Universidad del Desarrollo
Facultad de Gobierno

Documento de Trabajo

Nº9

"Theories of the evolution of cooperative behavior: A critical survey plus some new results"

Robert E. Rowthorn

Ricardo A. Guzmán

Carlos Rodríguez-Sickert

Enero de 2010

Facultad de Gobierno

Theories of the evolution of cooperative behaviour: A critical survey plus some new results

Robert E. Rowthorn [a]

Ricardo A. Guzmán [b]

Carlos Rodríguez-Sickert [b]

January 2009

Abstract

In this paper we survey the various theories of gratuitous cooperation (in favour of non-relatives and without repeated interaction), and we describe our own effort to integrate these theories into a self-contained framework. Our main conclusions are as follows. First: altruistic punishment, conformism, and gratuitous cooperation coevolve, and group selection is a necessary ingredient for the coevolution to take place. Second: people do not cooperate by mistake, as most theories imply; on the contrary, people knowingly sacrifice themselves for others. Third: in cooperative dilemmas conformism is an expression of preference, not a learning rule. Fourth: group mutations (e.g., the rare emergence of a charismatic leader that brings order to the group) are necessary to sustain cooperation in the long-run.

Keywords: Cooperation; altruism; altruistic punishment; conformism; group-selection; group mutations.

1. Introduction

A curious thing about of the human animal is that he is prone to sacrifice himself for non relatives who are unlikely to return the favour. This phenomenon, which we term *gratuitous*[1] *cooperation*, eludes traditional evolutionary explanations. A recent literature attempts to explain gratuitous cooperation applying the tools of several behavioural sciences: biology, cultural anthropology, evolutionary psychology, and economics. In this paper we survey the various ideas

[a] Corresponding author. Faculty of Economics, University of Cambridge, Cambridge, UK, and King's College, Cambridge, UK. E-mail: rer3@econ.cam.ac.uk.

[b] Escuela de Administración, Pontificia Universidad Católica de Chile, Santiago, Chile. E-mails: rnguzman@puc.cl (Guzmán), crodrigs@puc.cl (Rodríguez-Sickert).

[1] The adjective 'gratuitous' derives from the Latin word *gratuitus*, meaning free, freely given, spontaneous.

that float in this literature, and we describe our own effort to integrate these ideas into a coherent, self-contained framework.

The main conclusions of our work are as follows.

First: *conformism* (the tendency to imitate the most frequent behaviour), *altruistic punishment* (the tendency to punish free riders at a cost to oneself) and *gratuitous cooperation* coevolve. Group selection, the natural selection of competing groups, is a necessary ingredient in this evolutionary soup.

Second: gratuitous cooperation is not the result of a cognitive impairment. That is, people don't to cooperate by mistake. The cognitive-impairment assumption, which pervades the literature on gratuitous cooperation, is not only at odds with modern theories of the evolution of intelligence: it is an unnecessary assumption.

Third: conformism in cooperative dilemmas should be understood as a learned preference, instilled by parents in their children. In simple words, people are smart; if they conform, they do because they wish to follow the crowd.

Fourth: within groups, cooperation is a cycle of sudden rise, long lasting prosperity, and abrupt decadence. Occasional "group mutations" are needed to stabilise cooperation in the system as a whole; for example, the emergence of a charismatic leader who brings order to a group and leads it to war against other groups. Our simulations show that even if group mutations occur very infrequently, every 500 years or so, a culture of conformism, altruistic punishment, and gratuitous cooperation can persist in the long-run.

This chapter is organised as follows.

In section 2 we give a real life, large-scale example of gratuitous cooperation: the case of the Chernobyl liquidators. With this example we intend to persuade the reader that gratuitous cooperation is indeed gratuitous and deliberate; not, as some have argued, the result of confusion. It is also our intention to make the case for the empirical and social relevance of gratuitous cooperation; beyond the behavioural sciences' laboratory, and into the outside world.

In sections 3 to 7 we discuss previous attempts to explain gratuitous cooperation.

Finally, in sections 8 and 9 we describe our own work.

2. The Chernobyl dilemma

In 1986, six hundred thousand liquidators—firemen, soldiers, workers, medics, and many volunteers— entered the exploded Chernobyl Nuclear Plant. Exposing themselves to deadly radiation, they extinguished the fire, cleaned the area of radioactive debris, and built the concrete sarcophagus that now seals the reactor. Of the 40 firemen who were the first to deal with the disaster, the majority died within the first three months. The remaining ones are all dead today. About 60 thousand of the liquidators have died since 1986. Amongst the survivors, some 165 thousand are now chronically ill or disabled.

