in figure 2, but $m = 500$ rounds per generation; (b,d) as in figure 3 with $n = 20$. The frequency of a strategy is proportional to the area of the circle. Strategies with a frequency of less than 0.5% are not shown.

players; Tit For Tat strategists base their decisions on their own previous experience with the co-player, whereas discriminators use the experience of others. This is an essential advantage for a player who interacts with many co-players, but only a few times with each. (Such discriminators are also different from strategies based on 'standing'[27].)

We observe a frequency threshold: a minimum amount, x_{min}, of discriminators is necessary to ensure the establishment of cooperation. We also obtain the minimum number of rounds per generation which are necessary for the evolutionary stability of discriminators. In particular, cooperation through indirect reciprocity can only be stable if

$$q > c/b.$$

The probability to know the image of another player has to exceed the cost to benefit ratio of the altruistic act. This is remarkably similar to Hamilton's rule, which states that cooperation through kin selection works whenever the coefficient of relatedness between two individuals exceed the cost to benefit ratio[1,2]. In our case, relatedness is replaced by acquaintanceship.

In summary, cooperation based on indirect reciprocity works in the follwing way: a potential donor can choose whether to accept a certain cost in order to help another individual, or to avoid this cost. In the short term, avoiding the cost yields, of course, the higher payoff. In the long term, however, performing the altruistic act increases the image score of the donor and may therefore increase the chance of obtaining a benefit in a future encounter as a recipient. On the other hand, a discriminator who punishes low-score players by refusing them help pays for this by having his own score reduced. The decisive idea, relevant to human societies, is that information about another player does not require a direct interaction, but can be obtained indirectly either by observing the player or by talking to others. The evolution of human language as a means of such information transfer has certainly facilitated cooperation based on indirect reciprocity.

Methods

Defectors and discriminators.

Here we develop a simplified model for indirect reciprocity which can be fully understood in analytic terms. Consider two image scores: 0 for someone who has defected last round and 1 for someone who has cooperated last round. Thus the image score only depends on the last move of a player as a donor. Consider two types of players: discriminators, who only help players having image score 1, and defectors, who never help. Let us suppose that there is a probability, q, that discriminators have information about the image score of the potential recipient. In the absence of information, they assume an image score of 1 with probability p. (One can show that if indirect reciprocity works at all, then discriminators with larger p always outcompete the others[28]. Therefore we shall restrict ourselves in the following to the limiting value $p = 1$. The discriminator strategy in this case coincides with a variant of Tit For Tat which begins by defecting if the future co-player has been seen defecting in his last interaction[29] – a confirmation of Alexander's view that 'indirect reciprocity is a consequence of direct reciprocity occuring in the presence of others'[17]). For a defector, information about the image score does not matter. We denote by x_0, x_1, y_0 and y_1, respectively, the frequencies of discriminators with image 0 and 1, and the frequencies of defectors with image 0 and 1. The total frequency of discriminators is $x = x_0 + x_1$ and that of defectors $y = y_0 + y_1$. We have $x + y = 1$. A generation consists of several rounds of the game, during which x and y do not change. In each round all players are paired up, half of the players being donors, the other half recipients. The frequencies of players of image 0 or 1 change from round to round according to the difference equations:

$$\begin{aligned} x_0' &= [x_0 + x(1-\phi)q]/2 \\ x_1' &= [x_1 + x(1-q+q\phi)]/2 \\ y_0' &= [y_0 + y]/2 \\ y_1' &= y_1/2 \end{aligned}$$

Here $\phi = x_1 + y_1$ is the frequency of players with score 1. In each round the payoff to the individual types is $P(x_0) = [-c(1-q+q\phi) + bx(1-q)]/2$, $P(x_1) = [-c(1-q+q\phi)+bx]/2$, $P(y_0) = bx(1-q)/2$, $P(y_1) = bx/2$. The difference equation yields the expected payoff values $D_e(k)$ and $D_i(k)$ to defectors and discriminators in the k-th round: $D_e(k) = bx(1-q+q2^{-(k-1)})/2$, and $D_i(k) = D_e(k) + \{(1-q)(bqx-c)(1-qx)^{-1} - bq2^{-(k-1)} + q(b-c)(1-x)(1-qx)^{-1}[(1+qx)/2]^{k-1}\}/2$. We can either assume that the number of rounds per generation is constant, or that there exists a fixed probability w for a further round. In the latter case, the total payoff to defectors is $D_e = \sum_{k=1}^{\infty} w^{k-1} D_e(k)$, and similarly for discriminators. We find that

$$2(D_i - D_e) = (1-q)\frac{bqx-c}{(1-w)(1-qx)} + 2q[\frac{(b-c)(1-x)}{(1-qx)(2-w-wqx)} - \frac{b}{2-w}].$$

Modelling the change in frequency of discriminators and defectors from one generation to the next by the standard replicator equation[30], we find that defectors win if x is below a threshold value x_{min} given by $D_i = D_e$, whereas discriminators win if x is above this threshold. Discriminators are evolutionarily stable if and only if $x_{min} < 1$, i.e. if $D_i > D_e$ for $x = 1$. This can only happen for $q > c/b$, i.e. if the probability to know the image of the co-player exceeds the cost to benefit ratio, and if the average number of rounds, i.e. $1/(1-w)$, exceeds $(bq+c)/(bq-c)$. Note that for our numerical example of Figs. 1 and 2 , where $b = 1$ $c = 0.1$ and $q = 1$, we only need about 1.2 rounds per generation for cooperation to be stable against invasion by defectors.

The good, the bad, and the discriminating.

Clearly, indirect reciprocity only works when donors discriminate between individuals that have or have not helped others in the past. In order to understand the role of indiscriminate altruists, we add to the population of defectors and discriminators a fraction z of cooperators, who always give help, irrespective of their co-player's score. We can calculate the payoffs in each round as before. The cooperators' total expected payoff D_c differs from that of the defectors, D_e, by $[-c + (bwqx)/(2-w)]/[2(1-w)]$, whereas

$$D_i - D_e = \frac{(bqx-c)(1-q+qz)}{2(1-w)(1-qx)} - \frac{bq(x+y)}{2-w} + \frac{qy(b-c)}{(1-qx)(2-w-wqx)}.$$

The population is in equilibrium whenever $y = 0$ (no defectors) or $x = c(2-w)/bwq$. If x lies below the latter value, defectors win; if x exceeds it, then a mixture of discriminate and indiscriminate altruists gets established, depending on the initial value. This mixed state is proof against invasion by unconditional defectors, but in such a population both discriminate and indiscriminate altruists do equally well. Their frequencies will only be altered by random drift, not by selection. If the frequency of discriminators falls below $c(2-w)/bwq$ then defectors can invade and take over. Defectors in turn can be overcome by discriminators if their frequency fluctuates above x_{min}.

A universal constant of nature.

Let us now consider a situation where the image score can be any integer number between $-\infty$ and $+\infty$, but all players adopt the same strategy, $k = 0$. Denote by x_i the frequency of players with image score i. In the next round it is $x'_i = [x_i + x_{i-1}\phi + x_{i+1}(1-\phi)]/2$ where $\phi = \sum_{i=0}^{\infty} x_i$. If all players start with an image score greater than or equal to 0, then all players will cooperate in the first and all subsequent rounds. If all players start with an image score less than 0, then all players will defect in the first and all subsequent rounds. The situation becomes interesting if there is an initial distribution of image scores above and below 0. The question whether the system will ultimately converge to cooperation or defection is non-trivial. We find there is a maximum fraction of players with an initial image score below 0, such that the system ultimately converges to all-out cooperation. Numerical simulations show that this number is 0.7380294688360...

