euresis journal

euresis
Association for the Promotion of the Scientific Endeavour

Volume 4
Winter 2013

journal homepage
www.euresisjournal.org

Was Voltaire right? Social dilemmas and the evolution of the social contract

Karl Sigmund

Faculty of Mathematics, University of Vienna
Nordbergstrasse 15, A-1090, Vienna, Austria
Karl.Sigmund@univie.ac.at

Abstract

Free riders can exploit and therefore subvert joint enterprises. Empirical and theoretical research on mutual aid games indicates that the threat of punishment can curb free-riding. Since punishment is often costly, however, this raises an issue of second-order free-riding: indeed, the sanctioning system itself is a public good which can be exploited. Most investigations, so far, considered peer punishment: players could punish those who exploited them, at a cost to themselves. Only a minority considered so-called pool punishment. In this scenario, players contribute to a punishment pool before engaging in the mutual aid game, and without knowing who the free-riders are. This is a first step towards an institution forcing the members of the community to cooperate. Theoretical and experimental investigations show that peer punishment is more efficient, but pool punishment more stable. Social learning leads to pool punishment if sanctions are also imposed on second-order free-riders, but to peer punishment if they are not. Both types of coercion emerge only if the interaction is voluntary, rather than compulsory. This sheds light on Rousseau's opening sentence of his Social Contract: 'Man is born free, and everywhere men are in chains'.

1. Selfishness and solidarity

Voltaire wrote in his *Lettres Philosophiques*:

> *Assuredly, God could have created human beings uniquely interested in the welfare of others. In that case, traders would have been to India by charity, and the mason would saw stones to please his neighbor. But God designed things otherwise. It is through our mutual needs that we are useful to the human species: this is the grounding of every trade; it is the eternal link between men.*

Voltaire's opinion is shared by many thinkers. According to their view, mutual help and cooperation, which are so basic to human communities, and so ingrained in human nature, are grounded in self-interest. As Voltaire stated flatly: "It is impossible that a society can emerge and subsist without self-love..."

Voltaire was unaware of theoretical biology, but he was thinking in evolutionary terms, as the words "emergence" and "subsistence" testify. His term "self-love" (*amour propre*) seems close to the proverbial "selfish gene" of Darwinians. Indeed, the evolution of cooperation is one of the central topics in theoretical biology. Can the evolutionary approach help to our understanding of the social contract, of morality, virtues and institutions? To this vast question, evolutionary game theory has recently added intriguing elements [1, 2]. The first task that has to be addressed is to explain why the evolution of cooperation should raise any problem at all. Cooperation is so obviously a good thing that it seems hard to understand why two individuals, who both profit from cooperating, should ever fail to do so. The answer to this question is that frequently, each individual can profit even more by exploiting the other. If each yields to the temptation and defects, then both will forego the bonus of cooperation.

To be more specific, let us consider the so-called donation game, which can easily be implemented in an experiment. The experimenter asks two players to independently decide whether to send a gift to the other player or not. If player 'Alice' decides to send a gift to player 'Bob', then 'Alice' has to pay 5 euros to the experimenter, and the experimenter will give 15 euros to 'Bob'. The players cannot communicate with each other. They are in separate rooms, have never met each other and know that they will never see their co-player. After the experiment, they will go their separate ways, in complete anonymity.

Let us assume, more generally, that the donation, or "gift", confers a benefit b to the recipient, at a cost c to the donor, with $0 < c < b$. If both players cooperate, each receives a payoff $b - c > 0$, whereas both receive nothing if they both defect. This can be described by the following payoff matrix, which shows the payoff for 'Alice', i.e., player A (who can choose between the rows C and D, whereas the co-player can choose between the columns C and D:

	If B plays C	If B plays D
If A plays C	$b - c$	$-c$
If A plays D	b	0

Clearly, the socially desirable outcome is that of mutual cooperation. However, this outcome will be difficult to reach for players bent on maximizing their income. Indeed, it is inconsistent, in the sense that if the other player duly plays C, then one can improve the own payoff by playing D. Should the other player play D, then it is better to also play D. In each case, from the viewpoint of an individual player, it is always better to defect. Therefore, the only consistent outcome is for both players to play D: consistent in the sense that both players cannot improve their payoff by unilaterally switching strategy. *Unilaterally* is the essential qualifier: if the two player's decisions were correlated, for instance, if a rule stated that both had to choose the same move, then cooperation would be the obvious outcome. But the two players are individuals, i.e., independent actors with different preferences. Nothing correlates

their decisions. The experimental situation creates a conflict of interest.

