DOI: 10.2436/20.1501.02.87

Cent cinquanta anys després de *L'origen de les espècies*, de Darwin
(Arcadi Navarro i Carmen Segarra, ed.)

*Treballs de la SCB. Vol. 60 (2009) 213-229*

# THE EVOLUTION OF COOPERATION

Mauro Santos [1] i Eörs Szathmáry [2,3,4]

[1] *Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona.*
[2] *Collegium Budapest, Institute for Advanced Study, Hungary.*
[3] *Institute of Biology, Eötvös University, Hungary.*
[4] *Parmenides Center for the Study of Thinking, Germany.*

Adreça per a la correspondència: Mauro Santos. Departament de Genètica
i de Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona.
Campus de Bellaterra. 08193 Bellaterra. Adreça electrònica: *mauro.santos@uab.es*.

## RESUM

L'aproximació a l'evolució centrada en el gen o del *gen egoista* aparentment entra en conflicte amb l'observació que la cooperació és freqüent en les interaccions socials humanes i també es pot reconèixer en animals no humans. Sense cooperació no haurien pogut sorgir les unitats evolutives de nivells superiors. Aquí resumim el pensament evolutiu actual sobre com poden evolucionar la cooperació i l'altruisme. A més, discutim els resultats dels experiments de la teoria de jocs per estudiar les interaccions socials que indiquen que els humans no s'ajusten a les prediccions de l'equilibri de Nash «racional». Aquests resultats són de gran interès per als biòlegs i científics socials, especialment si es desitja tenir un marc de referència comú per entendre com sorgeix la sociabilitat.

**Paraules clau:** altruisme, cooperació, teoria de jocs, nivells de selecció, transicions evolutives principals.

## SUMMARY

The gene-centred or selfish-gene approach to evolution apparently conflicts with the observation that cooperation is commonplace in human social interactions, and can also be recognized in non-human animals. Without cooperation, higher-level units of evolution could not have emerged. Here we summarize current evolutionary thinking on how cooperation and altruism can evolve. We also discuss the results reached by game theoretic experiments for studying social interactions, which indicate that humans do not conform to Nash equilibrium ("rational") predictions. These results are of wide interest to biologists and social scientists, particularly if we want to have a common framework to understand how sociality arises.

**Key words:** altruism, cooperation, game theory, levels of selection, major evolutionary transitions.

## INTRODUCTION

The catchy phrase "survival of the fittest" is a metaphor widely used in popular literature to stress the idea that fierce competition prevails in the natural world, and has been taken (erroneously) as a short description of the Darwinian theory of evolution by natural selection. Although Darwin did think of natural selection as a process that allows the evolution of traits that directly and solely benefit the individual possessing them (e.g., sharp teeth, high speed, visual acuity, warning coloration, etc.), he also occasionally considered that characters may be selected for because they are advantageous to the "community" or the family. Thus, in the *Origin of Species* Darwin evokes the idea that certain instincts that lead to the death of the individual (e.g., the instinct that drives the bee to sting and thus to die) could have evolved by natural selection because they were "useful to the community" (Darwin, 1859). In the *Descent of Man* a similar idea is used to explain the development of certain moral virtues, such as courage, obedience and faithfulness (Darwin, 1871, p. 166):

> "It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an advancement in the standard of morality and an increase in the number of well-endowed men will certainly give an immense advantage to one tribe over another. There can be no doubt that a tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to give aid to each other and to sacrifice themselves for the common good, would be victorious over most tribes; and this would be natural selection".

Therefore, the typical notion laypeople have about Darwin's theory, in the sense that it provides a justification for any behavior that promotes selfishness and undermines moral standards, is simply wrong. Having said this, however, it must be acknowledged that Darwin's concept of selection was mainly individualistic and he never discussed how important the selection among groups ("tribes") could have been in shaping the history of life. The social insect case can be mentioned here because the advantage to the family is mentioned.

From the preceding paragraphs it is clear that the evolution of cooperation contrasts sharply with the concept of selection for traits that solely benefit the individual possessing them. Cooperation here is understood as a *behavior that involves an action performed by one individual that benefits one or more other members, embedded in a situation that poses a dilemma for the individual engaged in a social interaction*. Facing a social dilemma, when should a person cooperate? When should a person be selfish? Should we punish Russia for its hostile act against Georgia? (Admittedly, this is a bit of a problem since now we know that the Russian aggression was provoked by the Georgians and the initial reaction of many newspapers was to overlook that fact.)

## PLAYING THE GAME

An intuitive understanding in these types of situations can be obtained by using a particular game called the Prisoner's Dilemma. In its classical form, the version of this game goes as follows (e.g. Rice, 2004). Two people are suspected of having committed a joint crime. The suspects are confined to different rooms and cannot talk to each other. The police do not have sufficient evidence to convince a jury. The state attorney offers each of the suspects a deal: if one testifies for the prosecution and confesses ("defects") against the other and the other remains silent, the betrayer goes free and the silent accomplice receives the full sentence of ten years. If both confess, they will each receive seven years. If neither of them confesses, then both will go free after one year in prison for a minor charge because nothing can be proved. This dilemma is represented by the following matrix, where the rows refer to the action (i.e., remain silent or confess) chosen by one prisoner (say P1) and the columns to the action taken by the other prisoner (say P2):

| | | Prisoner P2 | |
|---|---|---|---|
| | | Remain silent | Confess |
| Prisoner P1 | Remain silent | P1 = –1,<br>P2 = –1 | P1 = –10,<br>P2 = 0 |
| | Confess | P1 = 0,<br>P2 = –10 | P1 = –7,<br>P2 = –7 |

Negative numbers obviously signify that a convict loses years of their life in jail according to each other's decision. How should the prisoners act? A formal approach to the problem is to use game theory, which has become the basic framework to study situations that involve strategic interactions. Assume for instance that you are prisoner P1. The game can be analyzed by considering your best strategy against the decision taken by prisoner P2. If he remains silent, it is best for you to confess and go free (P1 = 0, P2 = –10). If he confesses, you also better confess to avoid a longer sentence (P1 = –7, P2 = –7). Therefore, no matter what he does, the best strategy for you is to confess. A similar reasoning applies to prisoner P2, thus the best strategy is always to confess although it is clear that by doing so, both prisoners will suffer a sentence that is six years longer than that obtained if both remain silent. The strategy confess-confess defines a "Nash equilibrium": namely, a strategy profile such that each prisoner's play is the best response to that of the other. Cooperating (remain silent) is strictly dominated by defecting (confess).

