Heylighen F. (1992) : "Evolution, Selfishness and Cooperation", *Journal of Ideas*, Vol 2, # 4, pp 70-76.
Heylighen F. (1992) : "'Selfish' Memes and the Evolution of Cooperation", *Journal of Ideas* , Vol. 2, #4, pp 77-84.

Evolution, Selfishness and Cooperation

Francis HEYLIGHEN*

PO-PESP, Free University of Brussels, Pleinlaan 2, B-1050 Brussels, Belgium
E-mail: fheyligh@vnet3.vub.ac.be

Abstract. It is argued that replicators evolving through natural selection on the basis of fitness are intrinsically selfish. Though the synergy resulting from cooperation is generally advantageous, selfish or subsystem optimization precludes the reaching of a globally optimal cooperative arrangement. This predicament is exemplified by the "Prisoner's dilemma". Different proposals to explain the evolution of cooperation are reviewed: kin selection, group selection, reciprocal altruism ("tit for tat"), and moralism. It is concluded that the proposed mechanisms are either too limited in scope, unstable, or insufficiently detailed, and that the analysis must therefore go beyond the level of purely genetic evolution if human "ultrasociality" is to be explained.

KEYWORDS: sociobiology, cooperation, evolution, altruism, game theory, group selection, inclusive fitness, ethics.

I. Introduction

A fundamental problem in founding an evolutionary ethics is to explain how cooperation and altruism can emerge during evolution (Campbell, 1979). (Such an ethics forms one of the main parts of the evolutionary philosophy that is being developed in the Principia Cybernetica Project, cf. Heylighen, Joslyn & Turchin, 1991; Heylighen, 1991c). The evolutionary principle of "the survival of the fittest" seems to predispose individuals to selfishness. Yet all ethical systems emphasize the essential value of helping others. Everybody will agree that cooperation is in general advantageous for the group of cooperators as a whole, even though it may curb some individual's freedom. Highly developed systems of cooperation and mutual support can be found in all human societies. Yet we still do not have a satisfactory explanation of how such social systems have emerged. Therefore we also cannot determine how they would or should evolve in the future.

Perhaps the most fashionable approach to this problem is *sociobiology* (Wilson, 1975). Sociobiology can be defined as an attempt to explain the social behavior of animals and humans on the basis of biological evolution. For example, a lot of sexual behavior can be understood through mechanisms of genetic selection reinforcing certain roles or patterns. Yet the biggest problem of

[*] Senior Research Assistant NFWO (Belgian National Fund for Scientific Research)

social behavior, altruism and advanced cooperation, has not been adequately solved.

"Weak" altruism can be defined as behavior that benefits more to another individual than to the individual carrying out the behavior. "Strong" altruism denotes behavior that benefits others, but at one's own cost (Campbell, 1983). Both are common and necessary in those highly cooperative systems, which Campbell calls "ultrasocial". Ultrasociality refers to a collective organization with full division of labor, including individuals who gather no food but are fed by others, or who are prepared to sacrifice themselves for the defense of others. In the animal world, ultrasocial systems are found only in certain species of insects (ants, bees, termites), in naked mole rats, and in human society (Campbell, 1983). In spite of the many parallelisms between human society on the one hand, and insect or mole rat societies on the other hand, their development was caused by quite different mechanisms, as we will see.

The present paper will first analyse the evolutionary tendency toward selfishness, and the benefits and pitfalls of cooperation. Then a review will be offered of the different proposals made in sociobiology and related domains to explain the emergence of cooperation through natural selection. It will be concluded that none of the explanations is sufficient, though each of them has specific strengths. In a following paper (Heylighen, 1992b), a new model will be proposed which synthesizes all the advantages of the previous explanations, but without their disadvantages. This model will be based on the concept of a *meme* as replicating unit of cultural evolution. The present paper will mainly set the stage for that subsequent paper by reviewing the basic concepts and mechanisms necessary to understand the evolution of cooperation.

II. The natural selection of selfishness

We will assume that general evolution takes place through blind variation and natural selection (Heylighen, 1991a,b,c, 1992). This includes all processes of development and evolution, at the biological, as well as at the physical, chemical, psychological or social levels.

Natural selection can be defined as the survival or, more precisely, the selective retention or maintenance of the *fittest* system or configuration. Fitness corresponds in general to the probability of encountering the same or a similar system in the future. Systems have a high fitness if they are stable (they tend to maintain for a long time), and/or they leave many offspring when they disappear, that is if they have produced many other systems that can somehow be considered as copies or replicas of themselves. Such self-reproducing systems are called *replicators* (Dawkins, 1976; Csanyi, 1991; Csanyi & Kampis, 1985). Natural selection means that systems which have insufficient fitness, because they are unstable and do not produce offspring, are eliminated from the natural scene. This process of selective elimination occurs spontaneously and continuously.

The ever present variation, which implies that even stable systems do undergo small changes, or produce slightly different offspring, leads to a continuously renewed variety of configurations undergoing selection. Since at each stage or generation the least fit systems tend to be eliminated, the process of evolution leads to a generally increasing fitness of the remaining systems (at least relative to their competitors). The systems resulting from such a process will have maximal fitness as their *implicit goal*, in the sense that systems whose behavior is not directed at optimizing fitness simply won't maintain. Let us analyse this in more detail.

In general, systems that replicate need *resources* (building blocks, energy, space, ...) in order to build copies of themselves. Resources are normally limited. Since each replicator tries to produce a maximum of copies, it will also attempt to use these resources to the limit. However, if more than one replicator is using the same resources, there will be a situation of *competition* or conflict. Slight differences in fitness between the competitors will be exacerbated, since the more efficient replicator will gradually succeed in using more and more of the resources, leaving less and less for the less efficient one. In the long term, nothing will be left for the less fit one, with the result that only the fittest will survive.

Let us now define selfishness and altruism in these more abstract terms. Altruism means that a system performs actions that increase the fitness of another system using the same resources. Selfishness means that the system will only perform actions that increase its own fitness. Our analysis of evolution entails that naturally selected systems will not only be selfish, since they try to optimize their own fitness, but they will also tend to avoid altruism. Indeed, "helping" a competitor to increase his fitness entails that the competitor will be able to use more resources, and hence less resources will be left for one's own offspring. Thus, one's own fitness is indirectly reduced by altruistic behavior (in the case of strong altruism the reduction is even direct). Even worse, if the system would actively hinder or attack its competitor, this could increase its own fitness, since it would diminish the competitor's use of resources. (under the condition that the act of attacking and its risks, such as being wounded or killed, would not reduce the fitness more than what can be gained by thwarting the competitor.)

III. Cooperation and the prisoner's dilemma

The picture of naturally selected systems we have sketched here is rather grim. Yet it is not difficult to see that cooperation can have selective advantages. The above reasoning assumes that resources are finite, and that an individual system would be able to exhaust them on its own. In practice, the amount of resources that can be exhausted by any single system is only a small fraction of the total amount of resources potentially available. For example, a single wolf can only kill relatively small prey, such as rabbits or pheasants. A pack of wolves, on the other hand, is also able to kill large prey such as a moose or a deer. The amount of "reachable" resources, in the sense of meat available for feeding the wolves and thus keeping them alive and able to reproduce, is much larger in the second case. Cooperation, in this case among the wolves, can create a *synergy* which strongly extends the set of reachable resources.

This principle can be made more explicit by introducing some concepts from *game theory* (Maynard Smith, 1982; Axelrod, 1984). A game is an interaction or exchange between two (or more) actors, where each actor attempts to optimize a certain variable by choosing his actions (or "moves") towards the other actor in such a way that he could expect a maximum gain, depending on the other's response. One traditionally distinguishes two types of games. *Zero-sum* games are games where the amount of "winnable goods" (or resources in our terminology) is fixed. Whatever is gained by one actor, is therefore lost by the other actor: the sum of gained (positive) and lost (negative) is zero. This corresponds to the situation of competition we sketched in the preceding section.

Chess, for example, is a zero-sum game: it is impossible for both players to win (or to lose). Monopoly (if it is not played with the intention of having just one winner) on the other hand, is a non-zero-sum game: all participants can win property from the "bank". In principle, in monopoly, two players could reach an agreement and help each other in gathering a maximum amount from the bank. That is not really the intention of the game, but I hope I have made the distinction clear: in non-zero-sum games the total amount gained is variable, and so both players may win (or lose). The phenomenon of synergy sketched in the beginning of this section belongs to this category.

