COOPERATION AND REPUTATION

Prepared for the 2007 Annual meeting of the American Economic Association
January 5-7, Chicago, IL

Andreas Bergh*
Peter Engseld+

ABSTRACT

The standard method when analyzing the problem of cooperation using evolutionary game theory, is to assume that people are randomly matched against each other in repeated games. In this paper we discuss the implications of allowing agents to have preferences over possible opponents. We model reputation as a noisy observation of actual propensity to cooperate and illustrate how reputation based choice of opponents can explain both the emergence and deterioration of cooperation. We show that empirical and experimental evidence of cooperation is consistent with our hypothesis that people behave so as to minimize the risk of damaging their reputation as nice, cooperative persons.

* Dept. of Economics, Lund University and the Ratio Institute.
+ Dept. of Economics, Lund University.

1. INTRODUCTION

The problem of cooperation arises when the individual behavior consistent with short term self-interest leads to socially suboptimal outcomes. A central problem in all the social sciences is to understand the circumstances under which cooperation can arise and when it is likely to break down. In the vast literature, there are several ways to explain human cooperation. For example, consider the idea of cultural group selection as described by Henrich (2004), opening a special issue of Journal of Economic Behavior & Organization on Evolution and altruism. The comments on Henrich's paper in this issue illustrate the diversity of approaches that are used to explain human prosociality, but they all succeed only if they provide a mechanism through which agents with altruistic behavior to a sufficiently high degree will meet likeminded opponents in the games of life. In this paper we argue that reputation based choice of opponents is such a mechanism, and that those who dismiss explanations based on signaling and reputation are wrong in doing so.

We analyze the problem of cooperation using the Prisoner's Dilemma in an evolutionary setting, with the twist that rather than being matched randomly against each other, agents are allowed to form preferences over opponents in the game. In our model, cooperation can be understood as signaling that will affect individual reputation and thereby the type of persons that will be encountered in subsequent games. The theory of reputation based choice is formally developed in Engseld and Bergh (2005), and our goal here is to demonstrate that this approach can be used to understand when people will succeed in solving the problem of cooperation. In particular, we claim that reputation concerns can be used to make sense of the huge amount of empirical (including experimental) evidence regarding human prosociality and seemingly altruistic behavior.

The idea that prosocial behavior can be interpreted as a signal appears in several different disciplines, including the contributions by Nelson and Green (2003), Gintis et al. (2001), Posner (2000), Zahavi and Zahavi (1997) and Frank (1988). Our contribution is to analyze specifically the consequences of allowing reputation based choice of opponents in a

repeated Prisoner's Dilemma, and to compare the implications of reputation based choice with existing empirical evidence of human behavior.

In short, our theory is good news for the emergence and stability of prosocial behavior. Under broad circumstances a population with only defection is unstable, because reputation based choice allows invading nice agents to choose to play only with each other. However, the theory predicts that when reputation is noisy, in the sense that there is some discrepancy between observed reputation and actual behavior, populations with complete cooperation are also unstable: People will be able to get away with an occasional defection without a sufficiently negative effect on their reputation.

The paper proceeds as follows. In the next section we give a brief overview of the problem of cooperation and some of the different approaches that have emerged as possible solutions. In section three we describe the theory of reputation based choice formally developed in Engseld and Bergh (2005), henceforth denoted E&B. In section four, we contrast the predictions of the theory with empirical evidence of cooperation, and discuss the objections that are sometimes raised against reputation models. Section five concludes the paper.

2. THE PROBLEM OF COOPERATION AND HUMAN PROSOCIALITY

The problem considered in this paper is known under a variety of names. Bendor and Swistak (2003) refer to it as the problem of reconciling the rationality of individuals with the formal and informal institutions that individuals have created. An excellent overview of the problem and the state of research is still the Presidential address made by Elinor Ostrom (1998) at the American Political Science Association. Ostrom uses the term "social dilemma", said to occur "whenever individuals in interdependent situations face choices in which the maximization of short-term self-interest yields outcomes leaving all participants worse off than feasible alternatives." (p. 1) Well-known incarnations of the social dilemma are Hardin´s (1968) tragedy of the commons, and the Prisoner's Dilemma as analyzed for example by Axelrod (1984). Recent research tends to incorporate many similar situations in a

unifying structure. For example, Gintis (2003) simply refers to the problem of prosociality. In this paper, we use the terms prosocial behavior and the problem of cooperation.

In all different variants, the fundamental question in the problem of cooperation is one and the same: Why it is that people facing social dilemmas, sometimes act in a way that is socially optimal even when this behavior is associated with a substantial personal cost? A number of different answers have been suggested in a literature that currently spans over several disciplines, including (but not limited to) political science, theoretical biology, sociology and economics. We review some of the most quoted explanations below.

According to *Kin Selection* (Hamilton, 1964) the seemingly altruistic behavior of for example parents towards their children can be explained as a way for the parents to promote their own genetic future. Clearly, this explanation is insufficient because human prosociality typically involves interactions between persons who are genetically unrelated.

