The Cognitive and Communicative Demands of Cooperation

Peter Gärdenfors

Lund University Cognitive Science

Kungshuset, Lundagård

S-223 50 Lund, Sweden

Peter.Gardenfors@lucs.lu.se

ABSTRACT: I argue that the analysis of different kinds of cooperation will benefit from an account of the cognitive and communicative functions required for the cooperation. I investigate different models of cooperation in game theory – reciprocal altruism, indirect reciprocity, cooperation about future goals and conventions – with respect to their cognitive and communicative prerequisites. The cognitive factors considered include recognition of individuals, memory capacity, temporal discounting, anticipatory cognition and theory of mind. The communication considered ranges from simple signalling to full symbolic communication.

1. The span of cooperation in game theory

The evolution of cooperation is still an enigma. In surprisingly many situations, animals and humans do not behave as predicted by game theory. For example, humans cooperate more in prisoner's dilemma games than would be expected from a rationalistic analysis. Several explanations for this mismatch have been suggested, ranging from the claim that animals and humans are not sufficiently rational to the position that the rationality presumed in classical game theory is not relevant in evolutionary accounts. The purpose of this paper is to argue that the behaviour of various agents must be judged in relation to their cognitive and communicative capacities. In my opinion, these factors have not been sufficiently considered in game theory.

Printed from: *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz.* Ed. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson, 2007. www.fil.lu.se/hommageawlodek

There are many ways of defining cooperation. A broad definition is that is consists of joint actions that confer mutual benefits.[1] A more narrow definition concerns situations in which joint action poses a dilemma so that in the short run an individual would be better off not cooperating (Richerson et al. 2003, p. 358). In this paper, the focus will be on the more narrow definition.

Cooperation has been studied from two paradigmatic perspectives. Firstly, in traditional game theory, cooperation has been assumed to take place between individuals of the species *Homo oeconomicus*, who are ideally rational. *Homo oeconomicus* is presumed to have a perfect "theory of mind", that is the capacity to judge the beliefs, desires and intentions of other players. In a Bayesian framework, this amounts to assuming that everybody else are Bayesian decision makers. In the case of *Homo sapiens*, the theory of mind is well developed, at least in comparison to other animal species, but it is far from perfect.

Secondly, cooperative games have been studied from an evolutionary perspective. The classical ideas are found in Axelrod and Hamilton (1981) and Maynard Smith (1982). Here, the players are the genes of various kinds of animals. The genes are assumed to have absolutely no rationality and no cognitive capacities at all. However, their strategies can slowly adapt, via the mechanisms of natural selection, over repeated interactions and a number of generations. In the evolutionary framework, the "theory of mind" of the players is not accounted for or considered irrelevant.

The differences in cognitive and communicative demands of the two perspectives of cooperative games are seldom discussed. For example, in Lehmann and Keller's (2006) recent classification of models of the evolution of cooperation and altruism, only two parameters related to cognition and communication are included: a one period "memory" parameter m defined as the probability that "an

[1] A stronger criterion, focussed on human cooperation, is formulated by Bowles and Gintis (2003): "An individual behavior that incurs personal costs in order to engage in a joint activity that confers benefits exceeding these costs to other members of one's group." It should be noted that "joint" activity does not imply that the actions are simultaneous, but they can be performed in sequence.

individual knows the investment into helping of its partner at the previous round" and a "reputation" parameter q defined as the probability that "an individual knows the image score of its partner" (Lehmann and Keller 2006, p. 1367). Here, I want to argue that the analysis of different kinds of cooperative games would benefit from a richer account of the cognitive and communicative functions required for the cooperation. The cognitive factors that will be considered include recognition of individuals, memory capacity, temporal discounting, anticipatory cognition and theory of mind. The communication considered ranges from simple signalling to full symbolic communication.

2. Different kinds of cooperation

In this section, different kinds of cooperation will be outlined. I shall take my point of departure from the forms of cooperation that have been studied within game theory, in particular evolutionary game theory, but also from some themes from animal behaviour. The paradigmatic game in my analysis will be prisoner's dilemma (PD), mainly in its iterated form, although I will sometimes refer to other games (such as the "stag hunt" (Skyrms 2002)). A reason to focus on games of the PD type is that such games frequently arise in biological/ecological settings. A finding that generates much of the interest in studying these games is that biological players often cooperate considerably more than the total defection that is predicted by the strictly rationalistic analysis of the PD. A problem for a biocognitively oriented analysis is to identify the factors that promote cooperation. I shall argue that they require varying forms of cognition and communication.

I shall present the different kinds of cooperation roughly according to increasing cognitive demands. My list is far from exhaustive – I have merely selected some forms of cooperation where the cognitive and communicative components can be identified, at least to some degree. For several of the items on my list it is also possible to make finer divisions concerning the forms of cooperation.

2.1 Flocking behaviour

A cognitively low-level form of cooperation is *flocking behaviour*. Flocking is found in many species and its function is often to minimise predation. Simulation models of flocking (e.g. Reynolds 1987, Lorek and White 1993) show that the sophisticated behaviour of the flock can emerge from the behaviours of single individuals that follow simple rules. For instance, birds flying in a flock seem just to follow two basic rules: (a) try to position yourself as close as possible to the centre of the flock; (b) keep a certain minimal distance from your neighbours. These rules do not presume any cognitive capacities of the individuals beyond those of visual perception (other senses such as echolocation could in principle be used as well).