Cooperation is usually analysed in game theory by means of a non-zero-sum game called the "Prisoner's Dilemma" (Axelrod, 1984). The two players in the game can choose between two moves, either "cooperate" or "defect". The idea is that each player gains when both cooperate, but if only one of them cooperates, the other one, who defects, will gain more. If both defect, both lose (or gain very little) but not as much as the "cheated" cooperator whose cooperation is not returned. The whole game situation and its different outcomes can be summarized by table 1, where hypothetical "points" are given as an example of how the differences in result might be quantified.

Action of A \ Action of B	**Cooperate**	**Defect**
Cooperate	Fairly good [+ 5]	Bad [- 10]
Defect	Good [+ 10]	Mediocre [0]

Table 1: outcomes for actor A (in words, and in hypothetical "points") depending on the combination of A's action and B's action, in the "prisoner's dilemma" game situation. A similar scheme applies to the outcomes for B.

The game got its name from the following hypothetical situation: imagine two criminals arrested under the suspicion of having committed a crime together. However, the police does not have sufficient proof in order to have them convicted. The two prisoners are isolated from each other, and the police visit each of them and offer a deal: the one who offers evidence against the other one will be freed. If none of them accepts the offer, they are in fact cooperating against the police, and both of them will get only a small punishment because of lack of proof. They both gain. However, if one of them betrays the other one, by confessing to the police, the defector will gain more, since he is freed; the one who remained silent, on the other hand, will receive the full punishment, since he did not help the police, and there is sufficient proof. If both betray, both will be punished, but less severely than if they had refused to talk. The dilemma resides in the fact that each prisoner has a choice between only two options, but cannot make a good decision without knowing what the other one will do.

Such a distribution of losses and gains seems natural for many situations, since the cooperator whose action is not returned will lose resources to the defector, without either of them being able to collect the additional gain coming from the "synergy" of their cooperation. For simplicity we might consider the Prisoner's dilemma as zero-sum insofar as there is no mutual cooperation: either each gets 0 when both defect, or when one of them cooperates, the defector gets + 10, and the cooperator - 10, in total 0. On the other hand, if both cooperate the resulting synergy creates an additional gain that makes the sum positive: each of them gets 5, in total 10.

The gain for mutual cooperation (5) in the prisoner's dilemma is kept smaller than the gain for one-sided defection (10), so that there would always be a "temptation" to defect. This assumption is not generally valid. For example, it is easy to imagine that two wolves together would be able to kill an animal that is more than twice as large as the largest one each of them might have killed on his own. Even if an altruistic wolf would kill a rabbit and give it to another wolf, and the other wolf would do nothing in return, the selfish wolf would still have less to eat than if he had helped his companion to kill a deer. Yet we will assume that the synergistic effect is smaller than the gains made by defection (i.e. letting someone help you without doing anything in return).

This is realistic if we take into account the fact that the synergy usually only gets its full power after a long term process of mutual cooperation (hunting a deer is a quite time-consuming and complicated business). The prisoner's dilemma is meant to study short term decision-making where the actors do not have any specific expectations about future interactions or collaborations (as is the case in the original situation of the jailed criminals). This is the normal situation during blind-variation-and-selective-retention evolution. Long term cooperations can only evolve after short term ones have been selected: evolution is cumulative, adding small improvements upon small improvements, but without blindly making major jumps.

The problem with the prisoner's dilemma is that if both decision-makers were purely rational, they would never cooperate. Indeed, rational decision-making means that you make the decision which is best for you whatever the other actor chooses. Suppose the other one would defect, then it is rational to defect yourself: you won't gain anything, but if you do not defect you will be stuck with a -10 loss. Suppose the other one would cooperate, then you will gain anyway, but you will gain more if you do not cooperate, so here too the rational choice is to defect. The problem is that if both actors are rational, both will decide to defect, and none of them will gain anything. However, if both would "irrationally" decide to cooperate, both would gain 5 points. This seeming paradox can be formulated more explicitly through the following principle.

IV. The principle of suboptimization

When you try to optimize the global outcome for a system consisting of distinct subsystems (e.g. minimizing the total punishment for the system consisting of the two prisoners, or maximizing the amount of prey hunted for the pack of wolves), you might try to do this by optimizing the result for each of the subsystems separately. This is called "suboptimization". The principle states that suboptimization in general does not lead to global optimization (Machol, 1965, pp. 1-8). Indeed, the suboptimization for each of the prisoners separately is to betray the other one, but this leads to both of them being punished rather severely, whereas they might have escaped with a mild punishment if they had stayed silent. Similarly, the optimization for each of the wolves separately is to let the others do the hunting, and then come to eat from their captures. Yet if all wolves would act like that, no prey would ever be captured and all wolves would starve.

The *principle of suboptimization* can be derived from the more basic systemic principle stating that "the whole is more than the sum of its parts" (cf. Heylighen, 1992). If the system (e.g. the wolf pack) would be a simple sum or "aggregate" of its parts, then the outcome for the system as a whole (total prey killed) would be a sum of the outcomes for the parts (prey killed by each wolf separately), but that is clearly not the case when there is interaction (and in particular cooperation) between the parts.

As a last example, suppose you want to buy a new car, and you have the choice between a normal model, and a model with a catalyser, that strongly reduces the poisonous substances in the exhaust. The model with catalyser is definitely more expensive, but the advantage for you is minimal since the pollution from your exhaust is diffused in the air and you yourself will never be able to distinguish any effect on your health of pollution coming from your own car. Rational or optimizing decision-making from your part would lead you to buy the car without catalyser. However, if everybody would make that choice, the total amount of pollution produced would have an effect on everybody's health, including your own, that will be very serious, and certainly worthy the relatively small investment of buying a catalyser. The suboptimizing decision (no catalyser) is inconsistent with the globally optimizing one (everybody a catalyser). The reason is that there is interaction between the different subsystems (owners and their cars), since everybody inhales the pollutants produced by everybody. Hence, there is also an interaction between the decision problems of each of the subsystems, and the combination of the optimal decisions for each of the subproblems will be different from the optimal decision for the global problem.

Now that we have sketched: a) why natural selection tends to lead to selfishness; b) why cooperation between subsystems has definite selective advantages if the global or higher order system is conceived, the problem that remains is to explain how natural selection can move to the higher level, that is to say select on the basis of global optimality rather than suboptimality. Since evolution, as said, tends to progress by small local changes, it is not obvious how the transition to the global level could take place. We will now review different attempts to extend the purely local selection criteria working on individual systems, which try to explain the emergence of altruism.

V. Proposals for "altruistic" selection criteria

a. Kin selection

The most well-established generalization of "individual" selection is based on the so-called *inclusive fitness* (Hamilton, 1971). The fundamental idea is that in biological evolution it is not so much the survival and reproduction of individual organisms that matters, but the survival and reproduction of their genes. According to this view, genes are the true replicators, and organisms are merely their *vehicles* (Dawkins, 1976). Hence it are not the organisms that are "selfish", but their genes. Now since genes are shared by an individual and his offspring or kin, the fitness that is most important for selection is not that of the individual but that of the individual with the inclusion of that of his kin, insofar that this kin shares the same genes.

The problem is that you can never know which genes exactly are shared, for example, by two siblings. Yet there are simple statistical rules for estimating the total amount of shared genes. For example, siblings share 50 % of their genes; the same applies to parents and children; uncle and nephew, or grandparent and grandchild, share 25 %; cousins share 12,5 %, etc. Under these conditions, strong altruism can become advantageous in specific cases. For example, it is worthwhile to endanger your life in order to save the life of two brothers, or eight cousins. Indeed, if 8 times 12,5 % = 100 % of your genes would be saved by a potentially self-sacrificing action, the fact that you might die in the attempt becomes a risk worth taking in terms of inclusive fitness.

Such calculations of average benefit to the genes must take into account a number of complicating factors, depending on the context. For example, an individual who is too old to reproduce would optimize his fitness even by sacrificing his life for only one relative, since the genes killed in the action would anyway not be able to replicate any more. On the other hand, a parent may not find it worthwhile to care for a child, that is so weak that its chances for survival are anyway very small. Another difficulty is that it is not always obvious for an individual to recognize kin from strangers. That ambiguity is exploited by cuckoos who lay their eggs in nest of other birds, so that their offspring is fed and reared by foster parents that are genetically totally unrelated. Dawkins (1976) provides a lot of concrete examples of such considerations, and gives many correlated observations of animal behavior. These considerations quickly get very complex, but it is clear that this way of analysis gives a simple evolutionary justification for observed altruism towards offspring and kin (e.g. mother animals defending their children at the peril of their own life, or older siblings looking after younger ones).