Usually attributed to Wynne-Edwards (1962), G*roup Selection* is the idea that individual interaction takes place within groups, and groups where the members behave prosocially will do better than groups where the members behave selfishly. There is no general agreement regarding the role of group selection in human evolution. Furthermore, any theory of group selection needs to explain the absence of free riding behavior within groups. For a critique of group selection see for example Zahavi (2003), and for a nuanced defense see Sober and Wilson (1994).

Triver's (1971) idea of *reciprocal altruism* is that seemingly altruistic acts are eventually reciprocated in a beneficial way. In general, reciprocity requires some sort of enforcement mechanism to ensure that those who do not return favors are punished – otherwise it would be possible to free ride by enjoying the favors of others but never to reciprocate.[1] The question now becomes why people would engage in costly punishment of non-reciprocators, which is what Oliver (1980) calls a second order social dilemma.

Following Axelrod (1984), *tit-for-tat* became known as the best strategy in repeated games of Prisoner's Dilemma. The results rest on the fact that when the discounting of

[1] As pointed out for example by Frank (1988).

future benefits is sufficiently low, mutual cooperation is a Nash-equilibrium in the repeated Prisoner's Dilemma. This finding, however, is an insufficient answer to the problem of cooperation, because it says nothing about the problem of equilibrium selection: We know from the folk theorem that there are an infinite number of Nash-equilibria in repeated games of cooperation. Furthermore, as pointed out by Boyd and Lorberbaum (1987), a population of cooperating tit-for-tats can be invaded by nice but less retaliatory strategies, resulting in a population vulnerable to invasion by defecting strategies. Thus, tit-for-tat is not an evolutionarily stable strategy.

In economics, altruism is usually modeled by including in people's utility functions the well-being of others, or a preference for something assumed to be intrinsically good, such as fairness. Andreoni (1990) argues that altruism modeled this way is insufficient to explain observed behavior. His theory about *warm glow* rests on evidence suggesting that individuals derive some utility from the performance of certain prosocial actions, such as giving to charity.[2] It is our belief that when a theory essentially amounts to modifying the individual utility function to better describe observed behavior, we are in fact dealing with a description of human behavior and not a useful explanation of such behavior. Similar critique can be directed towards theories meant to explain behavior by modifying individual utility functions to include preferences for equality or fairness, see for example Fehr and Schmidt (1999).

Costly signaling (Smith and Bliege Bird, 2003) is closely related to the *Handicap principle* (Zahavi and Zahavi, 1997), and to the idea of *competitive altruism* (Roberts, 1998). According to these theories, prosocial behavior is in fact a signal about the sender's personal qualities. Ideas similar to signaling theory date back to Veblen's (1994 [1899]) analysis of conspicuous consumption. According to Veblen, people spend money on luxury goods to credibly signal their status. Analogously, because just saying that you are cooperative or trustworthy does not prove it, prosocial behavior can be seen as a signaling activity meant to transfer credible information about personal qualities. Spence (1973) showed that when signals convey

[2] Using a public good experiment, Palfrey and Prisbrey (1997) reject altruism in favor of heterogeneous warm glow effects.

information beneficial for the sender, there will be overinvestment in signaling activities. More recently, Posner (2000) presents a signaling story to explain how firms can signal their intentions of remaining in business for a long time for example by investing in expensive buildings and furniture. As pointed out by Smith and Bliege Bird (2003), costly signals are beneficial for the sender because of the information they transmit, and responding to signals in a way that benefits the signaler is simply the best move the responder can make given the available information.

Finally, we know that when the population is organized in some *social structure* that makes it more likely that cooperators end up meeting each other rather than being exploited by defectors, the probability that cooperation will emerge and be stable is higher, compared to random matching models. Examples include Stark and Bergstrom (1993), Boyd and Richerson (2002) and Skyrms (2004). But assuming there are such structures creates another question: What are the real life mechanisms that sustain such structures? Our theory about reputation based choice can be seen as an attempt to answer this particular question.

3. REPUTATION BASED CHOICE OF OPPONENTS IN THE PRISONER'S DILEMMA[3]

Consider a Prisoner's Dilemma game with the usual payoffs R (the reward for mutual cooperation), T (the temptation payoff for defection when the opponent cooperates), S (the sucker payoff) and P (the punishment for mutual defection). As customary, $S < P < R < T$, so that defect is a dominant strategy for each player, and $2R > S + T$, so that mutual cooperation is socially optimal. The game is repeated in an evolutionary environment and thus there is no discounting of future benefits. Our aim is to describe under what circumstances a cooperating population is evolutionarily stable in the sense that invading less cooperative strategies will earn a lower payoff than the cooperating incumbents.

