The Viability of Cooperation Based on Interpersonal Commitment

Back, István; Flache, Andreas

Published in:
Journal of Artificial Societies and Social Simulation

Publication date:
2006

István Back and Andreas Flache (2006)

The Viability of Cooperation Based on Interpersonal Commitment

Journal of Artificial Societies and Social Simulation vol. 9, no. 1
<http://jasss.soc.surrey.ac.uk/9/1/12.html>

For information about citing this article, click here

Received: 20-Jul-2005 Accepted: 25-Sep-2005 Published: 31-Jan-2006

Abstract

A prominent explanation of cooperation in repeated exchange is reciprocity (e.g. Axelrod 1984). However, empirical studies indicate that exchange partners are often much less intent on keeping the books balanced than Axelrod suggested. In particular, there is evidence for commitment behavior, indicating that people tend to build long-term cooperative relationships characterised by largely unconditional cooperation, and are inclined to hold on to them even when this appears to contradict self-interest.

Using an agent-based computational model, we examine whether in a competitive environment commitment can be a more successful strategy than reciprocity. We move beyond previous computational models by proposing a method that allows to systematically explore an infinite space of possible exchange strategies. We use this method to carry out two sets of simulation experiments designed to assess the viability of commitment against a large set of potential competitors. In the first experiment, we find that although unconditional cooperation makes strategies vulnerable to exploitation, a strategy of commitment benefits more from being more unconditionally cooperative. The second experiment shows that tolerance improves the performance of reciprocity strategies but does not make them more successful than commitment.

To explicate the underlying mechanism, we also study the spontaneous formation of exchange network structures in the simulated populations. It turns out that commitment strategies benefit from efficient networking: they spontaneously create a structure of exchange relations that ensures efficient division of labor. The problem with stricter reciprocity strategies is that they tend to spread interaction requests randomly across the population, to keep relations in balance. During times of great scarcity of exchange partners this structure is inefficient because it generates overlapping personal networks so that often too many people try to interact with the same partner at the same time.

Keywords:
Interpersonal Commitment, Fairness, Reciprocity, Agent-Based Simulation, Help Exchange, Evolution

Introduction

1.1

The most prominent explanation of endogenous cooperation in durable relationships is reciprocity under a sufficiently long ``shadow of the future'' (Axelrod, 1984; Friedman, 1971). In this view, actors engage in costly cooperation because they expect future reciprocation of their investment or because they feel threatened by future sanctions for non-cooperation (Falk et al., 2001; Fehr and Gächter, 2002; Fehr and Schmidt, 1999). Roughly, these analyses show that even in a competitive environment with changing exchange partners, strategies that reciprocate cooperation with cooperation and defection with defection, such as the celebrated ``Tit-for-Tat'', are by far more successful than strategies that aim to exploit their opponents. Evolutionary game theory has demonstrated that if exchange relations persist long enough, cheaters are outperformed by reciprocators. This is because reciprocators benefit from ongoing mutually cooperative exchanges, while cheaters gain at best a short term advantage at the outset of the exchange. This, however, cannot offset the long term losses caused by the early disruption of the exchange relationship.

1.2

This reciprocity explanation of cooperation has been suggested to apply to a number of domains ranging from business ties between organizations to interpersonal relationships. However, recent empirical studies of cooperative behavior, in particular in interpersonal relationships, indicate that often reciprocity may be much less strict and actors much less intent on keeping the books balanced than the original reciprocity argument suggests. A short excerpt from Nesse (2001) offers good examples:

> Perhaps the strongest evidence that friendships are based on commitment and not reciprocity is the revulsion people feel on discovering that an apparent friend is calculating the benefits of acting in one way or another. People intuitively recognize that such calculators are not friends at all, but exchangers of favors at best, and devious exploiters at worst. Abundant evidence confirms this observation. Mills has shown that when friends engage in prompt reciprocation, this does not strengthen but rather weakens the relationship (Mills and Clark, 1982). Similarly, favors between friends do not create obligations for reciprocation because friends are expected to help each other for emotional, not instrumental reasons (Mills and Clark, 1994). Other researchers have found that people comply more with a request from a friend than from a stranger, but doing a favor prior to the request increases cooperation more in a stranger than a friend (Boster et al., 1995).

Moreover, there is solid empirical evidence indicating that people have a tendency to build long-term cooperative relationships based on largely unconditional cooperation, and are inclined to hold on to them even in situations where this does not appear to be in line with their narrow self-interest (see e.g. Wieselquist et al., 1999). Experiments with exchange situations (Lawler and Yoon, 1996,1993; Kollock, 1994) point to ongoing exchanges with the same partner even if more valuable (or less costly) alternatives are available. This commitment also implies forgiveness and gift-giving without any explicit demand for reciprocation (Lawler, 2001; Lawler and Yoon, 1993). One example is that people help friends and acquaintances in trouble, apparently without calculating present costs and future benefits. Another, extreme example is the battered woman who stays with her husband (Rusbult et al., 1998; Rusbult and Martz, 1995).

1.3

Since the seminal work of Axelrod (1984), a range of studies have used evolutionary game theory to refine the strategy of strict reciprocity and adapt it to empirical criticism. One line of work focused on the advantages of ``relaxed accounting'' in noisy environments (e.g. Kollock, 1993; Nowak and Sigmund, 1993; Wu and Axelrod, 1995). Broadly, these experiments confirmed the hypothesis that uncertainty favors ``tolerant'' or ``relaxed'' conditionally cooperative strategies that do not always retaliate after defection of an opponent. Kollock (1993), for example, found that in noisy environments (with mistakes and miscues), strict reciprocity is prone to needless recrimination that can be avoided by looser accounting systems. However, these studies cannot address the empirical phenomenon of commitment to long term exchange partners, simply because they apply a repeated game framework in which there is no possibility to exit from an ongoing exchange in order to seek a new partner.

1.4

A number of authors have explored variations of Tit-for-Tat that combine looser accounting under uncertainty with selective

partner choice. Computational analyses of exit effects (Schüssler and Sandten, 2000; Vanberg and Congleton, 1992; Schüssler, 1989) put the role of the shadow of the future for emergent cooperation into perspective. The route to emergent cooperation that these studies uncover is commitment of cooperators to cooperators, with the consequence of exclusion of defectors from relationships with cooperative partners. This is based on the principle ``be cooperative but abandon anyone who defects.'' When enough members of a population adopt this strategy, cooperative players stay in stable relationships, leaving defectors with no one but other defectors to interact with. As a consequence, defectors perform poorly and conditional cooperation thrives even under anonymity conditions where unfriendly players can hide in a ``sea of anonymous others'' (Axelrod, 1984, 100) after they ``hit and run''. Considering more complex agent architectures, Schüssler and Sandten (2000) show that strategies that are to some degree exploiters may survive under evolutionary pressure but even then the most successful strategies will have the property of staying with a cooperative partner who turns out to be difficult to exploit. Other computational studies that include partner selection and arrive at similar conclusions are, for example, Yamagishi et al. (1994), or Hegselmann (1996) (cf. Flache and Hegselmann, 1999).

1.5

While previous work using evolutionary game theory could demonstrate the viability of relaxed accounting and commitment under certain conditions, it is doubtful whether this suffices to explain how humans may have acquired the deeply rooted emotions and behaviors related to interpersonal commitment that have been empirically observed.