References

1. Hamilton W D. The evolution of altruistic behaviour, *Am Nat* 97, 354-6 (1963)
2. Hamilton W D. The genetical evolution of social behaviour, *J theor Biol* 7, 1-16 (1964)
3. Williams G C. *Group Selection*, Aldine-Atherton, Chicago (1971)
4. Eshel I. On the neighborhood effect and evolution of altruistic traits, *Theor Pop Biol* 3, 258-277 (1972)
5. Wilson D S, Sober E. Reintroducing group selection to the human behavioural sciences, *Behavioural and Brain Sciences* 17, 585-654 (1994)
6. Trivers R. The evolution of reciprocal altruism, *Quarterly Review of Biology* 46, 35-57 (1971)
7. Axelrod R, Hamilton W D. The evolution of cooperation, *Science* **211**, 1390 (1981)
8. Axelrod R. *The Evolution of Cooperation*, (Basic Books, New York, 1984)
9. Nowak M A, May R M. Evolutionary games and spatial chaos, *Nature*, 359, 826-829 (1992)
10. Michod R E, Sanderson, M J. Behavioural structure and the evolution of cooperation, in P.J. Greenwood, P. Harvey and M. Slatkin (eds.) *Evolution: Essays in honor of John Maynard Smith*, 95-106, Cambridge UP (1985)
11. Peck J, and Feldman M. The evolution of helping in large, randomly mixed populations, *Am. Nat.* 127, 209-221 (1985)
12. Milinski, M. Tit for tat in sticklebacks and the evolution of cooperation, *Nature* **325**, 433-435 (1987)
13. May R M. More evolution of cooperation, *Nature* 327, 15-17 (1987)
14. Dugatkin L A, Mesterton-Gibbons M, Houston A I. Beyond the prisoner‘s dilemma: towards models to discriminate among mechanism of cooperation in nature. *TREE* 7, 202-5 (1992)

15. Nowak M A, Sigmund K. Tit for tat in heterogeneous populations, *Nature* **355**, 250-253 (1992)

16. Nowak M A, Sigmund K. Win-stay, lose-shift outperforms tit for tat, *Nature*, 364, 56-58 (1993)

17. Alexander R D. *The biology of moral systems*, Aldine de Gruyter, New York (1987)

18. Maynard Smith J. *Evolution and the Theory of Games* (Cambridge UP, 1982)

19. Wilson E O. *Sociobiology*, Harvard UP, Cambridge, Mass (1975)

20. Krebs J R, Davies N B. *An introduction to behavioural ecology*, Blackwell, Oxford (1987)

21. Buss J. *The evolution of individuality*, Princeton UP (1987)

22. Frank S A. The origin of synergistic symbiosis. *J theor Biol.* 176: 403-10 (1995)

23. Binmore K G. *Fun and Games: a Text on Game Theory*, Heath and Co, Lexington, Massachussetts, (1992)

24. Marler P, Evans C. Bird calls: Just emotional displays or something more? *Ibis* 138: 26-33 (1996)

25. Zahavi A, Zahavi A. *The Handicap Principle; a Missing Piece of Darwin's Puzzle*, Oxford UP (1997)

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

27. Sudgen R. *The economics of rights, cooperation and welfare*, Blackwell, Oxford (1986)

28. Nowak M A, Sigmund K. The dynamics of indirect reciprocity, submitted to *Journ. Theor. Biol.*

29. Pollock G B, Dugatkin L A. Reciprocity and the evolution of reputation, *Journal Theor. Biol.* 159, 25-37 (1992)

30. Hofbauer J and Sigmund K. *Evolutionary games and population dynamics*, Cambridge University Press (1998)

The Dynamics of Indirect Reciprocity

Martin A. Nowak
Karl Sigmund

Abstract

Richard Alexander (1987) has argued that moral systems derive from indirect reciprocity. We analyse a simple case of a model of indirect reciprocity based on image scoring (see Nowak and Sigmund, 1998). Discriminators provide help to those individuals who have provided help. Even if the help is never returned by the beneficiary, or by individuals who in turn have been helped by the beneficiary, discriminating altruism can be resistant against invasion by defectors. Indiscriminate altruists can invade by random drift, however, setting up a complex dynamical system. In certain situations, defectors can invade only if their invasion attempts are sufficiently rare. We also consider a model with incomplete information and obtain conditions for the stability of altruism which differ from Hamilton's rule by simply replacing relatedness with acquaintanceship.

1 Introduction

Altruistic behaviour is usually explained by inclusive fitness, group advantage, or reciprocity. The idea of reciprocal altruism, which is essentially economic, was introduced by Trivers (1971): a donor may help a recipient if the cost (to the donor) is less than the benefit (to the recipient), and if the recipient is likely to return the favour. This principle was explored in many papers, we mention only Axelrod and Hamilton (1981), Axelrod (1984), Sugden (1986), Boyd and Lorberbaum (1987), May (1987), Lindgren (1991), Nowak and Sigmund (1992, 1993), Nowak, May and Sigmund (1995), Sigmund (1995), Leimar (1997).

In his seminal paper of 1971, Trivers mentioned the further possibility of a 'generalised altruism', where the return is directed towards a third party. 'Individuals ... may respond to an altruistic act that benefits themselves by acting altruistically toward a third individual uninvolved in the initial interaction.' Trivers goes on to say: 'In a system of strong multiparty interactions, it is possible that in some situations individuals are selected to demonstrate generalised altruistic tendencies.' This possibility is further stressed in Triver's book on *Social Evolution* (1985), where it is speculated that a sense of justice may evolve 'in species such as ours in which a system of multi-party altruism may operate such that an individual does not necessarily receive reciprocal benefit from the individual aided but may receive the return from third parties.'

Richard Alexander greatly extended this idea, and coined the term of 'indirect reciprocity' (see Alexander, 1979 and 1987, and references quoted therein). In this case, one does not expect a return from the recipient (as with direct reciprocity), but from someone else. Cooperation is thereby channelled towards the cooperative members of the community. A donor provides help if the recipient is likely to help others (which is usually decided on the basis of experience, i.e. according to whether the potential recipient has helped others in the past). According to Richard Alexander (1986), indirect reciprocity, which 'involves reputation and status, and results in everyone in the group continually being assessed and reassessed', plays an essential role in human societies. Alexander argues (convincingly, to our mind) that systems of indirect reciprocity are the basis of moral systems.

The principles of direct reciprocity are usually studied by means of games (like the Prisoner's Dilemma) repeatedly played between the same two players. In this paper we investigate situations where the players engage in several rounds of the game, but with a negligible probability of ever encountering the same co-player again. This is, of course, an idealisation, and in human communities, both direct and indirect reciprocity occur together. In fact, Alexander stresses that 'indirect reciprocity is a consequence of direct reciprocity occurring in the presence of others'. But in order to better understand the mechanism of indirect reciprocity, we shall entirely eliminate direct reciprocity from our model.

In Nowak and Sigmund (1998), we analysed populations of individuals having the options to help one another or not. Following usual practice, we denote the benefit of the altruistic act to the recipient by b, the cost to the donor by c, and assume $c < b$. If the donor decides not to help, both individuals receive zero payoff. The payoff is in terms of incremental fitness.

Each player has an image score, s. The score of a potential donor increases by one unit if he or she performs the altruistic act; if not, it decreases by one unit. The image score of a recipient does not change. At birth, each individual has score 0. We consider strategies where potential donors decide to help according to the image score of the recipient. A strategy is given by an integer k: a player with strategy k provides help if and only if the image score of the potential recipient is at least k. Players who provide help must pay some cost, but they increase their score and are, therefore, more likely to receive help in the future. During their lifetime, individuals undergo several rounds of this interaction, either as donors or as recipients, but the possibility of meeting the same co-player again will be neglected in our model. At the end of each generation, individuals leave offspring in proportion to their accumulated payoff, which inherit the strategy of their parent (we assume clonal reproduction, as is usual in evolutionary games, see Maynard Smith, 1982).

In extensive computer simulations, Nowak and Sigmund (1998) showed that even for a very low number of rounds per generation, a cooperative regime based on indirect reciprocity can be stable. If one allows for mutations, then long-term cycling becomes likely. Populations of altruists discriminating according to the score of the recipient are undermined by indiscriminate altruists. Then, unconditional defectors invade, until discriminating cooperators return, etc. We also extended the model so that individuals would only witness a fraction of the interactions in their community, and therefore have incomplete information about their co-player's score.