The standard assumption when the problem of cooperation is analyzed using evolutionary game theory is that agents are randomly chosen to play with each other infinitely (random matching), or that each agent is paired against every other agent in the

[3] Formal proofs of the results presented in this section can be found in Engseld and Bergh (2005).

population (tournament matching). We assume instead that all agents have preferences over possible opponents, and that agents are matched up with their most preferred feasible opponent, based on mutual agreement. To understand how this works, think of a party consisting of an even number of singles. Preference based matching with mutual agreement is satisfied when everybody leaves the party with their most preferred partner, given that the partner agrees.

Our argument for leaving the standard assumptions of random/tournament matching is essentially based on realisticness. People do not play Prisoner's Dilemma with any person they randomly bump in to. Neither do people interact once with all other individuals in our society. In most cases we have the possibility to choose our opponents in the games of life. We choose which of our friends can be trusted with delicate tasks in which they will be tempted to defect upon us, and we choose what car dealer we use when we want to sell our old car – or buy a new one. Using the words of evolutionary game theory: Our strategies contain not only action rules regarding how to play the games of cooperation, but also preferences over opponents in these games.

Our model is based on two central concepts: propensity and reputation. By propensity we mean a measure of the actual probability that an agent with a certain strategy will cooperate in a given population. By reputation we mean a noisy observation of an agent's propensity. Thus, reputation gives a crude picture of how likely it is that a person will defect, and nothing else. A crucial feature of the model is that each agent sees only the reputation of her potential opponents. We assume that all agents are free to choose among all potential opponents..

A strategy in our model is a mapping from own propensity and the opponents' reputation to the action set which contains the two actions cooperate or defect, and to the set of all complete and transitive preferences over opponents. Strategies can be conditioned and/or mixed, which means that there are an infinite number of conceivable strategies – the strategy space is infinite. Some are easily described, such as 'play with anyone and always cooperate' or 'prefer agents with high reputation and play cooperate with probability 0.98',

but more complex strategies are also allowed. Because some specific strategies or types of strategies play a crucial role, it will be useful to name them as follows:

Smart Cooperation: The strategy that ranks opponents in order of preference from highest to lowest reputation, and always cooperates.
Smart Defection: Like Smart Cooperation, but always defects.
Smart Occasional Defection: Like smart cooperation, but plays cooperate with a probability close to but less than 1.
Naïve Cooperation: Any strategy that is either indifferent regarding the reputation of the opponent or prefers opponents with lower reputation, and always cooperates.

The model rests on the assumption that the adjustment process of propensities is much faster than the growth/learning process, which in turn is much faster than the process of mutations. This assumption means that in any given population, a certain strategy will induce a certain behavior which in turn will lead to a stable propensity. Equivalently, for every strategy mix, there is a corresponding propensity distribution. This allows us to examine evolutionary stability by checking if there is any strategy that could successfully invade the population of incumbents. Such invading mutants are assumed to have the propensity associated with the way their strategy behaves in the given population.

3.1 When reputation accurately reflects behavior

To understand the implications of reputation based choice, assume for a moment that there is no reputation noise: Anyone will be able to tell the difference between cooperators and occasional defectors – no matter how seldom defections occur. Under these assumptions, a population with almost complete cooperation is evolutionarily stable. To see this, note that an agent who defects in one period will earn a higher payoff in that period, but her reputation will be lower because of the defection, and it will be possible for other agents to avoid her in the future. A similar reasoning shows why any population with some degree of defection is evolutionarily unstable when reputation perfectly reflects past behavior: Smart

cooperators will be able to avoid occasional defectors, earn a higher average payoff and successfully invade population.

When there is no reputation noise, a cooperating population will be vulnerable to some evolutionary drift, similar to the drift that kills the stability of tit-for-tat, as described above. However, the drift in our model does not render cooperation unstable. To see this, consider a population of smart cooperators. In this population, naïve cooperators will earn just as much, and can thus neutrally invade the population. This invasion makes it possible for smart defectors to exploit the naïve cooperators, and earn more. However, when naïve cooperators end up being suckers, their fraction of the population will decrease and smart defectors will end up playing themselves more often – because the smart cooperators are able to avoid them. When smart defectors meet, they end up with the mutual defection payoff, and their share of the population will decrease. Thus, a fraction of naïve cooperators exploited by smart defectors guarantees the stability of a population dominated by smart cooperators (B&E, Proposition 1):

Result 1. When reputation perfectly reflects behavior, the population will consist almost entirely of smart cooperators and a small fraction of smart defectors and naïve cooperators.

3.3 When reputation is noisy

When reputation imperfectly reflects the propensity to cooperate, smart cooperators will sometimes mistakenly chose to play with occasional defectors, because they cannot successfully identify the occasional defector as having a lower propensity. If these identification mistakes are frequent enough, occasional defectors will reap the temptation payoff sufficiently often to earn more on average.

For an illustration, assume that the population consists of only two types, A and B, which exist in equal proportions (see figure 1). Type B cooperates more often than A, and the probability distributions show that when A-agents observe the reputation of B-agents, the observation will be the realization of a random variable. Assuming that reputation noise

is symmetric, reputation will underestimate propensity with 50 percent probability, and with the same probability overestimate it.