	Everyone else stays and helps	Everyone else flees
He stays and helps	B – C	–C
He flees the site	B	0

Table 1: The Chernobyl dilemma. The table displays the liquidator's material payoff in each scenario. If everyone else stays and helps with the exploded plant, the family of the liquidator is saved: he obtains a benefit of $B > 0$. The cost for a liquidator of staying and helping is $C > 0$.

The dilemma faced by a liquidator is shown in Table 1. If everyone else stays and helps with the exploded plant, the liquidator's family will be saved: better for him to run with his family and enjoy life than to expose himself to the radiation ($B > B - C$). And if everyone else flees the site of the disaster, the liquidator's family is doomed no matter what; there is nothing the liquidator can do about it. Run, and he and his family will have a better chance of surviving ($0 > -C$). Thus, a selfish liquidator will conclude it is best for him to flee regardless of what his fellow liquidators do. And if all liquidators are selfish, then all liquidators will flee.

And yet they stayed. And by staying, they averted a catastrophe of undreamed consequences.

The heroism of the liquidators is hard to explain from a biological point of view. Why should anyone kill himself (i.e., remove himself from the gene pool) for the sake of mankind? A gene that predisposed us to such an act should have gone extinct millennia ago. Kin selection, the evolution of selfish genes that move you to sacrifice yourself for blood relatives, is clearly not the answer to the conundrum.[2] The liquidators would have done their families a much better service by fleeing with them. Reciprocal altruism,[3] the exchange of favours in the course of time, neither is the

[2] Kin selection and the related principle of inclusive fitness were proposed by Hamilton (1964).

[3] Not really altruism, but long-term self interest. The notion is due to Trivers (1971).

answer. When you are due to die within three months, there is not much time to get repaid for any favour. The explanation must lie elsewhere.

The Chernobyl dilemma is an example of what is known as a *cooperative dilemma*, a situation in which a person must decide whether or not to subordinate his own interests to those of the group. And the behaviour of the liquidators represents an extreme example of gratuitous cooperation: the sacrifice of oneself for others to whom one is not genetically related, and who will have little or no chance of returning the favour.

Three themes recur in the literature that deals with the evolution of gratuitous cooperation: group selection, altruistic punishment, and conformism. We discuss each of these themes in what follows.

3. Group selection favours cooperative bands

The hypothesis of *group selection* can be traced back to Charles Darwin himself. In The Descent of Man, he speculates about the development of moral faculties in humans:

> "It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an increase in the number of well-endowed men and advancement in the standard of morality will certainly give an immense advantage to one tribe over another. A tribe including many members who from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to aid one another, and to sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection." (Darwin 1871, p. 166)

According to Darwin, tribes compete amongst themselves for survival, and those formed by cooperative people drive to extinction those formed by egoists.

Modern humans appeared on earth some 200 thousand years ago. And during the first 190 thousand years they lived on the basis of hunting and gathering. Every morning, the men separated into small parties and went hunting. On return to the camp, those men who had succeeded in catching an animal shared the meat amongst all the members of the band. And, since success in the hunt depended in large measure on luck, sharing was a form of insurance that guaranteed the availability of food (almost) every day (Gurven 2004).

The behaviour of our ancestors can only be described as gratuitous cooperation. Hunting is costly in reproductive terms: it consumes thousands of calories which are precious when one lives on the edge of starvation. A selfish hunter may plausibly blame bad luck if he returns to the campsite with empty hands. Fighting is also costly since there is a risk of death or serious injury. Nevertheless, our ancestors did their share of hunting and risked their lives fighting for the band. The theory of group selection identifies the intense competition between bands as the explanation of our ancestors' behaviour.

This account makes good sense except for one thing. It is not clear which must predominate: individual selection in favour of free riders (that is, those who do not cooperate with the band), or the selection of groups in favour of cooperative bands. The direction in which this tension is resolved depends, amongst other things, on the frequency of conflicts and migration rates between groups. It has been argued that migration rates among early humans were too high and conflict rates too low for group selection on its own to be effective (Williams 1966). Assuming this to be true, we are forced to discard group selection as a sufficient explanation for the existence of gratuitous cooperation.

4. Altruistic punishers may enforce cooperation, but only for a brief period of time

A second explanation for gratuitous cooperation is based on what is called *altruistic punishment* (Fehr and Gächter 2002). In most human societies, to be a free rider is not really free. The free rider may be exposed to social punishment: from disproval to ostracism to physical aggression. If the proportion of punishers is sufficiently high, free riding may end up as more costly than cooperating. And if that is the case, free riders will eventually disappear from the population or will become a very small fraction of it.

