Primer

Altruism

Karl Sigmund and Christoph Hauert

Math Inst. University of Vienna and Institute for Applied Systems Analysis, Laxenburg, Austria.

Altruistic actions are generally seen as 'noble'. Yet some 'lowly' organisms are apt to match the most heroic human acts of devotion and self-sacrifice. To use a widely-quoted example, consider *Dicrocoelium dendriticum*, also known as brainworm. These parasites spend some of their stages in the innards of cows, exit in the feces and, in the form of cercaria, are eaten by ants a few stages later. Once ingested, a gang of cercaria will break through the ant's stomach wall. One of them makes it to the brain of the ant, and causes it to climb on the tips of grass blades, thus exposing itself to be taken up by the grazing cattle. The other cercaria form cysts in the ant's body, ready to pursue their life-cycle within the cow that swallows them. But the one who made it to the brain — the 'brainworm' — dies without leaving offspring. It has effectively sacrificed itself for the survival of its gang. In humans, comparable feats would be the stuff of epic poetry.

Small wonder that evolutionary biologists feel challenged by such behaviour and see it as a high priority aim 'to take the altruism out of altruism'. To begin this task, they define the term 'altruism' in purely Darwinian terms, devoid of any moralistic undertones. An action performed by individual A and affecting individual B is said to be altruistic if it increases the fitness — the reproductive success — of the recipient B, and decreases the fitness of A. In this context, one may as well give names to the other possible scenarios: if the action increases the fitness of both A and B, one speaks of cooperation; of spite, if it decreases both fitnesses; and of selfishness if A's fitness is enhanced and B's fitness diminished. Both altruistic and spiteful traits lower the reproductive success of their bearers and seem at first inconsistent with the action of natural selection. Yet they abound.

The suicidal behaviour of the brainworm is a spectacular example of reproductive altruism. The other cercaria could not even reach their egg-laying stage otherwise. Further forms of reproductive altruism occur in the brood care of social insects. The worker castes consist of sterile individuals representing a dead end for the germ line. Nevertheless, it is their task to take care of the brood and even to commit suicide for the community's defence. In a wide variety of tropical bird species, unmated individuals — mostly males — help a breeding pair raise offspring. Similar alloparenting behaviour occurs among some fish and mammals, for instance cichlids, naked mole rats, or jackals.

The behaviour of foster-parents towards fledgling cuckoos falls also under the definition of reproductive altruism. Other frequently studied candidates for altruism are the grooming behaviour so widespread among animals, the alarm calls of birds and monkeys and the pronking — bouncing along on stiff legs — of gazelles. Yet another form of altruism is the restraint frequently observed by the winner of a contest. It lowers the winner's fitness as the loser may eventually return and win the next fight.

The most widespread reason for altruistic behaviour is doubtlessly kinship. The assistance provided to offspring by their parents has such an obvious value for the parents' own fitness that it actually hardly qualifies as altruism. But individuals have a genetic stake not only in their children, but also in their grandchildren, and indeed in all their relatives. This was already recognised by Darwin, although he could not ground it firmly in population genetics. A proper theory of kin selection was elaborated by W.D. Hamilton in the 1960s. Following Dawkins, it is usually explained from the gene's point of view. A gene increasing the propensity to help siblings, for instance, will promote individuals who are likely to bear copies of that gene and therefore favours its own spreading.

Hamilton's rule encapsulates this neatly: consider a gene causing its bearer A to perform an action towards a recipient B. This gene will increase in frequency if the relatedness between A and B, that is the probability that they share copies of that gene which are identical by descent, exceeds the ratio between fitness cost (to the bearer A) and fitness benefit (to the recipient B). In a phrase probably going back to J.B.S. Haldane, it 'pays' to die if this saves more than two brothers. The suicidal behaviour of the brain-worm, for instance, represents a trifling cost compared with the survival of the rest of the ingested cercaria, who are most likely sibs.

Hamilton's rule is only the tip of the iceberg of an elaborate theory based on the notion of inclusive fitness where one considers the effects of an action not only on the reproductive success of the actor, but also on that of all those affected, each weighted with the proper relatedness coefficient. Such computations are not easy, as they can lead to the trap of 'double counting', and are often plagued by the intricacies of inbreeding. An important theoretical tool is the Price equation, which expresses the increase in a gene's frequency in a neat formula describing selection within groups and selection among groups.

Kin selection led to remarkable progress in explaining social behaviour in communities with strong family ties, and allowed new levels of detail to be reached in studying conflicts of interest, for instance between brothers and sisters in hymenopteran insects like ants and bees, where males have no fathers and relatedness is not necessarily symmetric. Moreover, kin selection allows conflicts between generations, and in particular between parent and offspring to be addressed. Altruistic actions may come easily

to parents, but there remains the question of how to split the cost of the parental investment between mother and father, and how to divide the effort among the children, including those as yet unborn. Questions of this type have also greatly stimulated the study of life histories. How many seasons can a bird afford to serve its parents as a nest helper, for instance?

