In: Raimo P. Hämäläinen and Esa Saarinen, eds. 2007. Systems Intelligence in Leadership and Everyday Life.
Systems Analysis Laboratory, Helsinki University of Technology, Espoo.

CHAPTER 16

Emergence of Cooperation and Systems Intelligence

Otto Pulkkinen

Change of human systems for the better is often the result of cooperation. Many Systems Intelligent actions have the effect of uncovering the hidden potential for cooperation in a human context. An understanding of the nature and conditions for the emergence of cooperation therefore provides a useful background for the discipline of Systems Intelligence. Forms of cooperation can emerge in repeated interactions even between self-regarding parties with simple behavioural strategies, but both everyday experience and laboratory experiments indicate that humans have cooperative tendencies that cannot be explained with the model of material self-interest. A large portion of people seem to have intrinsic, non-material, social preferences that can be modelled as strong reciprocity. Interactions in heterogeneous groups with strong reciprocators and selfish individuals can result in systems with tipping points, where small changes of belief structures can lead to the actualization of hidden potential for cooperation. The heterogeneity has also important implications for the sustainability of collective action.

Introduction

Saarinen and Hämäläinen (2004) illustrate the Systems Intelligence embedded in the human capacity to change the hostile and seemingly immobile systems around us by the incident that was instrumental in shaping the civil rights movement in the U.S. in the fifties:

> When Rosa Parks refused to give her seat to a white man in a Montgomery city bus in 1955, most people had not heard of Rosa Parks, considered the bus system a technical matter, did not perceive the city of Montgomery as particularly significant, and would have considered irrelevant the question of a particular bus seat on a particular bus leg. But as Rosa Parks was arrested, the marginal incident snowballed, creating an avalanche that eventually reached epic proportions. Change was going to reshape the entire system of race distinction in the most powerful country of the world.

While Rosa Parks could not foresee the eventually huge consequences of her simple act, she certainly could predict the immediate cost for herself: she was arrested, taken to jail, and was later fined for disorderly conduct. Why did she, nevertheless, decide to stop tolerating the discrimination and face the punishment? Recalling the event later, Parks told "when that white driver stepped back toward us, when he waved his hand and ordered us up and out of our seats, I felt a determination cover my body like a quilt on a winter night" [1]. Clearly, her response was instinctive, emotional and deeply human.

Also the resulting systemic effects, including the bus boycott that followed were dependent on the individual decisions of thousands of people acting at the cost of a huge personal inconvenience. All these people were in effect making simultaneous, independent decisions to act for the larger common good despite the price they had to pay. Why did this kind of spontaneous cooperation emerge so quickly and effortlessly?

It is clear from everyday experience that people often benefit from acting together and sharing the results of the cooperation. In these cases the value of the results of the cooperation for the group is greater than the cost for each individual, but a purely self-regardingly rational human being would nevertheless be better off shirking the cost of cooperation and just benefiting from the results (for example still using the buses during the boycott and letting others to take care of changing the system). However, if everybody followed this logic, the cooperation would collapse. Therefore it seems that when this kind of group cooperation is sustained, it involves an altruistic element (willingness to accept a personal cost for the common good) (Gintis 2003a).

The fact that altruistic behaviour is apparently common among non-related people is, however, puzzling from the evolutionary point of view. Why is the propensity to cooperate even at a personal cost not wiped out from the human gene pool in an environment where individual genes fight for selective survival (Dawkins 1976)? This is the central problem to be explained by the research of cooperation. It is called by Gintis (2003a) the puzzle of prosociality.

Understanding the distinctively human aptitude for cooperation may also have huge practical consequences in the environments most of us live and work in. In particular, Ghoshal (2005) has strongly pointed out the effect of the models of human behaviour employed in academic research and education on the management practices of organizations. According to Ghoshal, prominence of the purely self-interested and rational *Homo Economicus* rooted in the models in neoclassical economics and now widely spread as the standard of human behaviour in economics textbooks has created a self-fulfilling prophecy. The "ideology-based gloomy vision" of human nature as well as the role and goals of firms is perpetuated in the education of business leaders and research with the result that greedy and opportunistic behaviour is seen as natural and appropriate. This mental model has, according to Ghoshal, in turn played its part in some recent cases of corporate misbehaviour.

On the other hand, an interest in reversing the circle of gloomy predictions and outcomes seems to grow within different academic disciplines (Ghoshal 2005). The study of positive psychology (Seligman 2002) and positive organisational scholarship (Cameron, Dutton and Quinn 2003) are welcomed by Ghoshal as the first steps in the direction of a more balanced science of management. In this context the study of the foundations of human cooperation summarized in this article can prove extremely fruitful. Also, some related, prominent research initiatives focusing on the role of trust in social dilemmas have been reported in (Ostrom and Walker 2003).

Cases such as that of Rosa Parks and the civil rights movement are definitely interesting as large-scale examples of the effects of Systems Intelligence in action. On the other hand, Systems Intelligence is also heavily involved in less noticeable, everyday settings. The examples of

[1] http://en.wikipedia.org/wiki/Rosa_Parks (accessed 17 January 2007).

Hämäläinen and Saarinen (2006) of the small systems of workplace and marriage depict instances of what could be called micro-cooperation: positive emergent phenomena in systems with relatively few participants. In micro-cooperation the stakes are often not high; contributions to common well-being can take the form of small gestures like smile, handshaking, rose-buying or taking an inquiring and encouraging attitude towards a presentation given by a colleague. Correspondingly, the ordinary systems of holding back often result from the breakdown of this subtle low-level cooperation. The fact that the stakes are often small (i.e. the immediate individual costs and benefits are often barely discernible) could perhaps explain the ease with which the cooperation opportunities are often overlooked and their systemic effects underestimated.