The fact that the rational pursuit of individual interest apparently drives both partners to an inferior profile helps account for the enduring interest in the Prisoner's Dilemma game. The conclusion appears to violate the gut feeling intuition that interactions among individuals usually lead to mutually beneficial outcomes; in other words, the two players in the Prisoner's Dilemma game should apparently cooperate to minimize their mutual loss (remain silent and both of them will spend only one year in prison), but this would not be a logical decision. What is the significance of all this to biology anyway?

It turns out that this problem is as old as the origin of life itself ["cooperators since life began" is the title of an insightful review by Queller (1997) in *The Quarterly Review of Biology*]. The evolution of biological complexity often requires the coordinate action ("cooperation") of different parts that function to ensure the survival and reproduction of the whole. Thus, earlier replicating molecules had to cooperate to form the first cells. During the emergence of multicellular organisms, single cells had to cooperate and today they exist only as parts of a larger individual. Eusociality (found primarily in social insects but also in some other organisms) is a level of social organization where members cooperate in brood care, and are separated into reproductive and non-reproductive castes. On a larger scale, humans cooperate in raising states and countries that they themselves engage in cooperative commercial and communication activities. Yet, our understanding of the origin and maintenance of cooperation at any level is fraught with difficulties. Genes in an organism sometimes "disagree" and, for example, can have a beneficial effect on females but a detrimental one on males. Cancers are selfish cell lineages, namely, clones of cells in a multicellular organism that have evolved a higher rate of division compared to other non-cancerous lineages. And the Achilles' heel in human cooperation was clearly appreciated by Thomas Hobbes when he wrote in the *Leviathan* (1651, chap. xvii):

> "For the laws of nature, as justice, equity, modesty, mercy, and, in sum, doing to others as we would be done to, of themselves, without the terror of some power to cause them to be observed, are contrary to our natural passions, that carry us to partiality, pride, revenge, and the like. And covenants, without the sword, are but words and of no strength to secure a man at all. Therefore, notwithstanding the laws of nature (which every one hath then kept, when he has the will to keep them, when he can do it safely), if there be no power erected, or not great enough for our security, every man will and may lawfully rely on his own strength and art for caution against all other men".

In other words, Hobbes argued that individuals would find it mutually beneficial to agree in order to restrain their "natural" tendencies toward deception. However, after the social contract any party would experience the incentive to violate the agreement and eventually would do so to enjoy an immediate benefit. The only satisfactory remedy to this situation of affairs, according to Hobbes, is the introduction of a coercive state with a monopoly on the use of force to which individuals would submit to voluntarily.

What humans normally do when faced with the Prisoner's Dilemma game? This issue can be explored by using experimental methods that allow the study of behavior in situations where different considerations may be controlled for. Notwithstanding the predictions suggested by game theory analysis, the results of various experiments that have been carried out so far show unambiguously that a majority of subjects (about 60%) achieved mutual cooperation. This is jointly better for the players than mutual defection, although defection is always individually superior. When Nash first heard about these results, he wrote the following note (Field, 2001, p. 5): "I would have thought them (*the players*) more rational". This comment raises the important issue about what it means for humans to be rational.

## EXPLAINING COOPERATION

Given the problem of cooperation, how can cooperative behaviors arise and be maintained? Conventional (individualistic) Darwinian theory seems unable to really explain why one individual should pay a cost in terms of fitness in order to benefit another individual. The reason is simple: an altruistic individual will leave less offspring than its selfish counterparts, so any inherited tendency to be altruistic will decrease in frequency and eventually disappear from the population. Apparently, the only way out to explain the evolution of cooperation is to restore the idea of selection between groups, as Darwin put forward. But in the 1960s a few scientists challenged most biologists' positive attitude towards the importance of group selection, and stressed that natural selection was intrinsically selfish and that cooperative acts could only evolve under restrictive conditions (Hamilton, 1964; Williams, 1966).

Before explaining the solution offered by Hamilton (1964) for the evolution of altruism, it pays to summarize here the dissatisfaction felt by some theoreticians with the idea of group selection (for a comprehensive account, see Wilson and Wilson, 2007). The following passage by Williams (1966, pp. 92-93) gives a flavor of the intellectual holes in the theory of group selection at that time:

> "This book is a rejoinder to those who have questioned the adequacy of the traditional (*individualistic*) model of natural selection to explain evolutionary adaptation… Many biologists have implied, and a moderate number have explicitly maintained, that groups of interacting individuals may be adaptively organized in such a way that individual interests are compromised by a functional subordination to group interests.
>
> It is universally conceded by those who have seriously concerned themselves with this problem […] that such group-related adaptations must be attributed to the natural selection of alternative groups of individuals and that the natural selection of alternative alleles within populations will be opposed to this development. I am in entire agreement with the reasoning behind this conclusion. Only by a theory of between-group selection could we achieve a scientific explanation of group-related adaptations. However, I would question the premises on which the reasoning is based. Chapters 5 to 8 will be primarily a defense of the thesis that group-related adaptations do not, in fact, exist. A group in this discussion should be understood to mean something other than a family and to be composed of individuals that need not be closely related".

