Indirect reciprocity, image-scoring and moral hazard

Hannelore Brandt and Karl Sigmund

September 21, 2004

'Give and it shall be given unto you'. But by whom? Luke (6.38) was not specific on that point. A helpul action, or a gift, can be returned by the recipient, in which case one speaks of direct reciprocation. But it can also be returned by a third party. Alexander (1987) called this 'indirect reciprocity', emphasising its reliance on status and reputation.

In a simple model, Nowak and Sigmund (1998) attached a binary score (*good* or *bad*) to each individual in the population. From time to time, two individuals meet randomly, one as donor, the other as recipient. At some cost $-c$ to the own payoff, the donor can help the recipient, i.e. increase the recipient's payoff by a benefit $b > c$. In that case, the donor's score will be *good* in the eyes of all observers, whereas the score of a 'donor' refusing to confer the benefit will be *bad*. A discriminating strategy of helping only those with a *good* score would channel benefits towards those who help, and discourage defectors. The question is whether such a strategy can evolve in the population, assuming that only strategies yielding a total payoff above average increase in frequency. This has attracted considerable attention, for two main reasons. One lies in the potential of indirect reciprocity for explaining the emergence, among humans, of cooperation between non-relatives. Alexander viewed this as the biological basis of morality; others saw in it a major motivation for language, gossip being a way of spreading reputations. The recent advent of e-commerce provides the other reason why understanding the assessment of reputations matters: the prevalence of anonymous one-shot interactions in global markets raises the issues of trust-building and moral hazard (Bolton et al 2002).

While economic experiments have strongly bolstered the concept of indirect reciprocity (see Wedekind and Milinski, 2000), the radically simplified model of Nowak and Sigmund has raised the skepticism of theoreticians (see Leimar and Hammerstein 2001). A discriminator who refuses to help recipients with a *bad* score receives a *bad* score, and risks to get no help in the next round. In this sense, punishing defectors by withholding help is costly. Can such a trait evolve? Would it not be advantageous to distinguish justifiable defections (against a *bad* recipient) from non-justifiable defections (against a *good* recipient), and attach a *bad* score only to the latter? This would constitute a non-costly form of punishment and greatly alleviate the discriminators' task. But such a distinction requires considerable cognitive capacities. Not only the recipient's past, but also that of the recipient's recipients etc must to be taken into account. If information spreads through rumour, rather than direct observation, the task may be alleviated, but the likelihood of misperception and manipulation grows. Conceivably, non-costly punishment cannot be realised; and many experiments show that humans, anyway, do not shrink from using costly punishment (Fehr and Fischbacher, 2003).

To return to theory, Ohtsuki and Iwasa (2004) analysed all conceivable strategies based on a binary score (some 4096 at first count) and found that eight of them are evolutionarily stable: all 'leading eight' differentiate between justifiable and non-justifiable defection. Nevertheless, the less sophisticated discrimination mechanism suggested by Nowak and Sigmund can promote cooperation if it leads to a stable mixture of discriminators and undiscriminate altruists. After all, a population need not be homogenous, although this is required for evolutionary stability. But Panchanathan and Boyd (2003) showed that in the presence of errors (or other causes for unintended defections, for instance lack of resources), such a mixture can be invaded by defectors. This blow was softened by Fishman (2004), who found that if the game extends over a constant number of rounds, the mixture of discriminating and undiscriminating altruists can repel defectors. But what is more likely, a constant number of rounds per lifetime or (as Panchanathan and Boyd assumed) a constant probability for a further round?

In fact, both assumptions appear unrealistic. Whereas in an experimental game all players may start at the same time and play their rounds synchronously, it seems plausible to assume that under natural conditions, players enter the population one by one, at random times, and interact asynchronously. The analysis of this model becomes even simpler, and boosts the conclusion of Panchanathan and Boyd.

Indeed, let denote by q the probability that a player knows the score of a randomly chosen co-player (either through direct observation or via gossip, through acquaintances) and that discriminators are trustful in the sense that if they have no information, they assume that their recipient's score is *good*. As simple calculation shows that whenever discriminating and undiscriminating altruists do equally well, defectors do just as well; which means that they will take over (see appendix, and figure). This is an extremely robust result, independent of the probability distribution of the number of rounds (which could also be constant, or infinite), and holding even if different strategies have different error probabilities, if discriminators are suspicious rather than trustful, or if they adopt the strategy of helping whenever the recipient's score is *good* or their own score is *bad*.

But there is a way out. It is based on an approach due to Mohtashemi and Mui (2004), who assumed in their model that whenever a donor provides help, the donor's set of acquaintances is added to the recipient's. We need not be so specific but only assume that a player's network of acquaintances grows with age. Then the probability q_n that a player in round n is informed about the recipient's image grows with n, i.e. $q_n > q_{n-1}$. It is easy to check that whenever the average level of information is sufficiently high, there exists a mixture of discriminating and undiscriminating altruists which is stable against defectors (see appendix, and figure).