This game is an example of a social dilemma: self-interested motives lead to self-defeating moves. This need not always be the case. In fact, such social dilemma situations are often neglected. The contemporary of Voltaire, Adam Smith, famously suggested that an "invisible hand" harmonizes the selfish actions of individuals. By striving to maximize their own revenue, they maximize the total good. To quote Adam Smith: "By pursuing his own interest, he frequently promotes that of the society more effectually than when he really intends to promote it." This view is widespread. In fact, Alexis de Tocqueville wrote (in the nineteenth century!) that "Americans show with complacency how an enlightened regard for themselves constantly prompts them to assist each other." But social dilemmas show that there exist situations where the invisible hand is not only invisible: it is gone.

The branch of mathematics dealing with conflicts of interest is *game theory*. In classical game theory, it is assumed that players are rational, and able to predict and optimize their payoff. But we can apply evolutionary game theory to the behavior of considerably less sophisticated agents. It is enough to assume that successful behavior is copied more easily. In a population of C-players and D-players randomly encountering each other, and interacting in a donation game, the D-players would always do better. According to evolutionary game theory, their behavior would therefore spread. The ultimate outcome would be the total elimination of cooperators. Thus the emergence of cooperation is not only a challenge for a population of rational agents, able to anticipate all possible moves of their opponents and to find the best possible answer. It also poses a major problem for evolutionary biology.

2. Kin selection

There are basically two classical approaches to overcome the problems associated with the evolution of cooperation, both known and studied since almost half a century. The first approach is named *Kin Selection*. It is essentially due to Willam D Hamilton, although precursors such as JBS Haldane and RA Fisher anticipated him (and Charles Darwin clearly was on the right track) [3, 4]. The basic observation is that most biological cooperation occurs within families, for instance in raising offspring. This is simply a corollary of the Darwinian "struggle for survival." Genes which help to promote their own spreading, by enhancing the survival and fecundity of the organisms carrying them, will clearly become more frequent than genes which do not. Parents programmed to help their children have necessarily an advantage in passing along their genetic program.

In exactly the same way, siblings programmed to help each other will also have a selective advantage. More precisely, a gene causing you to help your sibling will help to spread itself: for it is, with a high probability, carried by the sibling too. This helps in explaining, for instance, the phenomenon of nest helpers which is so frequent among bird species. Such

helpers stay at the parental nest for one or more seasons and help their parents in raising younger generations. Similarly, R.A. Fisher used the approach in explaining the prevalence of warning colors among distasteful caterpillars. The bright colors may seem at first sight a suicidal advertisement policy. But a bird that swallows such a lurid caterpillar will most likely never do it again. If this saves the life of the caterpillar's siblings (as will most probably be the case, since they travel in family groups), then the victim's death has not been a vain sacrifice, from the corresponding gene's point of view. The genes for bright colors will be passed on through the siblings, and can spread. The famous geneticist Haldane is supposed to have quipped, in a similar vein: "I am ready to lie down my life to save two of my brothers, or eight of my cousins." Why two? Why eight? Obviously, there is some theory behind it. It is based on the quantification of relatedness.

The relatedness between two players can be defined in various ways. Here, we simply assume that it measures relatedness by descent, also known as con-sanguinity: this is the probability r that a recently mutated gene, if it is carried by one player, is also carried by the other. Of course, any two humans are related, if we go back far enough. But we do not share all our genes. A mutation occurring in the body of my grandfather produces an allele which will be found with probability 1/2 in his children, and with probability 1/4 in his grandchildren. Under usual circumstances (in particular, no parental inbreeding), the coefficient of relatedness between two siblings is 1/2; between me and my nephew it is 1/4, and between two cousins it is 1/8. Relatives can be viewed in this sense as watered-down copies of oneself. The coefficient r measures the amount of dilution. The higher r is, the more the genetic interests coincide. The payoff matrix of the donation game described above now becomes

	If B plays C	If B plays D
If A plays C	$(b-c)(1+r)$	$br-c$
If A plays D	$b-cr$	0

This is obtained, roughly, by viewing any increase in your co-player's fitness as an increase of your own fitness, but discounted by the factor r. Clearly, if the coefficient of relatedness satisfies $r > c/b$, then the elements in the first row are larger than the corresponding elements of the second row. Hence, in this case, no matter what your co-player does, it is better to choose the first row, i.e., to cooperate. This is known as Hamilton's rule [5].

This "selfish gene" view, elaborated as the theory of *kin selection*, has been developed to a considerable extent, within the last fifty years. A more "modern" version of Hamilton's rule is based on relatedness by assortment, which replaces the coefficient of consanguinity by the difference between the conditional probabilities of encountering (as co-player) a cooperator, i.e., a C-player, depending on whether you are yourself a cooperator or a defector, i.e., a D-player. Again, this leads to Hamilton's rule, but this time for relatedness by assortment, rather than relatedness by descent.