Punishing free riders is also costly: the main cost being the risk that free riders will defend themselves or retaliate if punished. For this reason, punishment of free riders is called altruistic: punishers sacrifice themselves to protect society from the spread of free riding. If there are only a few free riders, the cost of being a punisher is small since it is rarely necessary to enforce cooperation. Even so, this small enforcement cost puts punishers at a reproductive disadvantage compared to people who cooperate without punishing. As a result, the proportion of pure co-operators will gradually increase at the expense of punishers.

As the number of punishers diminishes, the intensity of punishment declines and to be a free rider becomes steadily less costly. There will arrive a moment when the number of punishers is so

low that the cost of punishment is less than the benefit from not cooperating. At this point, free riders will obtain higher material payoffs than punishers and pure co-operators. The free riders' share in the group will therefore increase and they will eventually crowd out the other types.

Figure 1 illustrates the above sequence.

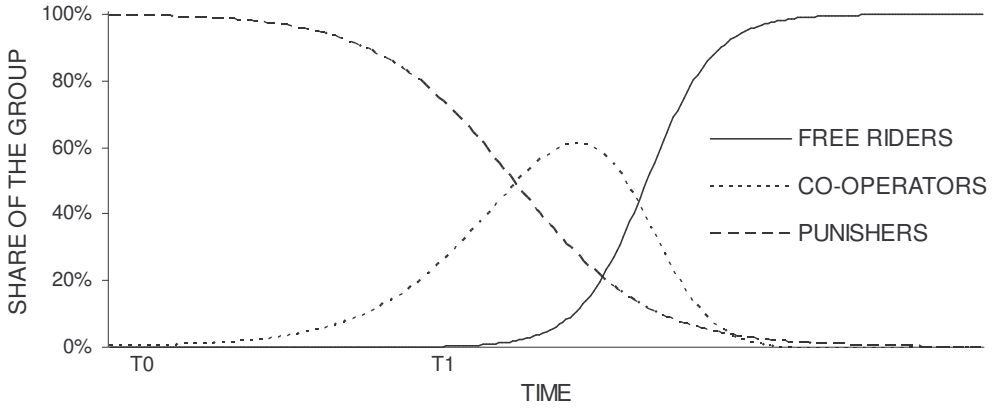

Figure 1: Free riders take over the group. The first pure co-operators appear in time T0. Starting from time T1, the number of punishers is so low that the cost of punishment is less than the benefit from not cooperating. From then on, free riders increase their share of the group until they eventually crowd out other types.

5. Altruistic punishment plus group selection can explain gratuitous cooperation in small groups

Robert Boyd, Peter Richerson, Samuel Bowles, and Herbert Gintis (2003) developed a model in which altruistic punishment is combined with group selection. In their model, bands of hunter-gatherers compete against each other, while within each group there is a competition between three types of individuals: free riders, who never cooperate in hunting; pure co-operators, who cooperate in hunting but do not punish free riders; and punishers, who cooperate and punish free riders. Inter-group competition penalises groups with a relatively high proportion of free riders. It also confers an advantage on groups with a lot of punishers, because such groups will tend to have relatively few free riders. By combining group selection and punishment, this model is able to explain why cooperative behaviour persists in groups as large as a band or as tribe.

The first problem with the model of Boyd et al. is that it cannot explain cooperation in large-scale societies, from the size of a village up to the size of a country. The fault is due to migration and to the Law of Large Numbers. Migration is like the flow in a communicating vase which

eliminates differences between groups. And the more homogeneous the groups are, the weaker will be the effect of group selection: if two almost identical groups enter into conflict, group selection will have, in practice, nothing to select.

In the extreme, two groups completely united by migration are like a single large group, and a world without frontiers is a world without group selection. Nevertheless, in small groups innovation can counteract the homogenising effect of migration. One single innovation can generate by itself enormous changes that break the homogeneity. For example, in a group formed by four free riders, if one of them is transformed into a co-operator, then cooperation leaps from 0% to 25%. In a world consisting of small groups, innovation will be constantly generating diversity, thereby ensuring that group selection continues to operate despite the homogenising effect of migration. In more numerous groups, in contrast, the Law of Large Numbers eliminates differences in group structure produced by random innovation. For example, suppose that the probability of a particular individual being transformed into a punisher is 1%. The Law of Large Numbers predicts that in a group composed of 10,000 free riders, roughly 100 of them, give or take a few, will transform themselves into punishers. Thus, in a world consisting of large groups, random innovation will generate very little diversity between groups and will not be sufficient to offset the homogenizing influence of migration. In the course of time, migration will cause these groups to become very similar to each other. When two of these groups enter into conflict with each other they will virtually indistinguishable and the force of group selection will therefore be negligible.