2.2 Ingroup versus outgroup

One simple way for a player to increase cooperation in an iterated PD is to divide the other individuals into an *ingroup* and an *outgroup*. Then the basic strategy is to cooperate with everybody in the ingroup and defect against everybody in the outgroup. Cognitively, this strategy only demands that you can separate members of the ingroup from the rest. In nature this is often accomplished via olfaction; for example, bees from a different hive smell differently and are treated with aggression. For a more advanced example, Dunbar (1996) speculates that dialects have evolved to serve as markers of the ingroup among humans. The fact that ingroup mechanisms are ubiquitous in the animal kingdom and their effects are so strong provides good reasons to suspect that the mechanism is at least partially genetically determined.

The spatial PD games studied by, among others, Nowak and May (1992), Lindgren and Nordahl (1994), Hauert (2001), Brandt, Hauert and Sigmund (2003), can be seen as a form of ingroup cooperation. In these games, players are organised spatially so that you only interact with your neighbours. The studies show that spatial structure promotes cooperation. Cooperators survive by forming spatial clusters, thereby creating ingroups that defect against players outside the cluster.

It seems fairly obvious that the evolutionary origin of ingroup formation is kinship selection. In a kin group, the outcomes of a PD game must be redefined due to the genetic relatedness of the individuals, and in many cases this makes cooperation the only evolutionarily stable strategy. In other words, a game that is a PD on the individual level, may be a perfectly cooperative game when the genes are considered to be the players. Kin groups can then form kernels from which larger cooperative groups can develop (Lindgren 1997, p. 351). In general, the ingroup requires some form of marker, be it just physical proximity, that helps distinguish the ingroup from the outgroup. However, evolutionary biologists (e.g. Zahavi 1975) stress that such markers should be hard to fake in order to exclude free riders from the ingroup.

Another way of generating an ingroup is by identifying a common enemy. West et al. (2006) present empirical evidence that when a group playing an iterated PD game is competing with another group, cooperation within the group will increase. Bernhard et al. (2006) present an anthropological study of two groups in Papua New Guinea that support the same conclusion. The downside of this mechanism, as already Hamilton (1975) pointed out, is that what favours cooperation within groups will also favour the evolution of hostility between groups. A parallel example from the world of apes concerns a flock of chimpanzees in the Gombe forest in Tanzania that become divided into a northern and a southern group. The chimpanzees in the two groups, who had earlier been playing and grooming together, soon started lethal fights against each other (de Waal 2005).

2.3 Reciprocal altruism

In an iterated PD, one player can *retaliate* against another's defection. Trivers (1971) argues that this possibility can make cooperation more attractive and lead to what he calls *reciprocal altruism.* In their seminal paper, Axelrod and Hamilton (1981) showed by computer simulations that cooperation can evolve in iterated PD situations. A Nash equilibrium strategy such as the well-known Tit-for-Tat can lead to reciprocal cooperative behaviour among individuals that encounter each other frequently. The cognitive factors required for this strategy are, at least, the ability to

recognize the individual you interact with and a *memory* of previous outcomes (see below for further specification of these criteria). *Trusting* another individual may be an emotional correlate that strengthens reciprocity and, following the lead of Frank (1988), such an emotion may have been selected for and thus become genetically grounded in a species that engages in reciprocal altruism.

If the players in an iterated PD choose their moves synchronously, then the size of the memory does not seem to affect which strategies are successful (Axelrod 1984, Lindgren 1991, 1997, Hauert and Schuster 1997). However for so called alternating iterated PD in which players take turns in choosing their moves, it seems that the best strategies depend on larger memory of previous moves (Frean 1994, Neill 2001).

Field studies of fish, vampire bats (Wilkinson 1984) and primates, among other species, have reported the presence of reciprocal altruism. Still, the evidence for reciprocal altruism in non-human species is debated and some laboratory experiments seem to speak against it (Stephens et al. 2002, Hauser et al. 2003). In line with this, Stephens and Hauser (2004) argue that the cognitive demands of reciprocal altruism have been underestimated (see also Hammerstein 2003). They claim that reciprocal altruism will be evolutionarily stable in a species only if (a) the temporal discounting of future rewards is not too steep; (b) they have sufficient discrimination of the value of the rewards to judge that what an altruist receives back is comparable to what it has given itself; and (c) there is memory capacity to keep track of interactions with several individuals.

Studies of temporal discounting reveal that the rates differ drastically between different species (for three examples, see Figure 1). Humans have by far, the lowest discount rate, which is a prerequisite for the anticipatory planning that will be presented below. An interesting problem, that should receive more attention within evolutionary game theory, is what factors (ecological, cognitive, neurological, etc) explain the discounting rate of a particular species.

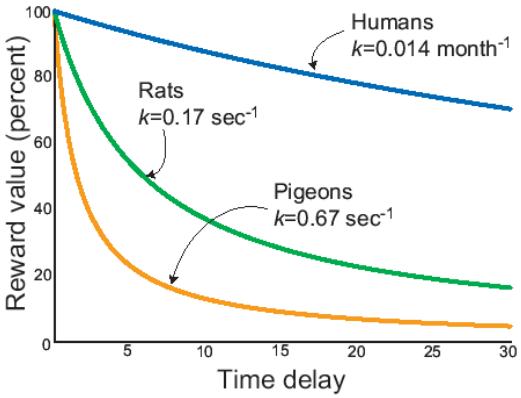

Figure 1: The discounting rates of different species describe how quickly a reward is devalued over time (from Stevens and Hauser 2004).

According to Stephens and Hauser (2004), the cognitive demands they present explain why reciprocal altruism is difficult to establish in other species than *Homo sapiens*. The debate on this issue will certainly continue and the cognitive demands will be scrutinized, but at least their challenge highlights the importance of a more detailed analysis of the cognitive underpinnings of reciprocal altruism.

