Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) 73–84

Ernst Fehr/Urs Fischbacher

Altruists with Green Beards

Abstract: If cooperative dispositions are associated with unique phenotypic features ('green beards'), cooperative individuals can be identified. Therefore, cooperative individuals can avoid exploitation by defectors by cooperating exclusively with other cooperative individuals; consequently, cooperators flourish and defectors die out. Experimental evidence suggests that subjects, who are given the opportunity to make promises in face-to-face interactions, are indeed able to predict the partner's behavior better than chance in a subsequent Prisoners' Dilemma. This evidence has been interpreted as evidence in favor of green beard approaches to the evolution of human cooperation. Here we argue, however, that the evidence does not support this interpretation. We show, in particular, that the existence of conditional cooperation renders subjects' choices in the Prisoners' Dilemma predictable. However, although subjects predict behavior better than chance, selfish individuals earn higher incomes than conditional cooperators. Thus, although subjects may predict other players' choices better than chance evolution favors the selfish subjects, i.e., the experimental evidence does not support the green beard approach towards the evolution of cooperation.

1. Introduction

In recent years, a large amount of evidence has accumulated which indicates that many people are willing to bear costs to reward others for cooperative acts or punish others for norm violations, even though they gain no individual net benefit from these acts (Dawes 1980; Güth et al. 1982; Ledyard 1995; Fehr/Fischbacher/Gächter 2002; Gintis et al. 2003). The combination of altruistic rewarding and altruistic punishment has been coined as strong reciprocity (Gintis 2001; Fehr/Fischbacher 2003). Recently, several papers have been published that build an evolutionary foundation for strong reciprocity (e.g., Gintis 2001; Henrich/Boyd 2002; Boyd et al. 2003). These papers typically assume that—in addition to the forces of individual within-group selection operating against the evolution of the altruistic trait—some sort of cultural evolution or gene-culture co-evolution interacts with selection at the group level to produce strongly reciprocal behavior. However, in principle, it might also be possible to generate an individual selection account of strong reciprocity. In this paper we deal with one particular individual selection theory that has been considered in the literature. A potential evolutionary explanation for strong reciprocity is to assume that individuals who reward and punish altruistically have observable characteristics ('green beards') that distinguish them from non-altruists (Frank 1987; 1988; Robson 1990; Amann/Yang 1998). It is then in any individual's self-interest to cooperate with a green beard because non-cooperation will be

punished. Thus, altruistic punishment evolves because the punishers directly benefit from their observable willingness to punish. Even in the absence of punishment opportunities, green beards favor the evolution of cooperation because individuals can condition their cooperation to the existence of a green beard. If an altruist meets a selfish individual without a green beard he defects, if he meets a green beard he cooperates. In this way, only the altruists themselves reap the benefits of altruistic behavior so that universal cooperation evolves. Although the green beard account of the evolution of strong reciprocity and human cooperation seems plausible at first sight, we take issue with this approach. Thus, we argue in the following why we regard the green beard theory as unsatisfactory.

2. Arguments Against the Green Beard Approach to Human Altruism

An obvious objection to the green beard approach is that it seems very unlikely that humans are capable of perfectly distinguishing altruists from non-altruists by means of observable characteristics. It takes at least a minimum effort to make this distinction, and judgments about other people are invariably associated with some uncertainty. This objection has, however, little force because the green beard argument can be made more realistic by assuming that it takes some effort or that there is only a positive probability of detecting a strong reciprocator (Frank 1988). However, the crucial assumption behind this approach is that there are no selfish mutants with green beards. As soon as one allows for such mutants, the argument breaks down because the mutants reap the same benefits as the altruists but do not bear the cost of altruistic acts. Therefore, any convincing model justifying the green beard argument must allow for such mutants. To our knowledge there are no such models. In prevailing evolutionary models of cooperation based on the green beard argument such mutants are ruled out by assumption (Frank 1988; Robson 1990).