By now, thousands of instances of altruism based on kin selection have been documented, and almost every social interaction trait is routinely analysed in terms of relatedness. As a side effect, such investigations, often using genetic markers, have drawn attention to mechanisms of kin recognition among animals. In many cases they prove to be amazingly efficient, but sometimes fail and leave the door open for exploitation, most spectacularly by the cuckoo and other brood parasites. Further examples of misdirected brood care are furnished by enslaved worker ants.

After so much progress with kin selection, it seems today difficult to conceive that helping behaviour among relatives could ever have been viewed as a challenge to Darwinism. The appearence of altruism was based on sloppy reasoning, such as ignoring inclusive fitness effects. This agrees well with the vernacular view, where nepotism is not seen as altruism, but as a subversive form of selfishness.

What remains a challenge is altruism among non-relatives. The founding event in this field was a paper by Trivers expounding reciprocal altruism. This occurs when altruistic acts, for which the cost is less than the benefit, are repeatedly exchanged between two individuals. Both, then, get a net increase in fitness. This type of altruism, summarized in the principle 'you scratch my back and I'll scratch yours', is based on economic rather than genetic ties, and seems much more fragile. The return of an altruistic move towards a relative is immediate, as the genetic correlation is already in place. But an eventual return through reciprocation can usually take place only later. Such altruistic acts are speculative investments in the future. It will often be easy, for the recipient, to cheat. This quandary is traditionally modelled by the Prisoner's Dilemma game, although usually based on simultaneous actions by two players. If both cooperate and help another, both get a reward; but if only one player helps the other, this player will incur a cost in fitness, and the other, the defector, will experience a benefit without having to pay for the return. No matter whether the other player cooperates or defects, it is best to defect in this game.

So how can reciprocal altruism evolve? Trivers showed that if the interaction was repeated sufficiently often, cooperative strategies based on reciprocation could persist: it would not pay to defect if there was a high probability for another round of the game, giving the other player an opportunity to retaliate (the 'shadow of the future'). Following a landmark paper by Axelrod and Hamilton, the theory of reciprocal altruism was extensively developed, often boosted by computer simulations modelling the evolution of strategies in the iterated Prisoner's Dilemma game based on selection-mutation chronicles of artificial populations.

The basic result here is that when the interaction is sufficiently likely to be long-lasting, a small cluster of players using retaliatory strategies can invade a population of defectors. A minority of defectors, on the other hand, cannot invade a population of stern retaliators. However, such a population can be weakened by the spreading of indiscriminate altruists, who may pave the way for the successful return of defectors exploiting their lack of a defensive mechanism.

Unfortunately, the empirical evidence for reciprocal altruism lags behind theory. A handful of examples have kept circulating for years: predator inspection by fish where for example two stickle-backs help each other if they jointly approach a predatory pike. Egg trading by hermaphrodite fish who switch between the role of the male and the more expensive role of the female several times during a sexual encounter. And mouth-to-mouth feeding by vampire bats; frequently, a hungry bat receives some of the cattle's blood regurgitated by a well-fed bat.

Today, the dearth of non-human examples of reciprocal altruism appears striking. Apparently, the requisites for retaliatory strategies — long series of interactions between the same two individuals, and the cognitive capabilities to identify partners and remember their actions — are hard to find outside primate communities. Among primates, however, many forms of reciprocal altruism are documented, for instance within coalitions of young males. In particular, the human tendency to reciprocation is evident in daily life, and has been verified by many experiments based on iterations of the Prisoner's Dilemma game. It must be noted in this context that the simplest evolutionary mechanisms can lead to a certain percentage of altruism — even if the Prisoner's Dilemma is not repeated — provided the players interact, not with randomly chosen individuals, but only with their nearest neighbors. Quite generally, altruism seems to be favoured by any form of population viscosity, even in the absence of reciprocation.

One speaks of direct reciprocation if the return of the help offered by a donor is provided by the recipient, and of indirect reciprocation if the return is obtained through third parties — 'give and ye shall be given'. Indirect reciprocation has been touted as the basis of all moral systems. Several behavioural rules have been proposed to implement indirect reciprocation, essentially based on status considerations. If the score of an individual increases whenever that individual performs an act of help, and decreases if the individual refuses to help, then a strategy of helping only those with a high score effectively channels altruism

towards other altruists, and hence discriminates against defectors.

Such strategies, and variants thereof, have been shown to operate in experiments with human groups. In computer simulations, however, such indirect reciprocation is often threatened by the emergence, first of do-gooders giving indiscriminately to all comers, and then of non-reciprocating exploiters. The reason is that discriminators refusing to help low-scorers thereby lower their own score, and hence their chance of being helped. Stabler regimes of discriminating altruism can be obtained through strategies which take into account whether a potential recipient's former refusals to help have been directed toward invertebrate exploiters or toward deserving altruists. This, however, requires considerable cognitive sophistication, and plenty of information about the past of the other group members.