2.4 Indirect reciprocity

Reciprocal altruism can be formulated as a slogan: "You scratch my back and I'll scratch yours." As we have seen, it is possible to make evolutionary sense of this principle. However, in humans one often finds more extreme forms of altruism: "I help you and somebody else will help me." This form of cooperation has been called *indirect reciprocity* and it seems to be unique to humans. Nowak and Sigmund (2005) show that, under certain conditions, this form of reciprocity can be given an evolutionary grounding. However, as we shall see, their explanation depends on strong assumptions concerning the communication of the interactors (which explains why indirect reciprocity is only found in humans).

In Nowak and Sigmund's (2005) definition of indirect reciprocity, any two players are supposed to interact at most once with each other. This is an idealising assumption, but it has the effect that the recipient of a defect (cheat) act in a PD cannot retaliate. Thus all

strategies that have been considered for the iterated PD are excluded. So the question is under what conditions indirect reciprocity can evolve as an evolutionarily stable strategy.

Nowak and Sigmund note that indirect reciprocity seems to require some form of "theory of mind". An individual watching a second individual (donor) helping a third (receiver) (or not helping a third in need) must judge that the donor does something "good" ("bad") to the receiver.[2] The form of intersubjectivity required for such comparisons is closely related to empathy (Preston and de Waal 2003, Gärdenfors to appear).

The key concept in Nowak and Sigmund's evolutionary model is that of the *reputation* of an individual.[3] The reputation of an individual *i* is built up from some members of the society *observing i*'s behaviour towards third parties and the observers *spreading* this information to the other members of the society (see Figure 2). In this way a level of reputation for *i* being a helper can be, more or less, known by all the members of the group. Gossip may be a way of achieving consensus about reputation. (Thus the function of gossip would not be a replacement of grooming as claimed by Dunbar (1996).) Then the level of *i*'s reputation is used by any other individual when deciding whether to help *i* or not in a situation of need.[4] It should be noted that the reputation is not something that is visible to all others, unlike status markers such as a raised tail among wolves, but each individual must keep a private account of the reputation of all others. And Semmann et al. (2005) demonstrate experimentally that building a reputation through cooperation is valuable for future social interactions, not only within but also *outside* one's own social group.

[2] This is in contrast to altruism towards kin, where an individual can experience as "good" that which improves its own reproductive fitness.

[3] A precursor to this concept is that of Sugden's (1986) "good standing".

[4] The ingroup behaviour considered in section 2.2 can be described as a limiting case of indirect reciprocity where all "in" individuals are treated as having good reputation and all "out" as having bad. Or the other way around: indirect reciprocity is a *flexible* way of determining the ingroup.

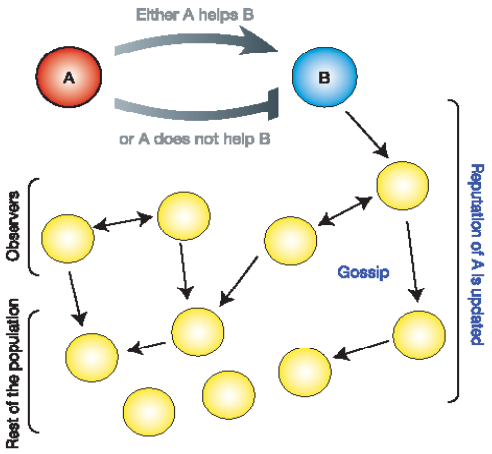

Figure 2: The mechanism of building a reputation (from Nowak and Sigmund 2005)

In these interactions it is important to distinguish between *justified* and *unjustified* defections. If a potential receiver has defected repeatedly in the past, the donor can be justified in defecting, as a form of punishment. However, the donor then runs a risk that his own reputation drops. To prevent this, the donor should communicate that the reason he defects is that the receiver has a bad reputation.

Nowak and Sigmund (2005) say that a strategy is first order if the assessment of an individual *i* in the group depends only on *i*'s actions. More sophisticated strategies distinguish between justified and unjustified defections. A strategy is called second order if it depends on the reputation of the receiver and third order if it additionally depends on the reputation of the donor. Nowak and Sigmund show that only eight of the possible strategies are evolutionarily stable and that all these strategies depend on the distinction between justified and unjustified defection.

The success of indirect reciprocity thus heavily depends on the mechanism of reputation. This means that indirect reciprocity, in the model presented by Novak and Sigmund (2005), requires several cognitive and communicative mechanisms. The cognitive requirements are (at least): (a) recognising the relevant individuals over time, (b) remembering and updating the reputation scores of

these individuals and (c) a form of empathy to judge whether a particular donor action is "good" or "bad". Apart from this, the communication system of the individuals in the group must be able to (d) identify individuals in their absence, e.g. by names, (e) expressions to the effect that "x was good to y" and "y was bad to x". There seems to be no animal communication system that can handle these forms. Thus it is no surprise that Homo sapiens is the only species exhibiting indirect reciprocity. The communication system need not be a language with full syntax, but a form of *protolanguage* along the lines discussed by Bickerton (1990) is sufficient. The communication can be symbolic, but it is possible that a system based on *miming* (Donald 1991, Zlatev, Persson and Gärdenfors 2005) would be sufficient.

The trust that is built up in reciprocal altruism is *dyadic*, that is, a relation between two individuals. In contrast, reputation is an emergent *social* notion involving everybody in the group. On this point, Nowak and Sigmund (2005, p 1296) write: "Indirect reciprocity is situated somewhere between direct reciprocity and public goods. On the one hand it is a game between two players, but it has to be played within a larger group."

As noted above, an individual in a donor situation should signal that the potential recipient has a "bad reputation" before defecting, in order not to lose in reputation because of misinterpretation from the onlookers. This is another form of communication that presumes (f) expressions of the type "y has bad reputation". In this context it should also be noted that Nowak and Sigmund's (2005) model considers only two kinds of reputation – "good" or "bad." It goes without saying that in real groups communication about reputation has more nuances.