In figure 1a, the reputation noise is small enough so that B will never mistakenly perceive A to have higher propensity. Thus B can successfully avoid A, earn more, and take over the population.

Fig. 1a

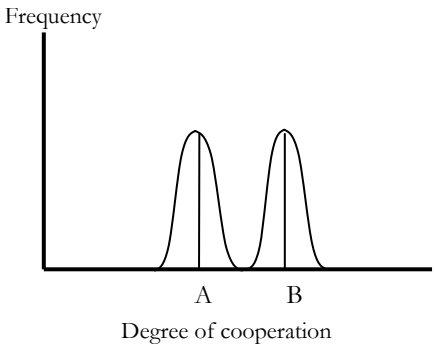
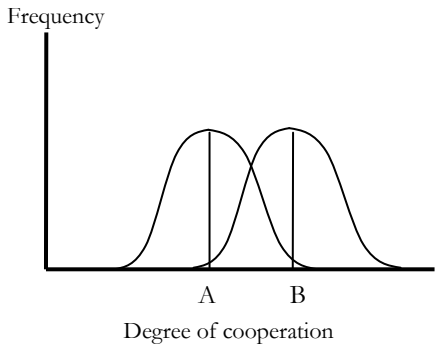


Fig. 1b



In figure 1b, the situation is different. Reputation noise is stronger, and with some probability, B will mistakenly perceive A to be more cooperative, choose to play with them in which case B gets the sucker payoff. The frequency of identification errors required for A-agents to earn more than B-agents on average depends on the relative size of the temptation payoff: The more can be gained from an occasional defection, the smaller is the frequency of identification errors required for a successful invasion by less cooperative agents. Note also that the frequency of identification errors will be higher, when the propensity difference between A and B is smaller, and when reputation is more noisy. The most important consequence of reputation noise is, however, independent of its size and does not depend on the payoff structure:

Result 2. When reputation is noisy, a population with only cooperation is not evolutionarily stable.

To see the intuition behind result 2, consider a population consisting only of smart cooperators. For a given payoff structure and a given level of reputation noise, it is possible for occasional defectors to invade the population: When reputation is noisy, the occasional defectors will earn more on average and the invasion will be successful. Intuitively, when reputation noise is lower, occasional defectors must defect less often – otherwise their reputation will reveal them, and their invasion will not be successful.

The presence of reputation noise will, however, not result in a complete collapse of all cooperation in the population. Suppose that a cooperating population is completely invaded by occasional defectors. The new population will be vulnerable to yet another invasion by occasional defectors, who defect slightly more often. After this second invasion is complete, it may be the case that the degree of cooperation has declined enough to ensure that invading smart cooperators will recognize each other sufficiently often to earn more on average, in which case the population can revert to complete cooperation and the process will start over.[4]

The extent to which defection must increase before the population can be successfully invaded by smart cooperators depends on the level of reputation noise. As reputation noise increases, smart cooperators will more often fail to find each other, and occasional defectors will earn more. If reputation is noisy enough, the reputation mechanism becomes impotent and complete defection will be evolutionarily stable.

This illustrates how the theory of reputation based choice provides an explanation of both why cooperation deteriorates, and why cooperation can suddenly arise even in populations with a high degree of defection.

The effect of decreased reputation noise is that nice agents will more often successfully end up playing each other, earning a higher payoff than less nice agents. This leads to the following result (E&B, proposition 2 & 3):

[4] A similar cyclicity result is described in Poulsen and Tinggaard Svendsen (2005). However, they only allow three types of behavior (always cooperate, always defect and a reciprocal strategy similar to tit-for-tat).

Result 3: The degree of cooperation is decreasing in the level of reputation noise.

Consider now the consequences of assuming that in some games of life, there is no free choice of co-players. Sometimes we involuntarily end up in cooperation games without having chosen our opponents. Does this substantially limit the cooperation-inducing effect of reputation based choice? The answer is no: Reputation is affected by behavior in all games, including one-shot interactions with strangers. The reputational effect of behaving nicely also in these situations may well make up for the short term costs associated with cooperative behavior.

4. REPUTATION BASED CHOICE AND EXISTING EMPIRICAL EVIDENCE

Evidence supportive of reputation based choice comes from a number of fields. First of all, there are numerous studies in which people have been playing actual games of Prisoner's Dilemma, see for example Rapoport, Guyer, and Gordon (1976) and Selten and Stoecker (1986). Typically, cooperation levels are relatively high even in finitely repeated games, where backwards induction suggests that there should be no cooperation at all. The same goes for games and situations which are essentially n-person Prisoner's Dilemma games, such as the collective action problem in handling the tragedy of the commons and the private provision of public goods. For example, Marwell and Ames (1979) showed that people contribute to public goods even when they have the opportunity to free ride. There are also results indicating that smaller groups are more successful in handling collective action problems and avoiding free riding - see Bandiera, Barankay, and Rasul (2004) and Wagner III (1995).