We can, incidentally, also use the opposite condition $q_n < q_{n-1}$, if we correspondingly suppose that the discriminators are distrustful, and refuse to help in the absence of information. We do not claim that this is a reason why persons whose social circle shrinks (the very old, for instance) tend to become suspicious. But both mechanisms, intriguingly, imply that people should become more tightfisted with age.

References

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

2. Nowak, M.A. and Sigmund, K (1998) Nature 393, 573-577

3. Bolton, G., Katok, E. and Ockenfels, A (2004) Management Science, in press

4. Leimar, O. and Hammerstein, P. (2001) Proc R Soc Lond B, 268, 745-753

5. Wedekind, C. and Milinski, M. (2000) Science 288, 850

6. Fehr, E. and Fischbacher, U. (2003) Nature 425, 785-791

7. Ohtsuki H, and Iwasa, Y (2004), JTB, in press

8. Panchanathan, K. and R. Boyd (2003) JTB 224, 115-126

9. Fishman, M.A. 2003 JTB 225, 285-292

10. Mohtashemi, M. and L. Mui (2003) JTB 223, 523-531

Electronic appendix

Let us denote by x, y and z the relative frequencies of indiscrimate altruists, defectors and discriminators in the population, by $1-r$ the probability of an unintended defection, by g the frequency of players with a *good* score (which, if the population is sufficiently large, can be taken to be stationary throughout one individual's lifetime), and by q the probability that a player knows the score of a randomly chosen co-player (either through direct observation or via gossip). Let us also assume, for convenience, that each player, in each round, interacts once as a donor and once as a recipient (always with different co-players, of course). Finally, let us posit that discriminators are trustful in the sense that if they have no information, they assume that their recipient's score is *good*. The payoff in the n-th round ($n > 1$) for an indiscriminate altruist is $P_x(n) = -cr + brx + br(1-q)z + br^2qz$, for a defector $P_y(n) = brx + br(1-q)z$, and for a discriminator

$$P_z(n) = -cr(1-q+qg) + brx + br(1-q)z + br^2qz(1-q+qg).$$

The last term in the sum, for instance, is obtained as follows: the discriminating recipient meets with probability z another discriminator, who, with probability q, knows the recipient's score. If that score is *good*, the recipient receives the payoff b with probability r (since $1-r$ is the probability that the intended donation fails). The score is *good* if the recipient, in the previous round, succeeded in an intended donation (probability r), either not knowing the co-player's score (probability $1-q$), or else knowing the co-player's score (probability q), which was *good* (probability g).

A straightforward computation shows that

$$P_z(n) - P_y(n) = [P_x(n) - P_y(n)](1-q+qg).$$

The same relation holds for the first round, and hence also for the total payoff values P_x, P_y and P_z. The replicator dynamics on the unit simplex S_3 is given by $\dot{x} = x(P_x - \bar{P})$ etc, where $\bar{P} = xP_x + yP_y + zP_z$ is the average payoff in the population. The fixed points are the corners of S_3 (where the population consists of one type only) and all the points on the segment with $z = c/brq$. Initial states with lower z-value will converge to the equilibrium with $y = 1$ (defectors only). Initial states with larger z converge to the equilibrium with $y = 0$ and $z = c/brq$. There, an arbitrarily small random perturbation can send the state to a lower z-value. Hence defectors will always become fixed in the population.

But if we assume that the probability to know a co-player's score is not a constant, but depends on age and is denoted by q_n in round n, then

$$P_z(n) = -cr(1-q_n+q_ng) + brx + br(1-q)z + br^2qz(1-q_{n-1}+q_{n-1}g),$$

where q, now, is the average of the q_n (i.e. if w_n is the probability to be in round n, then $q = \sum w_n q_n$). If $q_n > q_{n-1}$ for all n, then $q > s := \sum w_n q_{n-1}$.

We note that

$$P_z(n) - P_y(n) = P_x(n) - P_y(n) + r(1-g)[cq_n - zbrqq_{n-1}]$$

and hence

$$P_z(n) - P_x(n) = r(1-g)(cq_n - zbrqq_{n-1}).$$

For the total payoffs P_x,P_y, and P_z we obtain

$$P_x(z_{cr}) = P_z(z_{cr})$$

for $z_{cr} := c/brs$ (we note that $z_{cr} > c/brq$, and assume in the following that $c < brs$, i.e. $z_{cr} < 1$).

The relation $P_x(n) - P_y(n) = -cr + br^2qz$ implies that for $z = z_{cr}$ one has $P_x(n) - P_y(n) = cr(q-s)/s$ for $n > 1$ (and $= -cr$ for $n = 1$). It follows that for sufficiently small w_1 (i.e. a sufficiently large likelihood of having more than one round)

$$P_x(z_{cr}) > P_y(z_{cr}).$$

Hence there exists a mixture of discriminating and indiscriminating altruists only, $\mathbf{F}_{xz} = (1 - z_{cr}, 0, z_{cr})$, which cannot be invaded by the defectors. The resulting replicator equation is bistable: one attractor consists of defectors only, the other of a mixture of discriminating and indiscriminating altruists.