Here Williams punches in the so-called traditional view of group selection. Together with various theoretical analysis [in particular, Maynard Smith's (1964) "haystack model" and latter, more refined models (e.g. Eshel, 1972)], it was clear that traits that are disadvantageous to the individual, but that lower the probability of group extinction (recall Darwin's paragraph in the Introduction), can persevere only if the population is structured into extremely isolated groups. This result suggests that the likelihood of successful group selection is small.

Note that at the end of the preceding quotation Williams wrote, "A group […] should be understood […] to be composed of individuals that need not be closely related." This point is important and has been hidden in some recent theoretical analyses of group selection (see below). In 1932 and again in 1955, J. B. S. Haldane, one of the founders on the modern genetic theory of evolution, pointed out that an individual's genes could also be (indirectly) multiplied through the actions performed by such individual in favoring the differential survival and reproduction of collateral relatives (e.g., siblings, nieces, and cousins) to sufficient degree (Haldane, 1932, 1955). This initial insight was expanded by William D. (Bill) Hamilton (born August 1, 1936; died March 7, 2000), one of the most influential Darwinian thinkers of our time. Hamilton wondered: how natural selection can maintain (altruistic) traits that, even if advantageous for the population, apparently reduce the number of their carriers below their shares in the population? The answer he gave came from extending the classical population genetics theory to cover those situations where the altruistic behaviour is directed towards genetic relatives: an inherited behaviour is selected for in a population *if and only if* it results in an increase in the number of genes identical by descent (i.e., those gene copies that are shared between relatives according to Mendelian rules) to those of the individual that performs the behaviour. Thus, whenever the altruistic action is directed toward relatives the behavior can spread throughout the population by natural selection. The conditions under which the behavior increases are encapsulated in Hamilton's (1964) rule [we are sympathetic with Hawking's "fault" of including Einstein's celebrated equation $E = mc^2$ in his popular-science book *A Brief History of Time* (1988) and also include here the most famous inequality in evolutionary biology, which happens to have the same number of parameters]:

$$rb > c$$

where $r$ is the coefficient of relatedness (roughly speaking, the probability of sharing gene copies) between actor and recipient, $b$ is the fitness benefit (offspring gain) provided to the recipient, and $c$ is the reproductive cost to the actor for providing benefits. In colloquial language, the rule states that it pays to risk your life if by performing an action, you save the life of your two children. Any mother will surely agree with this, and the reasons for this are not only "psychological". Haldane (1955) put it is this way: "I will risk death to save my child from a raging river if the odds are at least two to one that I will succeed (because she shares roughly half my genes); but I will jump in the river to save my cousin only if the odds favoring success are seven in eight (because she has only one-eight of my genes). Trying to save my grandmother makes no sense at all because, being past child-bearing age, she can pass on none of my genes to the next generation."

Hamilton's rule to explain the evolution of altruistic traits became one of the central dogmas of the modern theory of evolution, and his approach is known as "kin selection" or "inclusive fitness theory" (albeit technically speaking these labels are not strictly equivalent). Hamilton's paper on "the genetical evolution of social behaviour" (1964) also became one of the most cited papers in all science (4,282 citations according to a recent search on Web of Science). Due to the combined influence of Hamilton's, Maynard Smith's and Williams' work, all previous explanations of social behaviors that have evolved because of their importance for the well-being of the group fell into widespread disrepute and were eradicated from the mainstream of evolutionary thinking. Group-selection became a taboo expression that has, however, not been expunged entirely from evolutionary theory, and remains vindicated in some quarters although with several meanings. For instance, Sober and Wilson have been passionate supporters of group selection and have attacked in their book *Unto Others* (1998) the widespread consensus that group selection is a negligible evolutionary process.

However, the traditional view of group selection advocated in particular by Williams (1966) and Maynard Smith (1964), is not what Sober and Wilson had in mind. A more recent view of group selection is that populations are frequently subdivided into small temporary groups, termed "structured-deme" or "trait-group" models (Wilson 1975, 1979), and that the subdivision can result in group-mediated evolutionary change (Hamilton, 1975; Wilson, 1975).

Figure 1 illustrates the "mating pool" mode of reproduction, where offspring genotypes at each generation are sampled from a common population and subdivided into temporary groups in which individuals representing a finite sample from the pooled distribution reproduce in proportion to their fitness. A single heritable trait (allele *A*) stimulates its bearer to provide a group benefit, and its eventual spread is conditional on the greater productivity of altruistic groups. But now things become (technically) complicated because different definitions of altruism have been used throughout the literature. Hamilton consistently defined altruism as a cooperative behavior that decreases the *absolute* fitness (i.e., involves an absolute cost in terms of offspring production) of the individual that performs it. However, various authors (including Sober and Wilson, 1998) refer to a given behavior as "altruistic" even if it decreases the *relative* fitness of the agent (i.e., the cost in terms of offspring production is relative to the local group and not to the population as a whole). This apparently harmless disparity in the definition of altruism has profound consequences on the eventual evolutionary fate of the behavior under consideration (also in the eventual role of group selection to explaining human cooperation; see Concluding Remarks). Sober and Wilson's argue that a behavior counts as altruistic if, first, within any group individuals that perform the behavior are less fit than those that do not; and second, the greater the proportion of individuals that perform the behavior in a given group, the greater the group's fitness (i.e., the greater the contribution of the group to the mating pool). These conditions do not imply that "altruism" reduces the donor's absolute fitness; namely, they do not imply that the behavior is really altruistic. This is an important point, but the reasons for it are unfortunately somewhat too technical to be explained here and the interested reader can refer to Okasha (2006, pp. 192-197) for a lucid discussion.