While these results are in line with the predictions of reputation based choice, the literature contains lots of additional empirical and experimental evidence supporting the idea that reputational concerns can explain seemingly altruistic behavior. In this section we review some of this evidence.

4.1 Evidence that reputation matters

First of all, reputation based choice suggests that individuals have a strong incentive to care about their reputation. One clear indication of this comes from the methodological literature on survey responses. In many situations there is a substantial discrepancy between what people say they do and what they actually do. A classic example is when Parry and Crossley (1950) showed that people lie about their donations to charity: About one third of their sample incorrectly claimed to have given – while no one falsely claimed the opposite. Citing the Parry and Crossley study, Nelson and Greene (2003) drive home the point of reputation based choice by concluding that the amount of lying discovered "does provide evidence that charity yields a reputational return in terms of more or better reciprocity partners". Arguing against explanations based on altruism, Nelson and Greene also note that "if people gave to charity solely for altruistic reasons, there would be no return to them from others believing that their charitable contributions were larger than they actually were." (p. 29)

The expected effects of reputation and anonymity are present in recent field experiments as well. Using data from Dutch churches, Soetevent (2004) shows that when churches collect money using baskets that allowed attendants' contributions to be visible to their direct neighbors in the church bench, contributions increased compared to when the amounts given could not be seen. Also, baskets induced switching to giving larger coins.

4.2 Reputation effects in the lab

A large body of evidence regarding seemingly altruistic behavior comes from laboratory experiments where subjects using real money are put in situations that require strategic interaction. In the context of the Prisoner's Dilemma game, Bohnet and Frey (1999) show that relaxing the strict anonymity conditions that are standard in most experiments, increases cooperation. Bolton, Katok, and Ockenfels (2005) show that providing information about a partner's immediate past action increases cooperation in an image scoring game designed specifically to test the influence of reputation information. These findings are in line with the

predictions from reputation based choice, and the effect appears in many other games as well – as will be shown below.

Economic experiments have revealed at least two patterns in human behavior (for an overview of the experimental evidence, see for example Fehr and Fischbacher, 2003). First of all, people are prone to play nice and generous even when short run rationality would imply a more selfish behavior. Secondly, people are prone to punish those who deviate from playing nice and generous.

A typical example of nice and generous behavior is the dictator game, where people voluntarily give away money to other subjects, even when they with certainty are allowed to keep the money for themselves – see for example Forsythe et al. (1994). To explain this using reputation based choice, we suggest that people benefit from having a reputation for being nice and generous, because it increases the possibility of future interactions with likeminded people. The prediction that such generosity is bigger and more common when the reputation effect is bigger has been verified experimentally, by for example Hoffman, McCabe, and Smith (1996) and Hoffman, McCabe, Shachat, and Smith (1994) who show that people behave less generously when social isolation increases and the reputation effect is smaller: When proposers in the ultimatum game (Güth, Schmittberger, and Schwarze, 1982) enjoy higher anonymity, their proposals are less generous.

Explaining punishing behavior and rejections in the ultimatum game is less straight forward. One possibility is that people can improve their reputation by punishing those who violate cooperative norms. We call this hypothesis cooperative punishing. Another possibility is that punishing produces a beneficial reputation for not accepting to be exploited. We call this hypothesis spiteful punishing.

The idea in cooperative punishing is to punish deviations from the socially optimal behavior. In economic experiments people regularly punish those who do not share random wealth, and those who do not respect property rights. To interpret this as cooperative punishment, it can be argued that sharing random wealth is socially efficient because it reduces consumption variance, as has been noted and documented in several food sharing

studies, for example Gurven (2004) and Kaplan and Hill (1985). From an individual's perspective, however, there is a temptation to violate sharing norms by never sharing but still accepting gifts from others. There are also efficiency reasons for respecting property rights: If people do not respect the property of others, everybody must spend resources on defending their own property, which is inefficient. However, the efficient outcome entails an incentive to steal from others – see Bos and Kolmar (2003) and the references cited therein. Cooperative punishing is puzzling, because in standard models punishing is not evolutionarily stable. This puzzle can be explained if punishing is reputation improving.

For an example of spiteful punishing, consider an ultimatum game and imagine that an agent has a reputation for punishing proposers who offer very small shares. If the proposer can detect this reputation before the offer is made, it will be optimal to offer just above the level where the responder will reject.[5] Seemingly, the responder rejects unfair offers, but the consequence may well be that he maximizes his own payoff. Note that punishing may have reputational consequences even when the action punished was directed against someone else, as in the third party punishment game (Fehr and Gächter, 2002).

Regardless of whether punishing is cooperative or spiteful, we expect the willingness to punish to be decreasing in anonymity. In general, this prediction has big support. Eckel and Wilson (2001) show that punishment is more common when there is a signal value to punishment. Also, Blount (1995) shows that people are less prone to punish unequal proposals when such signaling makes no sense because proposals are computer generated.