6. Conformists prevent the spread of free riding, but where does conformism come from?

The spread of free riding is a process of cultural evolution. A good part of the repertoire of human behaviour is transmitted through ideas called *memes* (Dawkins 1976). In a cooperative dilemma three distinct memes are competing: do not cooperate, cooperate without punishing, and cooperate plus punish free riders.

Let us suppose that originally everyone is a punisher. One fine day, a few people decide to innovate (innovation is the memetic equivalent to genetic mutation). Some innovators transform themselves into pure co-operators, others into free riders. Since punishers dominate the society, the punishment of free riders is ferocious. They rapidly realise that it is convenient for them to abandon non-cooperation and imitate a more successful behaviour, which could be to cooperate only, or to cooperate and punish free riders (imitation is the memetic equivalent to reproduction).

Explanations based exclusively on group selection and/or altruistic punishment suppose that people tend to maximise their material payoffs. Such a supposition is not at all realistic. On the contrary, much of the time people limit themselves to repeating what the majority does. This form of learning is called *conformism*.

In a cooperative dilemma, conformism translates into the strategy "cooperate if the majority cooperate, do not cooperate if the majority do not cooperate." If co-operators become the majority then everyone cooperates. This cooperative equilibrium is extremely stable through time. A small number of innovators will never induce the conformists to switch strategies.

Moreover, the larger the group of conformists, the more stable the cooperative equilibrium will be. Conformists will insist on cooperating unless more than half of the group innovates and free rides, or free rides by accident. If the number of conformists is large, the Law of Large Numbers makes it exceedingly unlikely that a substantial fraction of these conformists will simultaneously and independently innovate or accidentally free ride. An increase on the group size, which is bad for cooperation when people imitate profitable behaviour, helps to stabilise cooperation when people conform to the most common behaviour.

This story, just like the others, is unsatisfactory. It eliminates the question, "where does cooperation come from?" only to raise another: "where does conformism come from?"

7. The hitchhiking theory of conformism

It has been argued that conformism evolved in the context of decision problems that had nothing to do with cooperation. From there, conformism hitchhiked into cooperative dilemmas. This explanation for cooperative behaviour was first put forward by Herbert Simon (1990) and explored in detail by Gintis (2003) [see also Henrich 2004]. For example, instead of designing his own bow and arrow, a hunter may choose to imitate the prevalent design. This behaviour enhances the hunter's fitness because it reduces his expenses on "research and development." That same hunter may later cooperate in the hunt, just because all others do. This behaviour reduces the hunter's fitness. The supporters of the hitchhiking hypothesis argue that all that is needed for conformism to evolve is that it enhances individual fitness on average.

Conformism may be fitness enhancing because the conformist economises search costs by taking advantage of the cumulative knowledge of society which is embodied in the prevailing social practices (Feldman et al. 1996; Henrich and Boyd 1998). Three conditions are needed for

conformism to be effective. First, optimising must be expensive, demanding a large amount of time and resources. Second, at least part of the population must go it alone in the search for optimal behaviour. Otherwise, society will generate no new knowledge and conformists will have no ideas to copy (e.g., somebody has to invent the bow and the arrow). Third, the environment must be relatively stable. If not, the knowledge previously acquired by the group quickly becomes obsolete, in which case going it alone may be the most reasonable thing to do.

In his canonical exposition of the hitchhiking hypothesis, Gintis (2003) assumes the existence of an "internalization gene" which predisposes people to conform to social norms. The probability that such an individual will conform to any particular norm depends only on his level of exposure to this norm during some formative period. For example, if both parents conform to a certain norm, then a child with the internalisation gene will also conform to this norm. The probability of conformity is 50% if only one parent conforms to this norm. A child who carries the internalisation gene will become a hunter in later life if he is brought up amongst hunters. He will become an altruist if he is brought up amongst altruists. He will become an altruistic hunter if he is brought up amongst altruistic hunters and so on. The fitness of these diverse individuals will vary; but, for the internalisation gene to survive, the individuals that carry this gene must be (on average) at least as fit as the individuals who do not carry it. The loss in fitness that each of them suffers from behaving altruistically must be counterbalanced by the gain they get from following fitness-enhancing norms. If this is the case, the internalisation gene will survive and, as Gintis shows, for certain parameter values there will be a stable equilibrium in which altruistic behaviour is widespread.