In this paper we shall study analytically a class of simple models for indirect reciprocity, based on two score values only, which we denote by G (for 'good') and B (for 'bad'). We obtain some of the cycling behaviour seen in the computer simulations. Furthermore, we show that the probability q that a player knows the score of another player must exceed c/b, if indirect reciprocity is to work. This is an intriguing parallel to Hamilton's rule, the cornerstone of the kin-selection approach to altruism (Hamilton, 1963). Hamilton's rule states that the coefficient of relatedness must exceed c/b. In this sense, indirect reciprocity differs from kin selection in replacing relatedness with acquaintance. If the average number of rounds per lifetime exceeds $(bq + c)/(bq - c)$, then cooperation based on score discrimination is evolutionarily stable.

2 The basic model

For indirect reciprocity to work, some members of the group must assess the 'score' of other members, and discriminately channel their assistance toward those with a higher score. Of course, the group may also contain members who do not discriminate, and either always give help, or never. We shall denote the frequency of the former by x_1, and that of the latter by x_2. By x_3, we denote the frequency of the discriminators. These individuals assess their group members and keep track of their 'score'. If they only remember the last round, they distinguish between those who have helped and thereby acquired score G, and those who have withheld assistance, and acquired score B. Discriminators help only G-players.

We shall now assume that each generation experiences a certain number of rounds of interactions. In each round, every player is both in the position of a donor and in the position of a recipient. (This simplifies the calculations without changing the basic results. In Nowak and Sigmund, 1998, as well as in the last section of this paper, we assume that every player can be, with the same probability, a donor or a recipient.) In each of these roles, the player interacts with a randomly chosen co-player. If only few rounds occur, then the likelihood of meeting the same co-player twice is very small. The strategies which we consider take no account of this possibility.

In the first round, discriminators do not know the score of the potential recipient of their altruistic action. They have to rely on an *a priori* judgement, and assume with a certain 'subjective' probability p that they are matched with a G-individual. If they help, they acquire G-status and become possible beneficiaries of other discriminators in the next round. We first consider the case $0 < p < 1$, and later the case $p = 1$.

With g_n we denote the frequency of G-players in round n (it is convenient to set $g_1 = p$, the discriminators' initial guess). Clearly

$$g_n = x_1 + g_{n-1}x_3 \tag{1}$$

for $n = 1, 2, ...,$ so that by induction

$$g_n = \frac{x_1}{1-x_3} + x_3^{n-1}(p - \frac{x_1}{1-x_3}). \tag{2}$$

Hence g_n converges to $x_1/(x_1 + x_2)$, the percentage of cooperators among the indiscriminating players.

In order to compute the payoff, we have to monitor whether a recipient who meets a discriminating donor is perceived by the donor as a G-player. In the first round, this happens with probability p. From then on, it happens with probability 1 to the undiscriminate altruists (who have had occasion to prove their altruism), with probability 0 to the unconditional defectors (who are unmasked in the first round), and with probability g_{n-1} to the discriminators (since this is the probability that they have encountered a G-player and consequently provided help in the previous round).

In the first round, the payoff for an indiscriminate altruist is $-c+b(x_1+px_3)$ (he always provides help, and he receives help from the indiscriminate altruists as well as from those discriminators who believe that he has label G). The payoff for unconditional defectors is similarly $b(x_1 + px_3)$ and that for discriminators is $-cp + b(x_1 + px_3)$. Obviously, if there is only one round, unconditional defectors win.

In the n-th round ($n > 1$), the indiscriminate altruists receive payoff $-c+b(x_1+x_3)$, and unconditional defectors obtain bx_1. The proportion of G-scorers among the discriminators is g_{n-1} and their payoff is $-cg_n + b(x_1 + x_3)$. The other discriminators obtain $-cg_n + bx_1$.

Adding up, we receive as the discriminators' payoff in the n-the round $-cg_n + b(x_1 + x_3 g_{n-1})$, which by (2) is just $(b-c)g_n$.

If we assume that there exactly N rounds per generation, then the total payoff for indiscriminate altruists is

$$\hat{P}_1 = N[b(x_1 + x_3) - c] - (1-p)bx_3, \tag{3}$$

that for defectors is

$$\hat{P}_2 = Nbx_1 + bpx_3, \tag{4}$$

and that for discriminators is

$$\hat{P}_3 = (b-c)(g_1 + ... + g_N) + b(x_1 + px_3 - p). \tag{5}$$

It is easy to check that

$$g_1 + ... + g_N = \frac{1}{1-x_3}[(g_1 - g_2)\frac{1-x_3^N}{1-x_3} + Nx_1], \tag{6}$$

so that

$$\hat{P}_3 = (p - px_3 - x_1)(-b + \frac{b-c}{1-x_3}\frac{1-x_3^N}{1-x_3}) + \frac{N(b-c)x_1}{1-x_3}. \tag{7}$$

It is well-known that the structure of a game is unchanged if the same function is subtracted from all payoff functions (see, e.g., Hofbauer-Sigmund, 1998). It turns out that it is most convenient to substract $\hat{P}_2$. We then obtain as normalised payoff values $P_i := \hat{P}_i - \hat{P}_2$, the values $P_2 = 0$,

$$P_1 = (N-1)bx_3 - Nc \tag{8}$$

and

$$P_3 = \frac{x_1}{x_1 + x_2}P_1 + (p - \frac{x_1}{x_1 + x_2})\frac{bx_3 - c - (b-c)x_3^N}{1-x_3}. \tag{9}$$

For instance, if the game is stopped after the second round already, i.e. $N = 2$, then

$$P_1 = -2c + bx_3, \tag{10}$$

$P_2 = 0$, and

$$P_3 = -cp - cx_1 + (b-c)px_3. \tag{11}$$

3 The replicator equation for a constant number of rounds

This allows to investigate the evolution of the frequencies of the three types of players under the influence of selection. We can use either a discrete game dynamics monitoring the frequencies from generation to generation, or the continuous replicator dynamics (see Hofbauer-Sigmund, 1998)

$$\dot{x}_i = x_i(P_i - \bar{P}) \tag{12}$$

on the (invariant) simplex $S_3 = \{\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}) \in \mathbf{R^3} : \mathbf{x_i} \geq \mathbf{0}, \sum \mathbf{x_i} = \mathbf{1}\}$. Here, $\bar{P} = \sum x_i P_i$ is the average payoff in the population. We stick to the latter, somewhat more transparent dynamics, emphasising that it is obtained as a limiting case of the dynamics with discrete generations (see Hofbauer-Sigmund, 1998).

For simplicity, let us start with the case $N = 2$. If $b > 2c$, as we shall assume in the following, then there exists a unique fixed point $\hat{\mathbf{p}} = (p_1, p_2, p_3)$ in the interior of S_3, i.e. with all three types present. It is given by $P_1 = P_2 = P_3$, which yields (since $P_2 = 0$)

$$p_1 = p(1 - \frac{2c}{b}), \qquad p_2 = (1-p)(1 - \frac{2c}{b}), \qquad p_3 = \frac{2c}{b}\ . \tag{13}$$

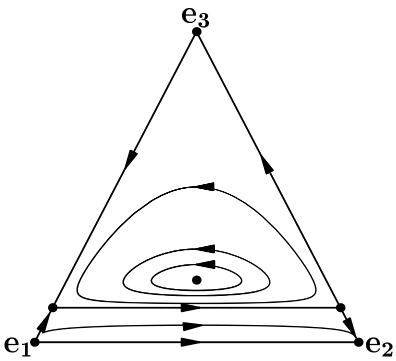

Figure 1. Phase portrait of the model described in Section 3. There are 3 strategies: cooperators, defectors and discriminators (corresponding to the three corner fixed points $\mathbf{e_1}, \mathbf{e_2}$ and $\mathbf{e_3}$, respectively). Discriminators help everyone with a good standing. In the first round, they help other individuals with a fixed probability, p We assume the game is played for 2 rounds; the payoff values are given by eqs (10,11). In the absence of discriminators, $x_3 = 0$, defectors win. In the absence of defectors, $x_2 = 0$, a stable equilibrium between cooperators and discriminators is reached. In the absence of cooperators, $x_1 = 0$, there is an unstable equilibrium between defectors and discriminators. If all 3 strategies are present, there is a seperatrix connecting the two boundary equilibria on the edges. If the initial frequency of discriminators is below a critical value, then defectors will win. If it is above this critical value, then we obtain neutral oscillations around a center.