Many experimental results are compatible with both hypotheses of punishing. Hoffman, McCabe, Shachat, and Smith (1994) find that people share less, and are less prone to punish unequal proposals in the ultimatum game, when the money is earned through a procedure rather than given like manna from heaven. Also, people spend resources on punishment of those who violate contracts (Fehr, Gächter, and Kirschsteiger, 1997) and those who overuse a common pool of resources (Ostrom, Walker, and Gardner, 1992).

[5] In fact, Tullberg (1999) has shown that offers in ultimatum games tend to be rational in this sence.

To conclude, we think that existing evidence support the idea that reputational effects can explain punishing, but further experimental testing is needed to discriminate between spiteful, cooperative or other types of punishing.

4.3 Evidence concerning the matching procedure

If the evidence that reputation matters is abundant, there is much less evidence available regarding the effect of being able to choose co players. Nevertheless, we have reason to believe that a free choice of co players would increase cooperation compared to experiments with random matching. One indication is that people seem to be able both to signal and detect cooperative intentions (Frank, 1988). Scharlemann, Eckel, Kacelnik, and Wilson (2001) find that facial expressions are interpreted as signaling whether subjects are cooperative and trustworthy. They also find some evidence that smiles and other facial expressions can elicit cooperation among strangers in one-shot games.

There is also some intuitive evidence that when pairs are formed based on cooperative intentions, cooperation will increase. McCabe et al. (2003) conducted a trust game where the subjects were regrouped so that trusting players were paired with trustworthy players. Unsurprisingly, the authors conclude that this sorting allows cooperation to emerge and protects cooperation from being invaded by defecting players, which is exactly the mechanism used in reputation based choice. Finally, there is also evidence that people do have preferences over opponents in strategic interactions: Engseld and Holm (2005) studied ultimatum proposals and dictator donations when proposers can choose the income and sex of the responder. It turns out that subjects preferred to send proposals to low-income responders, and that and that females were more popular responders than males.

4.4 How anonymous are anonymous experiments?

Henrich (2004) dismisses reputation explanations of cooperation because of we must be able to explain cooperation among anonymous people in one-shot encounters leads (p. 7). But in the reputation model described above, cooperation in one-shot encounters is no mystery: As noted in section 3, reputation is affected by all actions regardless of against

whom they are directed. For this reason, even if interaction takes place mainly within a limited group, actions in one-shot encounters with non-group members will affect reputation and thus subsequently whom you will meet within the group. Occasional kindness against strangers does have a reputational benefit.

The objection based on anonymity is more fundamental. If cooperation can be explained by reputation effects, why do we still see a substantial degree of cooperative and seemingly altruistic behavior in experiments with complete anonymity? To answer this question, we pose a counter-question: Are anonymous experiments really anonymous? Hoffman, McCabe, and Smith (1996) argue as follows (p. 655):

> *"In laboratory experiments we cannot assume that subjects behave as if the world is completely defined by the experimenter. Past experience is important in so far as beliefs are based on experience. The future is important in so far as people are accustomed to operate in an environment, in which there is ongoing social interaction, and in so far as subjects may be concerned about the extent to which their decisions have post-experimental consequences, or that others may judge them by their decisions."*

Because Hoffman et al. still see some prosocial behavior, even in their most anonymous setting, they conclude that "[t]his may reflect true utilitarian 'other-regarding' preferences. Alternatively, these subjects may be suspicious of our procedures to guarantee anonymity." (p. 658)

The last point here is worth expanding on. Suppose that subjects attach some positive probability to the possibility that the anonymity in a certain experiment is not guaranteed – after all, it has been known to happen that people cheat or lie, or that computers fail.[6] Even if the probability is very small, the expected reputation effect can easily overshadow the benefits from behaving selfishly, especially because the amount of money at stake in

[6] Also, one might question what it means that an experiment is anonymous. In their famous paper, Fehr and Gächter (2002:137) made the bold statement that their "design ruled out any kind of reputation formation". Careful reading of their footnote on methods, however, reveals that this particular experiment involved 10 sessions with 24 agents in each, picked from the same two universities, matched in groups of four agents, placed in computer booths and after each period informed about the behavior of the other three in the subgroup. We do not suggest that the design was flawed in any way, but we argue that it can not be excluded that some agents felt skeptical regarding their anonymity towards the computers, the conductors of the experiment, or that some participants could identify each other afterwards to compare payoffs or to discuss their actions.

experiments is typically relatively small. For this reason, we suggest that the alleged altruistic behavior in many experiments is simply individuals behaving so as to minimize the risk of ruining their reputation. If you have invested in your reputation as a cooperator, why risk it in an experiment where you are not 100 percent sure of anonymity?

If our explanation is valid, then it should be at least theoretically possible to produce selfish behavior by carefully ruling out all possible reputation effects. This was done by Cherry et al. (2002) who conducted dictator games where subjects did not have any contact before, during, or after the session, and where subjects where using earned wealth. Under these circumstances, 95 percent of the subjects behaved according to pure self-interest and kept their money. The authors conclude that strategic concerns as opposed to fairness appear to be the motivation for other-regarding behavior.