This is why de Vos and collaborators (de Vos and Zeggelink, 1997; de Vos et al., 2001; Zeggelink et al., 2000) extended theoretical models with assumptions from evolutionary psychology (Cosmides, 1989; Cosmides and Tooby, 1993). According to evolutionary psychologists, the way our mind functions today is the result of an extremely long evolutionary process during which our ancestors were subject to a relatively stable (social) environment. Individual preferences for various outcomes in typical social dilemmas stabilized in this ancestral environment and still influence the way we decide and behave in similar dilemma situations today.

1.6

To model a stylized ancestral environment, de Vos and collaborators designed a help exchange game in which members of a relatively small group need the help of others to survive a situation of distress from time to time. More precisely, in their model agents come into distress at random points in time and then ask other members of the group for help. They compared two major contestants in their simulations of the evolution of exchange strategies, a strategy they called ``keeping books balanced'' (KBB) and a strategy called ``commitment''. KBB corresponds to a strategy of strict reciprocity that is willing to help another actor but only as long as the favor is returned by the recipient as soon as possible. Otherwise, KBB will exit the relationship and seek new exchange partners. By contrast, commitment needs only a few successful initial help exchanges with a specific partner to become unconditionally cooperative to his partner further on. Broadly, de Vos and collaborators found that when both strategies need to compete against ``cheaters'' - i.e. actors who are unwilling to help but accept help from others - commitment is more viable than KBB under a large range of conditions. They conclude that in an environment where unpredictable hazards occur, KBB may be too quick to abandon exchange partners who get into trouble a second time before first reciprocating. As a consequence, a KBB player may often end up with no one willing to help it. A commitment player avoids this problem, because once commited to a cooperative partner it will not leave the partner in times of need and thus benefit from future help from this partner when itself comes into distress.

1.7

De Vos et al. tentatively conclude from their computational experiments that under conditions of the human ancestral environment, selection pressures may have shaped a tendency towards commitment and largely unconditional cooperation that contemporary humans may still have, even when the pressures that formed it are no longer present. However, it is clearly an important limitation of these studies that only three possible strategies, KBB, commitment and cheating, have been taken into account and confronted with each other in a tournament approach. As Binmore (1998) has argued forcefully, the outcome of computer tournaments and simulations of evolutionary dynamics strongly depends on the set of strategies that are initially present in a population. The small set of strategies used by de Vos and collaborators may hide two potentially severe problems for the viability of the strategy of commitment. The first problem is the unconditionality of the strategy's willingness to cooperate once it has been committed to a partner. This property obviously makes commitment highly vulnerable to exploitation by strategies who try to take advantage of its willingness to help. The second problem is that commitment may loose out in competition against more tolerant modifications of strict reciprocity. As the work by Kollock (1993) and others

suggests, such modifications may avoid the major weakness of strong reciprocity to disrupt potentially cooperative exchanges too readily when problems occur. At the same time, such strategies also are less exploitable than commitment, because they eventually avoid being exploited by a partner who steadfastly fails to reciprocate help.

1.8

To address whether and to what extent these two potential problems reduce the viability of commitment, we propose in this paper a method to considerably and systematically enlarge the strategy set used in the original analysis of the help exchange dilemma. The core idea is to represent strategies as a set of individual preference parameters, or traits with respect to possible exchange outcomes in a relationship. Agents in our model are boundedly and subjectively rational in the sense that they take decisions to cooperate, defect and change partners with the goal to maximize utility from their preferences. However, maximizing subjective utility based on individual preference values in our model does not necessarily lead agents to optimal exchange outcomes. We assume that individual preferences or strategies are subject to evolutionary pressure that selects for successful strategies based on the objective fitness consequences of the behavior resulting from the strategy. This approach is similar to the ``indirect evolutionary approach'' proposed by Güth and Kliemt (1998).

1.9

Our approach allows to systematically map a range of individual variation in decision making rules, e.g. variation in the extent of commitment or strictness of reciprocity. With this, we can carry out a stronger test of the viability of commitment than de Vos et al. (2001). We use our model to carry out two sets of simulation experiments designed to assess the viability of commitment in a larger set of potential competitors. For this, we take the original design of de Vos et al. as a starting point but systematically relax the assumption of unconditionality of cooperation in the first set of experiments. In the second set of experiments, we introduce and compare various degrees of relaxed accounting to reciprocity (``fairness'') strategies. In Section 2, we motivate and describe the model and our extensions. In Section 3, the computational experiments are reported. Section 4 contains conclusions and a discussion of our findings.

Model

2.1

Our model is based on a delayed exchange dilemma game, which is very similar to the one originally proposed by de Vos et al. (2001). The game is played by n agents in successive rounds. In the first round all agents are endowed with f_i fitness points. In the beginning of each round, Nature selects a number of agents with a given individually independent probability P_d who experience distress and thus become in need of help from other agents in order to preserve their fitness level. These agents who are struck by Nature are the initiators of interactions. They ask others for help which is either provided or not. Providing help costs f_h fitness points. Moreover, assuming that help giving is a time-consuming activity, each agent may only provide help once during one round; and only agents who are not distressed themselves may provide help. If a help request is turned down, the distressed agent may ask another agent for help but may not ask more than m agents altogether in the same round. If an agent does not manage to get help before the end of the round, it experiences f_d loss in fitness. If the fitness level of an agent falls below a critical threshold f_c, the agent ``dies'', i.e. it is eliminated from the agent society.

Modelling strategies

2.2

Agents in our delayed exchange dilemma face two different types of decision situations from time to time. If they are hit by distress, they have to select an interaction partner whom they believe most likely to be willing and able to help them. On the other hand, when they themselves are asked to provide help they have to decide whether to provide it and in case of multiple requests, whom to provide it to. Thus the mental model, or strategy, of an agent is represented as a combination of two substrategies: one for asking help and one for giving help.

2.3

In previous studies by de Vos and others, behavioral strategies of agents were defined in natural language in terms of a collection of condition-action rules (e.g. for agent a_i : *if* agent a_j helped me before when I asked *then* help him now) and then translated into a programming language. Even for simpler strategies several such decision rules had to be formulated, and this inherent arbitrariness limited the generalizability of the model.

Our most important addition to these models is that we integrate them into a utility-based framework and provide in this way an efficient method to cover a large range of different strategies. In our model, when an agent has to make a decision, it calculates utilities based on some or all of the information available to it without the ability to objectively assess the consequences of the decision on its overall fitness[3]. Moreover, we assume that actors are boundedly rational in the sense of being myopic, they evaluate the utility of an action only in terms of consequences in the very near future, i.e. the state of the world that obtains right after they have have taken the action. This exludes the strategic anticipation of future behavior of other agents. Since different agents calculate utility differently, there is variation in behavior. Some of the behaviors lead to better fitness consequences than others. In turn, more successful agents have better chances to stay in the game and to propagate their way of utility calculus to other agents, while unsuccessful ones disappear.

2.4

Recent advances in psychological research on interpersonal relationships point to the influence of subjective well-being experienced when making certain relationship-specific decisions (Karremans et al., 2003). Unlike many applications of evolutionary game theory, we define utility calculus such that agents derive an emotional utility from features of a relationship, in addition to materialistic costs and benefits of help exchanges. This emotional utility can be interpreted as feelings and emotions, such as togetherness, belonging, sense of safety, identity, pride etc., and the lack of it as loneliness, insecurity, shame etc. We concentrate our modeling efforts on describing and analyzing this additional utility as a function of the history of help exchanges in a relationship. One of our main goals is to determine whether utility calculus based on some form of commitment can lead to beneficial fitness consequences.