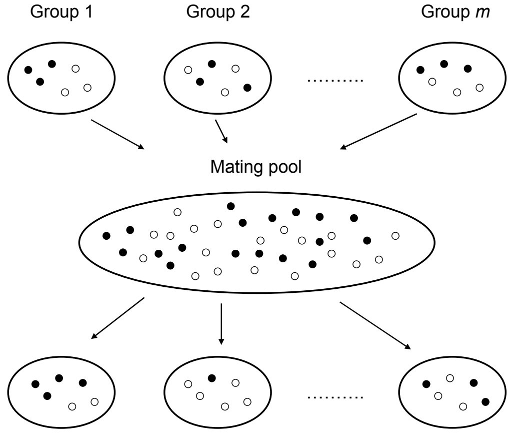

FIGURE 1. The "mating pool" mode of reproduction. At each generation, selection acts on groups of individuals (open and closed circles represent different genotypes) that contribute offspring to a common pool which again forms new groups before selection.

On the other hand, Hamilton's definition implies that a behavior is really altruistic and can evolve in a structured-deme model *if and only if* there is positive assortment for the benefits of altruism to fall preferentially on other altruists (i.e., altruists settle with altruists; Hamilton, 1975). Positive assortment guarantees altruist's mutual aid and Hamilton's inclusive fitness rule is recovered.

However, although assortative grouping may explain how altruism is sustained in the population it is not an explanation for its origin. It is important to avoid the "inverse genetic fallacy" (Field, 2001). Namely, that the inappropriate attribution of mechanisms may be sustaining cooperation to the explanation of its origin. The initial barrier to selection in Hamilton's model arises because positive assortment is negligible when altruism is initially very rare. The barrier must be passed by a different mechanism (see Santos and Szathmáry, 2008).

## LEVELS OF SELECTION

A particular illuminating formulation that can be used to understand the relative merits that individual and group selection mechanisms have on the evolution of cooperation is that of Price (1970, 1972). His method involves a simple algebraic expression that describes a population's evolution from one generation to another. Actually, Hamilton's (1975) paper of group selection was fostered by Price's idea, and included a section entitled "Levels of Selection" where Hamilton made use of Price's equation for the first time to formally study a hierarchical approach to the operation of natural selection (he assumed a mating pool mode of reproduction). This approach has become standard in evolutionary theory ever since and can be generalized to an indefinite number of hierarchical levels, an issue that underlines many disputes about group selection (the debate of group selection has been characterized by perennial disagreements over concepts and terminology).

In the last years, there has been an ongoing dispute on the equivalence of kin selection and levels of selection approaches for modeling social evolution. However, both approaches are indeed equivalent and the confusion, as shown by Bijma and Wade (2008), is partly caused by the fact that levels of selection models tend to hide the relatedness component of response to selection, whereas kin selection models tend to hide the multilevel selection component of response to selection. Thus, kin selection does not work without multiple groups of kin.

A key empirical breakthrough in evolutionary biology was to recognize that individual (selfish) organisms that adapt to the environment by natural selection originated as cooperative collectives that experienced a series of "transitions in individuality" (Buss, 1987). In the biological progression from molecule to cell to body to species, the units at each level tend to be nested into ever more inclusive units: a testament of the importance of cooperation in the evolution of life on Earth. Using a traditional view of natural selection, it is hard to understand how this hierarchical organization—with potential conflicts between the different units (genes, chromosomes, organelles, cells, etc.; see Burt and Trivers, 2006)—has evolved. From the growing body of work on "major evolutionary transitions" it has become clear that a multi-level selection scenario is central to explaining those transitions.

In their book *The Major Transitions in Evolution*, Maynard Smith and Szathmáry (1995, p. 3) wrote:

> "Our thesis is that the increase (*in biological complexity*) has depended on a small number of major transitions in the way in which genetic information is transmitted between generations. Some of these transitions were unique: for example, the origin of the eukaryotes from the prokaryotes […]. Other transitions, such as the origin of multicellularity, and of animal societies, have occurred several times independently".

And on page 8 they state:

> "The transitions must be explained in

> terms of immediate selective advantage to individual replicators: we are committed to the gene-centered approach outline by Williams (1966), and made still more explicit by Dawkins (1976)".

Does this mean that Maynard Smith and Szathmáry excluded (old) group selection? Not really! For Maynard Smith and Szathmáry the gene-centered view is a heuristic perspective—"I find the gene-centered approach both mathema-tically simpler and causally more appropriate, but this may merely reflect the fact that I prefer microscopic to holistic models: Maxwell-Boltzmann to classical thermodynamics, and Dawkins to Price's equation" (Maynard Smith, 2002, p. 523)—, not an empirical hypothesis concerning the path of evolution.

It is also illuminating here to recall what Williams wrote himself in 1996 (p. xii) in the preface to his 1966 reprinted book:

> "A few years after 1966, I was being given credit for showing that the adaptation concept was not usually applicable at the population or higher levels, and that Wynne-Edwards's thesis that group selection regularly leads to regulation of population density by individual restrains on reproduction was without merit. It also became fashionable to cite my work (sometimes, I suspect, by people who had not read it) as showing that effective selection above the individual level can be ruled out. My recollection, and my current interpretation of the text, especially of Chapter 4, indicates that this is a misreading. I concluded merely that group selection was not strong enough to produce what I termed *biotic adaptation*: any complex mechanism clearly designed to augment the success of a population or a more inclusive group. A biotic adaptation would be characterized by organisms' playing roles that would subordinate their individual interests for some higher value, as in the often proposed benefit to the species".

The question therefore is: do we have biotic adaptations in the evolution of biological complexity that need to be explained by group selection? Amazingly enough, the very origin of life may have required group selection!