Of course, the need for communication is dependent on the size of the group facing the PD situations. In a tightly connected group where everybody sees everybody else most of the time, there is no need for a reputation mechanism. One can compare with how the ranking within such a group is established. If I observe that x dominates y and I know that y dominates me, there is no need for (aggressive) interaction or communication to establish that x dominates me. However, in large and loosely connected groups,

these mechanisms are not sufficient, but some form of communication about non-present individuals is required. This form of "displacement" of a message does not exist in animal signalling, but is one of the criteria that Hockett (1960) uses to identify symbolic language.

There may exist still other mechanisms that influence the reputation of an individual. In many situations people are not only willing to cooperate but they also punish free riders. Punishing behaviour is difficult to explain because the punishment is costly and also the cooperative non-punishers benefit (Fehr and Gächter 2002). Thus punishment should decrease in frequency.[5] Barclay (2005) presents some evidence that the reputation of a punisher increases and that people are more willing to cooperate with punishers. In combination with the mechanisms presented above, this indicates that punishing behaviour is rewarded in the long run (see also Sigmund et al. 2001) and can stabilize cooperation in iterated PDs (Lindgren 1997).

2.5 Cooperation about future goals

Cooperation often occurs over an extended time span and then involves a form of planning. Gulz (1991) calls planning for present needs *immediate planning* while planning for future goals is called *anticipatory planning*. Humans can predict that they will be hungry tomorrow and save some food, and we can imagine that the winter will be cold, so we start building a shelter already in the summer. The planning of other animals concerns here and now, while humans are mentally both here and in the future. The central cognitive aspect of anticipatory planning is the capacity to *represent your future drives*.

Bischof (1978) and Bischof-Köhler (1985) argue that animals other than humans cannot anticipate future needs or drive states. Their cognition is therefore bound to their present motivational state (see also Suddendorf and Corballis, 1997). This hypothesis, which is called the Bischof-Köhler hypothesis, has been supported by the

[5] In line with Frank (1988), Fehr and Gächter (2002) say that negative emotions towards defectors are the proximate causes of altruistic punishment.

previous evidence concerning planning in non-human animals. However, recent results (Mulcahy and Call 2006) indicate that some of the great apes may be capable of anticipatory planning at least to some extent.

For most forms of cooperation among animals, mental representations of the goal are not needed. If the common goal is *present* in the actual environment, for example food to be eaten or an antagonist to be fought, the collaborators need not focus on a joint representation of it before acting. If, on the other hand, the goal is distant in time or space, then a *mutual* representation of it must be produced before cooperative action can be taken. For example, building a common dwelling requires coordinated planning of how to obtain the building material and advanced collaboration in the construction. In general terms, cooperation about future goals requires that the mental spaces of the individuals be coordinated.

The presence of mutual representations of a future goal will change a situation, which would be a PD without the presence of such representations, into a game where the cooperative strategy is the equilibrium solution. For example, if we live in an arid area, each individual (or family) will benefit by digging a well. However, if my neighbour digs a well, I may defect and take my water from his well, instead of digging my own. But if nobody digs a well, we are all worse off than if everybody does it. This is a typical example of a PD.

Now if somebody communicates the idea that we should cooperate in digging a communal well, then such a well, by being deeper, would yield much more water than all the individual wells taken together. Once such cooperation is established, the PD situation may disappear or at least be ameliorated, since everybody will benefit more from achieving the common goal. In game theoretical terms, digging a communal well will be a new equilibrium strategy. This example shows how the capacity of sharing detached goals in a group can strongly enhance the value of co-operative strategies within the group. The upshot is that strategies based on future goals may introduce new equilibria that are beneficial for all participants.

The cognitive requirements for cooperating about future goals involve crucially the capacity to represent your drives or wishes at a future time, that is, anticipatory planning. Even if there is presently plenty of water you will foresee the dry season and that you will be thirsty then. This representation, and not your current drive state, forms the driving force behind a wish to cooperate in digging a well.

What are the communicative requirements then? Symbolic language is the primary tool by which agents can make their inner representation known to each other. In previous work (Brinck and Gärdenfors 2003, Gärdenfors 2003, 2004, Osvath and Gärdenfors 2005), it has been proposed that there is a strong connection between the evolution of anticipatory cognition and the evolution of symbolic communication. In brief, the argument is that symbolic language makes it possible to *cooperate about future goals* in an efficient way. Again, it is not required that the symbolic communication involves any syntactic structures – protolanguage (Bickerton 1990) will do perfectly well.

An important feature of the use of symbols in cooperation is that they can set the cooperators free from the goals that are available in the present environment. Again, this requires that the present goals can be suppressed, which hinges on the executive functions of the frontal brain lobes. The future goals and the means to reach them are picked out and externally shared through symbolic communication. This kind of sharing gives humans an enormous advantage concerning cooperation in comparison to other species. I submit that there has been a co-evolution of cooperation about future goals and symbolic communication (cf. the "ratchet effect" discussed by Tomasello, 1999, pp. 37-40). However, without the presence of anticipatory cognition, the selective pressures that resulted in symbolic communication would not have emerged.

2.6 Commitments and contracts

Commitments and contracts are special cases of cooperation about the future. They rely on an advanced form of theory of mind that

allows for joint beliefs.[6] In general, joint beliefs form the basis for much of human culture. They make many new forms of cooperation possible. For example, to promise something only means that you intend to do it. On the other hand, when you *commit* yourself to a second person to do an action, you intend to perform the action in the future, the other person wants you to do it and intends to check that you do it, and there is joint belief concerning these intentions and desires (Dunin-Kepliz and Verbrugge 2001). Unlike promises, commitments can thus not arise unless the agents achieve joint beliefs and have anticipatory cognition. It has been argued (Tomasello 1999, Gärdenfors 2003, 2006) that the capacity for joint beliefs is only found in humans.