The background provided by the study of models and origins of human cooperation can contribute to the theory and practice of Systems Intelligence in two different ways. First, the understanding and recognition of the cooperative capacity as a fundamental human quality can help to form and foster mental models supporting Systems Intelligent action. Second, studying the evolutionary background of cooperation may help in the more theoretical endeavour of understanding the natural, ecological aspects of Systems Intelligence.

Iterated Prisoner's Dilemma and the Emergence of Cooperation

The first widely influential scientific explanation for the emergence of cooperation is closely related to the work of Robert Axelrod (1984). This provides experimental evidence and a description for the spontaneous emergence of cooperation in many kinds of evolutionary processes, whether natural, cultural or artificial. Although the case for human cooperation is much richer (as will be seen below), Axelrod's results introduce many of the basic concepts and conditions relevant also in systemic human interactions.

The prisoner's dilemma, PD in short, is one of the most famous and widely known concepts of game theory. It was invented in the early 50's by Merrill Flood and Melvin Dresher and soon formalized by Albert W. Tucker (Poundstone 1993, p. 8). The basic structure of the game is shown in TABLE 1 below in the normal form (Axelrod 1984, p. 8)[2].

The prisoner's dilemma (so named because of the story Tucker used in framing it) has been used extensively in research and literature because of its simplicity and capacity to capture the structure of many basic real-life encounters. The dilemma manifests itself in the fact that, assuming the players are rational and their only purpose is to maximize their own benefit[3], the only so-called Nash equilibrium (combination of strategies such that a player can not improve her payoff regardless of what the opponent chooses) is mutual defection[4]. The mutual defection, however, leaves both players worse off than mutual cooperation would. The "temptation" of a larger win and "fear" of getting nothing leave a self-regardingly rational player no real choice.

[2] In normal form a game is represented as a matrix showing the payoffs related to the combinations of the players' choices, i.e. strategies in the game. The payoff of the row player (player A here) is shown first. Therefore, for example, if player A chooses to defect and player B cooperates, A's payoff is 5 and B's 0. Note also that the dilemma is defined by the mutual relations of the payoffs, not the exact figures used in this example.

[3] In other words: the players behave according to the Homo Economicus model of traditional economics.

[4] To see this, consider the individual symmetric viewpoints of each player: if the opponent cooperates, my best choice is to defect and get the temptation payoff. If she defects, I have to do the same to avoid the sucker's payoff.

But when the same players will meet each other in a successive series of PD games, things get more interesting. In suitable circumstances repeated encounters may create a situation where mutual cooperation can emerge even between self-regarding players. This is the famous iterated PD studied in detail by Axelrod (1984). Axelrod organized a tournament for computer programs playing the repeated PD with the intention of gaining insight into the best strategies. The tournament consisted of matches of 200 successive PD rounds between two participants. Each program (implementation of a particular strategy) was to play a match against each participant (including itself) and, in addition, against a strategy making a random choice in every round. Axelrod received contributions from scientists in different fields, mostly related to game theory and the PD.

TABLE 1. The basic structure of the prisoner's dilemma.

		Player B	
		Cooperate	**Defect**
Player A	**Cooperate**	R=3, R=3 Reward for mutual cooperation	S=0, T=5 Sucker's payoff and temptation to defect
	Defect	T=5, S=0 Temptation to defect and sucker's payoff	P=1, P=1 Punishment for mutual defection

It is straightforward to show that in such a tournament no single strategy is optimal regardless of the opponents (Axelrod 1984, p. 15). The results of the tournament do not, either, prove anything conclusive, but provide some key insights into the properties of a good iterated PD strategy. The tournament was won by a strategy submitted by Anatol Rapoport from the University of Toronto called TIT FOR TAT (TFT for short). TFT was the simplest of the entrants in the tournament and is defined by two simple rules: 1) cooperate in the first round of the iterated PD, and 2) in the subsequent rounds imitate the opponent's last choice. Slightly baffled from the success of this almost trivial strategy among the much more sophisticated creations of game theory experts, Axelrod reported and analysed the results of the tournament in detail, published the report and announced a new tournament with basically the same rules. This time the number of submissions was much larger, Rapoport re-submitted TFT, and TFT won gain.

With these rather unexpected results Axelrod studied the emergence of cooperation further. These studies included the simulation of the evolutionary stability of the different strategies by hundreds of rounds of the computer-based iterated PD where the number of players with each strategy depends on their success in the previous rounds. TFT dominated also the evolutionary simulation in the sense that the population of TFT-playing program agents grew faster than that for any other strategy in every round of the simulation (Axelrod 1984, pp. 48–54). Axelrod defined the property of collective stability in the context of the evolutionary game: a strategy s is collectively stable if no other strategy can invade a population consisting of s-playing agents. He also showed that TFT is collectively stable if and only if the value of potential future payoffs related to the payoffs of the current round (the discount factor w) is larger than a minimum value which is a function of the four payoff parameters T, R, P, and S (Axelrod 1984, p. 59)[5]. This relates

[5] The discount factor is related to the probability of the game ending after the current round (which, in other words, has to be small enough for cooperation to be stable).

directly to the lasting value of Axelrod's contribution: in these specific conditions he was able to explain the plausibility of spontaneously emerging cooperation in the living nature, including humans.