Today every *autonomous* living system is cellular (prokaryotic or eukaryotic, uni- or multicellular) in nature, as advocated by the Schwann-Schleiden cell theory in the mid-nineteenth century. A working definition of "minimal life" system is a chemical supersystem comprising three systems: a metabolic network, template replication, and a boundary system (Gánti, 2003; Szathmáry *et al.*, 2005). A model that satisfies these criteria is the "stochastic corrector model" (SCM; Szathmáry and Demeter, 1987; Zintzaras *et al.*, 2002; Santos *et al.*, 2003), which describes the dynamics of independent replicators (genes: chromosomes were a latter development in the origin of life) encapsulated in a reproductive vesicle or compartment (protocell). The behavior of the (super-) system depends on a two-level selection dynamics (figure 2). First, there is within-compartment (among-replicator) selection, where replicators compete among them within compartments to increase their relative shares (i.e., "the immediate selective advantage" required for Maynard Smith and Szathmáry to explain the major evolutionary transitions). Second, between-compartment (among-protocell) selection on stochastically produced offspring variants after protocell fission can rescue the population from extinction (favoring cooperative molecules). Since protocells in this model are groups of independent replicating entities (genes), the SCM explicitly invokes group (among-protocell) selection. On biological grounds the ingredient of group selection was automatically guaranteed once compartments met the criteria for "minimal life". The SCM was advocated by Maynard Smith and Szathmáry (1995) since it opens great possibilities for continued evolution. For instance, it was specifically used to model the evolutionary origin of chromosomes (Maynard Smith and Szathmáry, 1993), to study how the danger of information decay in primitive genetic systems could have been overcome (Zintzaras *et al*., 2002), to investigate issues bearing on the origin of sex (Santos *et al*., 2003), and to test the potential effect of recombination to circumvent the crisis of information transmission in early evolution (Santos *et al*., 2004).

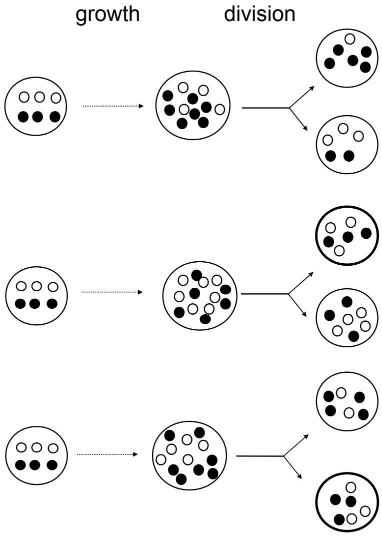

FIGURE 2. A "minimal life" model as depicted by the SCM. Different templates (labeled by open and closed circles) contribute to the well being of the compartments (protocells) in that they catalyze steps of metabolism, for example. During protocell growth templates replicate at differential expected rates, but stochastically. Upon division there is chance assortment of templates into offspring compartments. Stochastic replication and reassortment generate variation among protocells, on which natural selection at the compartment level can act and oppose (correct) internal deterioration due to within-cell competition.

The moral here is clear: Maynard Smith (1964) advanced the "haystack model" to show the unlikelihood of (old) group selection (although he did not dismiss it altogether), but later relied upon group selection to understand some major evolutionary transitions. In fact, in the preface to their book Maynard Smith and Szathmáry (1995) acknowledge that the (incomplete) similarity between the haystack and the SCM was one of the impetuses behind writing it. If the SCM can be taken seriously, we must conclude that group selection played a vital role in the origin of life. Obviously, this should not be taken to mean that (old) group selection is a widespread mechanism to explain cooperation. Perhaps the situation here is similar to that in the origin of eukaryotes: there is no doubt that mitochondria and chloroplasts are descended from endosymbiotic events that may have occurred very few times throughout the history of life, but the consequences of these very unlikely events have clearly been spectacular.

## EVOLUTIONARY GAME THEORY

Kin selection was not the only alternative explanation of cooperation that emerged out of discontent with traditional group selection. Another idea, which could explain altruism among non-relatives, was the evolutionary game theory introduced by Maynard Smith and Price (1973) although closely related ideas were developed by Trivers (1971) and Axelrod and Hamilton (1981). Evolutionary game theory means that the fitness of the individuals is not constant, but depends on the relative frequencies of various strategies in the population; in other words, the optimal thing for an individual to do critically depends on what others are doing at the same time (like in parlor games). The outcome of the game is related to reproductive success, and the payoffs determine fitness.

Trivers (1971) suggested that the evolution of cooperative behavior could be understood in terms of reciprocal aid given in repeated interactions between two partners. As discussed above, in a non-repeated Prisoner's Dilemma it is best to defect no matter which strategy is adopted by the other player. But if there are repeated encounters between the same two individuals, the idea "I help you and you help me" can lead to the evolution of cooperation. Axelrod (1984) made some important contributions to debates about the conditions under which cooperation can emerge under repeated iterations from his computer tournaments. He invited game theorists and other social and behavioral scientists familiar with the Prisoner's Dilemma to submit strategies for a multiplayer tournament. In the first of these contests each strategy was paired with itself, with a random strategy, and with each of the other submitted strategies: each strategy pair playing two hundred rounds. Strategies were scored according to the total number of points earned in all these pairings, with 3 points each for mutual cooperation, a 5/0 split where one defected and one cooperated, and 1 point each if both defected (i.e., a record of mutual cooperation throughout two hundred plays of pairing earned 600 points; while a record of 1,000 points could be attained by continuously defecting when the partner played cooperation in all plays, thus earning 0 points). How should a rational player have selected a strategy, assuming all other players are rational?

Because both players know that the game will end after two hundred rounds, we can use backward induction. Thus, there is no incentive to cooperate in the last round because the same analysis applies as for the one-shot Prisoner's Dilemma. Since we "know" what will happen on the last round, defection on the first-to-last round is also the "rational" approach to take. Since we also "know" this, defection is the correct move on the second-to-last round, and so for. The only strict Nash equilibrium is "always defect" (ALLD). However, the winning strategy, submitted by game theorist Anatol Rapoport, was "tit-for-tat" (TFT, which cooperates in the first stage and then does what the other player has done in the previous round). Why did this strategy perform so well? The reason is simple: even though ALLD wins TFT in individual pairings, the aggregate score of the former strategy is meager. When paired with TFT, ALLD wins on the first play but after that there is a continuous defection (TFT scores 199 and ALLD scores 204). When ALLD is paired with ALLD, both will receive a score of 200 at the end. When TFT is paired with TFT, their final score will be 600. Thus, the results of this tournament seem to suggest that cooperation can indeed evolve by direct reciprocity.