This point is a center. Indeed, one checks by a straightforward computation that the Jacobian at $\hat{\mathbf{p}}$ has trace 0 and determinant $2c^2p(1-p)(1-\frac{2c}{b})^2$. The eigenvalues are therefore purely imaginary.

On the boundary of the simplex S_3, we find five fixed points. In addition to the corners $\mathbf{e_1}, \mathbf{e_2}$ and $\mathbf{e_3}$ (where only one type is present), we find two mixed equilibria, namely

$$\mathbf{F_{23}} = (\mathbf{0}, \frac{\mathbf{b-2c}}{\mathbf{b-c}}, \frac{\mathbf{c}}{\mathbf{b-c}}) \tag{14}$$

and $\mathbf{F_{13}}$, which is obtained from $\mathbf{F_{23}}$ by exchanging the first and the second coordinate. In the absence of discriminators (i.e. on the edge $x_3 = 0$), the flow points from $\mathbf{e_1}$ to $\mathbf{e_2}$: defectors win. In the absence of defectors, i.e. for $x_2 = 0$, the flow on the edge $\mathbf{e_1 e_3}$ leads toward $\mathbf{F_{13}}$. In the absence of indiscriminate altruists, i.e. when $x_1 = 0$, the system on the edge $\mathbf{e_2 e_3}$ is bistable (see Fig.1).

Since $(b-c)x_3 = c$ is an invariant line (along this line, one has $\dot{x}_3 = 0$), it follows that there exists an orbit in the interior of S_3 which points along this straight line from $\mathbf{F_{13}}$ (its α-limit) to $\mathbf{F_{23}}$ (its ω-limit). The boundary of the triangle spanned by $\mathbf{e_3}, \mathbf{F_{13}}$ and $\mathbf{F_{23}}$ is a heteroclinic cycle: it consists of three saddle-points connected by three orbits.

Using the classification of phase portraits of the replicator equation due to Zeeman (1980) and Bomze (1983), we can conclude that the fixed point $\hat{\mathbf{p}}$ is surrounded by closed orbits filling the afore-mentioned triangle (in Bomze's notation, we obtain phase portrait 13). The time-averages of these orbits all converge toward the point $\mathbf{p}$. In the remaining part of the simplex S_3, all orbits converge to $\mathbf{e_2}$. If the frequency of discriminators x_3 is less than $c/(b-c)$, therefore, then defectors take over. If not, then the frequencies of the three types oscillate periodically. We note however that this situation is not persistent: a sequence of random fluctuations can lead to larger and larger oscillations, and finally cause the system to cross the separatrix line $(b-c)x_3 = c$ and end up with a regime of all-out defection.

We mention without proof that if there are $N > 2$ rounds, nothing much changes. The unique fixed point $\hat{\mathbf{p}}$ in the interior of S_3 has now the coordinates

$$p_1 = p(1-p_3), \qquad p_2 = (1-p)(1-p_3), \qquad p_3 = \frac{Nc}{(N-1)b}. \tag{15}$$

(The third equation follows because $P_1 = 0$, the first because for this value of p_1, one has $p - \frac{p_1}{1-p_3} = 0$.) Again, the eigenvalues at $\hat{\mathbf{p}}$ are pure imaginary; this fixed point is a center surrounded by periodic orbits. The points $\mathbf{F_{13}}$ and $\mathbf{F_{23}}$ now satisfy

$$x_3 + ... + x_3^{N-1} = c/(b-c) \tag{16}$$

(the equation for $\mathbf{F_{23}}$ is given by $P_3 = 0$, that for $\mathbf{F_{13}}$ by $P_1 = P_3$.) We note that the solution of (16) satisfies $x_3 > c/b$.

4 The prejudice p as an evolutionary variable

So far we have treated p, the prejudice of the discriminator, as a parameter. But p is likely to be an evolutionary variable. So let us consider a model where, in addition to the types used so far, with frequencies x_1, x_2 and x_3, we have another type of discriminator with a prejudice $\rho \neq p$. The frequency of this new type is denoted by x_4 (with $\sum x_i = 1$). Again we can describe the payoffs of the different types of players in the different rounds. In the first round, all players receive (as recipients) the payoff $b(x_1 + px_3 + \rho x_4)$, which we neglect henceforth, since it is the same for all; as donors, indiscriminate altruists pay $-c$, unconditional defectors 0, and the two types of discriminators $-cp$ and $-c\rho$, respectively. In the first round, it pays to have as low an opinion as possible concerning the score of the unknown partner. From then onward, the score is always G for the indiscriminate altruists, and never G for the unconditional defectors. The two types of discriminators have score G, in the second round, with probability p and ρ, respectively. It follows that in the second round, the frequency of G-players is $g_2 = x_1 + px_3 + \rho x_4$. For the n-th round, with $n > 2$, the frequency g_n of G-players satisfies the recurrence relation

$$g_n = x_1 + (x_3 + x_4) g_{n-1}. \tag{17}$$

In the second round, the payoff for p-discriminators is therefore given by $-cg_2 + b(x_1 + p(x_3 + x_4))$, and that for ρ-discriminators by $-cg_2 + b(x_1 + \rho(x_3 + x_4))$. In the n-th round ($n > 2$) the payoff is $-cg_n + bg_{n-1}$ for both types of discriminators. If there are altogether two rounds or more per generation, then the total payoff for the p-discriminators differs from that of the ρ-discriminators by $(p - \rho)(-c + b(x_3 + x_4)$. By the quotient rule for the replicator dynamics (see Hofbauer and Sigmund, 1998) it follows that

$$(x_3/x_4)^{\cdot} = (x_3/x_4)(p-\rho)(-c + b(x_3 + x_4)). \tag{18}$$

If the total frequency $x_3 + x_4$ of discriminators is sufficiently high (namely larger than c/b), then (18) shows that the ratio x_3/x_4 increases if and only if $p > \rho$. In particular, in a population where the p-discriminating type is established and defectors have gone to extinction, or are on their way to vanish (which means, as we have seen, that x_3 is larger than c/b), then the ρ-discriminating type can invade and take over, if and only if $\rho > p$. Thus we can conclude that if indirect reciprocity works at all, then it favours those discriminators having larger p-values, i.e. with a more positive prejudice in favour of an unknown partner. This leads to a trait-substitution sequence in the sense of Metz et al (1992): mutations introducing larger and larger p-values will successively take over under the influence of selection. The p-value will therefore grow, as an evolutionary variable, until it approaches its maximal value 1. We shall therefore restrict our attention to the limiting case $p = 1$.

¿From now on, a discriminator is a player who, in the first round, gives help, and from then on helps recipients with G-score only. (The first help can be viewed as an entrance fee to the club of G-players.) It should be stressed that discriminators are not Tit For Tat players. Tit For Tat is a very successful strategy for the iterated Prisoner's Dilemma, and consists in cooperating in the first round, and from then on doing whatever the co-player did in the previous round. Tit For Tat strategists base their decisions on their own previous experience with the co-player, whereas discriminators use the experience of others. Pollock and Dugatkin (1992), in their interesting paper on reputation, described this strategy as 'observer TFT'.

It should also be mentioned that this discriminator strategy is related to, but different from the so-called T_1-strategy in the book by Robert Sugden on *The Evolution of Rights, Cooperation and Welfare* (1986). The T_1-strategy is based on the concept of *good standing.* Every player is born with a good standing, and keeps it as long as he extends help to other players with good standing. If he does not, he loses his good standing. Sugden argues that such a strategy can work as a basis for an insurance principle within the population (in each round of his game, a randomly chosen player needs help, and all other players can contribute to it). We stress that a player can keep his good standing by refusing to help someone of bad standing, whereas in our model, he would lose his G-score whenever he refuses help, even if the potential recipient is a B-scorer. Sugden's T_1 strategy is more sophisticated, but like Contrite Tit For Tat, another strategy based on standing, it is vulnerable to errors in perception (see also Boerlijst et al, 1997).