Axelrod's Advice

Axelrod cooperated with biologist William D. Hamilton to extend the analysis of the evolutionary properties of TFT to very simple biological systems without human understanding and foresight (Axelrod 1984, pp. 88–105). But, being a political scientist by education, he was also interested in the practical implications of his results in the context of human social systems.

Axelrod highlights four properties of TFT as the basis of its success: the strategy is nice, provocable, forgiving and clear. Niceness means that in an encounter for another strategy, TFT is never the one to defect first. Provocability refers to the immediate retaliation after a defection by the opponent and discourages others from taking advantage of the cooperative tendencies. Forgiveness, i.e. the fact that TFT is ready to cooperate immediately when an opponent is, helps to restore the cooperation after mutual defection. Finally, the clarity and simplicity of the strategy enables others to recognize and understand it. This elicits long-term cooperation as others see and believe that TFT is ready to cooperate but can not be exploited.

Based on these insights Axelrod formulates recommendations for actors in real-life situations resembling the iterated PD (Axelrod 1984, pp. 109–123). Their relevance for the study of Systems Intelligence is that they provide a lesson in the basic laws of inducing cooperation. Although most human situations actually seem to be considerably richer in terms of motives for cooperation, Axelrod's maxims contain elements that are a part of successful cooperative strategies also in more complex systems involving repeated interactions.

(1) Do not be envious.

People and organizations involved in an iterated PD-type situation often confuse it with a zero-sum game, where the gains of one player equal the losses of the other. Companies in supply relationships, nations in trade arguments, and students in lab tests get involved in costly spirals of mutual retaliative defection because they erroneously assume that they have to do better than the other player in a particular two-party game in order to flourish. The iterated PD, however, is not a zero-sum game. Instead of the payoff of the other player, the correct reference for comparison is the best possible overall success of one's own. This is illustrated by the fact that while TFT clearly won both Axelrod's tournaments, it can never do better that its opponent in an individual game (because it is never the first to defect and never defects more often that the opponent).

(2) Do not be the first to defect.

Keeping in mind the qualifications related to the discount factor, being nice (not defecting) proved to be the distinctive feature of successful strategies in Axelrod's computer tournaments. He points out that the value of this property was underestimated by surprisingly many game theory experts. Strategies trying out sophisticated methods of exploiting others were repeatedly caught in mutual defection in cases where cooperation would have been possible. On the other hand, large part of the success of nice rules (including TFT) was their ability to initiate success together.

(3) Reciprocate both defection and cooperation.

Quick reciprocation of both cooperative and defecting actions has a central role in establishing cooperative behaviour in a group. Consider a population consisting entirely of the purely defecting strategies called ALL D. When a single actor with a different strategy is interacting with the population, it is never cooperated with and therefore can never flourish and invade the population. In other words, ALL D is always collectively stable. But the population can be invaded by a small group of cooperative strategies in circumstances where the benefits from their mutual cooperation can outweigh the occasional exploitation by a defecting ALL D. It turns out that the invading capability of a cooperative strategy is related to its capability to distinguish between ALL D (in order to minimize the exploitation) and itself (in order to benefit from cooperation). The strategies that are optimal in this sense are called by Axelrod (1984, p. 66) maximally discriminating. TFT has this property because of its reciprocity, i.e. because it will immediately cooperate with a copy of itself and never be exploited by ALL D after the initial round.

Thus the two-way reciprocity of a TFT-like strategy enables actors using it to establish cooperation in an adverse environment. On the other hand, once a nice, collectively stable strategy like TFT has established itself, it can not be invaded by ALL D even in groups[6] (Axelrod 1984, p. 67).

(4) Do not be too clever.

Unlike in a zero-sum game (such as chess), the other player in an iterated PD must not be regarded as someone who is out to defeat you. As you gain most by building a pattern of lasting cooperation, you must be aware that the other is watching you for sings of your intentions to cooperate and your own actions are likely to be echoed back. Therefore it pays off to be easily readable. TFT is very good in this respect: once the simple behavioural pattern is understood by the other player, it becomes clear that cooperation is the best option she has as long as it is reasonably probable that there will be a next round.

The Limits of Reciprocal Altruism

Axelrod's results together with some parallel research (e.g. Trivers 1971, Fudenberg and Maskin 1986) on cooperation provide a powerful explanation for human cooperation in small and stable groups. This model is often called reciprocal altruism[7] (Fehr and Fischbacher 2003). Reciprocal altruism is commonly held as a fundamental description of the importance of long-term relationships for the emergence of cooperation. There is unambiguous experimental evidence that people are more likely to cooperate in two-person interactions when future interactions are more probable (Andreoni and Miller 1993; Gächter and Falk 2002).

However, there's clearly much more to human cooperation than the bilateral reciprocal altruism of stable groups. In particular, it has been shown that the success of TFT-like strategies is very limited when the interactions take place between several individuals instead of just two. In an iterated n-person PD the only conditionally cooperative, evolutionarily stable strategy allows cooperation only if all other players cooperated in the previous round. Therefore the basin of attraction for emergent cooperation is very small, because the existence of a small number of non-cooperative participants suffices to prevent it (Boyd and Richerson 1988; Fehr and Fischbacher 2003).

[6] And for a nice strategy to be collectively stable, it must be provoked by (i.e., reciprocate) the very first defection of the opponent (Axelrod 1984, p. 62).

[7] The term altruism is used because cooperation in a PD-type situation involves giving up an immediate personal gain for longer term mutual benefit.