Many of the most remarkable examples of altruistic behavior occur among social insects, such as bees or ants, where the family ties are extremely tight. Typically, all workers in a bee hive or an ant hill are sisters, the daughters of one queen, who herself is often the product of intensive inbreeding. The degree of relatedness, in that case, is so high that one can view an insect state as a "super-organism." (Indeed, in an organism, the body cells all share the same genes, so that the degree of relatedness is 1).

Darwin, who had overlooked (like everyone else) the contemporary work by Mendel, did not have a clear idea of how inheritable traits could be passed on from one generation to the next. The notion of a gene was unknown to him. Nevertheless, he had as good a grasp of the principles of kin selection as was possible in his time. This can be seen in the following quote [6]:

> *One special difficulty [...] at first appeared to me insuperable, and actually fatal to my whole theory. I allude to the sterile females in insect-communities [who] differ widely [...] from both males and fertile females, and yet, from being sterile, cannot propagate their kind... Natural selection may be applied to the family, as well as to the individual.*

Kin selection can lead to impressive feats of altruism and self-sacrifice. For instance, so-called honey pots ants are worker ants spending all their life clinging to the wall of a subterranean chamber, their bodies exclusively used for storing nutrient. And worker bees are ready to sting intruders and thus to perform a suicide attack in order to defend their hive. It seems that such acts of self-immolation can only be explained by indirect fitness benefits, since neither honey pot ants nor kamikaze bees can have any direct descendants able to carry copies of their genes into future generations.

How much is the nature of human beings affected by kin selection? Aristotle, one of the earliest zoologist, classified humans with ants and bees as social animals, and we are certainly achieving comparable feats of solidarity. The role of bees in Napoleons heraldic propaganda is a well-known example recognizing the role of cooperation in human society. An important treatise written by Bernard Mandeville more than three hundred years ago was entitled "The Fable of the Bees" and compared the bustle of a modern city with that in a bee-hive. Some of the parallels between bees and humans are striking indeed: the division of labor, the ceaseless bustle and exchange, the hierarchical organization, etc. Nevertheless, it has become clear that human sociality is very different from that of hymenoptera (bees and ants) or termites. Basically, humans have not given up reproduction in favor of a few highly privileged individuals. Most humans can and do reproduce, in contrast to social insects where the job is delegated to specialized queens and consorts. Hence, the average relatedness in a town is only a tiny fraction of that in a hive.

Obviously, humans have a remarkable tendency to also cooperate with non-relatives. This does not mean to imply that nepotism is irrelevant for human societies. In fact, many problems are caused by exactly this tendency to favor close kin, and a large part of democratic

progress consists precisely in overcoming this. But by and large, close family ties are not required for cooperation.

Economic experiments show that players preferentially trust similar-looking co-players, and indicate that kin selection is at work, able to produce mechanisms which are not even conscious. The players are provided with pictures of their ostensible partners, and these pictures can be manipulated, through digital sorcery, to look like themselves to a greater or lesser extent. Invariably, similarity promotes cooperation [7]. Hence, familiarity enhances trust. Clearly, such cues for self-similarity can be promoted by cultural means. Many groups provide their members with characteristic uniforms, badges, tattoos, ties, hangouts, accents, musical tastes or slang idioms.

Under normal circumstances, many tend to view strangers as "honorary relatives," and cooperate. This can be exploited by a rhetoric propaganda which emphasizes fraternity, which depicts the own group as a band of brothers, and which calls the own country "fatherland" or "patria", and the native idiom "mother tongue". It is needless to add that these tricks can boost cooperation towards sinister ends: Mafiosi gangs term themselves "family."

3. Reciprocal altruism

Nevertheless, the widespread cooperation between non-related individuals cannot be attributed simply to the maladaptation of kin recognition mechanisms. It is obvious that altruistic acts between non-related individuals can often lead to direct benefits. Already Adam Smith, in his *Theory of Moral Sentiments*, emphasized "our propensity to trade, barter and truck" [8]. This leads to the second major approach, by evolutionary biologists, to explain the emergence of cooperation. It started with a seminal paper by Robert Trivers, and is generally designed as "Reciprocal Altruism" [9].

In its simplest embodiment, direct reciprocity can be defined as "the trading of altruistic acts in which benefit is larger than costs, so that over a period of time both parties enjoy a net gain." Darwin [10] anticipated this when he wrote that "the small strength and speed of man, his want of natural resources etc. are more than counterbalanced by his social qualities, which led him to give and receive aid from his fellow man."

In the simplest model, this can be described by an *Iterated Prisoner's Dilemma game*. The same two players meet in round after round. In that case, they are not obliged to choose the same option C or D in every round. They can use conditional strategies, and decide whether to cooperate or defect according to the past behavior of their co-player. The strategy coming most naturally to mind, in this context, is to reciprocate good with good, and bad with bad. Many experiments have shown that a large majority of humans are conditional cooperators, and want to play C if the co-player also choses C. But how can one

be sure that the co-player will play C? The problem is essentially one of trust. In the context of repeated games, players can base their decision on the past behavior of their co-player. The simplest such strategy is *Tit-For-Tat* (TFT): it consists in playing C in the first round, and from then on to use whichever move the co-player used in the previous round [11].