Though kin selection predicts that altruistic behavior will quickly diminish in inverse proportion to the degree of relatedness, it can still explain some extreme cases of "ultrasociality" where very large groups of organisms are cooperating (Campbell, 1983). The clearest case is the order of the hymenoptera, which includes wasps, bees and ants. These insects have the peculiar feature that males have only half of the chromosomes of females (haplodiploidy). This means that when a "queen" is fertilized by a single male, her female offspring will be more closely related to each other than to their mother, and effectively share 75 % of their genes (all the genes coming from the father will be the same). In that case the interest of the female "workers" is to have their mother produce more sisters, rather than to produce offspring of their own (Dawkins, 1976). Indeed, newly born sisters would share more genes with them than their eventual daughters, thus increasing their inclusive fitness. In practice workers will become infertile and spend their time caring directly or indirectly for the queen. This allows biologists to explain the very strong collective organization that is typical of ants nests or bee colonies. Workers will indeed be willing to sacrifice their life if that can help to save the nest, whose main function is to keep the queen producing offspring.

A similar mechanism, though on a different genetic basis, seems to work for termites and African naked mole rats (the only mammals known having this type of organization, Jarvis, 1981). In both cases, the "workers" are sterile, while the queen who is being fed and protected by the workers spends all her time producing offspring. In the case of termites it is hypothesized that a more than 50% sharing of genes between siblings may have resulted from very strong interbreeding, and the same may be true of mole rats. But is is clear that the nest type of organization relies on the maintained infertility of the workers. Otherwise the workers might be tempted to produce their own offspring, and thus come in competition with their queen mother, and competition is the end of group cooperation, as we will elaborate now. If for some reason the queen would die, often one of the workers would regain her fertility and take over the role of the queen.

b. Group selection

The most obvious, but also the least accepted, explanation for the development of altruism is selection at the level of the group. The argument is very simple: compare two groups of individuals, e.g. two packs of wolves. Suppose that one group is more cooperative, while the other consists of more selfish individuals. Now because of the principle of synergy, the cooperative group will be able to gather more resources, it will be more fit, and hence will be selected, whereas the non-cooperative group will be eliminated. Thus natural selection would promote cooperation.

The error with this reasoning is more subtle. Though it is true that individuals in an altruistic group will have better chances of survival, this applies to all members of the groups, including those who are not or less cooperative (because of blind variation there will always be slight differences in "cooperativity" among the group members). The more selfish ones will still have the advantages of the better cooperative organization, but will have less disadvantages since they spend less resources or take less risks in helping the other ones. The result is that they will be fitter than the altruists, and their genes will eventually replace the altruist genes in the gene pool of the group. In other words, cooperation in groups on a genetic basis tends to be self-destructive.

This may be clarified by introducing the concept of an *evolutionary stable strategy* (Maynard Smith, 1982; Axelrod & Hamilton, 1981; Dawkins, 1976). The strategy of the altruists, helping others even if they do not reciprocate the help, may lead to an increased fitness for the group but it is not stable, since it can be easily invaded by egoistic strategies that take advantage of the altruists' sacrifice, but without giving anything in return. Though the selfish strategies will globally lead to a decrease of fitness for the group, they are locally stronger than the altruist strategies. This is another expression of the principle of suboptimization: genetic evolution works at the level of the subsystem (the individual or the individual together with his kin), and what is optimal at that level will be selected, even though it is far from optimal at the level of the group. Campbell (1983, 1991) has summed up this predicament by the phrase "genetic competition among the cooperators": on the level of the genes rivalry continues, and that will eventually erode any cooperation on the level of the group.

c. Selection for reciprocal altruism

The evolutionary instability of the purely altruist strategy may be circumvented by a strategy of "conditional" altruism. Such an altruist would only help another individual if he expects the other one to return the favor. If the other one does not cooperate, the conditional altruist will stop cooperating, and hence will not incur the costs of spending resources from which his selfish companion would gain more than he does. In that sense, such a "reciprocal" altruist strategy (Trivers, 1971) may be stable against invasion from cheaters, while still keeping the advantages of synergy among those individuals that are willing to cooperate.

This idea was illustrated in a spectacular way by Axelrod (1984). Axelrod organized a tournament in which different game theorists were invited to submit a computer program which would implement the best strategy for winning a repeated prisoner's dilemma game. In the tournament two programs would play each other in a long sequence of prisoner's dilemma. The points they gained in each game were added. Each program would then play such a sequence against each other program. At the end the points gained in all sequences were added, and this allowed to designate the overall winner of the tournament.

Though the most complicated and cunning strategies were proposed by some of the most expert game theorists, the strategy that consistently won the tournament was extremely simple: "tit for tat". This strategy starts by cooperating. However when the opponent defects, "tit for tat" defects too. If afterwards, or from the beginning, the opponent starts to cooperate, "tit for tat" will reciprocate by cooperating. The characteristics of "tit for tat" (and of the other more successful strategies) can be summarized by three concepts:

1) the strategy is "nice": this means that it will never be the first to defect;

2) the strategy is "provocable": if the opponent defects, it retaliates by defecting too;

3) the strategy is "forgiving": as soon as the opponent cooperates again, the strategy forgets about the previous defection, and cooperates.

Niceness is advantageous because it opens the way to mutually beneficial cooperation. Retaliation is necessary in order to avoid being invaded by selfish profiteers. Forgivingness has the advantage of avoiding mutual rounds of retaliation. Indeed, suppose that an individual because of distrust, by way of test or just because of a misunderstanding would defect just once, then a non-forgiving strategy would continue to defect in reaction, and mutual cooperation could never emerge or be restored.

In a second study Axelrod generalized his game theoretic simulation to an evolutionary setting. In this setting, fitness was explicitly introduced by giving a strategy an amount of offspring proportional to the number of points it got in the previous tournament. The tournament was then repeated by playing all members of the new population of strategies (in which more successful strategies were now more numerous) against each other. Again the points of this tournament were used to produce a second generation of offspring. This generation played a following tournament, and so on. After many generations the less successful strategies would have been eliminated by natural selection, while the most successful ones would become more and more numerous. This setting is not equivalent to the previous one, since the fitness of the strategies depends on the opponents against which they play, and the field of opponents changes in the course of the simulated evolution. Hence a strategy that is successful in the original field of opponents may no longer be fit after the field has drastically changed. Yet Axelrod found out that it was still "tit for tat" that was most successful.

However, "tit for tat" does not turn out to be an evolutionary stable strategy in the strict sense. Indeed, once the field is dominated by "tit for tat", other strategies that are less retaliatory (or more forgiving) become as fit as "tit for tat" since there are no longer cheaters to take advantage from their unconditional altruism. However, once a sufficient percentage of strategies becomes too altruistic, selfish strategies can again gain in fitness by exploiting their altruism. The end result seems to be some kind of equilibrium mixture of strategies in which reciprocal altruists such as "tit for tat" dominate, but in which there may also appear small amounts of "nasty" (the opposite of "nice") strategies together with non-retaliating altruists.

In how far can these simulation results be generalized to real evolution? The most important restriction in the experiment seems to be that opponents interact with each other for a long, consecutive sequence of exchanges. This may be true for two individuals engaged in a close and stable relation. Reciprocal altruism may thus explain how a symbiotic relationship between two organisms can develop (e.g. a hermit crab and the sea anemone living on its shell) (cf. Axelrod & Hamilton, 1981). In the kind of situations involving a large number of individuals that interest us, on the other hand, it seems more likely that opponents will encounter each other only once, or now and then with long interruptions and exchanges with different opponents in between.

"Tit for tat" is only successful in an indefinitely repeated prisoner's dilemma. If there is only one transaction, no retaliation is possible afterwards, and rationality dictates that you should defect. If there is a finite number of transactions, it pays to defect during the last one, but if you expect your opponent to defect at the last one, you should also defect at the last-but-one, and hence your opponent would be wise to already defect at the last-but-two, and so on. Hence games with a fixed number (known by the participants) of transactions would lead to continuous defection. In practice, that does not seem to be the problem, since normally opponents do not know how often they will meet each other again.