5 Pyrrhic victories, or the advantage of rarely showing up

If we denote the frequency of discriminators by x_3, again, then the payoffs for indiscriminate altruists, unconditional defectors and discriminators are, in the first round, given by $-c+b(x_1+x_3), b(x_1+x_3)$, and $-c+b(x_1+x_3)$, respectively, and in the following rounds by $-c+b(x_1+x_3), bx_1$ resp. $-cg_n+b(x_1+x_3g_{n-1}) = (b-c)g_n$ where g_n is, as before, the frequency of G-players in round n and $g_1 = 1$. We now have by (2):

$$g_n = [x_1 + x_3^{n-1}x_2]/(x_1+x_2). \tag{19}$$

If there are exactly N rounds (with $N > 1$), then the total payoffs $\hat{P}_1$, $\hat{P}_2$ and $\hat{P}_3$ of indiscriminate altruists, unconditional defectors and discriminators, respectively, are given by

$$\hat{P}_1 = N[-c+b(x_1+x_3)], \tag{20}$$

$$\hat{P}_2 = Nbx_1 + bx_3, \tag{21}$$

$$\hat{P}_3 = (b-c)(g_1+g_2+...+g_N) - bx_2 \tag{22}$$

which yields

$$\hat{P}_3 = N(b-c) + x_2[-b + \frac{b-c}{1-x_3}(1+x_3+...+x_3^{N-1}-N)]. \tag{23}$$

Normalising such that $P_2 = 0$, this yields

$$P_1 = -Nc + (N-1)bx_3 \tag{24}$$

and

$$P_3 = P_1 + x_2[(N-1)b + \frac{b-c}{1-x_3}(1+...+x_3^{N-1} - N)]$$

and hence

$$P_3 = P_1 + [-(N-1)(bx_3+c) - cx_3(1+...+x_3^{N-2})]/(1-x_3). \tag{25}$$

Let us consider first the case $N=2$ of two rounds only. In this case, we have

$$P_1 = -2c + bx_3, \tag{26}$$

and

$$P_3 = P_1 + cx_2. \tag{27}$$

It follows immediately that the replicator equation admits no interior fixed point. The edge $\mathbf{e_1e_3}$ consists of fixed points: in the absence of unconditional defectors, both types do equally well. Along the edge $x_3 = 0$, the flow points from $\mathbf{e_1}$ to $\mathbf{e_2}$. On the edge $\mathbf{e_2e_3}$, there exists a fixed point $\mathbf{F_{23}}$, with $x_3 = c/(b-c)$. The restriction to this edge is bistable: in a competition between unconditional defectors and discriminators, discriminators win if and only if their initial frequency is larger than $c/(b-c)$. Since the average payoff $\bar{P}$ is equal to $x_1P_1 + x_3(P_1 + cx_2)$, if follows that at $\mathbf{F_{23}}$, the transversal eigenvalue $\dot{x}_1/x_1$ is given by $\frac{c(2c-b)}{b-c}$, which is negative. Hence $\mathbf{F_{23}}$ is saturated.

Along the fixed point edge $\mathbf{e_1e_3}$, the transversal eigenvalue $\dot{x}_2/x_2$ is equal to $-\bar{P}$ (i.e. to $2c - bx_3$). If we denote by $\mathbf{F}$ the point with $x_2 = 0$ and $\dot{x}_2/x_2 = 0$, i.e. with $x_3 = 2c/b$, then the points on the edge between $\mathbf{e_3}$ and $\mathbf{F}$ are saturated, and hence ω-limits of orbits in the interior of S_3, whereas all points on the segment between $\mathbf{F}$ and $\mathbf{e_1}$ are α-limits. If $(b-c)x_3 = c$ then $\dot{x}_3 = 0$ in the interior of S_3. It follows that the line l given by $x_3 = c/(b-c)$ is invariant. It corresponds to an orbit whose ω-limit is the saddle point $\mathbf{F_{23}}$ and whose α-limit, which we denote by $\mathbf{F_{13}}$, has coordinates $x_2 = 0$ and $x_3 = b/(b-c)$. This separatrix l divides the interior of the simplex S_3 into two regions. In one region, all orbits converge toward $\mathbf{e_2}$. In the other region, all orbits lead from the fixed point edge $\mathbf{e_1e_3}$ back to that edge; their α-limit is between $\mathbf{F_{13}}$ and $\mathbf{F}$, their ω-limit between $\mathbf{F}$ and $\mathbf{e_3}$; they surround $\mathbf{F}$. (See Fig. 2.) The equation also admits an invariant of motion: $x_1x_3^{-2}[-c + (b-c)x_3]$ (courtesy of Josef Hofbauer).

The interplay between the three strategies leads to a fascinating long-term dynamics. Depending of the initial condition, selection leads either toward a homogeneous regime of all-out defectors, or to a mixture of discriminators and indiscriminate altruists (with no unconditional defectors). In such a mixture, no type has a selective advantage. Random drift takes over, and the mixture fluctuates along the $\mathbf{e_1e_3}$-edge. From time to time, mutation can also introduce unconditional defectors. If such an invasion is attempted when the state lies between $\mathbf{e_3}$ and $\mathbf{F}$, it is promptly repelled. If it occurs while the state is between $\mathbf{F_{13}}$ and $\mathbf{e_1}$, then it succeeds and defectors take over. But if the invasion attempt occurs while the state lies between $\mathbf{F_{13}}$ and $\mathbf{F}$, then it knows a transient success only; the frequency of defectors increases at first, but then the proportion of discriminators grows at the expense of the indiscriminate altruists, and causes the defectors to vanish. The end result of this failed invasion attempt is, as before, a mixture of discriminators and indiscriminate altruists, but now with a much higher amount of discriminators, so that now it is able to stop any invasion attempt by defectors in the bud. Somewhat related examples of successful invasions which are ultimately self-defeating (Pyrrhic victories, so to speak) can be found in Mylius et al, 1998, where strategies are studied which are invasible yet unbeatable.

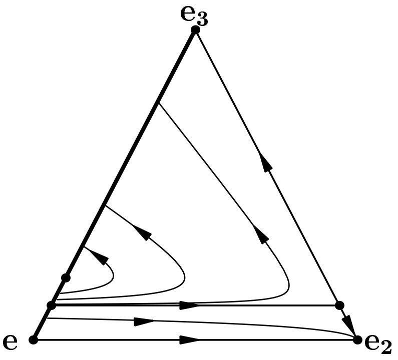

Figure 2. Phase portrait of the model described in Section 5 (eqs 26-27). As for figure 1, we consider cooperators, defectors and discriminators, but this time discriminators always help in the first round ($p = 1$). Again there is a seperatrix connecting two fixed points on the edges $\mathbf{e_1 e_3}$ and $\mathbf{e_2 e_3}$, but there is no fixed point in the interior of the simplex. Instead the whole edge $\mathbf{e_1 e_3}$ consists of fixed points, some of which are stable against invasion by defectors, while others are not. The overall dynamics of the system are as follows. Imagine a mixture between cooperators and discriminators. There is random drift along the edge $\mathbf{e_1 e_3}$. If there are sufficiently many discriminators then defectors cannot invade. There are two threshold levels of discriminators. If the frequency drops below the first value then defectors can invade, but will go extinct again leaving the system in a state with a higher frequency of discriminators. If the discriminator frequency fluctuates below the second value, then defectors can invade and take over. Hence, if defectors appear too often they cannot win. They only win when showing up rarely. This seems to be an interesting example for a more general, counter-intuitive principle where a mutation can only win if rare.

Of course, random drift can slowly lead the state back into the threatened zone. But if invasions by defectors occur frequently enough, these invasions will be attempted while the state is between $\mathbf{F_{13}}$ and $\mathbf{F}$, and hence the state will be led back into the invasion-proof zone. It is only if the frequency of invasion attempts by defectors is low that random drift along the fixed point edge $\mathbf{e_1 e_3}$ can lead the state across the 'gap' between $\mathbf{F_{13}}$ and $\mathbf{F}$ (whose width is $c(b-2c)/b(b-c)$). In this case the state enters into the segment between $\mathbf{F}$ and $\mathbf{e_1}$ where an invasion by defectors knows an irreversible success. Thus we see a remarkable phenomenon: a mutant that can succeed only if it occurs rather rarely!

Essentially the same situation holds when there are $N > 2$ rounds. The point $\mathbf{F_{23}}$ now has a coordinate x_3 which is given as the solution of the equation

$$x_3 + x_3^2 + ... + x_3^{N-1} = \frac{c}{b-c} \tag{28}$$

(see (16)). This is a value which, with increasing N, shifts from $c/(b-c)$ towards c/b. The point $\mathbf{F}$ has a coordinate x_2 given by $Nc/(N-1)b$. This is simply the limit of the interior fixed point $\hat{\mathbf{p}}$ given by (15), if p converges to 1.