If w is the probability that the same two players will engage in a further round of the game, then the expected number of rounds is given by $1/(1-w)$. If we consider only the two strategies TFT and All-D (which consists in defecting in every round), then the payoff matrix is given by

	If B plays TFT	If B plays $AllD$
If A plays TFT	$(b-c)/(1+w)$	$-c$
If A plays $AllD$	b	0

Indeed, if both players play TFT, they will cooperate, and thus earn the payoff $b-c$ in every round. If they both play AllD, by contrast, they will earn nothing at all. A TFT player will be exploited by an All-D player in the first round, but in subsequent rounds, both will defect. It is easy to see that if $w > c/b$, i.e., if the expected number of rounds is sufficiently large, then it does not pay, against a TFT-player, to play All-D: the advantage gained in the first round cannot make up for the handicap of turning the co-player into a defector. On the other hand, it does not pay, against an All-D player, to use TFT. In the repeated Prisoner's Dilemma game, there is not one strategy which is always the best, independently of the other player's decision. It is best to do whatever the other player does. This means that the evolutionary dynamics is bi-stable. If most players in the population use one of the strategies, then it is best to also adopt it. If w is close to 1, however, then the contest between TFT and All-D is rigged in favor of the former: its basin of attraction is much larger.

Nevertheless, the TFT strategy has some weaknesses. For instance, unconditional cooperators (i.e., All-C players) can enter a population of TFT-players by neutral drift. Indeed, both strategies do exactly as well, everybody cooperates in every round, and therefore, selection does not act in one direction or the other. This means that by sheer chance, or in other words by random drift, a sizable amount of All-C players can build up. But once this happens, then All-D players can invade, since they can exploit the All-C players. Another weakness of TFT-societies is that they are very vulnerable to errors. If two TFT-players interact and one of them defects by mistake, this will cause a long vendetta. If such errors are taken into account, the dynamics of a population consisting of TFT, All-C and All-D players is highly unstable. Either All-D eliminates the other two strategies, or else the frequencies of all three will endlessly undergo un-damped oscillations.

There are other conditional strategies which do not have the defects of TFT. This holds in particular for Win-Stay, Lose-Shift (WSLS). Players using that strategy start with a cooperative move and from then on repeat the former move if it yielded a positive income, but

switch to the other move if not. This can be viewed as the simplest learning mechanism: and it is striking to see it emerge spontaneously [12].

Needless to say, when applying the mechanisms of evolutionary game theory to humans, we do not assume that strategies are inherited from parent to offspring. For the hard-wired behavior of social insects, this assumption is reasonable enough, but for humans, it is absurd. Fortunately, we can use the machinery of evolutionary game theory even if strategies are transmitted, not through inheritance, but through social learning. If humans have a propensity to preferentially imitate more successful strategies, then we can apply the same dynamics to "memes" rather than genes, i.e., to ideas which can be transmitted from brain to brain. The role of "mutations", in this context, is provided by the random adoption of behavioral alternatives. Whether in the context of selection-mutation or of imitation-exploration, i.e., whether in biological or cultural evolution, we are led to the same process of trial-and-error, and hence to the same dynamics.

4. Generalized Reciprocity

So far, we have dealt with direct reciprocity. But the human tendency to cooperate is much more pervasive. Not every help is directed at a recipient able to return that help, and it is clear that such actions are beyond the realm of direct reciprocation. The story of the Good Samaritan is a case in point. At first glance, it seems to be beyond economic considerations. It may well be, however, that an act of help is returned, not by the recipient, but by a third party. The idea of an indirect, or "generalized" reciprocity can be found in Trivers' seminal paper already. In direct reciprocity, if 'Alice' provides help to 'Bob', then 'Bob' is supposed to return help to 'Alice'. In indirect reciprocity, if 'Alice' helps 'Bob', then the help can be returned to 'Alice' by some third party, for instance 'Charlie'. This seems a more subtle form of reciprocation. Direct reciprocity works on the principle that "I'll scratch your back if you scratch mine." Indirect reciprocity works on the principle "I'll scratch your back if you scratch somebody's." In direct reciprocity, I use my experience with someone. In indirect reciprocity, I also use the experience of others. This is cognitively much more demanding. But both direct and indirect reciprocity are forms of conditional cooperation: the willingness to assist those who are willing to provide assistance.