The above argument does not suffice to explain the emergence of conformity to costly altruistic norms. Something more is required: either cooperative dilemmas are sufficiently complex to justify the use of conformism over individual learning, or humans cannot easily distinguish between situations in which conformism is fitness enhancing from situations in which it is not. Both things are implausible. On the contrary: The cooperative dilemma is often simple, people are aware of its simplicity, and straightforward computations lead them to the conclusion that cooperation is a form of self-sacrifice. Nonetheless, they choose to conform to the altruistic norm. This forces us to reject the hitchhiking theory of conformism, and to conclude conformism in cooperative dilemmas is not a learning rule, but an expression of preference. We elaborate these ideas in what follows.

First, experiments in which people play some stylised version of the cooperative dilemma, such as the prisoner dilemma or the public good game, suggest people do not cooperate by mistake. Experimental subjects understand very well the logic of this dilemma; they know that free riding is

the optimal behaviour from a selfish perspective. Nevertheless, most of the subjects choose to conform to the majority: they cooperate if others cooperate, and they don't cooperate if others don't. This pattern of behaviour has been termed conditional cooperation (Fischbacher et al. 2004).

Arguably, the cooperative dilemma is just as easy to understand in real life as it is in the lab. Once again, consider the case of the Chernobyl Dilemma. Sarah Wallace (2007) interviewed many men who volunteered to be liquidators. Judging from their answers, their behaviour can hardly be attributed to misunderstanding. Here are two examples.

One of the liquidators interviewed by Wallace was a physics professor in Zhytomyr Oblast, who helped with the evacuation close to the nuclear plant. Wallace reports: "He would not change his decision to be a liquidator. He thinks that people need to help each other. He does not consider what he did to be very special." Ever since the disaster, the professor has suffered cardiovascular and neurological diseases. He attributes his condition to being exposed for so long to the radiation.

Another liquidator was a construction worker who in 1986 resided in Kyiv. Says Wallace: "He and his colleagues were asked to go to Chernobyl soon after the disaster. Their job was to seal the floor of the reactor with metal insulation and concrete so that irradiated matter would not seep into the groundwater. He agreed to go to Chernobyl out of patriotism. In his own words, if he did not do it, who would?" This man has had health problems since 1986, most likely due to the dose of 28.2 roentgens that he received while working at Chernobyl. During the interview, he declared that when he volunteered he was fully aware of the risks and that, if he could go back to 1986, he would volunteer again.

The possibility remains that people understand the cooperative dilemma, but still decide to conform to the altruistic norm because they do not trust their own judgment. That is, they may fail to distinguish situations in which conformism is fitness enhancing from situations in which it is not. Of course, this makes one wonder why the ability to discriminate between the two kinds of situation did not evolve. Gintis himself acknowledges this potential flaw in the hitchhiking theory (Gintis 2003, p. 416). Suppose there is a genetic mutation (for instance at another genetic locus) which allows the individual to distinguish between altruistic and fitness-enhancing behaviour and to eschew the former. Such a mutation will drive out the gene for blind conformism. If mutant individuals can discriminate perfectly between the two kinds of behaviour then altruism will eventually disappear. Individuals will follow the social norm only when it is in their self-interest and will otherwise ignore it. In this case, the hitchhiking explanation for altruism will break down

completely. If discrimination is partial, limited types of altruism may survive because individuals cannot distinguish them from fitness enhancing behaviour.

The extent to which humans are able to identify situations in which conformity is likely to reduce their fitness is an empirical matter. Gintis claims that humans have a very weak capacity to discriminate fitness reducing norms; and even if humans could discriminate, the degradation of the capacity to discriminate is much more likely than its evolution.

To evaluate this claim, we must first note that humans rarely have a direct concern with the fitness consequences of their actions. Indeed, they may be unaware of the biological concept of fitness. Instead, like other animals, they base their decisions on proxies, such as material rewards, amount of effort, risk of injury or death, etc. (in prehistoric times, when human nature evolved, such proxies were highly correlated with biological fitness). Humans often have quite an accurate knowledge of the likely costs and benefits of their actions, and when they weight up costs and benefits they are engaging in implicit fitness calculus. Once we recognise this, it becomes implausible to claim that conformism to cooperative norms is mainly the result of cognitive impairment. People understand that cooperation is a form of self-sacrifice (the mere existence of the term "self-sacrifice" reveals how well we understand the costly nature of cooperation). Even so we choose to conform to altruistic norms. Cognitive impairment cannot account for this. The only way to make sense of the facts is to conclude that conformism in cooperative dilemmas is an expression of preference.