However, the basic practical limitation is that of memory: the reciprocal altruist should not only remember how his opponent treated him during the last transaction (which may be a long time ago) he should also be able to recognize and distinguish all opponents with whom he has ever had transactions. This requirement does not seem to be realistic in large groups, such as human societies. Moreover, in such large systems, many encounters will take place for the first (and perhaps the last) time. In such cases reciprocal altruism does no good, and the "nice" individual who starts by cooperating may be cheated most of the time by others he will never see again. When I buy something in a shop in a city where I will never come back again, I do not expect to be cheated (though that is possible of course), even though I have no power to retaliate. In conclusion, reciprocal altruism seems still insufficient to explain the ultrasociality of human society.

d. Selection for moralism

The basic weakness of reciprocal altruism in explaining ultrasociality is that it starts from *dyadic* relationships, of the type "I'll scratch your back if you scratch mine". It is difficult to imagine how such one-to-one exchanges could be enlarged in order to form the basis for large collective organizations. Therefore we would like to see some evolutionary stable strategy that directs behaviour towards groups rather than towards other individuals.

It has been proposed that one such strategy is *moralism*, that is to say behavior that rewards or reinforces altruistic behavior by others, and punishes or inhibits cheating or defection. For example, Trivers (1971) has postulated selection for "moralistic aggression", and Lorenz (1975) speaks about "innate ethical sense". The advantage of moralism compared to pure altruism is that the costs of moralizing towards others are clearly less than those of being altruistic yourself (Campbell, 1983). However, if everybody around you is continuously moralizing, and ready to ostracize or even kill you if you do not behave altruistically, it becomes quite difficult to behave in a selfish way. Hence groups consisting of moralizing individuals will also tend to be altruistic, though the individuals are not motivated to be altruistic on their own. Moralism is also more stable than real altruism because cheaters will find it very difficult to invade a population that tends to make their life as difficult as possible. A disadvantage of moralism is that it encourages hypocrisy, that is to say behavior that does everything to look altruistic but is in fact selfish.

A basic weakness of the argument is that it is difficult to imagine how something as complicated as an "ethical sense" might develop through genetic evolution. It seems quite difficult to put down rules distinguishing moral or altruistic behavior from selfish behavior that are applicable to all situations and all individuals. We might perhaps imagine the evolution of simple behavior patterns such as aggressive reactions against the selfish wolf who does not want the others of the pack to share in the food, but that does not seem sufficient to explain ultrasociality.

Another weakness of the moralistic selection argument, is that moralism mainly functions to maintain an already functioning cooperation system, by reducing the fitness of those who do not obey the rules. However, the argument does not explain how the cooperative pattern, and the moralistic attitude that maintains it, may have developed from selfish behavior in the first place.

Conclusion

Natural selection produces systems with the implicit goal of optimizing their fitness. Though cooperation between individual systems would make global optimization possible, individual replicators are basically selfish, and hence it will be difficult for evolution to overcome their shortsighted strategies of local optimization. Yet we do observe cooperation, altruism, and ultrasociality in the animal world, and especially in our human society. Four extensions of purely individual, selfish selection have been reviewed, inspired by sociobiology and game theory, that each try to explain some of the observed altruistic behavior.

Kin selection is the least controversial model, but, except in extreme cases where very large groups of organisms share most of their genes (like in ant nests), the argument can only explain altruism towards to a small circle of close relatives. Group selection in se is no longer accepted by evolutionists, because of the instability of group strategies against individual strategies, though we can still imagine very specific circumstances in which it might have an influence (see e.g. Campbell, 1983). Reciprocal altruism, exemplified by the "tit for tat" strategy, is a quite attractive mechanism for explaining dyadic forms of cooperation, but it is not clear how it could be extended to the cooperation among large groups, in part because of the limitations of memory and repetition. Moralism is a way to make group altruism more stable, without much costs to the moralizers, but is not clear how such a type of behavior could have spontaneously evolved from selfishness.

It seems clear that such models based on genetic evolution are insufficient to explain the appearance of ultrasociality in people. In a following paper (Heylighen, 1992b), I therefore wish to bring the discussion to a different level, by focussing on a radically different type of replicators: memes. Though a meme, like all entities evolving through natural selection, can be called "selfish" (Dawkins, 1976), that selfishness at the cultural level will be argued to lead to cooperation at the level of the individuals below. This will allow me to show how all the weaknesses of these arguments can be evaded, while still keeping their strengths.

References

Axelrod R. & Hamilton W.D. (1981): "The Evolution of Cooperation", *Science* 211, p. 1390-1396.

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

Campbell D.T. (1979): "Comments on the Sociobiology of Ethics and Moralizing", Behavioral Science 24, p. 37-45.

Campbell D.T. (1983): "The Two Distinct Routes beyond Kin Selection to Ultrasociality: implications for the humanities and social sciences", in: *The Nature of Prosocial Development*, D. Bridgeman (ed.), (Academic Press, New York), p. 11-41.

Campbell D.T. (1991): "A Naturalistic Theory of Archaic Moral Orders", *Zygon* 26, No. 1, p. 91-114.

Csanyi V. & Kampis G. (1985): "Autogenesis: the evolution of replicative systems", *J. Theor. Biol.* 114, p. 303-321.

Csanyi V. (1991): *Evolutionary Systems and Society: a general theory*, (Duke University Press, Durham, NC).

Dawkins R. (1976): *The Selfish Gene*, (Oxford University Press, New York).

Hamilton W.D. (1971): "The Genetical Evolution of Social Behavior", in: *Group Selection*, Williams G.C. (ed.), (Aldine-Atherton, Chicago), p. 23-89.

Heylighen F. (1991a): "Modelling Emergence", *World Futures: the Journal of General Evolution* 31 (Special Issue on "Emergence", edited by G. Kampis), p. 89-104.

Heylighen F. (1991b): "Cognitive Levels of Evolution: pre-rational to meta-rational", in: *The Cybernetics of Complex Systems - Self-organization, Evolution and Social Change*, F. Geyer (ed.), (Intersystems, Salinas, California), p. 75-91.

Heylighen F. (1991c): "Evolutionary Foundations for Metaphysics, Epistemology and Ethics", in : *Workbook of the 1st Principia Cybernetica Workshop,* Heylighen F. (ed.) (Principia Cybernetica, Brussels-New York), p. 33-39.

Heylighen F. (1992): "Principles of Systems and Cybernetics: an evolutionary perspective", in: *Cybernetics and Systems '92*, R. Trappl (ed.), (World Science, Singapore). (in press)

Heylighen F. (1992b) : "'Selfish' Memes and the Evolution of Cooperation", *Journal of Ideas*.

Heylighen F., Joslyn C. & Turchin V. (1991) : "A Short Introduction to the Principia Cybernetica Project", *Journal of Ideas* 2, #1 p. 26-29.

Jarvis J.U.M. (1981): Eusociality in a Mammal: cooperative breeding in naked mole-rat colonies, *Science* 212, 571-573.

Lorenz K.A. (1975): "Konrad Lorenz Responds", in: *Konrad Lorenz: the man and his ideas,* Evans R.I. (ed.), (Harcourt Brace Jovanovich, New York).

Machol R.E. (1965): *System Engineering Handbook*, (McGraw-Hill, New York).

Maynard Smith J. (1982): *Evolution and the Theory of Games*, (Cambridge University Press, Cambridge).

Trivers R.L. (1971): "The Evolution of Reciprocal Altruism", *Quarterly Review of Biology* 46 (4), p. 35-57.

Wilson E.O. (1975): *Sociobiology: the New Synthesis*, (Harvard University Press, Cambridge).

Heylighen F. (1992) : "Evolution, Selfishness and Cooperation", *Journal of Ideas,* Vol 2, # 4, pp 70-76.
Heylighen F. (1992) : "'Selfish' Memes and the Evolution of Cooperation", *Journal of Ideas* , Vol. 2, #4, pp 77-84.

'Selfish' Memes and the Evolution of Cooperation

Francis HEYLIGHEN[*]

PO-PESP, Free University of Brussels, Pleinlaan 2, B-1050 Brussels, Belgium
E-mail: fheyligh@vnet3.vub.ac.be

Abstract. A new, integrated model for the evolution of cooperation is proposed, based on the concept of a meme, as replicating unit of culture. Meme evolution is much faster and more flexible than genetic evolution. Some basic selection criteria for memes are listed, with an emphasis on the difference between memetic and genetic fitness, and the issue of memetic units is discussed. The selfishness of memes leads to conformity pressures in cultural groups, that share the same meme. This keeps group cooperation conventions (ethical systems), resulting from reciprocal agreements, from being invaded by selfish strategies. The emergence of cooperative systems is discussed in general as a "metasystem transition", where interaction patterns between competing systems tend to develop into shared replicators, which tend to coordinate the actions of their vehicles into an integrated control system.