I, based on the evolutionary biologist Richard Alexander, who coined the term "indirect reciprocity", stressed that it "involves reputation and status, and results in everyone in the group continually being assessed and reassessed." He argued that it represents the "biological basis of moral systems" [13] (the title of his book). Indirect reciprocity requires a high degree of information. Whether 'Alice' provides or refuses help to 'Bob' can be either directly observed by third parties, or learned through gossip from others. This forms the basis for Charlie's decision on whether or not to help 'Alice', in turn. 'Charlie' in effect acts upon a moral judgment which determines whether 'Alice' deserves to be helped or not.

In the very simplest model, we can assume that every player has a binary reputation, which can be either G (for "good") or B (for "bad"). Individuals meet randomly, as potential donors or recipients, and the donors can confer a benefit b to the recipient at a cost c for themselves. Donors who provide help obtain reputation G, and those who refuse obtain B. The discriminating strategy consists in giving help to G-recipients, and withholding help from B-recipients. This discriminating strategy is what, in the simpler situation of direct reciprocity (repeated games between the same two players) corresponds to TitForTat. In indirect reciprocity, it may be that players meet only once. Reputation takes the place of repetition.

Let us consider a population consisting of discriminators, as well as AllC- and AllD-players. It turns out that if the degree of information, i.e., the probability q to know the other player's reputation, is larger than the cost-to-benefit ratio c/b, then the population will either evolve towards fixation of the all-out defectors playing AllD, or towards a mixture of discriminating and undiscriminating cooperators. This mixture can eventually be subverted by neutral drift, in a manner reminiscent, but not quite similar to the direct reciprocity case. If we assume that each player, in time, becomes better informed about the co-players, then the mixture of discriminating and undiscriminating altruists actually becomes a stable attractor for the evolutionary dynamics.

The discriminating strategy considered so far displays an element of paradox. Indeed, it is certainly useful to the society when cooperation is channeled towards cooperators, and defectors are kept at bay; but it is costly to the discriminator. By refusing to help a B-player, discriminators acquire themselves the B-label, and are therefore less likely to be helped in the next round. Clearly, it would be better to distinguish between justifiable and unjustifiable defection. But this requires more sophisticated rules for assessing what is good and what is bad.

The very rudimentary moral system considered up to now is called SCORING: according to this system, it is always bad to refuse help [14, 15]. The STANDING rule seems more reasonable: for this rule, it is bad to refuse help to a good player, but not to refuse help to a bad player. An even sterner version is named JUDGING: it views, additionally, any act of help directed towards bad players as a bad behavior.

We can classify the different types of assessment rules. A first order assessment rule simply takes into account whether help is given or not. A second order assessment rule takes moreover into account whether the recipient is good or bad. A third order assessment rule takes additionally into account whether the donor is good or bad. This leads to 256 value systems [16, 17]. Only eight of them are stable in the following sense. For a homogeneous population adopting such a value system, there exists a uniquely specified rule of action (prescribing when to help, depending on the donor's and the recipient's image) which leads to

cooperation and which cannot be invaded by any other rule of action, and in particular not by AllC or AllD. Two of the eight stable rules are of second order, none of first order.

So far, there exist only very preliminary theoretical results on the competition between several co-existing assessment rules. Empirically, however, it seems clear that in many societies, several assessment rules co-exist. In particular, experiments have consistently shown that SCORING, despite its drawbacks, is adopted by a substantial part of the players. It seems to be cognitively very demanding to adopt higher-order assessment rules, since such rules require information, not only on the recipient's past behavior, but also on that of the recipient's recipients, etc. It can be argued that a population adopting a higher order assessment rule could continuously update the images of all the players in a consensual process, but this seems to require an extraordinary amount of information exchange.

The competition of rudimentary "moral systems" under conditions which include the possibility of occasional errors in action or judgement turns out to be remarkably difficult to analyse. Indeed, the status of a given group-member will in general be different for observers using different assessment rules. If additionally, the assessment of a player is based on several actions of that player, or if there are more than two labels for a player's reputation (for instance, "good", "bad" and "indifferent"), the complexity of the moral system explodes. Formalizing ethics appears to be harder than formalizing logic. Practical philosophy defies mathematizing.