A final reason to doubt of the hitchhiking hypothesis is that it conflicts with modern theories of the evolution of intelligence. According to these theories, intelligence evolved precisely to deal with higher levels of social complexity (Dunbar 1993). And since the big brain is already there to assist us in social interaction, why should we turn it off while playing cooperative dilemmas? We are, indeed, surprisingly proficient at tasks having to do with social interaction; for example, detecting norm violators, or keeping account of indebted favours (Cosmides and Tooby 1992). Conformism as preference is entirely consistent with a high level of social intelligence.

8. Conformism, cooperation, and altruistic punishment coevolve; the mixture can explain gratuitous cooperation in large groups

In order to explain the evolution of conformism in cooperative dilemmas, we developed a model in which learning rules evolve alongside the evolution of behaviour (Guzmán et al. 2007).

Building on Boyd et al. (2003), we incorporated in our model two genetically-based learning rules that compete against each other: conformism to the most common behaviour, and payoff-dependent imitation (the imitation of more successful behaviours).[4] In our model, each person follows one and only one learning rule throughout his life, and uses that learning rule to choose amongst three behaviours: free ride, cooperate without punishing, and cooperate plus punish free riders. Payoff-dependent imitators have a reproductive advantage over conformists, since payoff-dependent imitators search for behaviours that maximise fitness, whereas conformist couldn't care less about fitness maximisation.

In our model, parents pass their genetically-based learning rules down to their children. But the inheritance of learning rules is not perfect: with a very small probability, the child of a conformist will grow up to be a payoff-dependent imitator, and the child of a payoff-dependent imitator will grow up to be a conformist. In order to bias the model against conformism, we did not impose a higher metabolic cost on payoff-dependent imitation.

Following Boyd et al., we simulated our model for conditions that approximate those in which early humans lived, and found that the mixture formed by altruistic punishment, conformism, and group selection was capable of sustaining gratuitous cooperation in groups of thousands. This result obtains even when there is a high rate of migration between groups, or when conflict amongst groups is infrequent. In our model, altruistic punishment, conformism and cooperation coevolve: none of them can develop fully without the presence of the others. This is illustrated in Figure 2, which shows the simulation results.

[4] When we wrote this paper, we hadn't yet adopted the view of conformism as a preference.

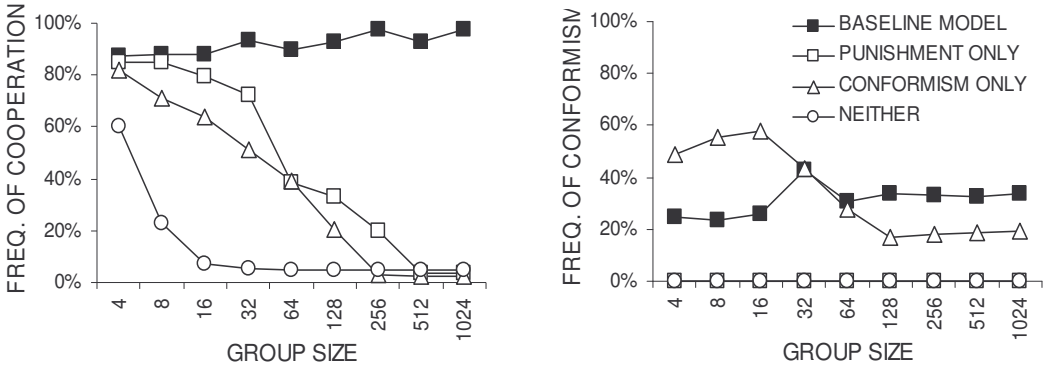

Figure 2: Cooperation in alternative models. Source: Guzmán et al. (2007).

What mechanisms underlie these results? First, group selection favours cooperative groups. Second, conformism combined with altruistic punishment immunises groups with a high proportion of punishers against the invasion of immigrant free riders. If the immigrant free rider is a conformist, he will automatically switch to being a punisher since this is the most common form of behaviour in his new group; if he is an imitator, he will imitate the most successful person that he comes across, which in a group dominated by punishers is likely to be a punisher. Moreover, if the local punishers are conformists, they will resist the temptation to stop punishing and becoming pure co-operators. Thus, in a group where a culture of punishment already exists, conformism discourages the invasion of free riders, either from without or from within, that would eventually wipe out cooperation.