Also, the limitation related to the likelihood of future interactions appears to be a serious one. In Axelrod's second computer tournament the discount factor w was set so that the median length of the games would be 200 rounds (Axelrod 1984, p. 42). But throughout the evolutionary history, humans have probably almost always had the option of stopping to interact with nonrelated individuals and getting away with a defection. This removes a key condition for the emergence of cooperation purely through reciprocal altruism.

Finally, and perhaps most relevantly, reciprocally altruistic cooperation is based entirely on the expectation of future gains. However, there is plenty of evidence from both everyday life and a large amount of laboratory experiments that humans actually cooperate and behave altruistically without any material incentives (Fehr and Fischbacher 2003). It has even been proposed that this non-materially motivated altruism is an illusion created by the fact that the cooperative capabilities of humans have evolved in environments supporting reciprocal altruism and we tend to systematically overestimate the future gains in current real-life interactions. However, this does not seem probable in the light of a large body of evidence from experiments where these effects have been systematically ruled out. Also, it seems that humans actually have very well developed cheating detection capabilities, which suggests that we are fine-tuned to an environment with short-lived interactions (Fehr and Fischbacher 2003).

Reputation-Seeking and Indirect Reciprocity

An important step in understanding the emergence of human cooperation is the introduction of the concept of reputation. Reputation-based models of cooperation relax the condition of stable long-term interactions required for reciprocal altruism. The possibility to acquire a reputation for being cooperative can both help an individual to receive cooperation even in short-term interactions with others and provide an incentive for cooperation.

Indirect reciprocity (Alexander 1987; Nowak and Sigmund 1998; Milinski et al. 2002) is a model of cooperation based on reputations. In an indirectly reciprocal social system, cooperation is directed towards individuals seen as valuable to the community. Help is provided to recipients that are likely to help others (which often means having a visible history of helping others). Therefore it also pays to advertise one's cooperative capacities. In particular, Nowak and Sigmund (1998) showed that in a simplified computer simulation model of indirect reciprocity, the emergence of cooperation depends, besides the frequency of interactions, the availability and reliability of information about the cooperative tendencies of others.

Living in this kind of systems requires sophisticated skills for the assessment of the status of others and for the analysis, planning and anticipation of social situations. These requirements may well have been a major force in the evolutionary shaping of our language and intelligence (Nowak and Sigmund 1998), contributing also to the development of Systems Intelligence. It is tempting to think that micro-reputations (almost unnoticeable beliefs regarding to the nuanced behaviour and attitudes of others) could develop and influence micro-cooperative situations.

Cooperation in the Laboratory: The Ultimatum Game

Another two-player game, the ultimatum game, has probably recently inherited the position of the prisoner's dilemma as the most widely used game-theoretic research tool. In an experimental ultimatum game a sum of money (say 10 €) is given to be divided by two players under conditions of anonymity (Gintis et al. 2005b). The task of one of the players, called a proposer, is to offer any portion of the total sum to the second player, called a responder. The responder (again anonymously) can choose whether to accept the offer. If she accepts, the sum is divided as proposed. Otherwise both players receive nothing.

Since the game is played only once and the players do not know the identities of each other, a responder interested only in her personal material payoff will accept any positive offer since the alternative is to get nothing. A similarly self-regarding proposer will propose the minimum possible amount allowed by the rules (say 1 €), which, according to a standard equilibrium analysis, will be accepted[8].

The experimental ultimatum game has been replicated in laboratory tests numerous times under varying conditions and sums of money. What actually happens is that a very small minority of the players behave in the self-regarding manner predicted by the equilibrium analysis. What we see instead is a form of cooperation emerging between the proposers and the responders. The proposers generally offer substantial portions of the money (50% of the total is generally the modal offer), and proposals below 30% of the total are frequently rejected. There is a great deal of individual variability in the results, with about quarter of test subjects behaving in a self-regarding manner. On average, however, the results are very similar in all the studies (Gintis et al. 2005b).

The results of the ultimatum game experiments therefore challenge to the traditional explanations of economic behaviour. They indicate a need for more sophisticated analysis and models of human preferences in cooperative situations.

The Public Goods Games and Group Cooperation

As discussed above, the explanatory power of reciprocal altruism also falls short when we are seeking an explanation for the emergence of group cooperation. A precise illustration of the nature of this shortcoming is given in laboratory settings by numerous reported examples of experimental games called public goods games.

A typical public goods game has several rounds, ten for example (Gintis et al. 2005b). The test subjects are fully explained all the rules and aspects of the game. In each round of the game, a test subject is grouped with several (for example, three) others. Strict anonymity is maintained. Each player is then given a set of point (for example, twenty) that can be changed to real money at the end of the game session. At the start of the round, each player places a fraction of her points to a common account and keeps the rest in her private account. The game administrator then tells each player how many points have been contributed to the common account and adds to each private account some portion of it (for example, 40%). So, if a player contributes in the first round all her 20 points, this will cost her 12 points but create a total benefit of 24 points (8 x 3) for the other players in the group. If everybody in the group did this, each player would have 32 (8 x 4) points after the first round.

It is easy to see that the only Nash equilibrium of the public goods game is a combination of strategies where each player contributes zero points in every round. Despite the fact that cooperation allows everybody to gain, a self-regarding player can always maximize her payoff by

[8] In game-theoretical terms, this is the only subgame perfect equilibrium of the game. A subgame perfect equilibrium is a strategy set representing a Nash equilibrium of every subgame (i.e. a game that consists of all the moves made by the players after a given point in the game) of the original game. Therefore it is a stricter equilibrium definition than the Nash equilibrium.

contributing nothing[9]. But in reality, cooperation again emerges. Only a small fraction of players conform to the self-regarding model. Most players begin by contributing about half of their endowment to the public account; then the level of contributions decays over the rounds and in the final rounds most players are behaving in a self-regarding way (Dawes and Thaler 1988).