2.5

In our delayed exchange game, agents have a very focused set of information available about their physical and social environment. They are aware of the fact that they got into distress, they follow the rules of the game (e.g. ask for help when in distress), and they remember previous encounters with other agents. This means that they know who and how often helped or refused them and who was helped or refused by them in previous rounds. The implicit assumption we make is that information about interactions between third-party agents is either not (reliably) available to the focal agent or is simply not taken into account in decision making.

2.6

We restrict the information available to agents from their earlier interactions to the following situation-specific decision parameters of an agent a_i for each interaction partner a_j ($i \neq j$):

Definition 1 (Situation-specific decision parameters)

$$EH_{ij} = \text{number of times i helped j (}\mathbf{e}\text{go }\mathbf{h}\text{elped),}$$

$$ER_{ij} = \text{number of times i refused j (}\mathbf{e}\text{go }\mathbf{r}\text{efused),}$$

$$AH_{ij} = \text{number of times j helped i (}\mathbf{a}\text{lter }\mathbf{h}\text{elped),}$$

$$AR_{ij} = \text{number of times j refused i (}\mathbf{a}\text{lter }\mathbf{r}\text{efused)}$$

2.7

As we mentioned above, agents face two different decisions situations. Accordingly, we define two independently calculated

subjective utilities that agents use in these two decisions. The *utility of donating* that agent a_i gains from helping agent a_j is defined as a function of the situation-specific parameters:

$$U_{ij}^D = U_m^D + eh_i^D \cdot EH_{ij} + er_i^D \cdot ER_{ij} + ah_i^D \cdot AH_{ij} + ar_i^D \cdot AR_{ij},$$

where U_m^D expresses materialistic costs of the interaction; $eh_i^D, er_i^D, ah_i^D, ar_i^D$ are *agent-specific parameters* (or traits) for donation of agent a_i that determine the weight of the situation-specific parameters in the total utility. In the actual implementation, every time an agent has to make a decision, there is also a probability P_e that the agent will make a completely random decision. This random error models noise in communication, misperception of the situation or simply miscalculation of the utility by the agent. Taking this random error into account increases the robustness of our results to noise in general[4].

2.8

For simplicity, we define the utility as a linear combination of situation-specific parameters weighted by agent-specific parameters.

The *utility of seeking* is defined in the same way, the only difference is that agents may put different weights on the situation-specific decision parameters than in the utility of donation:

$$U_{ij}^S = U_m^S + eh_i^S \cdot EH_{ij} + er_i^S \cdot ER_{ij} + ah_i^S \cdot AH_{ij} + ar_i^S \cdot AR_{ij},$$

Before agents make a decision, be it help seeking or help giving, they calculate the corresponding one of these two utilities for each possible help donor or help seeker. In case of help giving, they choose a partner with the highest utility, if that utility is above an agent-specific threshold U_i^t . If the utility of all possible decisions falls below the threshold utility, no help is given to anyone. Otherwise, if there is more than one other agent with highest utility, the agent chooses randomly.

2.9

As an addition to this rule, if an agent a_i is asked to donate by another agent a_j with whom a_i had no prior interaction (therefore all situation-specific parameters are 0), a_i assumes that $AH_{ij} = 1$. In other words, agents behave *as if* a successful interaction has already taken place between them. Suspicious (non-nice) strategies can be defined by choosing the utility threshold (U^t) parameter so that without any prior interaction the utility of seeking or donation is lower than the threshold utility.

2.10

In the case of help seeking, agents also choose a partner with the highest utility but there is no threshold, i.e. agents in distress will always ask someone for help.

2.11

Using these rules, a strategy S in our model is described by the way utility is calculated. In other words, a strategy can be fully described by the two times four agent-specific parameters and the utility threshold. [5] By specifying *ranges* for agent-specific parameters we can easily define classes of strategies which correspond to basic personality types. For example, we classify a strategy S as belonging to the group of Commitment-type strategies, if the fact that previously help was received from a certain partner, increases the utility an agent derives from donating help to or seeking help from that partner:

Definition 2 (Commitment) $ah^D, ah^S > 0$ *and* $er, ar = 0$

This means that an agent a_i of the Commitment-type derives more utility from choosing an agent a_j as an interaction partner, the more times a_j has helped a_i in the past. This is true for choosing from both a group of help seekers and from possible help givers. This also means that the cooperativeness of a Commitment type is unaffected by the fact whether their help was previously refused by an interaction partner.

2.12

If the utility threshold U^t is zero or negative and $ah^D > U^t$, the strategy starts by cooperating, if able to.[6] Otherwise, it behaves as ``Suspicious Commitment'', i.e. starts with defecting but after some cooperative moves of alter, it becomes cooperative.

2.13

For simplicity, in the following we assume a utility threshold of zero, if not mentioned otherwise. In the remainder of section, we show how a range of further strategies can be defined with our method. We assume a utility threshold of zero, if not mentioned otherwise. The strategy type of Defection can be modelled with the assumption that it derives zero or negative utility from donation under all conditions:

Definition 3 (Defection) $eh^D, er^D, ah^D, ar^D \leqq 0$ *and* $min(eh^D, er^D, ah^D, ar^D) < 0$

If the utility threshold U^t is positive and $ah^D < U^t$ the strategy always starts by defecting. Otherwise, it only starts defecting after some initial rounds of cooperation.

2.14

In general, we say that a strategy is a cooperator if at least one of its donation parameters is positive. In all other cases the strategy is a variant of the Defection type. Such a subset of Defection is AllD which never helps others but when it is in need it randomly choses others to ask for help:

Definition 4 (AllD)
Donation: $eh^D, er^D, ah^D, ar^D < 0$

Seeking: $eh^S = er^S = ah^S = ar^S = 0$

A much discussed strategy (type), especially in the experimental economics literature (see e.g. Fehr and Schmidt, 1999; Fehr et al., 2002) is Fairness.[7] It is based on the observation that people may be willing to invest in cooperation initially but will require reciprocation of these investments before they are willing to cooperate further. On the other hand people following the fairness principle are also sensitive to becoming indebted, therefore they will be inclined to reciprocate if they are in debt. In other words, their most important aim is to have balanced relationships. Again, translating this strategy class into our framework is straightforward.

Definition 5 (Fairness)
Donation: $eh^D < 0, er^D > 0, ah^D > 0, ar^D < 0$

Seeking: $eh^S > 0, er^S < 0, ah^S < 0, ar^S > 0$

2.15

Agents belonging to the Fairness class deduce more negative utility from helping if they helped their partner in the past or if the partner refused them before, and will deduce more positive utility from helping if the partner helped them or if they refused to help the partner earlier. The twist here is that the ones that are most likely to be selected for giving help to are different from those that are most likely to be selected for asking. Note moreover, that in case of a ``Truly Fair'' strategy, we would make the additional assumption about absolute values of traits such that $|eh| = |ah|$ and $|er| = |ar|$.

2.16

Suppose, for example, that an agent a_i receives two help requests at the same time, one from a_j , whom a_i has helped twice before but from whom a_i received help already three times. The other help request comes from a partner a_k who helped a_i three times and received help three times. A truly fair-minded person should in this situation help a_j and not a_k , and this is exactly what follows from our implementation, because in this case $U_{ij}^D = U_{ik}^D - eh^D$ and $eh^D < 0$.