9. Why do humans imitate?

Imitation is a form of social learning because it involves learning from others. The alternative to social learning is individual learning, which involves a rational assessment of available options and requires a higher level of intelligence. It is sometimes argued that social learning evolved because it economises on valuable brain power. Intelligence is related to brain size (Williams 2002), and the brain is a very costly organ to develop and maintain. Humans remain dependent on their parents longer than any other land animal because their brains are so large compared to their body mass. Having to support their offspring for so long imposes a huge fitness cost on human parents. The brain is also expensive in metabolic terms; it is by far the biggest energy guzzler in the human organism (Aiello and Wheeler 1995). Thus, brain evolution involves a trade-off between the benefits of better information processing, the high metabolic cost of brain tissue, and the inordinate

parental investment in big brained children (Niven 2005). Imitative learning may have a selective advantage over individual learning because it economises on brain power and hence on parental time and energy use.

The above is a plausible explanation for the evolution of imitative learning, but our previous work on conformism suggests an alternative explanation; one that does not depend on the economics of brain power, but relies instead on the force of group selection. To explore this possibility, we extended our original model to include, alongside conformism and pay-off dependent imitation, an additional learning rule: "best-response" (Rodriguez-Sickert et al., 2008). Best-responders are both selfish and hyper-rational. They always choose the most profitable behaviour. We assume that the extra intelligence required to achieve this result incurs no fitness cost for the best-responder. Thus, the issue of brain size plays no role in this model. Each individual can choose between one of three strategies: free ride, cooperate and punish free riders, and cooperate without punishing. These strategies spread through the group under the influence of the prevailing learning rules. As in the previous model, between-group and within-group selective pressures drive evolution.

We simulated this extended version of our model using the same parameters as in Guzmán et al. (2007). A hybrid pattern emerged: on average, there was a high share of conformists and a very low share of best-responders. Why? Best-responders speed up the invasion of free riding: they never punish free riders, and they automatically become free riders when punishers are few enough for free riding to be optimal. Thus, the presence of best-responders undermines the ability of the group to compete with rival groups, so group selection tends to eliminate groups of best-responders. Conformism and pay-off dependent imitation, on the other hand, thrive because they promote the individual self-sacrifice that is the key to success in inter-group competition. This result turned out to be quite robust to changes in the assumed rates of inter-group conflict and migration.

10. Work in Progress: charismatic leaders and the evolution of conformism as a preference

The group selection models described above share an important feature with the hitchhiking theory of conformism. Although the mechanism is different, they also rely on the fact that certain individuals (the payoff-dependent imitators) have a cognitive impairment which on occasion leads them to act unwittingly against their own self-interest. Such individuals are altruistic "by mistake." In our critique of the hitchhiking theory we argued that this assumption is unrealistic. Conformity to costly social norms is not in general a by-product of cognitive impairment but a product of

deliberate choice. In the context of a social dilemma, conformity to social norms should therefore be modelled as a preference and not as a learning rule.

Our current work explores the implications of this approach within the framework of a simple group selection model. In the new model, we ignore the influence of cognitive ability and focus exclusively on preferences.

There are two types of individual in the new model: *conformists* and *egoists.* These types are identical in terms of cognitive ability but have different preferences. The objective of the conformist is to follow the prevailing social norm as an end in itself, without regard to material pay-off. The objective of the egoist is to maximise his material pay-off. Both types are assumed to have complete and accurate knowledge about the current behaviour of other members of the group. This knowledge is acquired without cost. On the basis of this knowledge, they choose their behaviour for next period in accordance with their own preferences. The conformist chooses the most common behaviour that he observes; the egoist chooses the behaviour that would yield the highest pay-off given the current behaviour of other members of the group.[5] For simplicity, we assume that individuals are either pure conformists or pure egoists; although in reality human beings normally have mixed motives and do not belong to either extreme type.

Apart from an initial random error, children inherit their preferences from their parents. In this respect, the new model is similar to the other group selection models described above. There is, however, an important difference. In the new model the transmission mechanism from parents to children is cultural and not genetic. Individuals are genetically identical and their preferences are instilled in them by their parents during childhood. There are two reasons for assuming cultural rather than genetic transmission. In the first place, the error rate is much higher for cultural transmission than for genetic transmission, so the contagion of egoism is very fast. Therefore, cooperation will be more fragile in this new model than in the previous ones. In the second place, culturally transmitted preferences may be altered by events in later life. This opens the way to a group level shock, such as the appearance of a charismatic leader, which simultaneously transforms the preferences and behaviour of a large number of individuals within the same group.

---

[5] Egoists in the new model are identical in terms of knowledge and preferences to what were called "best-responders" in Rodriguez-Sickert et al. (2008). The reason for the change in terminology is to make clear the centrality of preferences. The term "best-responder" implies that maximising individual pay-off is "better" than conforming to the social norm. In reality, both conformists and egoists choose the best response available to them in light of their own preferences.