When questioned about their motives for decreasing their contribution, people typically bring up retaliation towards free-riders (Andreoni 1995). There is also direct experimental evidence supporting the interpretation that when subjects are allowed to "punish" noncontributors in a public goods game, they are willing to do it at a cost to themselves. For instance, Ostrom, Walker, and Gardner (1992) report a study, where subjects played a 25-round public goods game. Players could impose additional costs (fines) on others by paying a fee. If the players behave self-regardingly, no player ever pays the fee, nobody is ever fined for defecting and nobody contributes to the common pool. However, a significant amount of punishing actually took place.

Most of the original experimental ultimatum and public goods game studies were conducted in the U.S or Western Europe and using local students as test subjects. To find out whether the results reflect a universal human behavioural trait or are related to the culture of western university students, a group of researchers undertook a large cross-cultural study of behaviour in the ultimatum and public goods games (Henrich et al. 2001). Test subjects were recruited in 15 small-scale societies with a wide variety of economic and cultural conditions. The central finding of the study was that the self-regarding actor model is not supported by the results in any of the societies. In addition, there was a substantial amount of behavioural variation between the groups. The differences of economic organization and market integration within the cultures correlated strongly with the strength of cooperative tendencies in the games. In general, the game behaviour was consistent with economic patterns of the everyday life in the societies in question. Thus, the tendency for cooperative behaviour in the ultimatum and public goods game seems to be universally human, with the local culture influencing the detailed forms of the cooperative interactions.

Social Preferences: Fairness, Inequity Aversion and Strong Reciprocity

The evidence gathered during last years from the experimental ultimatum and public goods games presents a strong challenge to the model of humans as self-regarding payoff-maximizers that is often referred to as Homo Economicus. It clearly seems that besides material motivations, people are directed in their behaviour by other types of goals, often called social preferences (Fehr and Fischbacher 2005). Both in the experimental research laboratory and in everyday life, people simply seem to care about the well-being of others in a way that has been inaccessible for economics research.

Many researchers have started to include social preferences in theoretical models of human behaviour and to use the models in the analysis of organizational and economic phenomena (Gintis et al. 2005a). It is fundamental for the resulting models that (in contrast to earlier mainstream economics research) cooperative motives are not modelled as means to some other goals, but as ends in themselves (arguments in an individual's preference function). A variety of so-called prosocial emotions, including empathy, shame and guilt bias individual behavioural choices towards prosocial directions (Gintis 2003a). In Gintis' words: internalized prosocial norms are constitutive of the self. In fact, some recent research (Rilling et al. 2002) indicates that social

[9] This can be seen by backward induction: in the last round zero contribution clearly gives maximal payoff, when this happens the same holds for the previous round etc.

cooperation has a clearly identifiable neural basis. An interesting question for the research of Systems Intelligence would be whether similar, identifiable connections to neural activity could be found for Systems Intelligent behaviour.

Among the theoretical models developed for non-selfish behaviour are reciprocal fairness (Rabin 2001) that explicitly models different types of fairness motives in a game-theoretic framework, and inequity aversion (Fehr and Schmidt 1999, Bolton and Ockenfels 2000). A central point in Rabin's work is that in a prisoner's dilemma-type game people are actually willing to cooperate if they believe that the opponent will do the same. This leads to mutual cooperation being another equilibrium (besides mutual defection) and makes the individual's beliefs of each other crucial for the outcome of the interaction. Fehr and Schmidt develop a simpler and more easily analyzable model based on the notion that a fair share of people are inequity-averse, i.e. they value an equitable division of payoffs in itself. This leads to altruistic cooperation increasing the other's payoffs towards an equitable level. The other side of the inequity aversion model is envy – the strive to increase one's own payoffs until the equitable level is reached.

Both in the experimental research laboratory and in everyday life, people simply seem to care about the well-being of others in a way that has been inaccessible for economics research.

However, the seemingly by far the most significant model for social preferences and their role in the emergence of cooperation is that of strong reciprocity (Fehr and Fischbacher 2003; Gintis et al. 2005a). Quoting Gintis et al. (2005b):

> Strong reciprocity is a predisposition to cooperate with others, and to punish (at personal cost, if necessary) those who violate the norms of cooperation, even when it is implausible to expect that these costs will be recovered at a later date.

In particular, strong reciprocity is a model that is compatible with the observed behaviour in the ultimatum and public goods games. For the people involved in the experimental ultimatum games, the 50–50 outcome represents a fair split of the money. Responders reject proposals below 40% as a form of altruistic punishment for a non-fair behaviour. Proposers offer 50% because they are predisposed to being fair and cooperating, or at least 40% because they understand that non-fair proposals get rejected even in an anonymous one-shot game. This is supported also by the interesting result that if the offer in an ultimatum game is generated by a computer instead of a human, low offers are very rarely rejected (Blount 1995).

The decisive action of Rosa Parks discussed above can also be interpreted in terms of strong reciprocity. She felt strongly that the system that required her to stand up and move away from the seat in the bus was violating the norms of humanity and no longer working for the common good. With an instinctive, subjective certainty backed by strong prosocial emotions, she just stopped cooperating (and that way "punished" the violators of the universal norms) despite her understanding of the eventual personal cost.