2.17

Without making the assumption about absolute values, however, we are able to examine a larger class of ``Fairness-type'' strategies, such as ``Tolerant Fairness'' which increases credits (ah, er) more than it increases debts (ar, eh).

2.18

Note that U^t must be negative or zero for Objective Fairness, otherwise it requires more cooperation from its partner than itself is willing to perform. Another way of relaxing the strictness of Objective Fairness is to decrease U^t , which allows an asymmetry in favor of alter, in the amount of required reciprocation.

2.19

For analysing the individual rationality of cooperation we also define a trigger strategy, Grim Trigger. This strategy is the strictest form of cooperation, in that it permanently retaliates after its partner or itself defected and never cooperates again.

Definition 6 (Grim Trigger)
Donation: $eh^D = ah^D = 0, er^D < 0, ar^D < 0$

Obviously, our approach allows to generate a much larger range of strategies than we discussed above. For our present analysis, it suffices to use these strategy templates but we will explore a larger variety of possible behavioral rules in future work.

Evolutionary dynamic

2.20

The heart of our model is an evolutionary dynamic that ensures the *selection* of objectively successful strategies (preferences). The dynamic we apply in our simulation is based on the replicator dynamics (Taylor and Jonker, 1978). Broadly, the replicator dynamics dictates that after a generation of genotypes (strategies) replicates itself, each different genotype will be represented in the next generation according to its relative success compared to other genotypes in the current generation. This way, infeasible or self-harming preferences gradually become less widespread in the population, and give way to more ``rational'' preferences (see also Güth and Kliemt, 1998).

2.21

To ensure that the size of the group remains constant throughout a simulation run, we apply the replicator dynamics in the following way. Whenever an agent dies, we create a new agent whose probability of belonging to a strategy S is equal to the proportion of collective fitness that is held by the group of agents belonging to S at the time of the new agent's birth.

2.22

The evolutionary dynamic of our model is a strong simplification of the actual genetic reproduction that could have taken place in human evolutionary history. One argument for this simplification is to avoid the unnecessary overparameterization of our model. The central assumption we make is that better exchange outcomes of a strategy type translate into better chances for the propagation of that strategy. To capture this, there is no need to include individual level variables such as average and maximum number of children, age at giving birth etc., which are actually irrelevant for answering our research questions. Thus the great advantage of the replicator dynamics for our purposes is that it keeps the model of reproduction on the macro level. This also means that we only model selection of strategies but not mutation (see more under Discussion and Conclusions).

2.23

With our explicit model of evolution we improve upon previous work of de Vos et al. (2001) in a number of ways. In their study they did not explicitly model a replication dynamic but instead linked independent tournaments to each other in order to map evolutionary trajectories. More precisely, the authors assumed that in a sequence of evolution the final average distribution of strategies at the end of one generation taken accross a series of replications of that generation would also be the initial distribution in all replications in the next generation. This reduces repeatedly the distribution of individual populations to its average trajectory, which may entail a biased picture of the eventual distribution that arises. For example, unlike de Vos et al. (2001), we consider in our analysis also those simulation runs in which the entire population becomes extinct before a generation ends. These runs were originally disregarded by De Vos et al. This may have biased their results towards an overestimation of survival chances of Commitment because only replications in which commitment survived could have reached the end of a generation. Moreover, unlike previous work our model does not suffer from the specification of a ``cut-off'' parameter, i.e. there is no fixed number of rounds after which we stop our simulations.[8] In this way, we assure that the evolutionary dynamic reaches an equilibrium state where the population is homogenous, and avoid biasing our results towards strategies that may be only initially successful.

Results

3.1

De Vos et al. (2001) examined two cooperative strategies, Commitment and Keeping Books Balanced (KBB), both playing against defectors. They showed that Commitment, which is largely unconditionally cooperative to those previous interaction partners who gave help at least once, had better evolutionary success than the strictly reciprocal KBB under a large range of conditions. They tentatively interpret this result as evidence for the advantages of being unconditionally cooperative in an environment with scarce and uncertain opportunities to receive help. We argue that another conclusion may also be possible. It is plausible that being so unconditionally cooperative still makes Commitment more exploitable in comparison with conditional cooperators. The relative success of Commitment in comparison with KBB may rather be a result of KBB's disadvantegeous feature to disrupt relationships too readily when some mishap occurs. In order to test this possibility, we first conducted a simulation experiment to assess to what extent it makes a difference in Commitment's success against defectors, when various degrees of unconditionality are compared. We did this by comparing four different types of Commitment each playing against Defection.

3.2

We then turned to the possibility that a more tolerant fairness-based strategy may be more successful against defectors than Commitment or strict fairness (KBB), in an uncertain environment. To test this possibility, we compared Commitment with more tolerant versions of Fairness, and we also compared fairness strategies that vary in their degree of tolerance to each other.

3.3

Finally, while De Vos et al. measured and compared the individual success of cooperative strategies playing against defectors, they did not consider the possibility of an actual *evolutionary invasion* against Commitment by a conditional cooperator which is less vulnerable to exploitation by smart cheaters. We also provide results of this type below.

Simulation setup

3.4

Our goal with the simulation experiments was to compare different *cooperative* strategies with each other in terms of viability when there are initially some defectors in society. The most important indicator of a cooperative strategy's viability was its success in resisting this invasion of defectors under evolutionary pressures of selection and reproduction. More precisely, within each simulation run, we started out with a group of multiple strategies. We allowed this mixed group to play the game for an extended period under an evolutionary dynamic, until only one strategy was present in the population. Within one experiment, we independently repeated such simulation runs from their initial state n times[9], until standard errors of measured variables became sufficiently small in order to be meaningfully interpreted.

3.5

At first, we kept all environmental and model parameters constant and varied only strategy parameters from experiment to experiment. Even using our compact way of representing strategies (see section 2.1), we need to define strategies in a 9-dimensional space (using two times four weigth parameters and a threshold parameter). Assuming that parameters and the threshold can only take 5 possible values (i.e. -2, -1, 0, 1, 2), we are left with a strategy space of $5^9 = 1953125$ individual strategies. Fortunately, vast parts of this strategy-space yield similar behavior and thus can be classified under common concepts such as Defection, Fairness, Commitment, Trigger etc. (see strategy types described above). For example, multiplying all traits and the utility threshold of a strategy S with a positive value will yield a strategy S' that behaves identically to S . More generally, as long as a transformation on the trait parameters does not shift the level of utility below or above the threshold for any given situation-specific parameter and does not modify the ordering of alters, it has no effect on behavior. Therefore, in the analysis that follows we will not vary absolute values of single traits, only traits in proportion to each other.[10]

3.6

To assess the robustness of results derived from the simulation experiments we conducted exhaustive sensitivity tests for all sensibly variable parameters. We report interesting deviations from typical results in section 3.4 below. For a list of all parameters see Appendix B.

Initial parameters

3.7

To determine interesting initial parameters for the simulation experiments and to reduce the parameter space that must be explored, we conducted a game theoretical analysis of a simplified version of the dilemma. Our goal was to identify the set of conditions that makes the choice for agents between purposeful defection and (conditional) cooperation as difficult as possible. If cooperation places an excessively high burden on agents, or conversely, if cooperation entails no real sacrifices, the model would hardly yield any interesting insights.