KEYWORDS: memetics, cooperation, evolution, altruism, culture, selection, metasystem transition

I. Memes: the new replicators

In a previous paper (Heylighen, 1992b), I have reviewed different models proposed for explaining the evolution of cooperation in human society on a genetic basis, and concluded that none of them is sufficient. In the present paper I wish to extend this evolutionary analysis by moving to a different level, that of memes.

The theory of natural selection on the basis of fitness is in principle applicable to all replicating systems, not only to genes. Recently, a new type of replicator has been proposed as a unit of cultural evolution. *Memes* are defined as cognitive or behavioral patterns that can be transmitted from one individual to another one by learning and imitation (Dawkins, 1976; Moritz, 1991). Examples of memes in the animal world are most bird songs, and certain techniques for hunting or using tools that are passed from parents or the social group to the youngsters (Bonner, 1980). In human society, almost any cultural entity can be seen as a meme: religions, language, fashions, songs, techniques,

[*] Senior Research Assistant NFWO (Belgian National Fund for Scientific Research)

scientific theories and concepts, conventions, traditions, etc. The defining characteristic of memes as informational patterns, is that they can be replicated in unlimited amounts by communication between individuals, independently of any replication at the level of the genes. Storing a concept or a habit in memory after having encountered it through another individual, does not require any change of the DNA.

Of course, the capacity of the nervous system for learning is the result of evolutionary processes at the genetic level. Yet I will here not go into detail about why that capacity has been selected. The increased fitness resulting from a nervous system that is flexible enough to adapt its behavior to many new situations, seems obvious enough. If a useful type of behavior can be learned directly from another individual by communication or imitation, that seems like a most welcome shortcut for having to discover it by personal trial-and-error. More arguments for why the capacity for meme replication has evolved genetically can be found in most texts about the recently founded domain of *memetics* (Moritz, 1991). Memetics can be defined as an approach studying cultural evolution, which is inspired by Darwinian theories of genetic evolution (see e.g. Boyd & Richerson, 1985; Cavalli-Sforza & Feldman, 1981; Lumsden & Wilson, 1981; Csanyi, 1991).

Whatever was the exact origin of memes, once these new replicators appeared (and there can be no doubt that they did), we should expect the start of a new process of evolution, which is largely (though not completely) independent from the evolution of genetic replicators. Though that process will be subjected to the same basic principles of blind variation and natural selection on the basis of fitness, memetic evolution is basically a much more flexible mechanism. Genes can only be transmitted from parents (or parent in the case of asexual reproduction) to offspring. Memes can in principle be transmitted between any two individuals (though it will become more difficult the larger the differences in cognitive mechanisms and language are).

For genes to be transmitted, you typically need one generation, which for higher organism means several years. Memes can be transmitted in the space of hours. Meme spreading is also much faster than gene spreading, because gene replication is restricted by the rather small number of offspring a single parent can have, whereas the amount of individuals that can take over a meme from a single individual is almost unlimited. Moreover, it seems much easier for memes to undergo variation, since the information in the nervous system is more plastic than that in the DNA, and since individuals can come into contact with much more different sources of novel memes. On the other hand, selection processes can be more efficient because of "vicarious" selection (Campbell, 1974): the meme carrier himself does not need to be killed in order to eliminate an inadequate meme; it can suffice that he witnesses or hears about the troubles of another individual due to that same meme.

The conclusion is that memetic evolution will be several orders of magnitude faster and more efficient than genetic evolution. It should not surprise us then that during the last ten thousand years, humans have almost not changed on the genetic level, whereas their culture (i.e. the total set of memes) has undergone the most radical developments. In practice the superior "evolvability" of memes would also mean that in cases where genetic and memetic replicators are in competition, we would expect the memes to win in the long term, even though the genes would start with the advantage of a well-established, stable structure.

II. Memetic units

The main criticism that can be raised against the memetic approach is that memes are difficult to define. What are the elements or units that make up a meme? Does a meme correspond to a complete symphony, or to a symphonic movement, a melody, a musical phrase, or even a single note?

In order to model meme structure, we may use some concepts from cognitive science. Perhaps the most popular unit used to represent knowledge in artificial intelligence is the *production rule*. It has the form "if condition, then action". The action leads in general to the activation of another condition. In fact a production rule can be analysed as a combination of even more primitive elements: two *distinctions* (which discriminate between presence and absence of the condition and the action respectively) and a *connection* (the "then" part, which makes the first distinction entail the second one) (Heylighen, 1991d; see also Heylighen, 1990). For example, a meme like "God is omnipotent" can be modelled as "if a phenomenon is God (distinction of God from non-God), then that phenomenon is omnipotent".

Production rules are connected when the output condition (action) of the one matches the input condition of the other. This makes it possible to construct complex cognitive systems on the basis of elementary rules. Even remembered melodies might be modelled in such a way, as concatenations of production rules of the type "if C (musical note distinguished), then E (note produced and subsequently distinguished)", "if E, then A", and so on. (In fact, genes too are now being modelled using networks of "if... then" productions: a DNA string is activated by the presence of certain proteins (condition) to which it responds by producing specific other proteins (action), see Kauffmann, 1991).

It has been shown that production rules (or at least a simplified, binary representation of them, called "classifiers") can be used to build quite impressive computer simulations of cognitive evolution, using mutations, recombinations, and selection on the basis of "fitness" (Holland et al., 1986). Although these models do not as yet take into account distinct carriers, this looks like a very promising road to study memes formally and computationally.

Even if we would model memes as connected sets of production rules, we still have the problem of how many production rules define a single meme. If we call a religion or a scientific theory a meme, it is clear that this will encompass a very large number of interconnected rules. In practice it will be impossible to enumerate all rules, or to define sharp boundaries between the rules that belong to the meme and those that do not. However, that should not detract us from using memetic mechanisms in analysing evolution.

Indeed, Darwinian models of genetic evolution have certainly proven their usefulness, even though it is in practice impossible to specify the exact DNA codons that determine the gene for, say, blue eyes or altruism towards siblings.

As Dawkins (1976) notes, it is not necessary to be explicit about what are the constitutive elements of a gene, postulated to explain a particular characteristic or type of behavior. It is sufficient that we can distinguish the phenotypical effects of that gene from the effects of its rival genes (alleles). If we can determine the fitness resulting from these effects, taking into account the environment and the context of different, non-rival genes present in the genome, then we can make predictions about evolution.

The same applies to memes. If, for example, we observe that one meme (say Catholicism) induces its carriers to have more children than its competitors (say Calvinism and Anglicanism), and that the children tend to take over their memes from their parents, then, *all other things being equal*, we can predict that after sufficient time that meme will dominate in the population. Of course, in practice it is never the case that all other things are equal, but that is the predicament of all scientific modelling: we must always simplify, and ignore potentially important influences. The question is to do that as wisely as possible, and to maximally include relevant variables without making the model too complex.

III. Meme selection criteria

What is the relation between meme and gene selection? Both replicators have similar aims to the degree that they use the same vehicles: individual organisms. Everything that strengthens the vehicles should in general be good for the replicators, and hence both genes and memes should be selected on the basis of their support for increased survivability and reproducability of their carriers. However, the implicit goals of genes and memes are different to the degree that they use different mechanisms for spreading from one vehicle to another one. Memes will be positively selected mainly for increased communicability. Genes will be selected mainly for sexual reproducability. These different emphases may lead to direct conflicts.

For example, priests in many religions are prohibited to marry and to have children, in striking disagreement with genetic injunctions. Yet we can easily imagine that the religious meme of celibacy would have been selected because unmarried priests can spend more time and energy on "spreading the word", and hence replicating the meme.

An even more vivid example of countergenetic behavior, closely related to the issue of altruism, is that of martyrs, suicide teams, or kamikaze pilots, who are willing to give up their life in order to promote the spread of a meme: a religion, an ideology or a nation (i.e. a group defined by a common culture or ideology). In that case, the loss of one or a few carriers is compensated by the increased chances of survival for the other carriers or for the meme itself. For example, the suicide of an individual may attract the attention of other individuals to the meme he is carrying, and thus facilitate its spreading. A well-known example is Jan Palach, the Czech student who put himself to fire in order to protest the Soviet suppression of the "Prague Spring". In this case the meme would be the Czech version of "socialism with a human face".

Meme fitness will depend basically on three factors: 1) survivability of the carrier; 2) individual learnability of memes; 3) tendency of memes to spread or to be transmitted. The first criterion is similar to the one determining gene selection, so we won't go into much detail about it. Basically it states that, all other things being equal, memes leading to decreased probability of a carrier's survival, e.g. because they are bad for the health, or lead to dangerous or suicidal behavior, will tend to be eliminated. However, we must add that survival is to be interpreted on the level of the group of all carriers, rather than on the level of the individual, as is shown by the example above. This will be elaborated further on.