5. Morality and natural science

This leads to the question whether it makes sense, at all, to study morality by methods from the hard sciences. Many people, especially among those with a background in the humanities, are uneasy with the application of Darwin's theory, or mathematical models, to the evolution of moral norms. In their eyes, ethics is a taboo topic for natural science, since it has to do with values, rather than with empirical facts. The following quote [18] stems from Pope John Paul II:

> *Consequently, theories of evolution which [...] consider the mind as emerging from the forces of living matter, are incompatible with the truth about man.*

It is not only American creationists, but many European intellectuals who would essentially agree. They may accept Darwinism in all other aspects, but shrink from applying it to the so-called higher faculties of humans. The most distinguished "exceptionalist" was Alfred Russell Wallace, the man who almost scooped Darwin by co-discovering natural selection. Wallace [19] wrote:

> *Man's intellectual and moral faculties [...] must have another origin [...] in the unseen universe of Spirit.*

Darwin himself did never shrink from investigating the evolution of our moral sense. One of his folders, which he later entitled *"Old and useless notes on the moral sense"* dates from 1837, when he was still in his 'twenties. Darwin did certainly grasp the importance of reciprocation, as is clear from the quote: "We are led by the hope of receiving good in return to perform acts of sympathetic kindness to others." And when he wrote: "[Man's] motive to give aid [...] no longer consists solely of a blind instinctive impulse, but is largely influenced by the praise and blame of his fellow men," he had obviously understood that in contrast to social insects, human cooperation is to a large extent based on reputation [20].

Indirect reciprocity was an essential factor in human evolution, because it provided a selective pressure for social intelligence, human language, and moral faculties. This does not imply, of course, that moral rules are innate. Just as we do not inherit a particular language, but have an innate faculty to quickly acquire a language, so we are not born with a ready-made moral system, but have the faculty to adopt one at an early age.

It should be stressed that the models considered so far are extremely crude. It is obvious that they cannot reflect the wealth of social and psychological mechanisms at work in human cooperation. Nevertheless, even very crude recipes can be useful. A telling example was provided by the recent growth of *e-commerce*. Economic interactions of this type typically occur on a global scale, and between partners who are almost anonymous. Nevertheless, the extremely simple reputation mechanisms provided, for instance, in an *e-bay* exchange are enough, in general, to guarantee that partners do not cheat on each other [21, 22].

6. Mutual Aid and sanctions in larger groups

So far, we have only considered interactions between two players. These can, in general, be understood by a cost-to-benefit analysis. The ratio c/b has to be smaller than something – for instance, the coefficient of relatedness r, or the probability of another round w (the "shadow of the future"), or the degree of information q about other players. Many interactions occur in larger teams, however, and this raises additional difficulties. The mere concept of reciprocation, for instance, becomes more difficult. Whom do you reciprocate with, if your group contains both cooperators and defectors? A typical model for such a situation is given by the so-called *Mutual Aid game* [1]. The situation of the Mutual Aid game mimics situations where members in a group who fall on hard times (through illness or another type of mishap) are helped by the other group members; it depicts a kind of mutual assurance funds.

All N players in a group are asked to contribute some amount, knowing that this will be multiplied by a certain factor $r > 1$ and then divided equally among all other players. If all contribute the same amount, their return will be the r-fold of that amount. But each individual receives nothing from the own investment in return: it is obviously more profitable

to invest nothing, and exploit the contributions of the co-players. However, if the other players follow the same line of action, no one will contribute anything. In actual experiments, players often contribute a substantial amount, but then, from round to round, reduce their contributions gradually. They feel exploited by those who contribute less than they did, and try to retaliate by contributing even less. But this hurts the cooperative players too, who then reduce their contributions in turn, etc.

Obviously, the snag in this game is that exploiters cannot be treated differently from cooperators. If the game is modified so that between the rounds of the Public Goods contributions, players can punish or reward specific individuals, depending on the size of their contributions, then cooperation can often be stabilized at a high level. This targeted form of providing positive or negative incentives – the carrot and the stick – again relies on reciprocation.

Trivers [9] described this in his paper on reciprocal altruism. He wrote: "Altruistic acts are dispensed freely among more than two individuals, an individual being perceived to cheat if he dispenses less than others, punishment coming from the others in the system." In many-person interactions, cooperators can gang together to punish cheaters. This can also be implemented in a game lab by modifying the Mutual Aid experiment. In this new two-stage game, the first stage runs exactly as before. In the second stage, players can impose fines upon their co-players. These fines are collected by the experimenter. They do not land on the punisher's account. In fact, each punisher must pay a fee for the experimenter to collect the fine.

From the viewpoint of *homo economicus*, the analysis is easy. A player bent on maximising income should not punish, since this is costly. Hence nothing ought to happen in the second stage. Hence, the first stage will be unaffected. No punishment, no contributions, and no gains: the selfishly motivated inertia in both stages of the game leads to economic paralysis.

Gratifyingly, this does not happen in real experiments, which are usually slightly more sophisticated versions where players can choose between different levels of contribution and sizes of fines. In the absence of punishment, contributions slide downhill; with punishment, they quickly rise to almost hundred percent. This happens if the groups stay together, but most significantly even if they are newly formed between rounds, and players know that they will never meet a co-player twice. By inflicting punishment, they can conceivably educate and "reform" a defector. But punishers know that the future contributions of such neophyte co-operators will exclusively benefit others. Punishment appears as an altruistic act [23].