3.8

To approximate the conditions under which cooperation is rational at all in the delayed exchange dilemma, we calculated expected payoffs in a simplified version of the game using *trigger strategies*. A trigger strategy behaves so that as soon as its interaction partner or itself defects, it falls back into a period of unconditional defection. The most severe version of trigger strategies is Grim Trigger which never switches back to cooperation after its partner or itself defected. Even after its own unintended defection (i.e. due to being unable to help), Grim Trigger applies the most severe punishment possible in the game, permanent retaliation. If the sanction imposed by Grim Trigger cannot deter a rational player from unilateral defection, then no cooperative strategy can do so. As a consequence, there exists no Nash equilibrium - that is: a rational outcome - in which both players choose a conditionally cooperative strategy (see Abreu, 1988). The simplifications we make for the sake of the formal analysis are that we reduce the group size to two and that we omit the evolutionary dynamics and the possibility of death due to low fitness.

3.9

After solving this simplified dilemma situation (see Appendix A), we get a condition for the rationality of cooperation in the form of a relationship between the probability of distress, the cost for helping and the cost for not getting help:

$$f_h < f_d(1 - P_d)$$

We used this result to adjust the most important initial parameters of the model. In other words, we always varied the probability of distress, the cost of help and the cost of not getting help in a way that the above inequality remained true. The actual parameters that we used to draw figures below are: $P_d = 0.2, f_h = 1, f_d = 20, f_i = 100, f_c = 0, N = 25, P_e = 0.05, m = 2$. For the entire set of parameter ranges that we tested in the experiments refer to Appendix B.

The unconditionality of Commitment

3.10

We started our experiments with comparing four different versions of Commitment playing against defectors. The necessary and sufficient condition for a strategy S to be classified under commitment is that its ah^D and ah^S traits are positive, which means that agents belonging to S will be inclined to choose those alters for cooperation who have helped them in the past. In its simplest form, this is all that Commitment cares about, expressed by e.g. the following traits: $[0,0,1,0|1]$.[11] We will refer to this strategy as Weak Commitment from now on.

3.11

An important question about the behavior of Commitment is whether the fact that *ego helped alter before* (EH) should also increase ego's willingness to cooperate. Intuitively, such a preference makes an agent more vulnerable to exploitation and holds no obvious benefits for ego, if it has an effect at all. Therefore, the second version of Commitment we examine has a non-zero value on its eh parameter: $[1,0,1,0|1]$, we denote it Strong Commitment.

3.12

If we compare now how Weak Commitment and Strong Commitment play against Defection, we see that indeed the eh trait makes a difference. Whereas Weak Commitment managed to eliminate Defectors from the society in 5.1% of all simulation runs, Strong Commitment did so in 67.7%. These results are based on 2000 independent replications of the simulation for both conditions. What is important here are not the actual percentages but the relative success of Strong Commitment compared to Weak Commitment when playing against defectors. While the survival statistic will shift in favor of defectors when the cost of giving help (f_h) is increased or the cost of not getting help (f_d) is decreased, Strong Commitment remains superior to Weak Commitment.

3.13

Figure 1 shows how the fitness proportion of Weak and Strong Commitment agents playing against Defectors changes over time. In both cases Defection initially owns 20% of the total societal fitness (and group size), Commitment 80%. Each curve corresponds to an independent simulation run. All runs end with one strategy completely outnumbering the other but each run may be of different length. Shorter runs are complemented with dashed lines for clarity. What is important to observe is the proportion of curves ending in 0% compared to those ending in 100%[12]. Additionally, a black curve shows the average fitness proportion held by a strategy at each timepoint across multiple runs.

3.14

Note also how simulation runs tend to last much longer in the case of Strong Commitment. This indicates that it takes less time for Defection to push out Weak Commitment, than it takes Strong Commitment to push out Defection. Although Defection starts from a smaller proportion than its opponent in both cases, it clearly outpowers Weak Commitment in most runs. In the second case, by contrast, Defection hardly ever manages to climb to the fitness level of Strong Commitment.

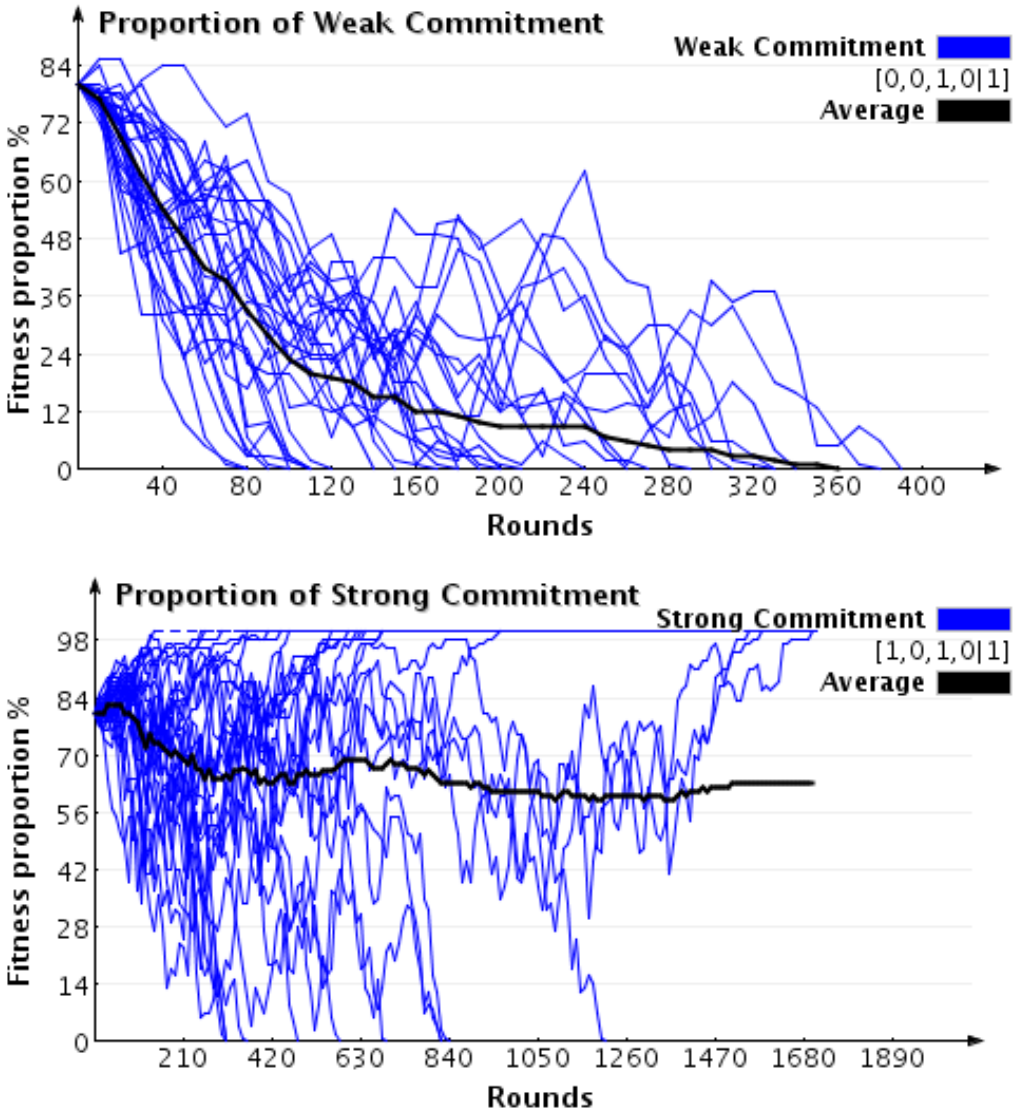

Figure 1: Weak and Strong Commitment playing against Defection

To assess the relative importance of the eh and the ah trait, we examined two ``mixed" versions of Commitment: $[1,0,2,0|1]$ and $[2,0,1,0|1]$. The former version represents a strategy that derives more utility from receiving help than from giving help to a particular partner, whereas the latter version derives more utility from giving than from receiving. Both had high survival statistics. The proportion of replications in which the corresponding Commitment strategy became universal in the simulated group was 64.2% and 72.4%, respectively. The results also hint at a stable positive effect of eh on survival success.