For similar reasons, *contracts* cannot be established without joint beliefs and anticipatory cognition. For example, Deacon (1997) argues that marriage is the first example of cooperation where symbolic communication is needed. It should be obvious that marriage is a form of contract. Even if I do not know of any evidence that marriage agreement was the first form of symbolic communication, I still find this example interesting in the discussion of early anticipatory cognition. A pair-bonding agreement implicitly determines which future behaviors are allowed and not allowed. These expectations concerning future behavior do not only include the pair, but also the other members of the social group who are supposed not to disturb the relation by cheating. Anybody who breaks the agreement risks punishment from the entire group. Thus in order to maintain such bonds, they must be linked to social sanctions.

2.7 Cooperation based on conventions

In human societies, many forms of cooperation are based on *conventions*. Conventions presume *common beliefs* (often called common knowledge). For example, if two cars meet on a gravel road in Australia, then both drivers know that this coordination problem has been solved by driving on the left hand side numerous times

[6] For an analysis of different levels of a ”theory of mind”, see Gärdenfors (2003), ch. 4 and Gärdenfors (to appear).

before, both know that both know this, both know that both know that both know this, etc, and they then both drive on their left without any hesitation (Lewis 1969). Many conventions are established without explicit communication, but, of course, communication makes the presence of a convention clearer.

Conventions function as virtual governors in a society. Successful conventions create equilibrium points, which, once established, tend to be stable. The convention of driving on the left hand side of the road will force me to "get into step" and drive on the left. The same applies to *language* itself: A new member of a society will have to adjust to the language adopted by the community. The meaning of the linguistic utterances emerges from the individuals' meanings.[7] There are, of course, an infinite number of possible "equilibrium" languages in a society, but the point is that once a language has developed, it will have strong explanatory effects regarding the behaviour of the individuals in the society.

Similarly, the existence of money depends on conventions. Money requires cooperation in the form of a mutual agreement to accept certain decorated pieces of paper as a medium of exchange. Being a member of a monetary society, or just visiting one, I must "get into step" and accept the conventional medium of exchange as legal tender.

Actually, there are many similarities between language and money as tools for cooperation. Humans have been trading goods as long as they have existed. But when a monetary system does emerge, it makes economic transactions more efficient. The same applies to language: hominids have been communicating long before they had a language, but language makes the exchange of knowledge more efficient. The analogy carries further: When money is introduced in a society, a relatively stable system of *prices* emerges. Similarly, when linguistic communication develops individuals will come to share a relatively stable system of *meanings*, that is, components in their inner worlds that communicators can exchange between each other.

[7] This form of emergence of a social meaning of language is analysed in detail in Gärdenfors (1993).

In this way, language fosters a *common structure* of the inner worlds of the individuals in a society (see Gärdenfors and Warglien 2006).

Money is an example of "social software" (Parikh 2002) that improves the cooperation within a society. In general, the *social ontology* that we construct increases the possibilities of establishing contracts, conventions and other social goods.

2.8 The cooperation of Homo oeconomicus

Homo oeconomicus, rational man, is assumed to have perfect reasoning powers. He is logically omniscient, a perfect Bayesian when it comes to probability judgments and a devoted utility maximiser. He also has a complete theory of mind in the sense that he can put himself totally in the shoes of his opponents (and vice versa) and see them as perfect Bayesians when reasoning about what would be the possible equilibrium strategies in the game he is facing. When he communicates, his messages have a logically crisp meaning. He cooperates, if, and only if, this maximizes his utility according to the description of the game.

The problem with *Homo oeconomicus* is that he seldom faces the complexities of reality, where the game to be played is not well defined and where its outcomes and the strategies available may fluctuate from minute to minute, where the utilities of the outcomes are uncertain, where reasoning time and memory resources are limited, where communication is full of vagueness and misunderstandings and where his opponents have all kinds of idiosyncracies and cognitive shortcomings, partly depending on their ecological conditions.

In game theory, it is almost exclusively strategies within a fixed game that have been studied. In nature, strategies that can be used in a variety of different but related games are more realistic. How should the rational man act when meeting an opponent in flesh and blood, with all kinds of biological constraints, when he knows that his opponent is not fully rational, but does not know what the exact limitations of the opponent are? Game theory without biology is esoteric; with biology it is as complicated as life itself (see e.g. Eriksson and Lindgren 2002).

3. The effects of communication

I want to make a distinction between, on the one hand, *communication* that involves a choice (conscious or not) by some individual of sending some signal and, on the other hand, mere *signalling* without choosing.[8] Thus a caterpillar, which by its bright colours signals that it is poisonous, is not communicating. Given this definition, communication becomes in itself a form of cooperative strategy. Of course, communication can be misused – what the sender communicates may be a deliberate lie that is intended to mislead the receiver.

When a communicator sends a message, the receiver may initially not be able to interpret the "meaning" of the message. However, in iterated interactions a message may develop a rather fixed meaning. It was Lewis (1969) who introduced a game-theoretic account of conventions of meaning. In this setting, signals transmit information as a result of an evolutionary process. A variety of computer simulations and robotic experiments (e.g. Steels 1999, 2004) have shown that a stable communicative system can emerge as a result of iterated interactions between artificial agents, even though there is nobody who determines any "rules" for the communication. A general finding of the experiments is that the greater number of "signallers" and "recipients" involved in communication about the same outer world, the stronger is the convergence of the reference of the messages that are used and the faster the convergence is attained.