On of the reasons for the influence of models of reciprocal altruism is that they provide a very clear explanation of the evolutionary mechanics of the cooperative behaviour. The apparent paradoxicality of the theories of social preferences in general and strong reciprocity in particular is related to the fact that such a straightforward evolutionary explanation is not available (Fehr and Fischbacher 2003)[10]. Some recent work proposes models based on cultural group selection (Boyd et al. 2005) or gene-culture coevolution (Gintis 2003b) as solutions to this puzzle. They are based on the idea that cultural norms supporting cooperation are sustained by altruistic punishment. If a sufficient number of altruistic (strongly reciprocal) punishers exist, cooperators gain an advantage over defectors who get punished. When cooperation is widely established, the cost incurred by altruistic punishers is very small, because actual punishment does not take place. Instead, the common belief in the reputation of the punishers is sufficient to sustain the cooperation (Fehr and Fischbacher 2003).

Heterogeneous Interactions, Equilibria and Tipping Points

Standard neoclassical economic theory mostly starts from the assumption that people have homogeneous (rational and self-regarding) preferences. The resulting models describe well some aspects of group-level economic behaviour, particularly those in competitive markets. However, as seen above, some experimental results can not be explained in this framework. It appears that a more plausible theoretical foundation involves more heterogeneity in human motivations, in particular the existence of a considerable portion of people having social preferences. It seems that many phenomena related to quick and large changes in human systems are generated by the interactions of people with different outlooks and involve also the effects of interaction structure, reputations and beliefs about others.

As a simple example of how subtle changes in the design of the structure of interactions can change the equilibrium of a system of people with heterogeneous preferences, consider a prisoner's dilemma played by a player behaving in a self-regarding manner and a strongly reciprocating player (Fehr and Fischbacher 2005). Assume also that the types of the players are common knowledge (everybody involved knows it and also knows that the others know). If the game is simultaneous (i.e. both players announce their moves without knowing the move of the other), the unique equilibrium of the game is still mutual defection (because the strong reciprocator knows the other player will defect and does the same). But if the game is sequential (the players move one after another) with the self-regarding player moving first, the first player, knowing the reciprocal behaviour of the second, effectively chooses between the outcomes (defect, defect) and (cooperate, cooperate). Being rational, she selects the latter, which therefore is the unique equilibrium.

One important insight to the effects of heterogeneity is related to the observed breakdown of cooperation in repeated public goods games (Fiscbacher et al. 2001; Fehr and Fischbacher 2003). Despite the fact a large number of strong reciprocators are involved, they cannot prevent the decay in contributions in the circumstances in question. Results derived within the analytical models for heterogeneous populations including strong reciprocity (Gintis 2003a) and inequity aversion (Fehr and Schmidt 1999) show that with fairly general assumptions, a minority of selfish individuals suffices to make the absence of cooperation the only equilibrium. An implication of this is that even if no cooperation can be observed in a social system, it is not possible to infer the absence of altruistic individuals. The strong reciprocators just withhold their cooperation if they believe that the others are not contributing. On the other hand, if they start to believe that others around them are likely to cooperate, they will respond by contributing in kind and creating a strengthening wave of cooperation. For the long-term maintenance of cooperation in a group setting, it is therefore vitally important to sustain the mutual belief in the cooperative outcome.

[10] An illustration of the phenomena of evolutionary instability is the breakdown of the cooperation described in the previous paragraph.

More generally, the heterogeneity of preferences leads to models of group behaviour where, instead of the single, gloomy equilibrium of universal defection, multiple equilibria exist (Gintis 2003a, Fehr and Schmidt 1999), including those of universal cooperation and mixed cooperation and defection. An interesting feature of the models with multiple equilibria is that they predict the existence of tipping points: critical combinations of system parameters such that a small change may cause a very large overall effect (for example, the emergence of universal cooperation). The tipping point effects describe the situations where small cooperative (of defecting) actions can change the nature of a system.

This is, then, what could have happened in the small revolution initiated by the action of Rosa Parks. The visibility (strengthened and highlighted by the Systems Intelligent actions of the community leaders) of her case, together with her reputation as a respectable and morally developed individual, induced some followers to cooperate regardless of the immediate personal cost. As the word was spreading, the common belief in the cooperative protest strengthened further and attracted others, even those with weaker cooperative propensities, to join. Finally even the minority of self-regarding individuals in the community probably saw that participating in the protest made sense in order to not be regarded as a free-rider.

Social and Organizational Implications

A considerable part of contemporary management and policy analysis is based on the earlier widely accepted assumption that all individuals pursue materially selfish goals. The theoretically predicted outcome in many types of collective-action situations is zero or very low level contributions to common good by the individuals. Consequently, the only perceived tools to overcome these Pareto inefficient equilibria are centrally designed and implemented, positive and negative material incentives. Centralized management or the state are viewed as substitutes for the shortcomings of individual behaviour and the presumed failure of community (Ostrom 2005).

However, as seen above, a large part of people in any setting are in fact likely to have intrinsic motivations for social behaviour that can be modelled as strong reciprocity. In addition, the proportions of different types of individuals are likely to change over time as the result of self-selection into different situations and changes in preferences. The resulting heterogeneity transforms many social cooperative dilemmas into different types of games with several equilibria. In particular, the interactions of strong reciprocators and rational egoists (self-regarding individuals) result in situations where it is not possible to rely exclusively on the intrinsic motivations of the participants, especially if cooperation needs to be sustained over time. In these cases the intrinsic motivation can be backed up by institutions that can enable the motivated individuals to solve collective-action problems while protecting them from free-riders (Ostrom 2005).