To explore the properties of the above model we have done some exploratory simulations. We assume that there is initially one cooperative group in which conformist-punishers are in the majority, and 19 groups of egoist-free riders. In any conflict with a group of another type, the group of conformist-punishers is likely to win and replace the other type of group by a clone of itself. In the course of time, through successive conquest, the number of conformist-punisher groups will increase until eventually all groups are of this type.

When all groups are equally cooperative, group selection becomes weaker and internal fitness dynamics take over (recall that group selection requires inter-group heterogeneity). Within each group, random error will produce some egoists who initially co-operate because they are in a small minority and wish to avoid punishment. Such individuals have a slight fitness advantage over co-operator punishers because they do not incur the cost of punishing accidental free riders. Within each group the proportion of egoists will gradually increase until there are so few punishers that free riding becomes profitable. The egoists will switch to free riding and social cooperation will fall sharply. If conformists are still in the majority they will continue punishing for a time, but they will gradually be replaced by egoist free riders. Eventually, egoist free riders will become a majority, which means that free riding will be the new norm of the group. At this point, conformists will switch their behaviour from punishment to free riding causing social cooperation to break down completely. Everyone in the group will free ride. This process occurs at a similar pace in all groups with the result that social cooperation breaks down across the entire system more or less at the same time.

At this point the system is stuck in a non-cooperative equilibrium. The probability of escaping from such equilibrium through independent random variation is zero for all practical purposes: aeons would have to pass before a group turned cooperative. The alternative escape route is through a "group mutation" which converts the majority of some group into conformist-punishers. Such a mutation could be the emergence of a charismatic leader who brings order to the group and mobilises his people to conquer other groups. If this happens, the system will go through the sequence described above. Conformism will spread through conquest until within-group cooperation becomes universal. This new equilibrium will slowly decay, but before cooperation breaks down completely, a new charismatic leader will emerge in some group. Since the ebb and flow of cooperation takes on average 150 years, cooperation will persist in the long-run even if the emergence of a charismatic leader is a very unlikely event. It is enough for each group to produce one Genghis Kahn every 500 years or so to prevent the spread of free riding and keep the wheels of cooperation turning.

References

Aiello L.C. and P. Wheeler. 1995. The expensive tissue hypothesis: The brain and the digestive system in human and primate evolution. *Current Anthropology* 36 (2): 199–221.

Boyd, R., H. Gintis, S. Bowles, P.J. Richerson. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America* 100 (6): 3531–3535.

Darwin, C. D. 1871 [1981]. *The Descent of Man*. Princeton: Princeton University Press

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Dunbar, R. 1993. Coevolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences*, 16 (4): 681–735.

Hamilton, W.D. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7 (1): 1–52.

Fehr, E. and S. Gächter. 2002. Altruistic punishment in humans. *Nature* 415 (6868): 137–140.

Feldman, M., K. Aoki and J. Kumm. 1996. Individual versus social learning: Evolutionary analysis in a fluctuating environment. *Anthropological Science* 104 (3): 209–231.

Fischbacher, U., S. Gächter and E. Fehr. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters* 71 (3): 397–404.

Gintis, H. 2003. The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. *Journal of Theoretical Biology* 220 (4): 407–418.

Gurven, M. 2004. To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences* 27 (4): 543–583.

Guzmán, R.A, C. Rodriguez-Sickert and R.E. Rowthorn. 2007. When in Rome, do as the Romans: The coevolution of altruistic punishment, conformism and cooperation. *Evolution and Human Behavior* 28 (2): 112–117.

Henrich, J. 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization* 53 (1): 3–35.

Henrich, J. and R. Boyd. 1998. The evolution of conformist transmission and between-group differences. *Evolution and Human Behavior* 19 (4): 215–242.

Niven, J. 2005. Brain evolution: Getting better all the time? *Current Biology* 15 (16): R624–R626.

Rodriguez-Sickert, C., R.E. Rowthorn and R.A. Guzman. 2008. The social benefit of slow learners. University of Cambridge, mimeo.

Simon, H. 1990. A mechanism for social selection and successful altruism. *Science* 250 (4988), 1665–1668.

Trivers, R.L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46 (1): 35–57.

Wallace, S. 2007. "Notes from Україна: A Blue Devil's Ukrainian experience," http://chernobyl-summer.blogspot.com/2007/07/chernobyl-liquidators-incredible-men.html.

Williams, G.C. 1966. *Adaptation and Natural Selection*. Princeton: Princeton University Press.

Williams, M. 2002. Primate encephalization and intelligence. *Medical Hypotheses* 58 (4): 284–290.