In the human context it is also important to keep in mind that humans are experts in deceiving others and that the abilities of ordinary humans to detect lies are not good. There is extensive literature showing that trained researchers are able to detect other people's lies with high probability using scientific methods, but that most humans are unable to detect lies better than chance (Ekman/O'Sullivan 1991). The inability of humans to detect lies holds for a wide variety of situations, including those where stakes are high. In the most convincing experiment (Ekman/O'Sullivan 1991), the researchers showed the participants videos of lying and non-lying subjects. They ensured that it was objectively possible, by applying scientific tools, to predict the liars with high probability by watching the videos. The data indicates that 70 percent of the subjects—including subjects who can be expected to be experts, such as federal polygraphers, crime investigators, judges and psychiatrists, are unable to detect lies significantly better than chance. Ockenfels and Selten (2000) conducted an experiment that explicitly tested the hypothesis of involuntary truth-signaling put forward by Frank (1988) in the context of a bargaining experiment. In this

experiment, two subjects had to bargain about 30 German Marks (DM) in a face to face interaction that lasted maximally 10 minutes. Some subjects had bargaining costs of DM 12, others had no such costs. Thus, if the subjects agreed on a 50:50 split of the DM 30, the net earnings of the subject with a bargaining cost was DM 3, while the subject without bargaining cost had net earnings of DM 15. Bargaining costs were randomly assigned to the subjects and the subjects were not allowed to directly inform the other party about their bargaining cost situation. However, in principle, bargaining cost could be involuntarily revealed by the bargaining tactics of the subjects or by nonverbal cues. After agreement had been reached, third parties, who observed the face-to-face bargaining between the bargaining parties, had to make guesses about the cost situation. In addition, each bargaining party had to make guesses about the opponent's cost situation. Ockenfels and Selten find, however, no evidence favoring the hypothesis of involuntary truth-signaling. Neither the bargaining parties themselves nor the third parties are able to guess the bargainers' cost better than chance.

These arguments, however, do not end the debate about the green beard theory of human cooperation. If it could be shown that humans are in fact capable of predicting whether they face an altruist or a selfish individual, we would have to explain why evolution made this possible. In the presence of such evidence we would have to search for evolutionary explanations for the apparent existence of green beards among humans. There is in fact interesting evidence indicating that subjects are actually capable of predicting other individuals' cooperation rate in Prisoners' Dilemma (PD) experiments better than chance if they are given the opportunity to communicate face-to-face and to make promises before the PD is played (Frank/Gilovich/Regan 1993; Brosig 2002). This evidence has been taken as support for the hypothesis that humans are able to distinguish true cooperators (i.e. altruists) from those who only pretend to cooperate (i.e., the non-altruists). If this interpretation were correct, the proponents of the green beard approach would have a strong argument in favor of the relevance of their theory.

There are, however, two objections which question this interpretation of the evidence. First, if subjects cannot make promises during the group discussion, they are unable to predict the opponent's behavior better than chance (Frank/Gilovich/Regan 1993). This fact even holds if the experimental subjects are acquaintances who have frequently interacted with each other in the past (Yamagichi/Kikuchi/Kosugi 1999). This result is quite remarkable because if subjects are acquaintances one would expect them to be able to detect the green beard if there is one. Thus, it may be the case that if subjects can make promises, their behavior is predictable better than chance simply because most of them keep their promises. Charness and Dufwenberg (2003) developed a proximate theory of guilt aversion in a recent paper. This theory predicts that guilt-averse subjects are far more likely to cooperate if they first have the opportunity to promise to cooperate in a PD. The psychological reason is that such subjects feel guilt if they break their promise.

Second, the predictability of the opponent's behavior may be due to the

existence of conditional cooperators and may have nothing to do with special human abilities to predict the other person's altruism. Moreover, as we will show below, the fact that players are on average able to predict the opponent's behavior in a PD better than chance may be evolutionarily irrelevant because the selfish players may nevertheless end up with a higher average payoff than the conditional cooperators.

Before we present this argument in detail, we have to clarify what we mean by conditional cooperation. Conditional cooperation means that a subject is willing to cooperate in the PD if the probability of the opponent's cooperation is sufficiently high. Figure 1–3 aids in illustrating some of the implications of conditional cooperation. The underlying situation in all three figures is the following: both players have an endowment of \$10 which they can keep or send to the other subject. If they send \$10 the other subject receives \$20, i.e., the experimenter doubles the amount sent. Thus if both keep the money, which is tantamount with mutual defection (DD), each subject earns \$10. If both send the money, which is tantamount with mutual cooperation (CC), each subject earns \$20. However, a subject is of course best off in monetary terms if she keeps the money while the opponent sends the money. The monetary payoff matrix of this game is presented in Figure 1. The first number in each cell indicates the money payoff of player 1, the second number indicates the money payoff of player 2. If subjects care only for the money, Figure 1 represents the relevant payoff matrix. Thus, these subjects are in a PD because unilateral defection is always better for them regardless of what the opponent does, but mutual cooperation makes both subjects better off relative to mutual defection.