So far we have considered interactions mostly between two individuals. Reciprocal altruism becomes considerably more difficult to sustain within larger groups. This is well evidenced in so-called public goods games, a staple of experimental economy. Four players receive twenty dollars each, and decide how much of the sum to invest in the common pool, keeping the rest to themselves. The experimenter then doubles the content of the pool and divides it evenly among the four players, irrespective of their contribution. If all players contribute fully, they double their payoff. But for each player, each dollar invested yields only a return of fifty cents, so that the selfish strategy is to invest nothing. In actual experiments, humans invest a considerable part of their endowment. After several rounds of this game, however, most contributions have dropped to zero. The reason seems to be that the only possibility of retaliating against players who contribute less than average is to reduce one's own contributions, which effectively unravels cooperation. The picture changes drastically if, after every round of the public goods game, players have the opportunity of inflicting fines against specific co-players. The fines do not go to the punishers, and hence punishment is an unselfish activity. On the contrary, imposing a fine is costly to the punisher. Nevertheless, one observes a great propensity to punish free riders, together with strong moralistic aggression. Players of such a public goods game obviously anticipate this: they contribute more, and if the game is repeated for a few rounds, the contributions actually increase.

Punishing strategies to maintain cooperative behaviour are well know in animal societies, for instance among wasps, naked mole rats and chimpanzees. If it serves to ensure future benefits to the punisher, costly punishment can be interpreted as a selfish act. However, recent experiments on humans have shown strong punishing behaviour even if the possibility of future benefits, through another round of the public goods game for example, is excluded. This is usually called altruistic punishment, although in the strict sense it falls under the definition of spite, as it lowers the fitness of both parties involved. Whether such socially beneficial instances of spite occur in non-human communities remains to be seen.

Experiments have revealed that in human groups, reward can play a role similar to punishment: if, between the rounds of a public goods game, two players are chosen randomly and one can donate a gift to the other, the probability that this will happen increases with the contributions of the recipient in the public goods game; and this stabilises the average contributions to the common pool on a high level.

In each case, altruistic actions are greatly boosted if they are broadcast within the group. According to the handicap principle, this can be viewed as a costly signal of an individual's fitness. The evolution of cognitive abilities and language — as evidenced in the role of gossip — must have fostered the score-keeping behind direct and indirect reciprocation, and facilitated the emergence of moralistic emotions such as sympathy, guilt, anger, conscience — our 'good nature', in the words of de Waal. Eventually, this may have led to the emotional appreciation of altruism as something precious and noble.

Key references

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

Axelrod, R. (1984). The Evolution of Cooperation. (Basic Books, New York).

Boyd, R. and Richerson, P.J. (1988). The Evolution of Reciprocity in Sizeable Groups. J. Theor. Biol. *132*, 337–356.

Clutton-Brock, T.H. and Parker, G.A. (1995). Punishment in animal societies. Nature *373*, 209–216.

Dawkins, R. (1989). The Selfish Gene. (Oxford University Press).

Dugatkin, L.A. (1997). Cooperation among animals: an evolutionary perspective. (Oxford University Press).

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans, Nature *415*, 137–140.

Frank, S.A. (1998). Foundations of Social Evolution. (Princeton University Press).

Hamilton, W.D. (1996). (2002). Narrow Roads of Gene Land. (vol I and vol II), Freeman, New York.

Leimar, O. and Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity, Proc. R. Soc. Lond. B *268*, 745–753.

Milinski, M., Semmann, D. and Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons'. Nature *415*, 424–426.

Nowak, M.A. and May, R.M. (1992). Evolutionary Games and Spatial Chaos. Nature *359*, 826–829.

Nowak, M.A. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. Nature *393*, 573–577.

Ridley, M. (1996). The Origin of Virtue. (Penguin, Harmondsworth).

Sigmund, K., Hauert, Ch. and Nowak, M.A. (2001). Reward and Punishment. Proc. Natl. Acad. Sci. U.S.A. *98*, 10757–10762.

Sober, E. and Wilson D. S. (1999). Unto Others: The Evolution and Psychology of Unselfish Behavior. (Harvard University Press).

Trivers, R. (1985). Social evolution. (Benjamin Cummings, Menlo Park, California).

Trivers, R.L. (1971). The evolution of reciprocal altruism. Q. Rev. Biol. *46*, 35–57.

de Waal, F. (1996). Good natured. (Harvard University Press).

Wedekind, C. and Milinski, M. (2000). Cooperation through image scoring in humans, Science *288*, 850–852.

Wilson, E.O. (1975). Sociobiology. (Harvard UP).

Zahavi, A. and Zahavi, A. (1997). The Handicap Principle. (Oxford UP).

Acknowledgements

C. H. acknowledges the support of the Swiss National Science Foundation.