But there are two problems with Axelrod's original tournaments. The first is that in real-world situations, unlike in an error-free digital universe, individuals incur in mistakes (Nowak and Sigmund, 1994). In the presence of mistakes, two TFT players can achieve a low payoff. A single mistake moves the game from mutual cooperation to alternating between cooperation and defection, and a second mistake can lead to mutual defection (TFT cannot correct mistakes). But even if we stick to the error-free digital universe, a second problem arises when interpreting the results. Both Hamilton's inclusive fitness and Trivers' reciprocal altruism as in the iterated Prisoner's Dilemma rely fundamentally on the same principle: the positive assortment of helping behaviors (Fletcher and Zwick, 2006). This can perhaps be easily understood using Hamilton's (1975) mating pool mode of reproduction (figure 1).

Assume the simplest situation of randomly formed temporary groups, with 2 individuals that interact in the one-shot Prisoner's Dilemma game and then are pooled back to the global population. The process is repeated again for successive "generations" (encounters) of the game. Assume a population with the two strategies, ALLD and TFT, initially at the same frequency. As indicated earlier, the outcome of the game is related to reproductive success, and the payoffs (the same as those given by Axelrod) determine fitness. At "generation 1" three different groups are formed: ALLD-ALLD, ALLD-TFT, and TFT-TFT at frequencies 0.25, 0.5, and 0.25, respectively. It is easy to see that after the first interaction the average payoff of ALLD is 0.5 × 1 + 0.5 × 5 = 3 (i.e., in ALLD-ALLD groups both partners will be scored 1 and in ALLD-TFT groups ALLD will be scored 5), and that of TFT is 0.5 × 0 + 0.5 × 3 = 1.5. The scores at "generation 2" (and at later generations) are somewhat more difficult to calculate because TFT is a conditional cooperator. It means that after the first encounter with TFT it will continue cooperating, but after the first encounter with ALLD it will defect. Without going into the details, a simple computer program shows that eventually ALLD replaces TFT, so defection prevails. This is exactly what Hamilton (1975) concluded: with random group formation (random encounters at each generation) altruism cannot progress, and positive selection of altruism is only possible when altruists tend to settle with altruists. Note that under purely positive assortment, only two groups are formed at "generation 1": ALLD-ALLD and TFT-TFT at frequencies 0.5 and 0.5, respectively. It then becomes clear that TFT wins because it will have a higher score than ALLD. By letting TFT-TFT play for two hundred rounds in a row, Axelrod's computer tournaments guaranteed positive assortment and cooperation succeeded. In summary, reciprocal altruism does not seem to be fundamentally different from inclusive fitness or multilevel selection.

## WHY DO HUMANS COOPERATE?

Kin selection can explain why cooperation in the animal world is largely associated with family life (although a number of authors have questioned the primarily relevance of the relatedness component in the evolution of eusociality; Gadagkar, 2001; Wilson and Hölldobler, 2005). However, human behavior is unique in that cooperation occurs in large societies composed of many unrelated individuals. It is unlikely that reciprocal altruism might explain close cooperation in moderately large groups of a few hundred or a few thousand people (not to speak of millions in our large-scale societies) since most encounters between unrelated individuals will happen only once or a very few times during a lifespan. For instance, even in the relatively small network of scientific co-authorship (collaboration), there is a "small world" component—namely, the tendency for scientists to publish with only a few fellows (Albert and Barabási, 2002). Human societies are thus an interesting special case to study the evolution of cooperation.

It is essential to summarize here the contribution of game theory to the design of experiments beyond the Prisoner's Dilemma aimed at understanding how individuals behave when engaged in strategic interactions. Perhaps the single most important point to make is that both *altruistic cooperation* and *altruistic punishment* (helping others at a cost to oneself) are often observed, particularly when individuals face a social dilemma. The reasons why we observe extensive cooperation in humans are the subject of some speculation, but "they come down to the plausible insight that human social life is so complex, and the rewards for prosocial behavior so distant and indistinct, that adherence to general rules of propriety, including the strict control over such deadly sins of anger, avarice, gluttony, and lust is individually fitness-enhancing" (Gintis, 2008, p. 50). In all the games described below subjects are anonymous to each other, they are college students who are recruited by bulletin board and newspaper announcement, and they are paid real money. They are instructed to fully understand the rules and the payoffs of the games. A more comprehensive account can be found in Gintis (2008, chap. 3).

### Ultimatum game

A player called the "proposer" is given a sum of money, say €10, and is instructed to give the second player, called the "receiver", any amount from €1 to €10. The responder can either accept or reject the offer. The catch is that if the responder rejects the offer, neither receives anything. The Nash equilibrium is to accept the minimum amount, so a "rational" proposer will give €1 and a "rational" receiver will accept the offer. What happens in fact is that proposers routinely offer responders a substantial share (50% being the modal offer), and responders frequently reject offers below 30%. These experiments have been run in various countries around the world, including some where the amount of money at stake is substantial, and the results are similar. Apparently most people prefer nothing to something that is perceived as unfair, and punish the proposer accordingly.

### Dictator game

Here a player called the "dictator" receives a sum of money but, in contrast to an ultimatum game, after offering, the dictator keeps whatever amount of money he does not choose to give to the recipient. Obviously, a "rational" dictator will give nothing to the recipient. The results show that although some subjects will indeed keep all of the money, not all will. Actually, a frequent outcome in some recent experiments was to split the money down with offers varying between 50% and 0%, with about 20% of subjects keeping all the money. These results provide clear confirmation of altruistic and "irrational" (from a game theoretical analysis) behavior among humans.