An important aspect is that the addition of communication changes the space of strategies in a game. In game-theoretic settings, it has been thought that signals that cost nothing are of no significance in games that are not games of pure common interest. And evolutionary biologists have emphasized signals that are so costly that they cannot be faked (Zahavi 1975). In contrast to this, Skyrms (2002) shows that in the "stag hunt" game and in a bargaining game, costless signals have dramatic effects on the dynamics compared to the same games without signals. The presence of signals creates new equilibria and change the stability properties and basins of attraction of old

[8] This is in contrast to Hauser (1996), who uses "communication" to also include all forms of signals.

equilibria. The signals that are used in the new equilibria may not end up having a unique "meaning" but they can be polymorphic (see Skyrms (2002), section 4). The upshot is that even if communication is not necessary for some of the forms of cooperation listed above, adding communication to a game may drastically change its structure.

The effect was first pointed out by Robson (1990) in relation to a population of defectors in an evolutionary setting of a PD. A signal that is not used by this population can be used by a mutant as a "secret handshake" that function as a sign that somebody belongs to the ingroup. Mutants would cooperate with those who share the secret handshake and defect against the others. They would then do better than the original defectors and invade the population. Without signals available, pure defection would be an evolutionarily stable strategy in a PD, but with cheap signalling, it is no longer so.

In traditional game-theoretical analysis, the "moves" in a game are assumed to be fixed and it has been assumed that the players either cannot communicate at all, or that they have full access to linguistic communication. However, from an evolutionary perspective these extremes are not very realistic. In many natural settings, the interactors can communicate in some way, and there is almost always a possibility to create a new signal that in combination with previous choices immediately multiplies the number of possible "moves" in a game.

As mentioned above, signals used in a game do not come with a given "meaning" but the role of a signal is settled during the evolution of a game – if ever, since signals may end up in equilibria with polymorphic uses. Taking these factors into account means that a game-theoretical analysis along traditional lines becomes immensely more complicated. Already the simple communication strategies analysed by Skyrms (2002) bear witness of this.

An animal can communicate in various ways without employing much cognitive capacity and a system of signals can reach a communicative equilibrium also without exploiting any cognitive forces of the participants. However, as soon as we bring in the theory of mind and the strong tendency to assign "meanings" to signals that

one finds in humans, the game-theoretical situation becomes more complex again.

To give but one example of the effect of cooperation, Mohlin and Johannesson (to appear) experimentally investigated the effects of communication in a dictator game where the dictator decides how much of an allotted amount to give to a recipient. They compared one-way written communication from an anonymous recipient with no communication and they found that communication increase donations by more than 70 percent. In order to eliminate the "relationship effect" of communication, another condition with communication from a third party was tested. In this third situation, the donations were about 40 percent higher than in the condition with no communication, which suggests that the impersonal content of the communication affects donations.

Charness and Dufwenberg (2006) provide a possible rationale for the effects of this "endogenous communication": The dictator suffers from guilt if he hurts the recipients relative to what they believe that they will get. By taking the guilt into account, in what is called "guilt aversion", the dictator thus relies on his belief about the beliefs of the recipients, which is a clear case of an advanced theory of mind. Communication may affect these beliefs and in this way influence the allocation to the recipient. On this account, it is reasonable to expect that communication from a bystander will have smaller effects than a direct message from a recipient, albeit anonymous.

4. Conclusion

The somewhat eclectic list of different forms of cooperation and their cognitive and communicative prerequisites that were presented in Section 2 can be summarized in the following table:

Type of cooperation	*Cognitive demands*	*Communicative demands*
Flocking behaviour	Perception	None
Ingroup-outgroup	Recognition of group member	None
Reciprocal altruism	Individual recognition, memory, slow temporal discounting, value comparison	None
Indirect reciprocity	Individual recognition, memory, slow temporal discounting, value comparison, minimal theory of mind (empathy)	Proto-symbolic (mimetic) communication
Cooperation about future goals	Individual recognition, memory, anticipatory planning, value comparison, more advanced theory of mind	Symbolic communication (protolanguage)
Commitment and contract	Individual recognition, memory, anticipatory planning, joint beliefs	Symbolic communication (protolanguage)
Cooperation based on conventions	Theory of mind that allows common knowledge	None, but enhanced by symbolic communication
The cooperation of *Homo oeconomicus*	Full theory of mind, perfect rationality	Perfect symbolic communication

Table 1: The cognitive and communicative demands of different forms of cooperation.

The different kinds of cooperation presented here are ordered according to increasing cognitive demands. They do not exhaust the field of possibilities and there is most certainly a need to make finer divisions within the forms presented here. Nevertheless, the table clearly shows that the cognitive and communicative constraints are important factors when analysing evolutionary aspects of cooperation.

It is only humans who clearly exhibit reciprocal altruism, indirect reciprocity, cooperation about future goals and cooperation based on conventions.[9] This correlates well with the cognitive factors that are required for these forms of cooperation.

The analysis, or a more detailed version of it, can be used to make predictions about the possible forms of cooperation that can be found in different animal species. For example, if a monkey species can be shown to recognize members of its own troop and to remember and evaluate the results of some previous encounters with individual members, but if two monkeys cannot communicate about a third individual, then the prediction is that reciprocal altruism between the members of the species can be expected, but not indirect reciprocity. Thus a causal link from cognitive capacities to cooperative behaviours can be established. Conversely if a species exhibits a particular form of cooperative behaviour, then it can be concluded that they have the required cognitive or communicative capacities.