The second criterion examines how easy it is for an individual to acquire and retain a given new meme. First, memetic patterns should be stable in memory, that is to say they should not be easily forgotten or confused with other memes. The learnability depends on how simple and explicit the meme is, but also on the genetic and cultural predisposition to learn a meme. Behavior patterns that are inconsistent with genetically determined instincts or behaviors will be much more difficult to assimilate. For example, though behaviorists have shown that it is very easy to condition a pigeon to discharge food by pecking on specific buttons or handles, it is extremely difficult for a pigeon to learn how to get food by *not* pecking (Holland et al., 1986). Genetic evolution has clearly favoured the association of pecking with getting food, and it is very difficult to undo this inherited bias. Similarly, certain genetically useful behaviors are very easy to learn. For example, rats can be taught to avoid a particular type of food after being subjected just once to feelings similar to food poisoning (Holland et al., 1986).

We may conclude that successful memes will tend to be similar to genetically favoured behavioral patterns, or at least neutral to genetic biases. In cases where there is a direct contradiction, like in the example of celibacy above, the meme will tend to function at a different level of abstraction than the "wired-in" genetic mechanism. For example, the celibacy meme does not deny the fact that priests or nuns undergo sexual attraction, it rather tries to sublimate these feelings to a more abstract "love of God", and thus divert them from their original function.

The same dependence on bias or predisposition to learn applies to memes that are added to an already existing store of memes. New memes that contradict already well-established memes will be much more difficult to assimilate to the cognitive system because of the tendency to avoid cognitive dissonance (Festinger, 1957). On the other hand, memes that reinforce or are reinforced by previously acquired memes will be added quite smoothly to the memetic store.

Where the second criterion examined the passive acceptance of memes offered to a potential new carrier, the third criterion examines the active tendency of a meme to "infect" carriers. For example, a melody that induces its carriers to continuously sing or hum it, will more easily spread to others who hear the singing, than a beautiful image that tends to be enjoyed in silence. Some memes, such as religions, incorporate the idea that the carrier should actively seek to convert other people to the meme. If this "*contagiousness*" of a meme is strong enough, the meme may spread in spite of it not fulfilling the previous selection criteria. For example, there have been religious sects that stimulated their followers to commit suicide, in contradiction with criterion one. Though such memes normally won't survive in the long term, they may be quite successful in the short term, and that is one of the major dangers of memetic replication mechanisms.

It is this contagiousness that most strikingly differentiates meme selection from gene selection. Genes, indeed, are not contagious. You can never get a gene, that you did not already have, from someone else. At most you can produce a child that shares some of your genes with some genes of another person. The issue of contagiousness becomes the more important the better the available communication media. It is striking how quickly clothing or dancing fashions spread over the whole globe nowadays. In more primitive societies, the factor of contagion would be limited by the relatively small group of people with which the individual would interact.

Until now studies of cultural evolution have mainly looked at the influence of memes (or "culturgens") on genetic fitness (see e.g. Lumsden & Wilson, 1981), or on the genetic bias underlying cultural patterns. However, the present analysis shows that memes have a fitness of their own, that is in a number of respects independent of, if not in contradiction with, genetic fitness. That is what makes a memetic analysis of evolutionary processes fundamentally different from a sociobiological one, in spite of the many parallelisms. These differences will become more explicit when we turn our attention to the problem of cooperation.

IV. The memetic evolution of cooperation

Now that we have examined the basic mechanisms of memetic evolution, and of the evolution of cooperation (in a previous paper, Heylighen, 1992), we can try to integrate them. Memes use limited resources for replication: memory space in the carriers. Hence memes will tend to compete for that space, and rival memes (memes that cannot coexist within that space, because they are cognitively dissonant, such as two different religions, or two mutually contradicting hypotheses) will tend to fight each other. In other words, memes are as selfish as any other replicator, and won't do anything that might help a rival meme to increase its fitness.

However, when we now look at the level of the vehicles, all memes have the implicit goal of making their carriers more fit, since an increased number of carriers signifies that there will be more memory space available for meme replication. If, as we have argued previously, cooperation among the carriers tends to increase the overall fitness of the group of carriers, it will be in the interest of memes to promote that cooperation. Moreover, since cooperation requires communication, and since meme spreading critically depends on communication, the "motivation" of memes to bolster cooperation should be even stronger than that of genes.

However, it is not sufficient to argue that a meme supporting strong cooperation would have a large selective advantage; we also need to explain how such a meme could have developed through small evolutionary steps, and be stable against invasion by memes promoting selfishness of their carriers.

Therefore we will memetically reinterpret the mechanisms proposed for the genetic evolution of altruism.

The mechanism of kin selection can be extended to memes by redefining inclusive fitness as the fitness of a meme taking into account all its carriers. All individuals carrying the same meme can be viewed as relatives or kin insofar as this meme is concerned. Hence, the tendency to be altruistic towards offspring or close kin that follows from genetic evolution can be generalized to altruistic tendencies towards members of the same memetic or cultural group. The explanation for ultrasociality in the social insects on the basis of genetic inclusive fitness can be readily transposed to the emergence of ultrasociality in humans on the basis of *memetic inclusive fitness*.

V. Memes and group conformism

In fact this means that in memetic evolution there is selection at the level of cultural groups, if we define a group as that set of individuals carrying the same meme. Consider two groups, characterized by different memes. Suppose that one meme fosters altruism whereas the other one encourages its carriers to be selfish. The altruistic group, as argued previously (Heylighen, 1992), will be more productive and hence have a selective advantage over the selfish one.

The selective advantage does not mean that the less fit group will be physically eliminated during the competition with the cooperative group. On the level of memes, elimination can happen when the selfish group gives up its belief in the meme for selfishness and adopts the meme for altruism from the other group. This conversion of one group by another one may happen by direct physical force, for example because the altruist group is better organized to win a war between the groups, and can thus subdue the other one. It can also happen in a peaceful way because the less successful group simply imitates the more successful one.

The basic argument against group selection is that group strategies can be easily invaded. However, that is not the case for memetic strategies. Indeed, suppose that a "mutant" meme promoting selfishness would appear in an altruist group. According to genetic reasoning, its carrier would be more fit than an altruist one, since he can profit from the altruism of the other carriers without paying the corresponding costs. However, the fitness of a meme is different from the fitness of its carrier. Though the selfish carrier might have enhanced fitness in the sense that he gets more food or other resources, the fitness of the meme he carries depends on how easily other members of the group can be converted to it.

Now memes are selfish, which means that they have the implicit goal of thwarting all rival memes that compete for the same memory space. Hence the majority meme in the group will tend to consolidate the memory space it already occupies in the majority of carriers. A likely mechanism for that might be that the different carriers continuously reinforce each other's belief by communication and imitation. The carrier of the mutant meme, on the other hand, is alone and does not get any reinforcement from his fellows. He will find it very difficult to convert any of them to his non-conformist ideas, since the influence on any individual of a majority of conformists will be much stronger than that of a single dissident. This tendency of the majority meme to impose itself on minorities leads to intra-group homogeneity, as confirmed by Boyd and Richerson's (1985) mathematical model of cultural evolution (see also Campbell, 1983, 1991).

In that sense, memetic strategies tend to have a self-stabilizing effect, which makes it difficult for mutant memes to invade well-established groups, except when the group as a whole takes over a meme from another, apparently more successful group, in which case the majority-minority argument does not apply. The evolutionary stability of memetic strategies does not mean that memetic evolution won't be able to proceed because of conformist pressure, though. The resource of memory space is sufficiently rich to accommodate the appearance of many new memes that are not dissonant or in direct contradiction with the dominating ones. For example, though one cannot at the same time be selfish *and* altruistic, it is perfectly possible to be simultaneously religious and artistic, or to believe in Euclidean geometry and the goodness of man. After some time the new memes may have gained so much in importance, that the old ones are forgotten, and so it becomes possible to accommodate memes that are in contradiction with the previous ones. However, it seems unlikely that a meme for cooperation might "fade away" like that, since its fitness increasing effect on its carriers won't fade that easily.