However, if subjects' preferences are shaped by inequity aversion (Fehr/Schmidt 1999) or reciprocal fairness (Rabin 1993; Dufwenberg/Kirchsteiger 2004; Falk/Fischbacher, in press), Figure 1 does not represent the subjective payoffs of the subjects. Instead, Figure 2 or Figure 3 captures the relevant payoffs. In these figures, the first number in a cell represents the subjective payoff of player 1, the second number indicates the subjective payoff of player 2. One crucial feature of Figures 2 and 3 is that each player prefers the mutual cooperation (CC) outcome over the unilateral defection (DC) outcome. If CC occurs in Figure 2, for example, player 1 has a subjective payoff of 20 while in case of DC the subjective payoff for player 1 is only 8. This feature has important implications because defection in Figures 2 and 3 is no longer a dominant strategy that is better regardless of what the opponent does. Instead, if a player believes that the other player chooses C, it is in the best subjective interest of the player to also choose C. Therefore CC is an equilibrium. However, DD is also an equilibrium because if a player expects that the opponent defects it is in the best interest of the player to also defect. In this sense, the payoff matrices of Figures 2 and 3 capture the essence of conditionally cooperative preferences. The fact that fairly strong evidence indicates that a considerable share of the subjects exhibits preferences for conditional cooperation (Hayashi et al. 1999; Kyonari/Tamida/Yamagichi 2000; Fischbacher/Fehr/Gächter 2001) is crucial for our purposes.

		Player 2	
		Cooperate (C)	Defect(D)
Player 1	Cooperate (C)	20, 20	0, 30
	Defect (D)	30, 0	10, 10

Figure 1: Monetary payoff in the prisoners' dilemma

		Player 2	
		Cooperate (C)	Defect(D)
Player 1	Cooperate (C)	20, 20	-5, 8
	Defect (D)	8, -5	10, 10

Figure 2: Subjective payoffs of exploitation-averse conditional cooperators in the Prisoners' Dilemma

		Player 2	
		Cooperate (C)	Defect(D)
Player 1	Cooperate (C)	20, 20	-15, 8
	Defect (D)	8, -15	10, 10

Figure 3: Subjective payoffs of strongly exploitation-averse conditional cooperators in the Prisoners' Dilemma

The second crucial feature of Figure 2 and 3 is that subjects are exploitation-averse: the subjective payoff if a player cooperates and the opponent defects is -5 in Figure 2 and -15 in Figure 3 indicating that unilateral cooperation is an aversive experience that goes beyond the mere loss of money. The difference between Figure 2 and 3 depicts the degree of exploitation aversion. Subjects are weakly exploitation-averse in Figure 2 while they are strongly exploitation-averse in Figure 3. The degree of exploitation aversion is important because a subject who is strongly exploitation-averse requires a high probability of opponent cooperation before the subject is willing to cooperate, while a weakly exploitation-averse subject is willing to cooperate at a lower level of opponent's cooperation probability.

The third crucial feature in Figure 2 and Figure 3 is that subjects prefer the mutual defection (DD) outcome over the unilateral defection (DC) outcome. Thus, if player 1 knows that he is going to defect (because, say, the probability that the other player cooperates is not perceived to be sufficiently high) he does not want the other player to cooperate. It can be shown that players who are strongly inequity-averse exhibit such preferences. Guilt aversion also implies such preferences because if the DC outcome occurs the player feels guilt, a psychologically aversive experience, while no guilt occurs in the DD outcome.

.

3. The Evolutionary Irrelevance of the Predictability of Other Players' Choices

In this section we present a model that captures the essence of a PD experiment in which subjects have the opportunity to make promises before they choose their actions. We will show that subjects are indeed able to predict the choices of their opponent better than chance, but that this does not help the cooperative subjects earn higher material payoffs than do the selfish subjects. In fact, the selfish subjects reap the highest payoff in the equilibrium of this game. Our argument is based on the following two-stage experiment. In stage one, each player simultaneously sends a signal c or a signal d. The signal c represents the non-binding promise to cooperate in the subsequent PD, the signal d indicates the non-binding promise to defect. The promises are non-binding in the sense that the players are free to disregard their previous promises when they subsequently play the PD. It is important to emphasize that what we call a PD here is a game whose *material* payoff structure is given by PD payoffs, such as the material payoffs in Figure 1. At the end of stage 1, both players observe the opponent's signal. In stage two, the subjects play the PD, i.e. they simultaneously choose the action C or D. The signaling stage can be interpreted as a stylized representation of communication opportunities that are made available to the subjects before they play the PD. At the end of this stage, each player has formed an opinion about the other player's intention in the subsequent PD.