5 CONCLUDING DISCUSSION

Under reputation based choice with at least some degree of reputation noise, the best strategy is to be an occasional defector seeking to play with high reputation opponents. When observability is high, defections should be rare. When anonymity is high, more frequent defections will be beneficial. As we have shown, this simple intuition can be used to make sense of the overwhelming body of empirical evidence of seemingly altruistic behavior.

The theory of reputation based choice has some interesting implications suitable for further research. We need to know more about how reputation is transmitted in everyday life. A number of ways in which people can signal a good reputation is examined in the book appropriately titled *Signaling Goodness* (Nelson and Greene, 2003).

There are also some potential policy implications from our approach. Cooperation can be fostered by providing institutions where people have incentives to care about their reputation. Such incentives are present if people have the possibility to choose their partners in strategic interactions, and they will be stronger when transparency is higher and behavior is observed by more.

For example, reputation based choice can be part of an explanation why some political institutions seem to be correlated with higher quality government. The problem of

corruption has the properties of a cooperation game, where the socially efficient outcome – no corruption – is unstable because of the big temptation in bribery. Gerring and Thacker (2004) found that more political competition, openness and transparency is correlated with lower corruption. In general, we can understand the effect of transparency as lowering the level of reputation noise when citizens observe the behavior of governmental institutions and each other. This suggests that the higher degree of transparency in North European countries is a possible explanation of why corruption in these countries is lower than in the rest of Europe – see Della Porta and Mény (1997).

To conclude, we have shown that a large amount of evidence of human altruism is in fact compatible with the hypothesis that agents behave so as minimize the risk of ruining a good reputation. We strongly believe that this is an encouraging result: Human cooperation is not paradoxical – it has evolved, and will continue to do so.

References

Andreoni, J. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving." *Economic Journal* 100:464-477.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2004. "Cooperation in Collective Action: A review of the Evidence and Some New Results." *Mimeo*.

Bendor, Jonathan and Piotr Swistak. 2003. "The Rational Foundations of Social Institutions: An Evolutionary Analysis." in *Politics from Anarchy to Democracy*, edited by I. Morris, J. Oppenheimer, and K. Soltan: Stanford University Press.

Blount, Sally. 1995. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes* 63:131-144.

Bohnet, Iris and B. S. Frey. 1999. "The Sound of Silence in Prisoner's Dilemma and Dictator Games." *Journal of Economic Behavior and Organization* 38:43-57.

Bolton, Gary E. , Elena Katok, and Axel Ockenfels. 2005. "Cooperation among strangers with limited information about reputation." *Journal of Public Economics* forthcoming.

Bos, D and M Kolmar. 2003. "Anarchy, efficiency, and redistribution." *Journal of Public Economics* 87:2431-2457.

Boyd, Robert and Jeffrey Lorberbaum. 1987. "No Pure Strategy is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game." *Nature* 327:58-59.

Boyd, Robert and P. J. Richerson. 2002. "Group Beneficial Norms Spread Rapidly in a Structured Population." *Journal of Theoretical Biology* 215:287–296.

Cherry, Todd L, Peter Frykblom, and Jason F Shogren. 2002. "Hardnose the Dictator." *American Economic Review* 92:1218-21.

Della Porta, Donatella and Yves Mény. 1997. "Democracy and Corruption in Europe." in *Social change in Western Europe*. London: Pinter.

Eckel, Catherine and Rick K. Wilson. 2001. "Why Fairness? Facial expressions, evolutionary psychology, and the emergence of fairness in simple bargaining games." *Paper prepared for WOW2 -- the 25th Anniversary Workshop on the Workshop*.

Engseld, Peter and Andreas Bergh. 2005. "Choosing Opponents in Prisoners' Dilemma: An Evolutionary Analysis." Presented at the winter meeting of the Econometric Society in Boston, January 2006. Working Paper No. 2005:45, Dept. of Economics, Lund University.

Engseld, Peter and Håkan J. Holm. 2005. "Choosing Bargaining Parties - An experimental study on the impact of information about income and gender." *Forthcoming in Experimental Economics*.