This is an astonishing outcome. Without sanctions, mutual aid is not realised. With sanctions, it is, although selfish reckoning prescribes that costly punishment should not be delivered. In the absence of institutions, players are willing "to take the law into their own hands." This enforces cooperation in many-player interactions between unrelated individuals

–- a remarkable trait of human societies, and surely an essential factor in our evolutionary history. The investigation of the interplay between mutual assistance and social enforcement is a booming enterprise. Economists use experimental games to study the effects of positive and negative incentives –- reward and punishment –- on our propensity to collaborate. Anthropologists visit small-scale societies to measure the culture-dependence and universality of pro-social norms. Psychologists study the often sub-conscious cues eliciting emotions which lead to helping behaviour or moralistic aggression. Neurologists use magnetic resonance techniques to correlate social dilemmas with brain activities. Political scientists attempt to improve governance of institutions promoting collective actions. Trans-disciplinary dialogues are in full swing, although communication sometimes needs semantic abetment.

The mathematical explanation of the emergence of costly sanctioning raises interesting questions. In particular, since punishing others is costly, players ought to be tempted by so-called "second-order free riding." This means to leave it to others to punish exploiters, and to free-ride on the punisher's efforts. This could, of course, be overcome by also punishing the second-order free-riders. But this raises the possibility of third-order free riding, etc. Moreover, it has been shown by economic experiments that usually, the punishment of second-order free riders is little used, and has only small effects. There are two explanations for the emergence of the human propensity to punish exploiters. One is based on reputation. It is clear that if some individual is known to react aggressively against those who infringe norms of solidarity, then one should think twice about free-riding in an interaction which involves this individual. In this sense, it pays to broadcast one's commitment to punish defectors. Anger is loud.

Most economic experiments involving punishment in mutual aid games or related social interactions are conducted in terms of anonymity, and nevertheless show a widespread propensity to punish norm-breakers. How can this be explained? A first approach would suggest that we are simply witnessing an example of mal-adaptive behaviour. For thousands of generations, humans have lived in small groups, bands of hunter-gatherers or village-dwellers, where everyone had considerable knowledge of everyone else. We are simply not adapted to anonymity, and this is likely to affect our behaviour. Conscience has been famously described as "the nagging feeling that someone might be watching." We are back to reputational concerns: people will talk.

Evolutionary game theory has also helped in developing another approach. Let us assume that the participation in a joint enterprise, such as the mutual aid game, is not compulsory. Players, then, have the possibility of non-participation: they do not contribute to the communal benefit, and do not profit from it. After all, the participation in a mutual aid game (or "insurance game") is a kind of speculation: it is beneficial if most others contribute, but not otherwise. Let us assume that such non-participants can obtain a payoff on their own (through some autarkic activity) which lies somewhere between the payoff of a mutual aid

game which succeeds, and a mutual aid game which fails.

In this case, there exist four strategies. Players can (a) participate, but not contribute (these are the free riders); (b) participate, contribute, but not punish exploiters (these are the second-order free riders); (c) participate, contribute and punish exploiters (these are the pro-social individuals); and (d) not participate at all (the loners). The payoff for these strategies depends on the composition of the population. The composition can change with time, depending on the payoffs; indeed, if we assume social learning, players will from time to time switch to another strategy, with a propensity for adopting a strategy with a higher payoff. In this case, both theoretical analysis of the resulting stochastic process and computer simulations show that in the long run, the pro-social strategy prevails. Intriguingly, if we remove strategy (d), i.e., if we assume that the game is compulsory, then the pro-social strategy will not emerge. Instead, defectors will reign [24].

The same result occurs for many different versions of the learning process or the strategic interaction. The option of abstaining from the game acts as a catalyst for pro-social behaviour. So far, we have assumed that players engage themselves in sanctioning others. They take, in a sense, the law into their own hands. In civilised society, however, it is usually a police-like institution which is supposed to punish free riders. This can also be modelled by a simple scenario. Again, it consists of two stages. But this time, players first can contribute to a punishment pool (the rudimentary form of a police station), and then, in the second stage, they play the mutual aid game. All those who fail to contribute to the punishment pool or to the Mutual Aid will be punished, with an intensity which is proportional to the content of the punishment pool. This is called pool punishment, in contrast to the former peer punishment. In a sense, pool punishment is even more costly than peer punishment, because players have to contribute to the pool even if no need for punishment arises. If all players contribute to the Mutual Aid, there is no occasion to mete out sanctions. Nevertheless, the police will have to be paid. Self-justice, by contrast, is only costly if it is actually used. Again, evolutionary game theory shows that voluntary participation greatly promotes the likelihood for the emergence of pool punishment [25].