We assume that there are three types of players: selfish players, weakly exploitation-averse conditional cooperators, and strongly exploitation-averse conditional cooperators. The percentage of conditional cooperators (weakly plus strongly exploitation-averse) is given by σ; hence the percentage of selfish players is given by 1-σ. Among the players who exhibit conditionally cooperative preferences, there is a share μ of weakly exploitation-averse players and a share 1-μ of strongly exploitation-averse players. Thus, $\mu\sigma$ percent of the subjects are weakly exploitation-averse and (1-μ)σ percent are strongly exploitation-averse. The game is played once by two players who are randomly selected from the population of players. The players are assumed to know σ and μ. We would like to emphasize, however, that the basic argument (i.e. that selfish players do better than conditional cooperators) does not hinge on these knowledge assumptions nor does it depend on the rationality assumptions we make for this game.

We assume that the weakly exploitation-averse subjects are willing to take the risk of cooperation once they meet the signal combination cc while the strongly exploitation-averse subjects, who value the CD outcome very negatively, are never willing to take the risk of cooperation because they are afraid the selfish players will exploit them. Under these assumptions we prove the following proposition in the appendix: There is a Bayesian Nash equilibrium in which

- The selfish player and the weakly exploitation-averse conditional cooperator send the signal c. The strongly exploitation-averse conditional cooperator sends the signal d.

- The selfish player and the strongly exploitation-averse conditional cooperator defect regardless of the signal combination at the end of stage 1.

- The weakly exploitation-averse conditional cooperator cooperates if the signal combination cc occurs. Otherwise he defects.

Only the weakly exploitation-averse conditional cooperator chooses C if the signal combination cc prevails in this equilibrium because he is willing to take the risk of being paired with a selfish player who always defects. The selfish player cheats his opponent because he announces the signal c while all the while intending to defect. This is the reason why the signal c is not fully reliable. Therefore, the strongly exploitation-averse subject will never cooperate, even if the opponent signaled *c*, because this subject is afraid of facing a selfish subject who never cooperates in the one-shot PD. Moreover, recall from Figure 3 that these subjects prefer the mutual defection outcome DD over the outcome DC in which they defect and the opponent cooperates. Therefore, these subjects have an interest in signaling *d* at the signaling stage to inform the opponent that they will defect. If, they were instead to signal *c,* they might induce their opponent to choose C, if the opponent happens to be a weakly exploitation-averse subject.

A decisive feature of the equilibrium above is that the strongly exploitation-averse subjects can credibly tell the truth with their signal because they prefer DD over DC. Given their preferences, it is in their interest to tell the truth! Thus all the other subjects have reason to believe that a subject who signals *d* will in fact play D. As a consequence, a subject who signals *c* can only be a selfish player or a weakly exploitation-averse player. This means that the probability of facing a weakly exploitation-averse player, after observing the *cc* signal combination is given by $\mu\sigma/(1\text{-}\sigma+\mu\sigma)$ and the probability of facing a selfish player is given by $(1\text{-}\sigma)/(1\text{-}\sigma+\mu\sigma)$. Rational players will thus believe that after the signal cc the actions C and D are played according to these probabilities. We can now show that players who have the beliefs that correspond to the equilibrium mentioned above and who play this equilibrium predict the opponent's action better than chance. We show this in Figure 4 which presents the correlation ρ between a players' beliefs about the opponent's action and the actual action of the opponent for different assumptions about σ and μ. If this correlation is positive, the players in the PD are able to predict the opponent's action better than chance. We compute ρ by assuming that the players implement a random mechanism after the signal combination cc which generates with probability $\mu\sigma/(1\text{-}\sigma+\mu\sigma)$ the prediction that the opponent cooperates and with probability $(1\text{-}\sigma)/(1\text{-}\sigma+\mu\sigma)$ a prediction that the opponent defects. Note that this prediction is based on the observation of the signals (or the inferences made) at the signaling stage. The players predict, in particular, that if the opponent signals *d* he will in fact play D.