The institutional rules crafted in many robust, self-organized common-property regimes are compatible with the general conditions for the sustainability of cooperation as well as reciprocity (Ostrom 2000). They tend to increase the probability of long-term, repeated interactions among the participants. Furthermore, appropriation rights tend to be designed so that the actions of an individual are visible to others and thus a reputation in the community will be built quickly.

From the viewpoint of sustained cooperation, especially important are the phenomena of crowding out and crowding in (Ostrom 2005). Institutional systems can crowd out (i.e. diminish or drive out of existence) behaviours based on intrinsic motivations when individuals feel that their self-determination or self-esteem has been hurt by their design. Crowding in is possible when the systems support the individuals. In particular (Frey and Jegen 2001):

- External interventions crowd out intrinsic motivation when they are perceived as controlling.
- External interventions crowd in intrinsic motivation if they are perceived as supportive; in this case individuals feel that they are given more freedom to act and their self-determination is enlarged.

Conclusions

Human cooperation takes many different forms and depends on diverse motivations and environmental factors. The reciprocal altruism represented by the TFT strategy in Axelrod's computer tournaments gives the simplest explanation for spontaneously emerging cooperation. Although the explanation seems simplistic in the face of the full richness of human social systems, it is encouraging because it shows how cooperation can in suitable circumstances emerge even in the most hostile and seemingly inhuman environments. Axelrod (1984, pp. 73–87) describes as an extreme example the "live-and-let-live" system that spontaneously and against the leaders' explicit orders emerged between the enemies in the World War I trench warfare.

Encouraging advice based on Axelrod's results can be given to an individual wishing to foster cooperation even if the system around her seems unresponsive and intriguing. With some qualifications, it pays to be nice and forgiving, i.e. ready for cooperation from the beginning and quick to let bygones be bygones. Being quick to reciprocate and making your behaviour clear and credible for everybody helps to sustain a cooperative environment. The key qualification is that interactions between the same parties need to be continued for a fairly long time for all this to work.

Possibility for long-term repeated interactions is in general very beneficial for the emergence of human cooperation. Regardless of the intrinsic motivations of the interacting people, it is very useful to try to increase both the frequency of communication between them and the duration of the period during which it happens. A team in a business organization is more likely to cooperate efficiently towards a common objective if internal meetings and discussions are frequent and lively and the team stays together for a long time.

However, both in normal life and laboratory experiments many people seem to contribute to common goals for the sake of their intrinsic social values, often in cases when it is totally unrealistic to expect any personal gain in return. Similarly, some people are willing to punish or discipline others at their own cost when the others are seen to violate common norms. When social systems are formed by people with differing levels of these predispositions, they may have several cooperative equilibria. Specifically, altruism can be hidden in the sense that although everybody is acting selfishly, a small change in the beliefs of the individuals regarding others may enable the emergence of cooperation. On the other hand, a small portion of selfish participants typically suffices to make the cooperation difficult to sustain by intrinsic motivation alone over longer periods. In these cases, a carefully designed regime may help. In particular, many successful decentralized regimes rely on making sure the participants interact repeatedly and that their actions are visible to everybody. This makes reputation formation easier; reputations, on the other hand provide an extra incentive for co-operation.

Systems Intelligence believes in and relies on human cooperation, both on large (as in human rights movements) and small (as in marriage or work environment) scales. In practice, the sudden dynamics of social tipping points are often enabled and large-scale cooperation empowered by Systems Intelligent behaviour.

Systems Intelligence is seen to be an instinctive, natural and evolved human capability (Saarinen and Hämäläinen 2004). Therefore the research on systems of actual human cooperation and human properties such as strong reciprocity contribute also to the understanding of the roots and application of Systems Intelligence. The recent results described above, highlighting the previously theoretically underappreciated richness of human cooperative behaviour support the claims for Systems Intelligence as a fundamental social capability. Specifically, understanding the evolution of cooperation in human societies can also increase the understanding of the evolution of Systems Intelligence. Ultimately, this could even lead to an understanding of the neural basis of Systems Intelligent behaviour.

References

AKERLÖF G.A. 1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, vol. 97, pp. 543–569.

ALEXANDER R.D. 1987. *The Biology of Moral Systems*. New York: Aldine de Gruyter.

ANDREONI J. 1995. Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review*, vol. 85, no. 4, pp. 891–904.

ANDREONI J. AND J.H. MILLER. 1993. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, vol. 103, pp. 570–585.

AXELROD R. 1984. *The Evolution of Cooperation*. New York: Penguin Books.

BOYD J. AND P.J. RICHERSON. 1988. The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, vol. 132, pp. 337–356.

BLOUNT S. 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, vol. 63, no. 2, pp. 131–144.

BOLTON G.E. AND A. OCKENFELS. 2000. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, vol. 100, pp. 166–193.

BOYD R., H. GINTIS, S. BOWLES, AND P.J. RICHERSON. 2005. The evolution of altruistic punishment. In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, Gintis H., Bowles S., Boyd R., and Fehr E., eds., Cambridge: The MIT Press, pp. 215–227.

CAMERON K.S., J.E. DUTTON, AND R.E. QUINN, EDS. 2003. *Positive Organizational Scholarship: Foundation of a New Discipline*. San Fransisco: Berrett-Koehler.

DAWES R.M. AND R.H. THALER. 1988. Cooperation. *Journal of Economic Perspectives*, vol. 2, no. 3, pp. 187–197.