At most the apparently easier life of a selfish dissident may induce other carriers to imitate him. But we have seen (Heylighen, 1992) that moralism is a quite effective weapon against profiteering. Hence we may expect the dominant meme to develop such a moralistic attitude. That would also be sufficient to diminish the dissident's genetic fitness, since, apart from undergoing possible direct aggression, he will lose access to collective resources and to mates. Thus memes will even succeed in transcending the problem of the genetic competition between the cooperators. That should not surprise us, if we remember that competition between memes and genes is normally won by the more flexible memes.

Our criticism of the argument on the basis of moralism, namely that a complete ethical system seems too complex to evolve by genetic selection, does not apply to memetic evolution, which is much faster, and which adapts more readily to abstract models of the world. There remains the criticism that the ethical system should be able to develop by small steps. It is here that our analysis of reciprocal altruism may be useful.

VI. Memetic spreading of reciprocity

Reciprocity, as shown by Axelrod's simulation (Axelrod, 1984; see previous paper: Heylighen, 1992), is a quite simple strategy that easily invades and outcompetes any other strategy (except continuous defection, see further) in a "prisoner's dilemma" type of evolutionary setting. When two individuals, after sufficient interactions, have reached a stable cooperative relationship or pact, based on reciprocity, that agreement or convention can be viewed as a *meme with two carriers.* If there is communication, that same meme can be transferred to a third and a fourth carrier, and so on, who would thus come to join the convention. Indeed, an individual who observes an existing reciprocal relationship between two other individuals, and who notices the advantages following from their mutual cooperation, would be tempted to imitate their behavior. If a certain behavioral pattern tends to be imitated, that makes it a meme by definition. If that meme, in addition to it being contagious, also furthers the genetic fitness of its carriers, we may conclude that it has a high memetic fitness, and thus will tend to replace rival memes with a lower fitness.

This mechanism does not exist in genetic evolution: though the tendency to use "tit for tat" strategies may be inherited, the equilibrium pattern resulting from a repeated sequence of "tit for tat" transactions between two specific individuals, cannot be genetically transmitted to any offspring. Hence any cooperation agreement reached will have to be renegotiated for every new individual, with the concomitant risks of being taken advantage of, if the other individual is not "nice". With memes, on the other hand, such agreements can readily spread around the population, if they are seen to increase the participating individuals' fitness.

The limitations on memory that make reciprocity difficult for large groups, can also be evaded by memetic mechanisms. Indeed, a meme can easily evolve mechanisms for making members of the same cultural group easy to recognize. Individuals belonging to the same culture or ethnical group will usually distinguish themselves by clearly perceivable attributes or behavior. "Thus the Luo of Kenya knock out two front teeth of their men, while the adjacent Kipsigis enlarge a hole pierced in their ears to a two-inch diameter" (Campbell, 1991). If such signs allow you to identify a member of your group, you can expect that he will also follow the group's agreement on reciprocity, and hence you can trust that he will cooperate, without you having to renegotiate a pact. If he does not, you can still alarm the other members of your group, and he will be subjected to moralistic aggression. In that sense it is to the advantage of both the group and separate individuals to wear the appropriate attributes. If they do not wear the attributes, they won't receive altruistic treatments from other group members. If they do wear the attributes, but do not exhibit the appropriate cooperative behavior, they put their life at risk.

We must make one more note about the issue of reciprocity. In fact there are two evolutionary stable strategies in Axelrod's setting: a mixture of "tit for tat" related strategies, and "always defect". When all individuals in a group are totally selfish, no single mutant reciprocal altruist can increase his fitness, since whatever cooperative or defective moves he proposes, no one will ever enter into a cooperation, and so he can only lose by being cooperative. So we must explain how an initial population of "tit for tat" players might have appeared. However, that is not difficult if we go back to kin altruism (Axelrod & Hamilton, 1981). Two brothers, say, would be genetically predisposed to behave altruistically towards each other. That predisposition would form a good basis for initially cooperative moves towards different individuals, even if they are not all close kin. That might be sufficient to start a "tit for tat"-like exchange in small, kin-based groups.

The memes for reciprocal altruistic agreement, and their moralistic extensions which go beyond pure reciprocity (e.g. the "turn the other cheek" meme), would have found an adequate breeding ground in such inherited dispositions. Indeed, many ethical systems explicitly refer to the ideal of "fraternity", and sometimes members of the same cultural group (e.g. monks or Freemasons) are supposed to call each other "brother". Though these are not brothers in the biological sense, the meme attempts to harness the innate tendency to behave altruistically towards kin and to use it for purposes different from the increase of genetic inclusive fitness. This is similar to the reorientation of sexual feelings to divine love that we have noted earlier.

VII. The emergence of cooperation as a metasystem transition

The evolutionary road towards cooperation we have sketched in this and the previous paper is long and winding. Yet it is possible to discern a clear progression from pure selfishness, to kin-restricted limited altruism, to "tit for tat" based dyads, to multi-individual reciprocal agreements, to moralism and group ideologies, and finally to the complex ultrasocial systems of cooperation characterizing present society.

These levels of cooperation might be paralleled by general levels of evolutionary complexity. We would imagine pure selfishness to characterize primitive organisms such as plants, amoebae, or molluscs, who seem to completely ignore other members of their species, except as obstacles or possible prey. Even many species of fish will eat their own offspring if they have the opportunity, though some species have a strongly developed brood care. Kin altruism would start somewhere with the insects, reaching an extreme in the social insects, and apply to most vertebrates in varying degrees. At what stage reciprocal altruism appears is more difficult to judge. Reciprocity within groups requires at least a certain level of memory and perceptual skills. But it seems clear that meme-based altruism is typical for human groups able to use language. With the capacity for language appears the capacity to rapidly spread complex memes, and that gives memes a definite advantage over genes in directing further evolution. In recent times, the memes that seem to be dominating are those that tend to make the ideal of altruism or brotherliness universal, ignoring the distinctions created by older memes such as languages or religions. We will not go into detail about why that is happening but note that the evolutionary tendency towards more and more far-reaching or inclusive cooperation seems to continue, albeit with many ups and downs.

Such an evolution towards stronger integration of subsystems, allowing optimization at the global level, is exemplified by Turchin's concept of a *metasystem transition* (MST; see Turchin, 1977; Heylighen, 1991a,b). Such an evolutionary transition is characterized by the appearance of a control system at the metalevel, steering and optimizing the actions of the subsystems at the level below. Turchin proposes the following sequence of basic metasystem transitions, leading from a level of organization comparable to that of amoebae to the level of present humans:

control of position = movement
control of movement = irritability (simple reflex)
control of irritability = (complex) reflex
control of reflex = associating (conditional reflex)

control of associating = human thinking
control of human thinking = culture

Though this sequence does include the emergence of culture and society, all previous MST's seem totally independent of any of the issues of cooperation and competition we have examined. Turchin's pre-cultural MST's take place *within a single organism*, and hence no competition is involved. I have argued (Heylighen, 1991b) that such MST's might be better conceptualized as increases of (internal) variety coupled with emergence of control, rather than as integrations of (independent) subsystems coupled with emergence of control, the way Turchin defines them.

However, Turchin gives another example of an MST, independent of his basic sequence, that seems much more closely related to the emergence of cooperation: the emergence of multicellular organisms from unicellular ones. Here too we could imagine that individual cells were originally in a situation of competition for the available resources, whereas cells in an organism have a kind of "ultrasocial" organization with full division of labour and cells sacrificing themselves for the survival of other cells (e.g. cell of the immune system fighting intruders, or cells of the skin that continuously die off because of friction).

Very little seems to be known about how that organization has emerged during evolution. An orthodox biologist would probably argue that there is no real problem since all cells of a multicellular organism have the same genes, and hence it would have been in their "inclusive fitness" interest to further each other's survival. Yet the basic problem is that of knowledge and communication: how can a cell know that certain of the cells it is competing with share the same genes, and hence should be treated altruistically? We have seen that even fishes and birds are sometimes ignorant about their own offspring. So how could we expect a single cell to be smart enough to recognize its "genetic allies"?

A basic conclusion of our general analysis is that you need communication before you can have cooperation, that is to say *information must be shared.* In organisms cooperating because of kin selection, the medium of communication is sexual reproduction: genetic "messages" are transmitted through special cells (sperm or egg cells) that cannot survive or develop independently of the process of sexual reproduction. It is this shared information that creates the connection between parents and offspring or, indirectly, between relatives. In primitive multicellular organisms, such as algae, the "communication" might be based on simple spatial contiguity (sharing of the same location). In complex organisms where there is differentiation of functions among cells, you need special chemical signals to coordinate the different cells (this can be modelled by "genetic networks", Kauffmann, 1991). In sociocultural systems, the basic medium supporting cooperation is meme spreading.