This cycle of invasions is related to a phenomenon found in the numerical simulations by Nowak and Sigmund (1998), which are based on a more sophisticated model of indirect reciprocity where scores can take all integer values (see Fig. 3).

6 Random numbers of rounds

Let us now assume, not a constant number of rounds per generation, but rather a constant probability w for a further round. The total number of rounds per generation is then a geometrically distributed random variable with mean value $1/(1-w)$. The payoffs are of

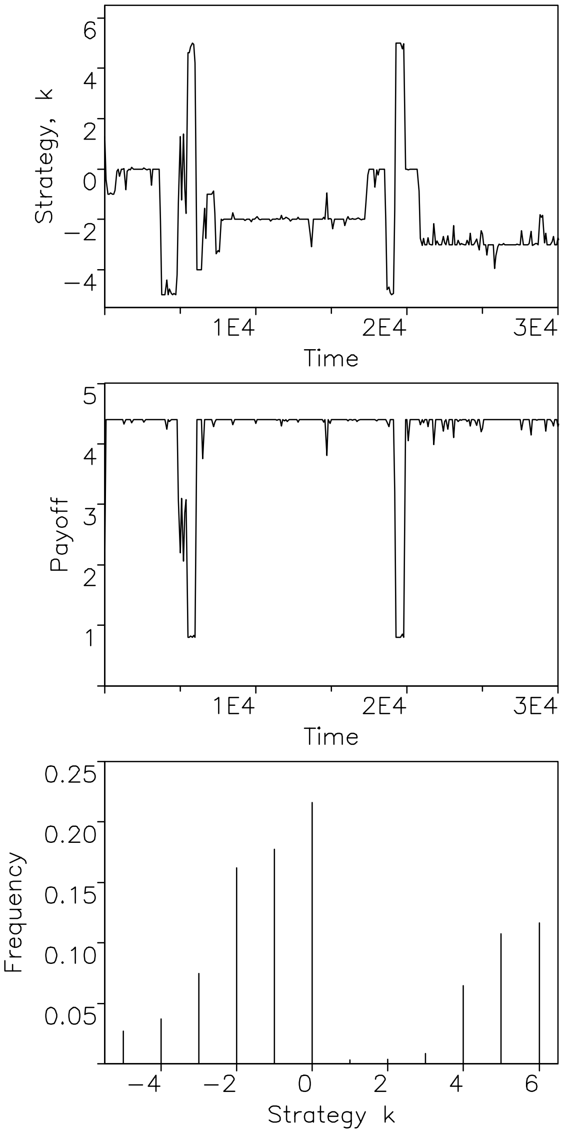

Figure 3. Cycling behaviour in the model by Nowak and Sigmund (1998). Donor-recipient pairs are formed at random. The score of a newborn is 0, it increases by one unit whenever the individual provides help and decreases by one unit if the individual refuses to help. A strategy is given by an integer j. An individual with strategy j provides help to all potential recipients with score at least j. Players with strategy $j = 0$ can be viewed as discriminate altruists. Players with a low j (for instance $j = -3$) are *de facto* indiscriminate altruists, because they help every co-player; indeed, if players experience only two or three rounds per lifetime, there will be no players with score less than -3. Players with a high j (for instance $j = 4$), on the other hand, are defectors; they will never provide help. Numerical simulations show how populations of discriminate altruists are eventually undermined by indiscriminate altruists (the average j-value drops), that defectors cash in (the average j-value sharply increases) and that this brings discriminators to the fore again (the average j-value drops back to 0). (a) The average j-value of the population. (b) The average payoff per individual, per generation. (c) Frequency distribution of strategies sampled over many generations ($t = 10^7$). Parameter values: $b = 1$, $c = 0.1$ (to avoid negative payoffs we add 0.1 in each interaction); $m = 300$ rounds per generation.

the form $A_1 + wA_2 + w^2A_3 + ...$, where A_n is the payoff in the n-th round. Then, by using the first paragraph of section 4,

$$\hat{P}_1 = \frac{1}{1-w}[-c + b(x_1 + x_3)], \tag{29}$$

$$\hat{P}_2 = \frac{1}{1-w}bx_1 + bx_3 = \frac{b(x_1 + x_3) - wbx_3}{1-w}, \tag{30}$$

$$\hat{P}_3 = (b-c)(g_1 + wg_2 + w^2g_3 + ...) - bx_2. \tag{31}$$

Writing $g := g_1 + wg_2 + w^2g_3 + ...$, we see that $g = 1 + w(x_1 + g_1x_3) + w^2(x_1 + g_2x_3) + ...$, and hence that

$$g = 1 + \frac{wx_1}{1-w} + x_3wg. \tag{32}$$

Therefore

$$g = \frac{1 - w + wx_1}{(1-w)(1-wx_3)} \tag{33}$$

and thus

$$\hat{P}_3 = -bx_2 + \frac{(b-c)(1-w+wx_1)}{(1-w)(1-wx_3)}. \tag{34}$$

It is convenient again to normalise the payoff values such that $P_2 = 0$. In this case

$$P_1 = \frac{wbx_3 - c}{1-w} \tag{35}$$

and

$$P_3 = \frac{(b-c)(1-w+wx_1)}{(1-w)(1-wx_3)} - bx_2 - bx_3 - \frac{bx_1}{1-w} \tag{36}$$

which yields

$$P_3 = \frac{(1-w+wx_1)}{1-w}(\frac{b-c}{1-wx_3} - b), \tag{37}$$

and thus finally

$$P_3 = \frac{1-w+wx_1}{1-wx_3}P_1. \tag{38}$$

In contrast to the case of a fixed number of rounds, we now obtain a line l of fixed points in the interior of S_3, given by $x_3 = c/wb$ (we assume from now on that $w > c/b$). The edge $\mathbf{e_1e_3}$ consists of fixed points too. On the edge $\mathbf{e_1e_2}$ the flow leads towards $\mathbf{e_2}$, and on the edge $\mathbf{e_2e_3}$ we have a bistable competition, with threshold point $\mathbf{F_{23}}$ given by the intersection with the fixed point line l. This line l acts as separatrix. It divides S_3 into two regions, in one region the ratio x_1/x_2 decreases and in the other it increases. All orbits in the former region converge to $\mathbf{e_2}$ and lead to a population of unconditional defectors; in the other region, all orbits converge to the fixed point edge, and hence lead to a mixture of discriminators and indiscriminate altruists.

7 An analogy with the Prisoner's Dilemma game

Although the dynamics of indirect reciprocity given by (29)-(32) is based on a model which is quite distinct from the repeated Prisoner's Dilemma game, it yields a remarkably similar dynamics. Indeed, let us consider the Prisoner's Dilemma (PD) game, where each of the two players has, in each round, two options: to play **C** (to cooperate) or **D** (to defect). The payoff matrix is given by

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} \tag{39}$$

where $T > R > P > S$, i.e. the reward R for mutual cooperation is larger than the punishment P for joint defection, but a unilateral defector receives the highest payoff T (the temptation) and a unilateral cooperator the lowest payoff S (the sucker's payoff). Let us assume that in each generation, each player is matched with one randomly chosen co-player for a variable number of rounds. Again, we assume that the probability for a further round is constant and given by some $w < 1$. Let us assume that the population contains only three types of players, the unconditional cooperators, the unconditional defectors, and the Tit For Tat players. Let x_1, x_2 and x_3 be their respective frequencies. The expected payoffs are (as is well known, see for instance Nowak and Sigmund, 1987)

$$\hat{P}_1 = \frac{1}{1-w}[R(x_1+x_3)+Sx_2] \tag{40}$$

$$\hat{P}_2 = \frac{Px_2+Tx_1}{1-w} + (T+\frac{wP}{1-w})x_3 \tag{41}$$

$$\hat{P}_3 = \frac{R(x_1+x_3)}{1-w} + (S+\frac{wP}{1-w})x_2. \tag{42}$$

If we normalise these payoff values, such that $P_2 = 0$, and if we set, as is natural, for the temptation by unilateral defection $T = b$, for the reward by mutual cooperation $R = b - c$, for the punishment of bilateral defection $P = 0$ and for the cost of being suckered $S = -c$, then the payoffs in the PD model become

$$P_1 = \frac{bwx_3 - c}{1-w} \tag{43}$$

and

$$P_3 = \frac{bwx_3 + cwx_2 - c}{1-w} = P_1 + \frac{cwx_2}{1-w}, \tag{44}$$

which behaves like the dynamical system with $N = 2$. In fact, for $w = 1/2$ it is exactly the same system. (If however $w = (N-1)/N$ for $N > 2$, then the equations do not agree with the dynamics given by (10)-(11); we also note that the system (29)-(32) with a random number of rounds is different, and in particular contains higher order terms.)