Unconditionality and AllC

3.15

One doubt that might have surfaced in the heedful reader about the characteristics of our model is that the more cooperative a strategy is, the more successful it will become due to the relatively low costs of cooperation. In order to show that this is not true, we provide the results for the strategy AllC playing against Defection.

3.16

AllC is the upper end of cooperativeness: it always chooses to cooperate. All other strategies are either equally or less cooperative than AllC. If more cooperativeness implied higher survival chances, AllC should be the winner of all.

3.17

This is not what we see in the results: Defectors managed to overthrow AllC in 88.6% of all simulation runs (see also Figure 2). Comparing the survival rates of AllC with those of any version of Commitment except Weak Commitment, it is obvious that AllC is less successful. The weakness of AllC is caused partly by its partner selection behavior. Since AllC is blindly cooperating it is also blind in choosing its interaction partners. When deciding who to give help to and who to ask help from, AllC is indifferent between all possible partners.[13]

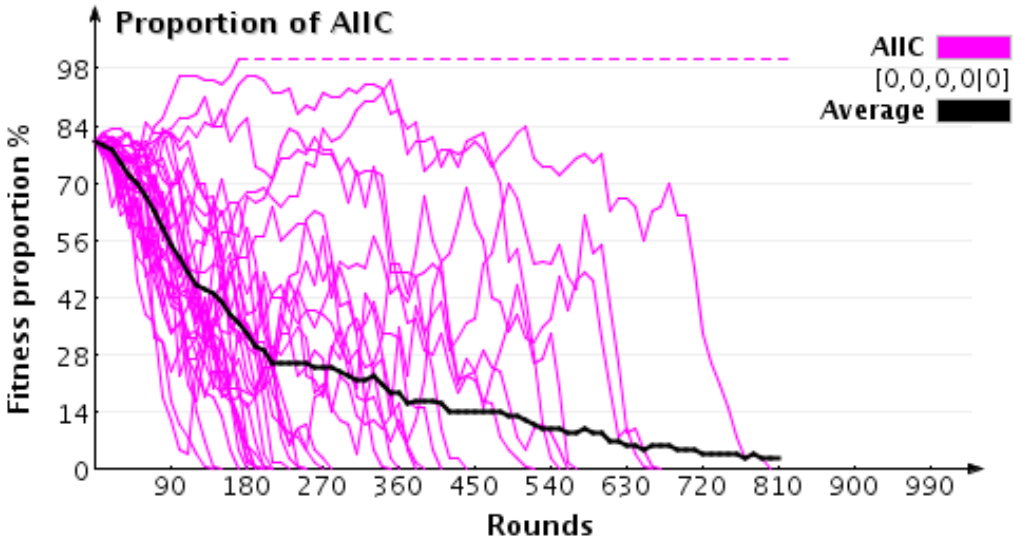

Figure 2: AllC playing against Defection

The main weakness of AllC compared to Commitment is its lack of an explicit partner selection strategy. Both versions of Commitment are more likely to help those others who have helped them before. Accordingly, a player who tries to exploit Commitment will always be less likely to get help than somebody else who cooperated with AllC, all other conditions being equal. Due to its random partner selection method, AllC is the upper end of not only cooperativeness but of *unconditionality* as well.

Commitment in comparison with fair conditional cooperators

3.18

The classic fair type (known e.g. from Fehr and Schmidt, 1999; Fehr et al., 2002) keeps a close watch on the balance with its opponent. A fair ego calculates the balance with regard to help donation to a particular alter as follows. Whenever ego helps alter or alter refuses ego, ego subtracts a unit from the balance with alter and whenever alter helps ego or ego refuses alter, ego adds one unit to the balance. The balance with regard to help seeking is calculated in exactly the opposite way. That is, the balance is most favorable with respect to the agent who *owes ego* the most. In terms of our model, this strategy is defined as $[-1,1,1,-1|1,-1,-1,1|-1]$. We will refer to this as Objective Fairness.

3.19

If we examine how this objective version of Fairness plays against Defection, we see that it has a 2.5% chance to survive, which is lower than the worst we saw for Commitment. The failure of Objective Fairness (Figure 3) can be explained with the large number of rejections out of unwillingness to help. These rejections are due to asymmetries in the number of times helping partners become distressed: e.g. if ego becomes distressed too often compared to alter (i.e. before he can reciprocate help from alter), alter will no longer provide help for ego. This result is consistent with previous game theoretical and simulation analyses that pointed to the disadvantages of strict reciprocity in uncertain environments (e.g. Kollock, 1993).

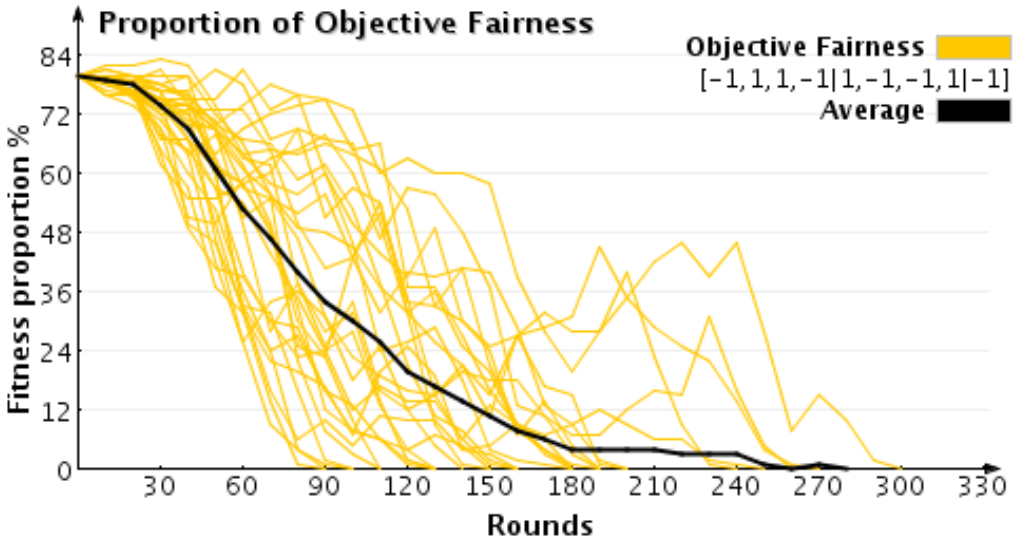

Figure 3: Objective Fairness playing against Defection

Objective Fairness can be straightforwardly modified in such a way that it becomes more tolerant against temporary fluctuations in the frequency of needing help. For our simulation, we define a corresponding strategy of Tolerant Fairness as $[-1,2,2,-1|2,-1,-1,2|-1]$.

3.20

The survival statistics of Tolerant Fairness (see also Figure 4) are clearly better than those of Objective Fairness. Tolerant Fairness stayed standing against Defection in 13.2% of all simulation runs. This result, however, is still worse than the result of Strong Commitment.