We might conceive the development of communication in the following way. Systems operating in the same environment (and especially those sharing the same resources) will interact. Since naturally selected systems are by default selfish, those interactions will tend to be competitive, if not directly conflictual.

However, every process of variation tends to reach a stable configuration after a sufficient time (Heylighen, 1992), and, hence, we may expect to see a stable interaction pattern emerge. That pattern might be cooperative, or, more probably, it may merely limit the damage of direct conflicts, by making the systems restrain from certain actions that would lead to losses for all of them.

Now, since such a pattern becomes part of the environment to which the systems must adapt, they will tend to evolve internal models ("vicarious selectors", Campbell, 1974; Heylighen, 1991a,b, 1992) of that pattern. Since the models of different individuals represent basically the same pattern, we might say that they share knowledge or information about that pattern. Even though the models themselves may be structured quite differently, their external effects will be the same, since they will select for the same type of stable interaction. Hence it becomes possible to distinguish a new abstract system or information structure, that is shared between initially independent or competing systems.

That shared information might now develop a dynamics of its own, that is to say it may start to spread and replicate, undergoing variation and selection. The spread of shared information is advantageous to the competing systems since it eliminates the risks involved in renegotiating a stable interaction pattern with new systems that do not already share the convention. Once shared information starts to evolve autonomously, the systems that share it become "vehicles" for its further spreading. The selfish interest of the shared replicator is to have its vehicles cooperate more and more effectively. This cooperation between vehicles may develop so much that it forms a basis for a higher level, integrated system.

The complete evolutionary sequence would be something of the following form: competition, communication, stable interaction patterns, internal models of pattern, shared models, shared replicators, cooperation promoted by shared replicators, integration with the shared replicators as coordinators. In Turchin's terminology, the shared information will become a *control* for the systems of the level below, coordinating, monitoring and directing their cooperation. Hence the process of development of cooperation between initially competing subsystems through the development of shared replicators can be seen as a true metasystem transition.

We can distinguish the following MST's of this type:

1) cooperation between cells, leading to a multicellular organism. The shared information resides in the network of connections between genes, which determines which genes are switched on or switched off in each cell type (Kauffman, 1991). The shared replicator is the whole multicellular organism, which is reproduced independently of the replication going on at the level of its cells;

2) sexual intercourse is a way of communicating genetic information, leading to genes shared among members of the same family or species. The "cooperation" among individuals consists in their mating and family interaction patterns, that further the reproduction of their genes. The "integrated system" is the species, that can be defined as a reproductively closed population;

3) at the level of human society, the cooperation is supported by memes as shared replicators. The integrated system, which Turchin calls "superbeing", is the culture as a whole.
4) In fact we might even consider the emergence of cooperation between pieces of DNA and enzymes, within the integrated system of a cell, as a most primitive MST, that would characterize the origin of life. Similar to the multicellular organism, the cell organization itself can be viewed as the replicator, shared by all individual replicators consisting of single pieces of DNA. A virus is an example of a cheater, that takes unfair advantage of that cooperation in order to have its own DNA selfishly replicated.

These MST's, that have many things in common, seem quite different from the sequence of intra-organism MST's such as the emergence of the capacity to learn, or the capacity to move. The difference seems to be that the latter, insofar as there is an integration of subsystems, do not involve competing subsystems (different muscles or neurons in the same organism do not seem to be involved in competition). Hence we could expect such MST's to be faster, easier and more profound than the former type, where competition between selfish entities is to be overcome, and where there is always a possibility of intrusion by cheaters. If we do not take into account cellular or multicellular integration, integration at the level of the species and of the culture does indeed seem much more superficial, and its evolution seems to be more irregular.

Perhaps another major difference between the two MST types is that the *intra-organism* MST already starts with some kind of control structure, and that the development of a higher control occurs in mutual feedback with an increase of variety at the level below (according to Turchin's "law of the branching growth of the penultimate level" (1977), see also Heylighen, 1991b). The *between-competitors* MST, on the other hand, starts with a large variety of independent subsystems, and has to build a control, in the form of shared information, from scratch, taking into account that any preliminary control regime that is not sufficiently stable can be invaded by selfish strategies taking unfair advantages.

Conclusion

It is clear that the whole issue of how competing subsystems can start to cooperate and thus become (partly or completely) integrated into a globally optimizing supersystem is very complex. Many questions about cooperation, shared information, and higher levels of control still have to be answered. Yet I think it is equally obvious that these problems are of the utmost importance if we wish to understand our own further evolution, as individuals, as a species, as a culture, or as parts of the global world system (Heylighen, 1991c). In particular, we must look for an answer to the question whether evolutionary development will take place basically *between* individuals, developing in the form of Turchin's "superbeing", or *within* individuals, leading to what I have called a "metabeing" (Heylighen, 1991b).

These answers will be especially needed if we wish to develop a new ethics, based on evolutionary insights, that might help us to cope with the problems of our present society (Heylighen, 1991c). The analysis of the evolution of cooperation from the viewpoint of selfish memes, as contrasted to more traditional studies focusing on either genes, individuals, or society as a whole, is definitely helpful as a heuristic to discover new mechanisms, that may simplify previously intractable seeming problems. But the real hard work has merely started.

References

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

Bonner J.T. (1980): *The Evolution of Culture in Animals*, (Princeton University Press, Princeton).

Boyd R. & Richerson P.J. (1985): *Culture and the Evolutionary Process*, (Chicago University Press, Chicago).

Campbell D.T. (1974): "Evolutionary Epistemology", in: *The Philosophy of Karl Popper*, Schilpp P.A. (ed.), (Open Court Publish., La Salle, Ill.), p. 413-463.

Campbell D.T. (1983): "The Two Distinct Routes beyond Kin Selection to Ultrasociality: implications for the humanities and social sciences", in: *The Nature of Prosocial Development*, D. Bridgeman (ed.), (Academic Press, New York), p. 11-41.

Campbell D.T. (1991): "A Naturalistic Theory of Archaic Moral Orders", *Zygon* 26, No. 1, p. 91-114.

Cavalli-Sforza L.L. & Feldman M.W. (1981): *Cultural Transmission and Evolution: a quantitative approach*, (Princeton University Press, Princeton).

Csanyi V. (1991): *Evolutionary Systems and Society: a general theory*, (Duke University Press, Durham, NC).

Dawkins R. (1976): *The Selfish Gene*, (Oxford University Press, New York).

Festinger L. (1957): *A Theory of Cognitive Dissonance*, (Harper, New York).

Heylighen F. (1990): "Classical and Non-classical Representations in Physics I", *Cybernetics and Systems* 21, p. 423-444.

Heylighen F. (1991a): "Modelling Emergence", *World Futures: the Journal of General Evolution* 31 (Special Issue on "Emergence", edited by G. Kampis), p. 89-104.

Heylighen F. (1991b): "Cognitive Levels of Evolution: pre-rational to meta-rational", in: *The Cybernetics of Complex Systems - Self-organization, Evolution and Social Change*, F. Geyer (ed.), (Intersystems, Salinas, California), p. 75-91.

Heylighen F. (1991c): "Evolutionary Foundations for Metaphysics, Epistemology and Ethics", in : *Workbook of the 1st Principia Cybernetica Workshop*, Heylighen F. (ed.) (Principia Cybernetica, Brussels-New York), p. 33-39.

Heylighen F. (1991d): "Structuring Knowledge in a Network of Concepts", in : *Workbook of the 1st Principia Cybernetica Workshop*, Heylighen F. (ed.) (Principia Cybernetica, Brussels-New York), p. 52-58.

Heylighen F. (1991e): "A Cognitive-Systemic Reconstruction of Maslow's Theory of Self-Actualization", *Behavioral Science*.(in press)

Heylighen F. (1992) : "Evolution, Selfishness and Cooperation", *Journal of Ideas*.

Holland J.H., Holyoak K.J., Nisbett R.E. & Thagard P.R. (1986): *Induction : processes of inference, learning and discovery*, (MIT Press, Massachusetts).

Kauffman S.A. (in press, 1991): *Origins of Order: self-organization and selection in evolution*, (Oxford University Press, Oxford).

Lumsden, Charles, and Wilson, Edward (1981): *Genes, Mind, and Culture: the Coevolutionary Process*, (Harvard University Press, Cambridge).

Moritz E. (1990): "Memetic Science: I - General Introduction", *Journal of Ideas* 1, p. 1-23.

Turchin V. (1977): *The Phenomenon of Science. A cybernetic theory of human evolution*, (Columbia University Press, New York).