Fehr, Ernst and Urs Fischbacher. 2003. "The nature of human altruism." *Nature* 425:785-791.
Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415.
Fehr, Ernst, Simon Gächter, and G. Kirschsteiger. 1997. "Reciprocity as a contract enforcement device: experimental evidence." *Econometrica* 65:833-860.
Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics* 14:815-868.
Forsythe, R. , J. Horowitz, N.E. Savin, and M. Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior* 6:347-369.
Frank, Robert H. 1988. *Passions within Reason.* New York: Norton.
Gerring, John and Strom C. Thacker. 2004. "Political Institutions and Corruption: The Role of Unitarism and Parliamentarism." *British Journal of Political Science* 34:295-330.
Gintis, Herbert. 2003. "Solving the Puzzle of Prosociality." *Rationality and Society* 15:155-187.
Gintis, Herbert, Eric Alden Smith, and Samuel Bowles. 2001. "Costly Signaling and Cooperation." *Journal of Theoretical Biology* 213:103-119.
Gurven, Michael. 2004. "Reciprocal altruism and food sharing decisions among Hiwi and Ache hunter–gatherers." *Behavioral Ecology and Sociobiology* 56:366-380.
Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* 3:367-388.
Hamilton, W. D. 1964. "The genetical theory of social behavior." *Journal of Theoretical Biology* 7:1-32.
Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162:1243-1248.
Henrich, J. 2004. "Cultural group selection, coevolutionary processes and large-scale cooperation." *Journal of Economic Behavior and Organization* 53:3-35.
Hoffman, E., Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* 7:346-380.
Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review* 86:653-660.
Kaplan, H and K Hill. 1985. "Food sharing among Ache foragers: tests of explanatory hypotheses." *Current Anthropology* 26:223-245.
Marwell, G. and R. E. Ames. 1979. "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem." *American Journal of Sociology* 84:1335-1360.
McCabe, Kevin, Mary L. Rigdon, and Vernon L. Smith. 2003. "Sustaining Cooperation in Trust Games." *Mimeo.*
Nelson, Phillip J. and Kenneth V. Greene. 2003. *Signaling Goodness: Social Rules and Public Choice.* USA: University of Michigan Press.
Oliver, Pamela. 1980. "Rewards and punishments as selective incentives for collective action: Theoretical investigations." *American journal of sociology* 85:1356-1375.
Ostrom, Elinor. 1998. "A behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Addres, American Political Science Association, 1997." *The American Political Science Review* 92:1-22.
Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Convenants with and without a Sword - Self-Governance is Possible." *American Political Science Review* 86:404-417.
Palfrey, Thomas R. and Jeffrey E. Prisbrey. 1997. "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *American Economic Review* 87:829-846.
Parry, Hugh and Helen Crossley. 1950. "Validity of Responses to Survey Questions." *Public Opinion Quarterly* 14:61–80.
Posner, Eric A. 2000. *Law and Social Norms.* Cambridge: Harvard University Press.
Poulsen, Anders and Gert Tinggaard Svendsen. 2005. "Social capital and endogenous preferences." *Forthcoming in Public Choice.*
Rapoport, A., M.J. Guyer, and D.G. Gordon. 1976. *The 2×2 Game.* Ann Arbor, MI: University of Michigan Press.
Roberts, Gilbert. 1998. "Competitive Altruism: From Reciprocity to the Handicap Principle." *Proceedings: Biological Sciences* 265:427-431.
Scharlemann, Jörn P. W., Catherine C. Eckel, Alex Kacelnik, and Rick K. Wilson. 2001. "The value of a smile: Game theory with a human face." *Journal of Economic Psychology* 22:617-640.
Selten, R. and R. Stoecker. 1986. "End behavior in sequences of finite repeated prisoner's dilemma supergames." *Journal of Economic Behavior and Organization* 7:47-70.
Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure.* USA: Cambridge University press.
Smith, E.A. and R. Bliege Bird. 2003. "Costly signaling and prosocial behavior." in *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life*, edited by H. Gintis, S. Bowles, R. Boyd, and E. Fehr. Cambridge: MIT Press.
Sober, E. and D.S. Wilson. 1994. "Reintroducing group selection to the human behavioral sciences." *Behavioral and Brain Sciences* 17:585-654.
Soetevent, Adrian R. 2004. "Anonymity in Giving in a Natural Context -- A Field Experiment in Thirty Churches." *Forthcoming in Journal of Public Economics.*
Spence, M. A. 1973. "Job market signaling." *Quarterly Journal of Economics* 87:355-374.

Stark, Oded and Theodore C. Bergstrom. 1993. "How altruism can prevail in an evolutionary environment." *American Economic Review* 83:149-156.
Trivers, R. L. 1971. "The evolution of reciprocal altruism." *Quarterly Review of Biology* 46:35–57.
Tullberg, Jan. 1999. "The Ultimatum Game Revisited." *Working papers series in Business Administration, 1999:2, Stockholm School of Economics, revised 10 Jan 2002*.
Wagner III, John A. 1995. "Studies of individualism-collectivism: Effects on cooperation in groups." *Academy of Management Journal* 38:152-173.
Veblen, Thorstein. 1994 [1899]. *The Theory of the Leisure Class*. New York.: Dover. (orig. publ. by Macmillan, NY).
Wynne-Edwards, V.C. 1962. *Animal Dispersion in Relation to Social Behaviour*. New York, NY.: Hafner Publishing Company.
Zahavi, A. 2003. "Indirect selection and individual selection in sociobiology: my personal view on theories of social behavior." *Animal Behavior* 65:859-863.
Zahavi, A. and A. Zahavi. 1997. *The Handicap Principle*. New York: Oxford University Press.