8 A model with incomplete information

Even in small groups, where everyone knows everyone else, it is unlikely that all group members witness all interactions. Therefore each player has a specific perception of the image score of the other players. The same player can have different image scores in the eyes of different individuals. Furthermore, it is unrealistic to assume that episodes as donor and recipient alternate in a well synchronised way. Some individuals will be more often in a position to give help than others.

We shall therefore assume from now on that in each round, a given individual is with probability 1/2 either a donor or a recipient. If there are only few rounds, it is quite possible that a given individual is never a donor. This is more in line with the stochastic simulations in Nowak and Sigmund (1998). We extend the previous two-score model by assuming that with probability q a given individual knows the score of a randomly chosen opponent. A discriminator who does not know the score of the co-player will assume with probability 1 that this score is G. If g_n denotes, as before, the frequency of G-scorers in the population, and $x_{1G}(n), x_{2G}(n)$ and $x_{3G}(n)$ are the frequencies of indiscriminate altruists, unconditional defectors resp. discriminators in round n, then clearly $x_{1G}(n) = x_1$ and $x_{2G}(n) = (1/2)x_{2G}(n-1)$, since a defector is with probability 1/2 in the role of a donor and then unmasks himself. Therefore

$$x_{2G}(n) = \frac{x_2}{2^{n-1}} \ . \tag{45}$$

The score of a discriminator remains unchanged if he is a recipient. If he is a potential donor, he will either know the co-player (with probability q) and help if the co-player has score G (as happens with probability g_n), or else he will not know the co-players score, and help (this happens with probability $1-q$). Altogether, this yields

$$x_{3G}(n) = (1/2)x_{3G}(n-1) + (1/2)x_3(1-q+qg_n). \tag{46}$$

Since $g_n = x_{1G}(n) + x_{2G}(n) + x_{3G}(n)$, it follows that

$$g_n = sg_{n-1} + (x_1 + (1-q)x_3) \tag{47}$$

with $s = (1+qx_3)/2$. This recurrence relation implies (together with $g_1 = 1$) that

$$g_n = (\frac{1+qx_3}{2})^{n-1}\frac{x_2}{1-qx_3} + \frac{x_1+(1-q)x_3}{1-qx_3}. \tag{48}$$

The payoff for the indiscriminate altruists in round n is

$$\hat{A}_1(n) = -(c/2) + (b/2)(x_1 + x_3). \tag{49}$$

The payoff P_2 for the unconditional defectors depends on their score. Those with score B receive $b(x_1 + (1-q)x_3)/2$ and those with score G in addition $qbx_3/2$, so that

$$\hat{A}_2(n) = (b/2)[x_1 + (1-q)x_3 + x_3q(x_{2G}(n)/x_2)] \ . \tag{50}$$

Finally, a discriminator receives $[-c(qg_n + 1 - q) + bx_1 + (1-q)bx_3)]/2$ if he has score B, and in addition $bqx_3/2$ if he has score G, so that we obtain

$$\hat{A}_3(n) = -(c/2)(qg_n + 1 - q) + (b/2)(x_1 + x_3) - (b/2)qx_3[1 - (x_{3G}(n)/x_3)]. \tag{51}$$

Normalising by subtracting $\hat{A}_2(n)$, this yields

$$A_1(n) = -(c/2) + (b/2)qx_3(1 - 2^{-(n-1)}) \tag{52}$$

and

$$A_3(n) = -(c/2)(1-q) + (q/2)(b-c)g_n - (b/2)qx_1 - (b/2)q(x_2 + x_3)2^{-(n-1)}. \tag{53}$$

If we assume that $w < 1$ is the probability for a further round, then the total payoff for unconditional defectors is $P_2 = 0$, that for indiscriminate altruists is

$$P_1 = \frac{1}{2(1-w)}[-c + \frac{bwqx_3}{2-w}] \tag{54}$$

and that for discriminators is

$$P_3 = \frac{(bqx_3 - c)(1-q+qx_1)}{2(1-w)(1-qx_3)} - \frac{bq(x_2+x_3)}{2-w} + \frac{q(b-c)x_2}{(1-qx_3)(2-w-wqx_3)}, \tag{55}$$

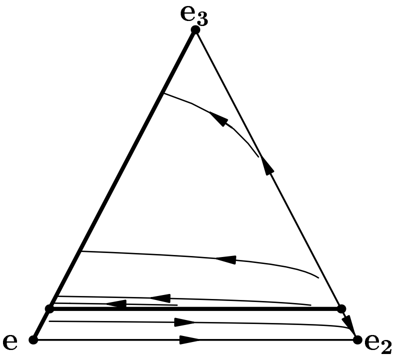

Figure 4. Phase portrait of the model described in Section 6 (eqs 29-31). We consider the same situation as for figure 2, but this time there is not a fixed number of rounds, but a probability, w, of a next round. The separatrix becomes a line of fixed points. The edge $\mathbf{e_1 e_2}$ is also a line of fixed points. Again there are two regions in phase space. If there are sufficiently many discriminators then defectors become eliminated, if the frequency of discriminators drops below a critical Level then defectors take over.

and hence

$$P_3 = P_1 + \frac{qx_2}{1-qx_3}[\frac{c-bqx_3}{2(1-w)} + \frac{b-c}{2-w-wqx_3}]. \tag{56}$$

It is obvious that $P_1 = 0$ holds iff

$$x_3 = \frac{c(2-w)}{bwq}. \tag{57}$$

A straightforward computation shows that for this x_3-value, $P_3 = 0$. Hence the fixed points of the corresponding replicator equation are (apart from the vertices of the simplex S_3) the edge $\mathbf{e_1 e_3}$ and the line l given by $bwqx_3 = c(2-w)$. This line divides the interior of S_3 into two regions: in one region, all orbits converge to $\mathbf{e_2}$, in the other region, towards a point on the $\mathbf{e_1 e_3}$-edge which depends on the initial value. This is exactly as in section 5 (see Fig.4).

Of course this holds only if the value of x_3 is less than 1, i.e. if $w(c+bq) > 2c$, in other words if the expected number of rounds, i.e. $(1-w)^{-1}$, satisfies

$$1/(1-w) > (bq+c)/(bq-c). \tag{58}$$

If we consider only the two strategies defector and discriminator, then discriminator can be evolutionarily stable only if

$$q > c/b. \tag{59}$$

This looks exactly like Hamilton's rule for altruism through kin selection, except that the coefficient of relatedness, k, is replaced by the probability to know the co-player's score, q.

9 Discussion

Several authors, starting with Trivers himself, have stressed that reciprocal altruism need not be restricted to dyads of interacting individuals (see Trivers, 1971, Boyd, 1988, Dugatkin et al , 1992, May, 1987, Axelrod and Dion, 1988, Binmore, 1992, and chapter 7 of Sugden, 1986, for instance.)

There are several ways to model generalised or indirect reciprocity. Alexander, who elaborated on the importance of this notion, did not fully specify the mechanisms involved, but mentioned several possibilities. One conceivable form of reward (see, e.g., Alexander, 1987, p.94) consists in having the success of the group contribute to the success of his own descendants, which is simply group selection in the modern sense, see Wilson and Sober (1994). One other form has been investigated by Boyd and Richerson (1989): individual A helps B, who helps C, who helps D, who finally returns the help to A. Thus individuals are arranged in closed, oriented loops, reminiscent of the hypercycles in the theory of Eigen and Schuster (1979) on catalytic loops of selfreplicating molecules. Boyd and Richerson investigate two strategies: upstream Tit For Tat (A keeps helping B if D keeps helping A) and downstream TFT (A keeps helping B if A observes that B keeps helping C). They find that the second type is much more efficient than the first, but that it is also more difficult to perform. (It should be noted that for two-member loops, both strategies reduce to Tit For Tat.) Boyd and Richerson conclude that this type of indirect reciprocity is less likely to evolve than pairwise reciprocity, and is only effective for relatively small, closed, long-lasting loops.