Despite the fact that peer punishment is more efficient than pool punishment, it is less stable. The reason seems to be that if everyone contributes to the Mutual Aid funds, it is not possible to spot second-order defectors in the peer-punishment scenario. By contrast, those who do not contribute to the punishment pool are just as easily spotted as those who do not contribute to the Mutual Aid. This is very reminiscent of the theory of the social contract. In fact, the very first sentence of Jean Jacque Rousseau's book says that "Man is born free, and everywhere he is in chains." This is often misquoted as "Man is born free, but everywhere he is in chains," as if there was an opposition between the two half-sentences. Evolutionary game theory suggests that this is not the case. It is just because men are free that they can commit themselves to mutual coercion, for mutual aid.

The role of institutions as tools providing incentives for human cooperation has been studied by many political, social and economic theorists. Such institutions (as long as they are not corrupted) provide external forces to guide us to cooperation. They are a device which seems, to a large extent, unique to the human species. But most of us would say, from introspection, that they do not engage in solidarity and fairness solely in order to avoid punishment, or to reap praise. They do it because they feel good about it: this is the proverbial "inner glow" provided by virtuous actions. Virtue is, in this sense, another device to guide us to cooperate, and very likely also unique, to a large extent, to the human species. Recent experiments have shown that humans act more generously if they have less time to think [26]. Our virtue is faster than our calculations. This calls to mind a bonmot by Talleyrand: "Never trust your first impulse: it is good."

It could well be that institutions and virtues have evolved jointly, and even that our propensity for developing virtue is based on a biological evolution which has followed the cultural evolution of sanctioning institutions. Humans display qualities such as obedience, and teachability. Moreover, we are particularly gifted at not only learning, but teaching. This can be viewed as the outcome of self-domestication.

References

1 Sigmund, K. (2010) The calculus of Selfishness (Princeton University Press)
2 Nowak, M.A., R. Highfield (2011) SuperCooperators: Why We Need Each Other to Succeed (Simon & Schuster)
3 Hamilton, W.D. (1996) Narrow Roads of Geneland: Collected Papers I (Freeman: New York)
4 Fisher, R.A. (1930) The Genetical Theory of Natural Selection (Oxford, Clarendon Press)
5 Dawkins, R. (1976) The Selfish Gene (Oxford University Press)
6 Darwin, C. (1859) The Origin of Species (London, John Murray)
7 Krupp, D.B., L.M. DeBruine, P. Barclay (2008) "A cue for kinship promotes cooperation for the public good," Evolution and Human Behavior, 29, 49-55
8 Smith, A. (1759) The Theory of Moral Sentiments (Glasgow)
9 Trivers, R. (1971) "The evolution of reciprocal altruism," Quart. Rev. Biol., 46, 35
10 Darwin, C. (1871) The descent of man, and selection in relation to sex (London, John Murray)
11 Axelrod, R. (1984) The Evolution of Cooperation (Basic Books: New York)
12 Nowak, M.A., K. Sigmund (1993) "Win-stay, lose-shift outperforms tit for tat," Nature, 364, 56
13 Alexander, R.D. (1987) The Biology of Moral Systems (Aldine de Gruyter: New York)

14 Wedekind, C., M. Milinski (2000) "Cooperation through image-scoring in humans," Science, 288, 850

15 Nowak, M.A. and K. Sigmund (1988) "The dynamics of indirect reciprocity," Nature, 393, 573

16 Nowak, M.A. and K. Sigmund (2005) "Evolution of indirect reciprocity," Nature, 437, 1292

17 Ohtsuki, H. and Y. Iwasa (2006) "How should we define goodness? Reputation dynamics in indirect reciprocity," J. Theor. Biol., 239, 435

18 Johannes Paul II (1997) "Message of the Pope to the Pontifical Academy," in Quarterly Review of Biology, 72, 397-399

19 Wallace, A.R. (1905) My Life: A Record of Events and Opinions (Chapman & Hall: London)

20 Darwin, C. (1872) The expression of emotions in animals and man (John Murray)

21 Whitfield, J. (2012) People will talk: The surprising science of reputation (John Wiley: New Jersey)

22 Beal, A. and J. Strauss (2009) Radically Transparent: Monitoring and Managing Reputations Online (John Wiley & Sons: New York)

23 Sigmund, K. (2007) "Punish or Perish? Retaliation and collaboration among humans," Trends Ecol. Evol., 22, 593-600

24 Hauert, C., A. Traulsen, M.A. Nowak, H. Brandt and K. Sigmund (2007) "Via freedom to coercion: the emergence of costly punishment," Science, 316, 1905-1907

25 Sigmund, K., H. de Silva, C. Hauert and A. Traulsen (2010) "Social learning promotes institutions for governing the commons," Nature, 466, 861-863

26 Rand, D.G., J.D. Greene and M.A. Nowak (2012) "Spontaneous giving and calculated greed," Nature, 489, 427-430