### Public goods game

This is an *n*-person game that typically involves a group of subjects each of which is given an initial sum of money, say €10, and told they can either keep it ("private account") or deposit it in a group ("common") account. The experimenter then increases the common account by, say, doubling it and then equally distributes the resulting amount of money among the group. The best solution for the group is for everyone to put €10, enabling each subject to double the initial quantity. But the unique Nash equilibrium for each player is to put nothing, in which case everyone ends up with the original amount. Each individual following the Nash strategy hopes to keep his or her private account and free ride on the voluntary contributions of the rest. Contributing to the common account is risky since, in the worst case of only one contributor, this (altruistic) subject will end up with substantially less money than the initial €10. When the game is played for a number of rounds, the results show that the average initial contribution of the individuals to the common account continually decreases along the time period involved (say ten rounds), clearly indicating that cooperation declines. However, when cooperators are given the opportunity to punish defectors, cooperation is sustained or even increases during the game. These experiments give some credibility to the suggestion that altruistic punishment may be an important ingredient to sustaining cooperation in human societies (e.g., Fehr and Gächter, 2002; Boyd *et al.*, 2003; Rockenbach and Milinski, 2006), although there is a huge cross-societal variation (see Herrmann *et al.*, 2008).

## CONCLUDING REMARKS

Strong reciprocity theorists advocate that most people do genuinely care about the welfare of others, and we agree. However, the role of group selection in human evolution is open to strong criticisms. Darwin's potential solution for social groups to function as an adaptive evolutionary unit (Darwin, 1871, p. 166; see Introduction) is a two-sided coin. The "nice" face is within-group cooperation (altruism), but the other side is hostility towards individuals from other groups ("parochialism"). Darwin thus recognized war as a powerful evolutionary force that might foster social solidarity among fellow members of one's group. Hamilton's speculation about how this could occur were discussed in the same paper in which he introduced Price's approach to understand the relative merits that individual and group selection mechanisms have on the evolution of cooperation (Hamilton, 1975); also known as Hamilton's "Fascist paper" (Hamilton, 1996, pp. 316-317):

> "Robert Trivers was later to refer to the article that I contributed to Fox's volume as my 'Fascist paper'. I believe he was referring not mainly to his own impression but rather to what others were saying about it and particularly to one strong response by a noted anthropologist, S. L. Washburn, in which, singling my paper out of the whole volume, he called it 'reductionist, racist, and ridiculous'. Washburn obviously didn't like the whole thing but focused his objection on some of my more speculative paragraphs concerning the warlike propensities of pastoral people and the possible involvement of these in current trends in the histories of Old World civilizations…
>
> Be that as it may, whether my paper really is racist or absurd I will leave to the reader's judgment and just comment that I have hardly changed my opinions and certainly haven't with regard to what Washburn indicated as the most offending passage".

We are clearly moving on shaky grounds here, but cannot oversee the empirical importance of both altruism and hostility to members of other groups. A recent game-theoretic analysis suggests that under conditions likely to have been experienced by late Pleistocene and early Holocene humans, both altruism and parochialism could have evolved jointly by promoting group conflict. This coevolution apparently helps explain why group boundaries have such a powerful influence on human behavior (Choi and Bowles, 2007). Choi and Bowles are cautious enough and on page 640 stress that "we have not shown that a warlike genetic predisposition exists, only that should one exist, it might have coevolved with altruism and warfare in the way we have described".

We will not question here the results of this work, or Bowles' (2006) empirically-based support for Darwin's argument that a possible explanation for the evolution of human altruism is that groups with more altruists survive when groups compete among them. It should be stressed, however, that the definition of altruism they used is relative to the local group, and not to the population as a whole. This leads to the confusing situation where cooperation can be favored because it provides a direct benefit to the cooperator. This issue is related to the previously discussed difference between Hamilton's *vs*. Sober and Wilson's definitions of altruism. A detailed analysis of this problem and how the previous conclusions can potentially change would be extremely useful.

Our final remark relates to the tricky question at the end of the introduction: Should we punish Russia for the hostile act against Georgia? To some commentators Georgia's invasion could lead the World to a similar state of affairs than those during the Cold War. It thus might be interesting to realize here that superpower relations at that time were considered by some as a sequential rather than a simultaneous move in a Prisoner's Dilemma game: either player had the option of moving first in an eventual nuclear conflagration (the defect-defect Nash equilibrium). In this case, the best option was a first strike. Few people are aware that John von Neumann (the father of game theory) and Bertrand Russell both advocated a first strike: "we attack at one o'clock (it is now 12:59), and the enemy is either too devastated, too demoralized, or too rational to strike back (defect-cooperate). But if for whatever reason … the attack provokes retaliation, nuclear exchange (defect-defect) is still preferable to an unprovoked attack (cooperate-defect)" (Field, 2001, p. 169). Would this outcome have been without a dramatic cost? One cannot avoid thinking that sometimes it may be a relief to know that politicians are irrational!

## ACKNOWLEDGMENTS

We thank Enrico Rezende and Lluís Serra for helpful comments on an earlier draft. MS is supported by grants CGL2006-13423-01/BOS from the Ministerio de Ciencia y Tecnología (Spain) and 2005SGR 00995 from Generalitat de Catalunya to the Grup de Biología Evolutiva, ES is supported by the National Office for Research and Technology (NAP 2005/KCKHA005) in Hungary. MS and ES are supported by funding of the INCORE project under the Sixth Research Framework Programme of the European Union. The views expressed in this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

## BIBLIOGRAPHY

Albert, R.; Barabási, A.-L. (2002). «Statistical mechanics of complex networks». *Rev. Mod. Phys.*, 74: 47-97.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Axelrod, R.; Hamilton, W. D. (1981). «The evolution of cooperation». *Science*, 211: 1390-1396.

Bijma, P.; Wade, M. J. (2008). «The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection». *J. Evol. Biol.*, 21: 1175-1188.

Bowles, S. (2006). «Group competition, reproductive levelling, and the evolution of human altruism». *Science*, 314: 1569-1572.

Boyd, R.; Gintis, H.; Bowles, S.; Richerson, P. J. (2003). «The evolution of altruistic punishment». *Proc. Natl. Acad. Sci. USA*, 100: 3531-3535.