Figure 4 shows that if the share of conditional cooperators is zero, ρ is also

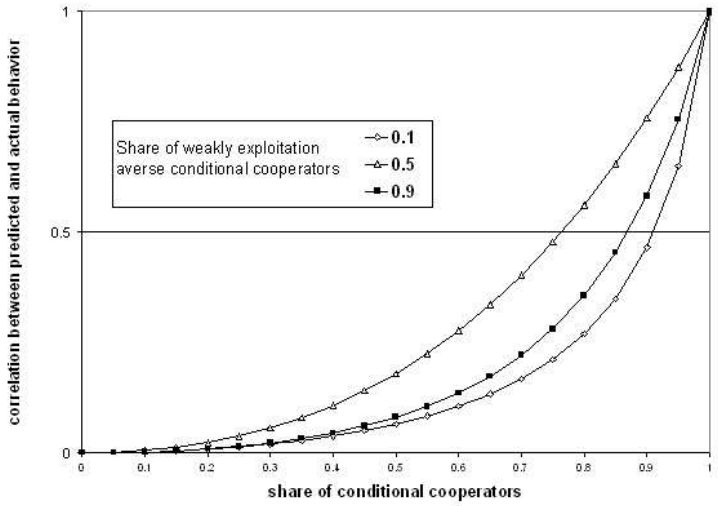

Figure 4: Correlation between predicted and actual behavior of the opponent

zero. However, the correlation becomes positive for positive values of σ and if σ increases, ρ also increases. If σ equals one, so does ρ, regardless of how many players are strongly exploitation-averse. Thus, Figure 4 indicates that the signaling stage enables the players to predict better than chance. Intuitively, the reason for this ability to predict comes from the credibility of the signal of the strongly exploitation-averse players because as shown above, it is in the interest of these players that their opponents defect because they don't want to mislead their opponents.

What are the material payoffs of the players in the Bayesian Nash equilibrium? To represent these payoffs, let Π_{CC} be the material payoff of mutual cooperation, Π_{DD} denotes the material payoff of mutual defection, Π_{CD} denotes the material payoff if a player cooperates and the opponent defects and Π_{DC} is the material payoff if a player defects and the opponent cooperates. Note that the PD payoffs are given by the inequalities $\Pi_{DC} > \Pi_{CC} > \Pi_{DD} > \Pi_{CD}$. With this notation the expected material payoff of the selfish player is given by

$$\mu\sigma\Pi_{DC} + (1-\sigma)\Pi_{DD} + (1-\mu)\sigma\Pi_{DD}, \tag{1}$$

because the selfish player meets a weakly exploitation-averse player (who cooperates) with probability $\mu\sigma$, meets a selfish player (who defects) with probability (1-σ) and meets a strongly exploitation-averse subject (who defects) with probability (1-μ)σ.

The expected material payoff of the weakly exploitation-averse player is given by

$$\mu\sigma\Pi_{CC} + (1-\sigma)\Pi_{CD} + (1-\mu)\sigma\Pi_{DD} \tag{2}$$

The expected material payoff of the strongly exploitation-averse player is given by

$$\mu\sigma\Pi_{DD} + (1-\sigma)\Pi_{DD} + (1-\mu)\sigma\Pi_{DD} \tag{3}$$

It is not difficult to see that the payoff structure of the material PD ($\Pi_{DC} > \Pi_{CC} > \Pi_{DD} > \Pi_{CD}$) implies that the selfish player always has a higher expected payoff than either of the conditional cooperators as long as $\mu\sigma$ is positive. The expected payoff difference between a selfish and a weakly exploitation-averse subject is given by $\mu\sigma(\Pi_{DC} - \Pi_{CC}) + (1-\sigma)(\Pi_{DD} - \Pi_{CD})$ which is positive because both terms in parentheses are positive. The difference in expected payoff between a selfish subject and a strongly exploitation-averse subject is also positive because the selfish subject exploits the weakly exploitation-averse subject (with probability $\mu\sigma$) while the strongly exploitation-averse subject always earns Π_{DD} and never exploits anybody. Therefore, although the existence of strongly exploitation-averse conditional cooperators enables the subjects to predict better than chance, the existence of weakly exploitation-averse players also permits the selfish players to earn more than both types of conditional cooperators. The mere fact that players are able to predict better than chance is of no help for the evolution of cooperation.