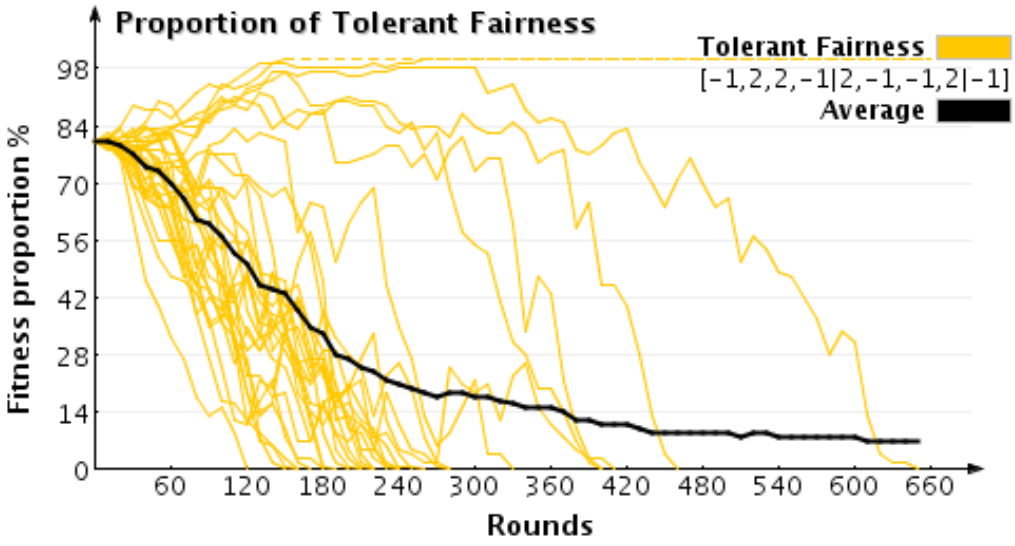

Figure 4: Tolerant Fairness playing against Defection

Explanation: the importance of strong ties

3.21

To understand the variation in the success of different cooperative strategies we first tested how viable they are when they play without defectors, in a homogeneous society composed of only one cooperative strategy. Using identical parameter settings and group size we found that there is still significant variation in success, even between unconditionally cooperative strategies.

3.22

When Commitment-type agents play against each other, being largely unconditionally cooperative, the most frequently observed type of refusal is when an agent is not able to help another agent. Hence, *whether* strategies cooperate or not cannot explain variation in success of different commitment-types. However, *whom* players select to cooperate with or to request cooperation from turns out to be decisively different between versions of Commitment.

3.23

But why should partner selection matter when everybody else is playing Commitment and will never refuse to help out of unwillingness? There are two problems faced by an agent even in this homogeneously friendly world of fellow cooperators. One is if help is sought from an agent who is distressed himself and cannot help; the other is if multiple agents ask the same agent to help. A collectively ideal strategy works so that help requests are evenly distributed among agents who are not distressed. It turns out that commitment comes very close to being such an ideal strategy, due to a phenomenon of *dyadization*.

3.24

To understand dyadization, let us consider what happens in the first few rounds. According to individually independent random events with probability P_d , a fraction of the whole society $N \cdot P_d$ becomes distressed. Since nobody had an interaction before, the distressed agents will all chose randomly whom they ask for help. For a distressed agent the probability of loosing fitness by the end of the run is composed of two parts (assuming that ego can ask only one alter for help):

$$P_{fitnessloss} = P_1 + P_2,$$

where P_1 is the probability that another distressed agent also asked and got help from alter, and P_2 is the probability that alter is also distressed. Let us assume now that $eh = 0$ for all agents. If an agent a_i who was not distressed and helped in the previous round becomes distressed now, that agent a_i will face a population of equally preferred others to ask for help. Therefore, a_i will have to choose randomly, facing the same probability P_1 that the other unlucky ones faced in the previous round. Now, if $eh > 0$, all agents who gave help before will have a preference for those they helped, and thus P_1 will be reduced in the second round.

3.25

To put it more generally, in later rounds, if everybody simply chooses the partner they had the most interactions with, it is likely that there will be little collisions between help requests. This leads to the formation of increasingly strong ties, and this is what we referred to as dyadization. Graphing the social network of Weak and Strong Commitment clearly shows the difference in the extent of dyadization (see Figure 5; the darker a tie the more interactions have taken place between the two agents). In the case of Strong Commitment, we see few but stronger ties (``close friendships''), while in the case of Weak Commitment we see many but weaker ties. A more careful look reveals that most Strong Commitment agents have exactly one strongest tie (``best friend''), while this is less true for Weak Commitment and even less for Fairness (Figure 6). What we see in a network of Fairness players are homogeneously weak relationships and nodes with a larger degree in terms of the number of ties.

3.26

All networks are graphed after 200 rounds which means that the added strength of all ties (network density) in the networks is roughly[14] equal.

3.27

To precisely measure dyadization of a network, we first compute ``individual dyadization'' for each node in the network by dividing the strength of the strongest tie of a node by the total strength of all ties of that node. The strength of a tie is obtained by counting the number of interactions that have taken place along that tie. We average the resulting values over all agents to

obtain the dyadization measure of the network. A network dyadization of 1, or perfect dyadization, means that every node is connected to exactly one other node. The dyadization after 200 rounds averaged across 2000 runs for a network of Weak Commitment players is 0.29, whereas it is 0.39 for Strong Commitment. The dyadization measure for Tolerant Fairness is 0.23.

Figure 5: Emergent social network of Weak and Strong Commitment players after 200 rounds

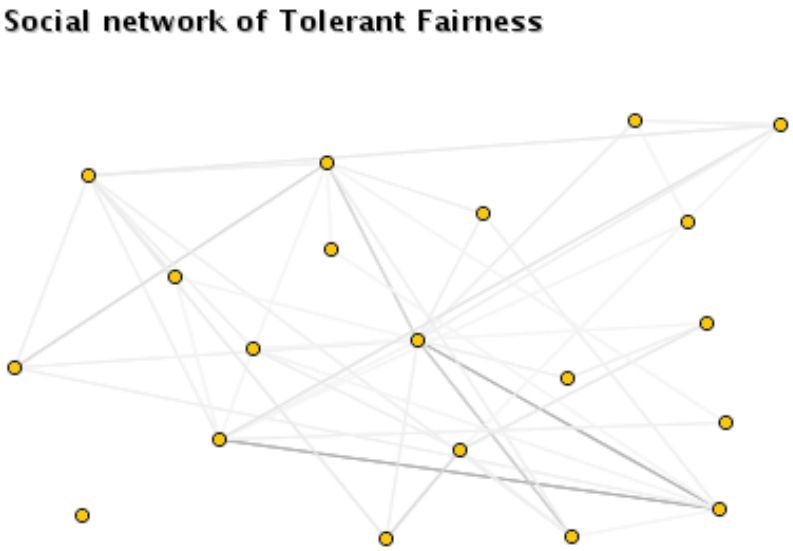

Figure 6: Emergent social network between Tolerant Fairness players after 200 rounds

Invasion of cooperators on cooperators

3.28

In a final test of the relative viability of Commitment compared to Fairness we let Strong Commitment play against Tolerant Fairness in the presence of Defectors. Commitment and Fairness both started with equal initial proportions (40% each) while Defectors were in minority (20%). Although Commitment did better (46.3%) than Fairness (2.7%), it was Defection who won (51.0%) this tournament. Apparently, the cooperative group in this case was considerably weakened by the occasional lack of cooperation of Fairness players (see Figure 7).