Similarly, the analysis can also be turned into a tool for predictions about the evolution of cognition in hominids. For example, Osvath and Gärdenfors (2005) argue that the first traces of anticipatory planning are found during the Oldowan culture 2.5 million years ago. Thus it is possible that the forms of cooperation about future goals that depend on this form of cognition had been established already then. And conversely, if the archaeological record of some hominid activities provides evidence for certain types of cooperation, the prediction is that the hominids also had developed the cognitive and communicative skills necessary for that form of cooperation, which

[9] As mentioned above, it is open to some controversy whether other species engage in reciprocal altruism.

in turn may generate predictions concerning their brain structure and other physiological factors.

Acknowledgements

I would like to thank Barbara Dunin-Keplicz, Jan van Eijck, Martin van Hees, Kristian Lindgren, Erik Olsson, Mathias Osvath, Rohit Parikh, Erik Persson, Wlodek Rabinowicz, Robert Sugden, Rineke Verbrugge, Annika Wallin and members of the seminars in Cognitive Science and Philosophy at Lund University and of the project group on Games, Action and Social Software at the Netherlands Institute for Advanced Studies for very helpful comments on drafts of the paper. I also want to thank the Netherlands Institute for Advanced Studies for excellent working conditions during my stay there in October 2006. This paper has been written as part of the EU project Stages in the Evolution and Development of Sign Use (SEDSU).

References

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

Axelrod, R. and Hamilton, W. D. (1981): "The evolution of cooperation", *Science 214*, 1390-1396.

Barclay, P. (2005): "Reputational benefits for altruistic punishment", *Evolution and Human Behaviour* 27, 325-344.

Bernhard, H., Fischbacher, U. and Fehr, E (2006): "Parochial altruism in humans", *Nature* 442, 912-915.

Bickerton, D. (1990): *Language and Species*, The University of Chicago Press, Chicago, IL.

Bischof, N. (1978): "On the phylogeny of human morality", in G. Stent (ed.), *Morality as a Biological Phenomenon*, Abakon, Berlin, 53-74.

Bischof-Köhler, D. (1985): "Zur Phylogenese menschlicher Motivation", in L. H. Eckensberger and E. D. Lantermann (eds.), *Emotion und Reflexivität*, Urban Schwarzenberg, Vienna, 3-47.

Brandt, H., Hauert, C. and Sigmund, K. (2003): "Punishment and reputation in spatial public goods", *Proceedings of the Royal Society in London B* 270, 1099-1104.

Brinck, I. and Gärdenfors, P. (2003): "Co-operation and communication in apes and humans", *Mind and Language* 18, 484-501.

Bowles, S. and Gintis, H. (2003): "The origins of human cooperation", in P. Hammerstein (ed.), *The Genetic and Cultural Origins of Cooperation*, MIT Press, Cambridge, MA, 429-443.

Charness, G. and Dufwenberg, M. (2006): "Promises and partnership", *Econometrica* 74, 1579-1601.

Deacon, T. W. (1997): *The Symbolic Species*, Penguin Books, London.

De Waal, F. (2005): *Our Inner Ape*, Riverhead, New York, NY.

Donald, M. (1991): *Origins of the Modern Mind*, Harvard University Press, Cambridge, MA.

Dunbar, R. (1996*): Grooming, Gossip and the Evolution of Language*, Harvard University Press, Cambridge, MA.

Dunin-Kepliz, B. and Verbrugge, R. (2001): "A tuning machine for cooperative problem solving", *Fundamenta Informatica* 21, 1001-1025.

Eriksson, A. and Lindgren, K. (2002): "Cooperation in an unpredictable environment", in R. K Standish, M. A. Abbass, and H. A. Bedau (eds.), *Proceedings of Artificial Life VIII*, MIT Press, Cambridge, MA, 394-399.

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

Frank, R. (1988): *Passions within Reason: The Strategic Role of the Emotions*, Norton, New York, NY.

Frean, M. R. (1994): "The prisoner's dilemma without synchrony", *Proceedings of the Royal Society of London B* 257, 75-79.

Gärdenfors, P. (1993): "The emergence of meaning", *Linguistics and Philosophy 16*, 285-309.

Gärdenfors, P. (2003): *How Homo Became Sapiens*, Oxford University Press, Oxford.

Gärdenfors, P. (2004): "Cooperation and the evolution of symbolic communication", in K. Oller and U. Griebel (eds.), *The Evolution of Communication Systems*, MIT Press, Cambridge, MA, 237-256.

Gärdenfors, P. (to appear): "Evolutionary and developmental aspects of intersubjectivity", to appear in H. Liljenström and P. Århem (eds.), *Consciousness Transitions: Phylogenetic, Ontogenetic and Physiological Aspects*, Elsevier, Amsterdam.

Gärdenfors, P. and Warglien, M. (2006): "Cooperation, conceptual spaces and the evolution of semantics", in P. Vogt et al. (Eds.): *EELC 2006*, LNAI 4211, Springer Verlag, Berlin, 16 – 30.

Gulz, A. (1991): *The Planning of Action as a Cognitive and Biological Phenomenon*, Lund University Cognitive Studies 2, Lund.

Hamilton, W. D. (1975): "Innate social aptitudes of man, an approach from evolutionary genetics", in R. Fox (ed.), *Biosocial Anthropology*, Malaby Press, London, 133-157.

Hammerstein, P. (2003): "Why is reciprocity so rare in social animals? A protestant appeal", in P. Hammerstein (ed.), *The Genetic and Cultural Origins of Cooperation*, MIT Press, Cambridge, MA, 83-93.

Hauert, C. (2001): "Fundamental clusters in spatial 2x2 games", *Proceedings of the Royal Society of London B* 268, 761-769.