4. Summary and Conclusions

We have argued that the green beard approach does not provide a satisfactory solution for the evolution of human cooperation because it rests on problematic assumptions about the absence of mutants that mimic the observable phenotypic features of cooperators (or altruistic punishers) and it has little empirical support. It remains to be seen whether it is possible to construct rigorous individual selection theories for the evolution of human cooperation that are plausible alternatives to the group selection approaches mentioned in the beginning. It is important to emphasize in this context that human cooperation occurs in relatively large groups, making the application of intuitions that work in two-person interactions to the case of public goods problematic (i.e., multilateral interaction). Reciprocal altruism, which works fine in the case of bilateral interactions has, for example, little force in the case of multilateral, interactions (Boyd/Richerson 1988). Thus, it seems that approaches based on some kind of reputation formation mechanism – costly signaling (Gintis/Smith/Bowles 2001) or indirect reciprocity (Milinski/Semmann/Krambeck 2002; Panchanathan/Boyd 2004) – are the most promising candidates. However, we should not forget that reputation games typically have multiple within-group equilibria and it is difficult

to argue in the absence of some kind of selection between groups that those within-group equilibria will be selected which involve high levels of cooperation.

Bibliography

Amann, E./D.-L. Yang (1998), Sophistication and the Persistence of Cooperation, in: *Journal of Economic Behavior and Organization 37*, 91–105

Boyd, R./P. J. Richerson (1988), The Evolution of Reciprocity in Sizable Groups, in: *Journal of Theoretical Biology 132*, 337–356

— /H. Gintis/S. Bowles/P. J. Richerson (2003), The Evolution of Altruistic Punishment, in: *PNAS 100*, 3531–3535

Brosig, J. (2002), Identifying Cooperative Behaviour: Some Experimental Results in a Prisoner's Dilemma Game, in: *Journal of Economic Behavior & Organization 47*, 275–290

Charness G/M. Dufwenberg (2003), *Promises and Partnership.* Discussion Paper, University of Arizona

Dawes, R. M. (1980), Social Dilemmas, in: *Annual Review of Psychology 31*, 169–193

Dufwenberg, M./G. Kirchsteiger (2004), A Theory of Sequential Reciprocity, in: *Games and Economic Behavior 47*, 268–298

Ekman, P./M. O'Sullivan (1991), Who Can Catch a Liar? In: *American Psychologist 49*, 913–920

Falk, A./U. Fischbacher (in press), A Theory of Reciprocity, in: *Games and Economic Behavior*

Fehr, E./K. M. Schmidt (1999), A Theory of Fairness, Competition, and Cooperation, in: *Quarterly Journal of Economics 114*, 817–868

— /U. Fischbacher/S. Gächter (2002), Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms, in: *Human Nature 13*, 1–25

— /S. Gächter (2002), Altruistic Punishment in Humans, in: *Nature 415*, 137–140

Fischbacher, U./S. Gächter/E. Fehr (2001), Are People Conditionally Cooperative? Evidence from a Public Goods Experiment, in: *Economic Letters 71*, 197–404

Frank, R. (1987), If Homo Economicus Could Choose his own Utility Function, Would he Want one with a Conscience?, in: *American Economic Review 77*, 593–604

— (1988), *Passions within Reason. The Strategic Role of the Emotions*, New York

Frank, R. H./T. Gilovich/D. T. Regan (1993), The Evolution of One-Shot Cooperation: An Experiment, in: *Ethology and Sociobiology 14*, 247–256

Gintis, H. (2000), Strong Reciprocity and Human Sociality, in: *Journal of Theoretical Biology 206*, 169–179

— /E. A. Smith/S. Bowles (2001), Costly Signaling and Cooperation, in: *Journal of Theoretical Biology 213*, 103–119

— /S. Bowles/R. Boyd/E. Fehr (2003), Explaining Altruistic Behavior in Humans, in: *Evolution and Human Behavior 24*, 153–172

Güth, W./R. Schmittberger/B. Schwarze (1982), An Experimental Analysis of Ultimatum Bargaining, in: *Journal of Economic Behavior & Organization 3*, 367–388

Hayashi, N./E. Ostrom/J. Walker/T. Yamagishi (1999), Reciprocity, Trust, and the Sense of Control—A Cross-Societal Study, in: *Rationality and Society 11*, 27–46

Henrich, J./R. Boyd (2001), Why People Punish Defectors—Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas, in: *Journal of Theoretical Biology 208*, 79–89