Burt, A.; Trivers, R. (2006). *Genes in conflict. The biology of selfish genetic elements*. Cambridge: The Belknap Press of Harvard University Press.

Buss, L. W. (1987). *The evolution of individuality*. Princeton: Princeton University Press.

Choi, J.-K.; Bowles, S. (2007). «The coevolution of parochial altruism and war». *Science*, 318: 636-640.

Darwin, C. (1859). *On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life*. London: Murray.

— (1871). *The descent of man, and selection in relation to sex*. 2 vol. London: Murray.

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.

Eshel, I. (1972). «On the neighbor effect and the evolution of altruistic traits». *Theor. Pop. Biol.*, 3: 258-277.

Fehr, E.; Gächter, S. (2002). «Altruistic punishment in humans». *Nature*, 415: 137-140.

Field, A. J. (2001). *Altruistically inclined? The behavioral sciences, evolutionary theory, and the origins of reciprocity*. Ann Arbor: The University of Michigan Press.

Fletcher, J. A.; Zwick, M. (2006). «Unifying the theories of inclusive fitness and reciprocal altruism». *Am. Nat.*, 168: 252-262.

Gadagkar, R. (2001). *The social biology of Ropalidia marginata*. Cambridge: Harvard University Press.

Gánti, T. (2003). *The principles of life*. Oxford: Oxford University Press.

Gintis, H. (2008). *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton: Princeton University Press. [In press]

Haldane, J. B. S. (1932). *The causes of evolution*. London: Longmans.

— (1955). «Population genetics». *New Biology*, 18: 34-51.

Hamilton, W. D. (1964). «The genetical evolution of social behaviour, I & II». *J. Theor. Biol.*, 7: 1-52.

— (1975). «Innate social aptitudes of man: an approach from evolutionary genetics». In: Fox, R. [ed.]. *Biosocial Anthropology*. London: Malaby Press, 133-153.

— (1996). *Narrow roads of gene land. The collected papers of W. D. Hamilton. Vol. 1. Evolution of social behaviour*. Oxford: W. H. Freeman.

Hawking, S. W. (1988). *A brief history of time. From the Big Bang to black holes*. New York: Bantam Books.

Herrmann, B.; Thöni, C.; Gächter, S. (2008). «Antisocial punishment across societies». *Science*, 319: 1362-1367.

Hobbes, T. (1651). *Leviathan, or the matter, forme, and power of a common-wealth ecclesiastical and civil*. Indianapolis: Hackett Publishing, 1994.

Maynard Smith, J. (1964). «Group selection and kin selection». *Nature*, 201: 1145-1147.

— (2002). «Commentary on Kerr and Godfrey-Smith». *Biol. Philos.*, 17: 523-527.

Maynard Smith, J; Price, G.R. (1973). «Logic of animal conflict». *Nature*, 246: 15-18.

Maynard Smith, J; Szathmáry, E. (1993). «The origins of chromosomes I. Selection for linkage». *J. Theor. Biol.*, 164: 437-446.

— (1995). *The major transition in evolution*. Oxford: Oxford University Press.

Nowak, M. A.; Sigmund, K. (1994). «The alternating prisoner's dilemma». *J. Theor. Biol.*, 168: 219-226.

Okasha, S. (2006). *Evolution and the levels of selection*. Oxford: Clarendon Press.

Price, G. R. (1970). «Selection and covariance». *Nature*, 227: 520-521.

— (1972). «Extension of covariance selection mathematics». *Ann. Hum. Genet.*, 35: 485-490.

Queller, D. C. (1997). «Cooperators since life began». *Q. Rev. Biol.*, 72: 184-188.

Rice, S. H. (2004). *Evolutionary theory*. Massachusetts: Sinauer Associates.

Rockenbach, B.; Milinski, M. (2006). «The efficient interaction of indirect reciprocity and costly punishment». *Nature*, 444: 718-723.

Santos, M.; Szathmáry, E. (2008). «Genetic hitchhiking can promote the initial spread of strong altruism». *BMC Evol. Biol.*, 8: 281.

Santos, M.; Zintzaras, E.; Szathmáry, E. (2003). «Origin of sex revisited». *Orig. Life Evol. Biosph.*, 33: 405-432.

— (2004). «Recombination in primeval genomes: a step forward but still a long leap from maintaining a sizeable genome». *J. Mol. Evol.*, 59: 507-519.

Sober, E.; Wilson, D. S. (1998). *Unto others. The evolution and psycology of unselfish behavior*. Cambridge: Harvard University Press.

Szathmáry, E.; Demeter, L. (1987). «Group selection of early replicators and the origin of life». *J. Theor. Biol.*, 128: 463-486.

Szathmáry, E.; Santos, M.; Fernando, C. (2005). «Evolutionary potential and requirements for minimal protocells». *Top. Curr. Chem.*, 259: 167-211.

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

Wilson, D. S. (1975). «A theory of group selection». *Proc. Natl. Acad. Sci. USA*, 72: 143-146.

— (1979). «Structured demes and trait-group variation». *Am. Nat.*, 113: 606-610.

Wilson, D. S.; Wilson, E. O. (2007). «Rethinking the theoretical foundation of sociobiology». *Q. Rev. Biol.,* 82: 327-348.

Wilson, E. O.; Hölldobler, B. (2005). «Eusociality: Origin and consequences». *Proc. Natl. Acad. Sci. USA*, 102: 13367-13371.

Williams, G. C. (1966). *Adaptation and natural selection. A critique of some current evolutionary thought*. Princeton: Princeton University Press.

— (1996). Preface to *Adaptation and natural selection. A critique of some current evolutionary thought*. Princeton: Princeton University Press, 1966.

Zintzaras, E.; Santos, M.; Szathmáry, E. (2002). «"Living" under the challenge of information decay: the stochastic corrector model vs. hypercycles». *J. Theor. Biol.*, 217: 167-181.