Hauert, C. and Schuster, H. G. (1997): "Effects of increasing the number of players and memory size in the iterated prsioner's dilemma: a numerical approach", *Proceedings of the Royal Society of London B* 264, 513-519.

Hauser, M. D. (1996): *The Evolution of Communication*, MIT Press, Cambridge, MA.

Hauser, M. D., Chen, M. K., Chen, F. and Chuang, E. (2003): "Give unto others: Genetically unrelated cotton-top tamarin monkeys preferentially give food to those who altruistically give food back", *Proceedings of the Royal Society of London B* 270, 2363-2370.

Hockett, C. F. (1960): "The origin of speech", *Scientific American* 203(3), 88-96.

Lewis, D. (1969): *Convention: A Philosophical Study*, Harvard University Press, Cambridge, MA.

Lindgren, K. (1991): "Evolutionary phenomena in simple dynamics", in C. G. Langton, J. D. Farmer, S. Rasmussen and C. Taylor (eds.), *Artificial Life II*, Addison-Wesley, Redwood City, CA, 295-312.

Lindgren, K. and Nordahl, M. G. (1994): “Evolutionary dynamics of spatial games”, *Physica* D 75, 292-309.
Lindgren, K. (1997): “Evolutionary dynamics in game-theoretic models”, in B. Arthur, S. Durlauf, and D. Lane (eds.), *The Economy as an Evolving Complex System II*, Addison-Wesley, Reading, MA, 337-367.
Lorek, H. and White, M. (1993): “Parallel bird flocking simulation”, http://citeseer.ist.psu.edu/lorek93parallel.html.
Maynard Smith, J. (1982): *Evolution and the Theory of Games*, Cambridge University Press, Cambridge.
Mohlin, E. and Johannesson, M. (to appear): “Communication: content or relationship?”, manuscript, Department of Economics, Stockholm School of Economics, Stockholm.
Mulcahy, N. J. and Call, J. (2006): “Apes save tools for future use”, *Science* 312, 1038-1040.
Neill, D. B. (2001): “Optimality under noise: higher memory strategies for the alternating prisoner’s dilemma, *Journal of Theoretical Biology* 211, 159-180.
Nowak, M. A. and May, R. M. (1992): “Evolutionary games and spatial chaos”, *Nature* 359, 826–829.
Nowak, M. A. and Sigmund, K. (2005): “Evolution of indirect reciprocity”, *Nature* 437, 1291-1298.
Osvath, M. and Gärdenfors, P. (2005): “Oldowan culture and the evolution of anticipatory cognition”, *Lund University Cognitive Studies* 122, Lund.
Parikh, R. J. (2002): ”Social software”, *Synthese* 132, 187-211.
Preston, S. D. and de Waal, F. (2003): “Empathy: Its ultimate and proximal bases”, *Behavioral and Brain Sciences* 25, 1-72.
Reynolds, C. W. (1987): “Flocks, herds, and schools: A distributed behavioral model”, *Computer Graphics* 21(4), 25-34.
Richerson, P. J., Boyd, R. T. and Heinrich, J. (2003): “Cultural evolution of human cooperation”, in P. Hammerstein (ed.), *The Genetic and Cultural Origins of Cooperation*, MIT Press, Cambridge, MA, 357-388.
Robson, A. J. (1990): “Efficiency in evolutionary games: Darwin, Nash and the secret handshake”, *Journal of Theoretical Biology* 144, 379–396.

Semmann, D., Krambeck, H.-J. and Milinski, M. (2005): “Reputation is valuable within and outside one’s own social group”, *Behavioral Ecology and Sociobiology* 57, 611-616.

Sigmund, K., Hauert, C. and Nowak, M. (2001): “Reward and punishment”, *PNAS* 98, 10757-10762.

Skyrms, B. (2002): “Signals, evolution and the explanatory power of transient information”, *Philosophy of Science* 69, 407–428.

Steels, L. (1999): *The Talking Heads Experiment*, Laboratorium, Antwerp.

Steels, L. (2004): “Social and cultural learning in the evolution of human communication”, in K. Oller and U. Griebel, (eds.): *The Evolution of Communication Systems*, MIT Press, Cambridge, MA, 69-90.

Stephens, D. W., McLinn, C. M. and Stevens, J. R. (2002): “Discounting and reciprocity in an iterated prisoner’s dilemma”, *Science* 298, 2216-2218.

Stevens, J. R. and Hauser, M. (2004): “Why be nice? Psychological constraints on the evolution of cooperation”, *Trends in Cognitive Science* 8, 60-65.

Suddendorf T. and Corballis M. C. (1997): “Mental time travel and the evolution of human mind”, *Genetic, Social and General Psychology Monographs* 123, 133-167.

Sugden, R. (1986): *The Economics of Rights, Co-operation and Welfare*, Blackwell, Oxford.

Tomasello, M. (1999): *The Cultural Origins of Human Cognition*, Harvard University Press, Cambridge, MA.

Trivers, R. L. (1971): “The evolution of reciprocal altruism”, *Quarterly Review of Biology* 46, 35–57.

West, S. A, Gardner, A., Shuker, D. M., Reynolds, T., Burton-Chellow, M., Sykes, E. M., Guiness, M. A. and Griffin, A. S. (2006): “Cooperation and the scale of competition in humans”, *Current Biology 16*, 1103-1106.

Wilkinson, G. S. (1984): “Reciprocal food sharing in the vampire bat”, *Nature* 308, 181–184.

Zahavi, A. (1975): “Mate selection - a selection for a handicap”, *Journal of Theoretical Biology* 53, 205-214.

Zlatev, J., Persson, T. and Gärdenfors, P. (2005): “Bodily mimesis as the ‘missing link’ in human cognitive evolution”, *Lund University Cognitive Studies* 121, Lund.