Ledyard, J. (1995), Public Goods: A Survey of Experimental Research, in: J. Kagel/A. Roth (eds.) *Handbook of Experimental Economics*, Princeton, 111-194

Kiyonari, T./S. Tanida/T. Yamagishi (2000), Social Exchange and Reciprocity: Confusion or a Heuristic, in: *Evolution and Human Behavior 21*, 411–427

Milinski, M./D. Semmann/H. J. Krambeck (2002), Reputation Helps Solve the 'Tragedy of the Commons', in: *Nature 415*, 424–426

Ockenfels, A./R. Selten (2000), An Experiment on the Hypothesis of Involuntary Truth-Signalling in Bargaining, in: *Games and Economic Behavior 33*, 90–116

Panchanathan, K./R. Boyd (2004), Indirect Reciprocity can Stabilize Cooperation without the Second-Order Free Rider Problem, in: *Nature 432*, 499–501

Rabin, M. (1993), Incorporating Fairness into Game Theory and Economics, in: *American Economic Review 83*, 1281–1302

Robson, A. (1990), Efficiency in Evolutionary Games: Darwin, Nash and Secret Handshake, in: *Journal of Theoretical Biology 144*, 379–396

Yamagishi, T./M. Kikuchi/M. Kosugi (1999), Trust, Gullibility, and Social Intelligence, in: *Asian Journal of Social Psychology 2*, 145–161

Appendix

In this appendix we prove that the strategies described in Section 3 of the paper constitute a Bayesian Nash equilibrium. For convenience we repeat the strategies here:

- The selfish player and the weakly exploitation-averse conditional cooperator send the signal c. The strongly exploitation-averse conditional cooperator sends the signal d.
- The selfish player and the strongly exploitation-averse conditional cooperator defect regardless of the signal combination that is observed at the end of stage 1.
- The weakly exploitation-averse conditional cooperator cooperates if the signal combination cc occurs. Otherwise he defects.

Recall that we assumed that the weakly exploitation-averse conditional cooperators are willing to take the risk of being exploited once they observed the signal combination cc. This assumption can be expressed more formally if we denote U_{XY} as the utility of a player if the player chooses $X \in C, D$ and the opponent chooses $Y \in C, D$:

$$\frac{\mu\sigma}{1-\sigma+\mu\sigma}U_{CC} + \frac{1-\sigma}{1-\sigma+\mu\sigma}U_{CD} > U_{DD} \tag{1}$$

The left hand side of this inequality is the expected utility of cooperation after the signal combination cc has been observed, which is given by the probability of facing a weakly exploitation-averse conditional cooperator, $\mu\sigma/(1\text{-}\sigma+\mu\sigma)$, times the utility from CC, plus the probability of facing a selfish subject, $(1\text{-}\sigma)/(1\text{-}\sigma+\mu\sigma)$, times the utility of the CD outcome. The right hand side of this inequality is the utility of mutual defection. It is assumed that the weakly exploitation-averse conditional cooperator's preferences obey this inequality. We also assume that the reverse inequality holds if the player is a strongly exploitation-averse conditional cooperator. We call this inequality (A2).

Next, we show that that the players' behavior in PD stage is optimal. It is obvious that the selfish players defect. Since the strongly exploitation-averse conditional cooperator sends defect, he 'knows' that in equilibrium, the opponent defects irrespective of the opponent's type. Therefore, it is also optimal for the strongly exploitation-adverse player to defect regardless of the signal combination observed in stage 1. The weakly exploitation-averse conditional cooperator also defects if her partner's signal is *d*. If the opponent sent the signal c, the probability that this is a weakly exploitation-averse conditional cooperator equals $\mu\sigma/(1\text{-}\sigma+\mu\sigma)$. Therefore, inequality (A1) guarantees that it is in a weakly exploitation-averse player's interest to cooperate if he receives the signal *c*.

Next we show the individual optimality of the signals. Because sending the signal *d* never induces the opponent to cooperate, the selfish players will send the signal *c*. If a strongly exploitation-averse conditional cooperator would send *c* and receive the signal *c*, she would nevertheless defect because her preferences obey inequality (A2). Thus, by sending *c*, she risks exploiting the opponent (if he is a weakly exploitation-averse player) implying a lower expected utility than U_{DD}. Therefore, she prefers to send *d*. Finally, the weakly exploitation-averse cooperator has an incentive to send the signal *c* because of inequality (A1).

q.e.d.