In a sense, this indirect reciprocity is still quite direct, and the social networks in human groups (or primates, for that matter – see de Waals, 1996) are much more fluid than the 'long-lasting loops' indicate. Alexander (1987) envisions a more diffuse mechanism when he stresses (p.85) that 'the return [of the beneficence] may come from essentially any individual or collection of individuals in the group', and emphasised the importance of assessment and status. We have tried to model this in Nowak and Sigmund (1998) by means of 'scores' assigned to each group member. If the model is reduced to the minimum (two scores only), we obtain the discriminator strategy.

The same strategy has been reached, through a different approach, in Pollock and Dugatkin (1992), who termed it Observer Tit For Tat. They studied it in the context of the repeated Prisoner‘s Dilemma, which is the usual framework for analysing direct reciprocity. Pollock and Dugatkin allowed the players to occasionally observe a co-player before starting the repeated interaction. If the future co-player was seen defecting in his last interaction, then Observer Tit For Tat prescribes to defect in the first round. Pollock and Dugatkin were mostly interesting in comparing this strategy with the usual Tit For Tat, but they also found that it could hold its own against defectors when no degree of future interaction with the current partner was presumed. They also obtained a condition similar to (53), but without modelling the different rounds in an individual's lifetime, and in particular without (52). The approach by Pollock and Dugatkin is truly remarkable. They did not aim at a model of indirect reciprocity, but actually investigated what Alexander would view as its prerequisite, namely 'direct reciprocity occurring in the presence of interested audiences' (Alexander, 1987, p.93), and came out with what we believe is the simplest strategy under which indirect reciprocity can be implemented – an unintended support for the correctness of Alexander's intuition.

The success of a discriminating player is somewhat hampered by the fact that whenever he refuses to help a B-scorer, he loses his G-score. A more sophisticated strategy has been studied by Sugden (1986) in a context which is only slightly different. In Sugden's model, in each round a randomly chosen player needs help, and each of the other players can provide some help (thus the needy player can get as payoff $(m-1)b$, where m is the group size). Sugden's T_1 strategy is based on the notion of standing: a player is born with good standing, and keeps it as long as he helps needy players who are in good standing. Such a player can therefore keep his good standing even when he defects, as long as the defection is directed at a player with bad standing (this in contrast to the discriminator strategy).

We believe that Sugden's strategy is a good approximation to how indirect reciprocity actually works in human communities: it offers, as Sugden remarks, a workable insurance principle. But as stressed in Boerlijst et al (1997) in connection with Contrite Tit For Tat, strategies based on standing are prone to be affected by errors in perception. If information is incomplete, then a player observed while withholding his help may be misunderstood; he may have defected on a player with good standing, or punished someone with bad standing. An eventual error can spread. The discriminator rule is less demanding on the player's capabilities, and still works. We expect that in actual human communities, indirect reciprocity is based on more complex reckonings, and believe that this should be amenable to experimental tests.

Finally, we mention that according to Zahavi (1995), Arabian babblers 'compete with each other to invest in the interests of the group, and often interfere with the helping of others'. This jostling for the position of the helper cannot be explained in terms of group selection, kin selection or direct reciprocation. However, if helping raises one's score and therefore one's fitness, this type of competition can easily be understood: indirect reciprocity based on image scoring provides a simple explanation.

Acknowledgement: We wish to thank Martin Posch, Immanuel Bomze, Josef Hofbauer, Robert May and Alex Kacelnik for helpful discussions. The FWF grant P11144 and the Wellcome Trust are gratefully acknowledged.

References

Alexander, R.D. (1979) *Darwinism and Human Affairs*, Univ. Washington Press, Seattle.

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

Axelrod, R. (1984) *The Evolution of Cooperation*, reprinted 1989 by Penguin, Harmondsworth.

Axelrod, R. and Hamilton, W.D. (1981) The evolution of cooperation, *Science* **211**, 1390-6.

Axelrod, R. and Dion, D. (1988) The further evolution of cooperation, *Science* **242**, 1385-90.

Binmore, K.G. (1992) *Fun and Games: a Text on Game Theory*, Heath and Co, Lexington, Massachussetts.

Boerlijst, M., Nowak, M.A. and Sigmund, K. (1997) The Logic of Contrition, *Journ. Theor. Biol.*, **185**, 281-293.

Bomze, I. (1983) Lotka-Volterra equations and replicator dynamics: A two dimensional classification. *Biological Cybernetics*, **48**, 201-211.

Boyd, R. (1988) Is the repeated Prisoner's Dilemma a good model of reciprocal altruism?, *Ethology and Sociobiology* **9**, 278-305.

Boyd, R. and Lorberbaum, J.P. (1987) No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma, *Nature*, **327**, 58-9.

Boyd, R. and Richerson, P.J. (1989) The evolution of indirect reciprocity, *Social Networks* **11**, 213-36.

de Waals, F. (1996) *Good natured: the origins of right and wrong in humans and other animals*, Harvard UP, Cambridge, Mass.

Dugatkin, L.A., Mesterton-Gibbons, M. and Houston, A.I. (1992) Beyond the Prisoner's Dilemma: towards models to discriminate among mechanisms of cooperation in nature, *TREE* **7**, 202-5.

Eigen, M., and Schuster, P. (1979) *The hypercycle: A principle of natural selforganization.* Berlin-Heidelberg: Springer.

Hamilton W.D. (1963) The evolution of altruistic behaviour, *American naturalist* **97**, 354-6.

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

Leimar, O. (1997) Repeated games: a state space approach, *Journ. Theor. Biol.*, **184**, 471-98.

Lindgren, K. (1991) Evolutionary phenomena in simple dynamics, in *Artificial life II* (ed. C.G. Langton et al), Santa Fe Institute for Studies in the Sciences of Complexity, Vol. X, 295-312.

May, R.M. (1987) More evolution of cooperation, *Nature* **327**, 15-17.

Maynard Smith, J. (1982) *The Theory of Games and evolution*, Cambridge UP.

Metz, J.A.J., Nisbet, R.M. and Geritz, S.A.H. (1992) How should we define fitness for general ecological scenarios?, *Ternds Evol. Ecol.* **7**, 198-202.

Mylius, S., Doebeli, M. and Diekmann, O. (1998) Can initial invasion dynamics correctly predict phenotypic substitutions?, preprint.

Nowak, M.A. and Sigmund, K. (1987) Oscillations in the evolution of reciprocity, *JTB* **137**, 21-26.

Nowak, M.A. and Sigmund, K. (1992) Tit for tat in heterogeneous populations, *Nature* **355**, 250-2.

Nowak, M.A. and Sigmund, K. (1993) Win-stay, lose-shift outperforms tit-for-tat, *Nature*, **364**, 56-8.

Nowak, M.A. and Sigmund, K. (1998) Evolution of indirect reciprocity by image scoring, *Nature*, in press.

Nowak M.A., May R. M., and Sigmund K. (1995) The Arithmetics of Mutual Help, *Scientific American* **272**, 76-81.

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

Sigmund, K. (1995) *Games of Life*, Penguin, Harmondsworth.

Sugden, R. (1986) *The evolution of rights, co-operation and welfare.* Blackwell, Oxford.

Trivers, R. (1971) The evolution of reciprocal altruism, Quarterly Review of Biology **46**, 35-57.

Trivers, R. (1985) *Social Evolution*, Menlo Park. CA, Benjamin Cummings.

Wilson, D.S. and Sober, E. (1994) Re-introducing group selection to human behavioural sciences, *Behavioural and Brain Sciences* **17**, 585-654.

Zahavi, A. (1995) Altruism as a handicap – the limitations of kin selection and reciprocity, *Journ. of Avian Biology* **26**, 1-3.