DAWKINS R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

FEHR E. AND U. FISCHBACHER. 2003. The nature of human altruism. *Nature*, vol. 425, pp. 785–791.

FEHR E. AND U. FISCHBACHER. 2005. The economics of strong reciprocity. In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, Gintis H., Bowles S., Boyd R., and Fehr E., eds., Cambridge: The MIT Press, pp. 151–191.

FEHR E. AND K.M. SCHMIDT. 1999. A theory of fairness, competition and cooperation. *The Quarterly Journal of Economics*, vol. 114, pp. 817–868.

FISCHBACHER U., S. GÄCHTER, AND E. FEHR. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, vol. 71, pp. 397–404.

FREY B.S. AND R.M. JEGEN. 2001. Motivation crowding theory. *Journal of Economic Surveys*, vol. 15, no. 5, pp. 589–611.

FREY B.S. AND F. OBERHOLZER-GEE. 1997. The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, vol. 87, pp. 746–755.

FUDENBERG D. AND E. MASKIN. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, vol. 54, no. 3, pp. 533–554.

GHOSHAL S. 2005. Bad management theories are destroying good management practices. *Academy of Management Learning & Education*, vol. 4, no. 1, pp. 75–91.

GINTIS H. 2003a. Solving the puzzle of prosociality. *Rationality and Society*, vol. 15, no. 2, pp. 155–187.

GINTIS H. 2003b. The hitchhiker's guide to altruism: Genes, culture and the internalization of norms. *Journal of Theoretical Biology*, vol. 220, no. 4, pp. 407–418.

GINTIS H., S. BOWLES, R. BOYD, AND E. FEHR, EDS. 2005a. *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge: The MIT Press.

GINTIS H., S. BOWLES, R. BOYD, AND E. FEHR. 2005b. Moral sentiments and material interests: Origins, evidence, and consequences. In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, Gintis H., Bowles S., Boyd R., and Fehr E., eds., Cambridge: The MIT Press, pp. 3–39.

GÄCHTER S. AND A. FALK. 2002. Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics*, vol. 104, no. 1, pp. 1–26.

HENRICH J., R. BOYD, S. BOWLES, C. CAMERER, E. FEHR, H. GINTIS, R. MCELREATH, M. ALVARD, A. BARR, J. ENSMINGER, K. HILL, F. GIL-WHITE, M. GURVEN, F. MARLOWE, J.Q. PATTON, N. SMITH, AND D. TRACER. 2005. 'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, vol. 28, pp. 795–855.

HÄMÄLÄINEN RAIMO P. AND ESA SAARINEN. 2006. Systems intelligence: A key competence in human action and organizational life. *Reflections: The SoL Journal*, vol. 7, no. 4, pp. 17–28. Reprinted in *Systems Intelligence in Leadership and Everyday Life*, Raimo P. Hämäläinen and Esa Saarinen, eds., 2007, Espoo: Systems Analysis Laboratory, Helsinki University of Technology.

MILINSKI M., D. SEMMANN, AND H-J. KRAMBECK. 2002. Reputation helps solve the 'tragedy of the commons'. *Nature*, vol. 415, pp. 424–426.

NOWAK M.A. AND K. SIGMUND. 1998. Evolution of indirect reciprocity by image scoring. *Nature*, vol. 393, pp. 573–577.

OSTROM E. 2000. Collective action and the evolution of social norms. *Journal of Economic Perspectives,* vol. 14, no. 3. pp. 137–158.

OSTROM E. 2005. Policies that crowd out reciprocity and collective action. In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, Gintis H., Bowles S., Boyd R., and Fehr E., eds., Cambridge: The MIT Press, pp. 253–275.

OSTROM E. AND J. WALKER. 2003. *Trust and Reciprocity: Interdisclipinary Lessons from Experimental Research.* New York: Russell Sage Foundation.

OSTROM E., J. WALKER, AND R. GARDNER. 1992. Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, vol. 86, no. 2. pp. 404–417.

POUNDSTONE W. 1993. *Prisoner's Dilemma*. New York: Anchor Books.

RABIN M. 1993. Incorporating fairness into game theory and economics. *American Economic Review*, vol. 83, no. 5. pp. 1281–1302.

RILLING J.K., D.A. GUTMAN, T.R. ZEH, G. PAGNONI, G.S. BERNS, AND C..D. KILTS. 2002. A neural basis for social cooperation. *Neuron*, vol. 35, pp. 395–405.

SAARINEN ESA AND RAIMO P. HÄMÄLÄINEN. 2004. Systems intelligence: Connecting engineering thinking with human sensitivity. In *Systems Intelligence: Discovering a Hidden Competence in Human Action and Organisational Life*, Raimo P. Hämäläinen and Esa Saarinen, eds., Espoo: Systems Analysis Laboratory Research Reports A88, Helsinki University of Technology, pp. 9–37. Reprinted in *Systems Intelligence in Leadership and Everyday Life*, Raimo P. Hämäläinen and Esa Saarinen, eds., 2007, Espoo: Systems Analysis Laboratory, Helsinki University of Technology.

SELIGMAN M.E.P. 2002. *Authentic Happiness*. New York: The Free Press.

TRIVERS R.L. 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology*, vol. 46, no. 1, pp. 35–57.

Internet References

WIKIPEDIA. *Rosa Parks*. http://en.wikipedia.org/wiki/Rosa_Parks (accessed 17 January 2007).

Author

The author works as a senior system design manager in the Nokia Corporation.

otto.pulkkinen@nokia.com