3.29

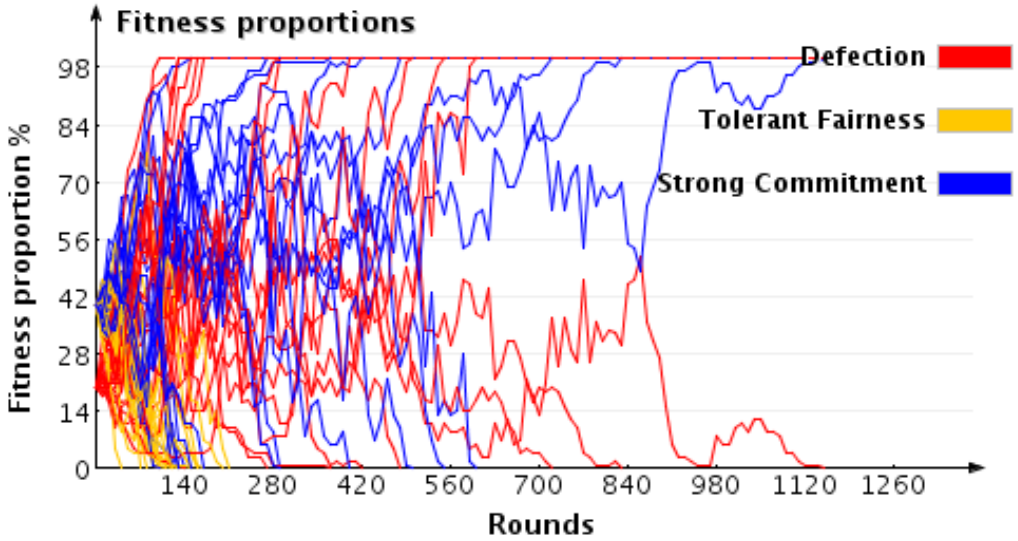

Figure 7: Fairness and Commitment playing against Defectors

Examining the network structure of all agents (Figure 8) we see again the characteristic strong friendships between Commitment players. Defectors, on the other extreme, clearly attempt to interact with as many others as possible, in search of partners to be exploited. Moreover, we also see some strong ties between Commitment and Fairness agents. What happens in these relationships is that the Commitment player becomes attached to the Fairness player after some initial rounds of helping. The problem for the Commitment agent arises if the relationships becomes unbalanced - Commitment will keep asking its Fairness ``friend'' for help even in the face of repeated refusals. This clearly points to a weakness of Commitment.

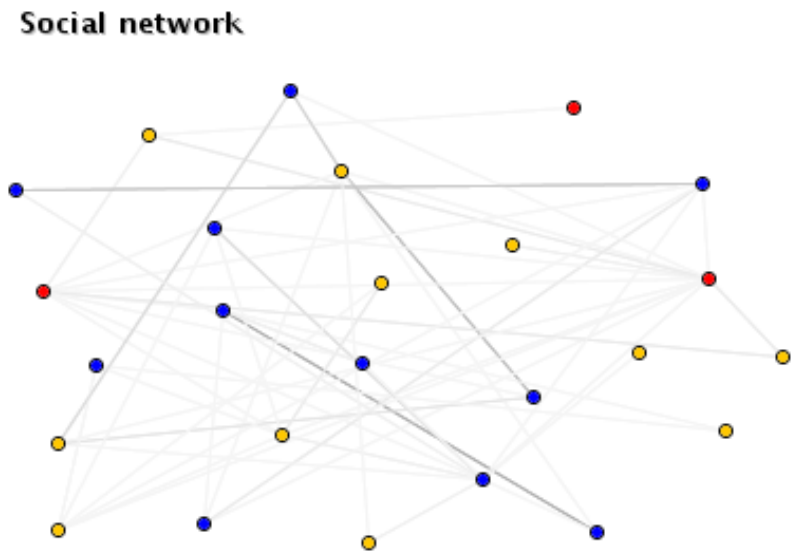

Figure 8: Network of Fairness (yellow), Commitment (blue) and Defection (red) playing together (after 200 rounds)

Sensitivity to initial parameters

3.30

To assess the robustness of the results reported above, we examined their sensitivity to the choice of initial parameters.

The cost of helping, the cost of not getting help and the probability of distress

3.31

As we pointed out earlier, the ratios of the three parameters, cost of helping (f_h), cost of not getting help (f_d) and probability of distress (P_d) are essential to determine the individual rationality of cooperation. Increasing f_h in proportion to f_d results in better survival chances of defectors. According to the analysis of the simplified dilemma (see Section 3.1), increasing the probability of distress makes the conditions under which conditional cooperation is individually rational more restrictive. This is not what we see in the simulations. Increasing P_d from 0.05 to 0.2 actually benefits *cooperators*, especially Commitment. This result is consistent with what was found by de Vos et al. (2001). The explanation is that Commitment players find each other sooner and strengthen their relationships faster, the harsher the environment is, i.e. the larger the probability of distress is.

Initial distribution and group size

3.32

In those simulation experiments reported above, where a group of cooperators played against a group of defectors, we started from an initial population of 20 cooperators against 5 defectors. Decreasing the initial population size to 10 or increasing it to 50, keeping the intial ratio of cooperators to defectors constant, did not result in notable deviations from our results.

3.33

Changing the initial proportions of cooperators and defectors did change percentages of survival to some extent but it did not reverse our qualitative conclusions. More precisely, although we found that in all experiments increasing the initial proportion of defectors led to lower survival chances for cooperative strategies, the survival chances for strong commitment were still higher than those of weak commitment or fairness strategies. Decreasing the initial proportion of defectors did not change results qualitatively either. We found that the evolutionary dynamic enabled even a smaller but superior invading strategy to

take over the entire population. We can conclude that the choice of these parameters does not affect our results qualitatively.

Probability of decision making error

3.35

We found no unexpected results when varying the error of decision making between 0.0 to 0.5: the larger the error was, the smaller the difference became between the behavior of different strategies as they all approached a completely random strategy. Defectors suffered most from a high level of decision making error: although they became more marginally cooperative, their choice of when and with whom to interact was completely haphazard.

Number of subrounds and initial fitness

3.36

By increasing the number of subrounds (m), agents in distress have a higher chance to find a helping partner within one round. The general effect we expected was that survival becomes easier for all agents, and that connections build up faster as agents encounter more alters during the same number of rounds. It is intuitively not obvious, however, who benefits more from a second (third etc.) chance - cooperators or defectors? Rerunning the simulations, keeping all parameters unchanged except varying m between 1,2 and 3, we found that it is defectors who have better chances to push out cooperators, the larger m is.

3.37

Whereas more subrounds give more chances to get help within one distress period, higher initial fitness keeps the agent alive across multiple distress periods. We expected the same general effect for changing the initial fitness parameter (f_i) as for changing the subrounds parameter because an increase in either parameter results in increased survival chances of all agents.

3.38

Varying the initial fitness parameter f_i , we found an interesting non-linearity. Increasing the initial fitness from 50 to 100 resulted in defectors becoming more successful against both conditional and unconditional cooperators. Further increasing f_i , however, resulted in defectors becoming relatively even less successful against Commitment players than at lower initial fitness.

The general result of Commitment being more successful than Fairness when playing against defectors remained constant throughout this test as well.

Discussion and Conclusion

4.1

Our results suggest that strategies following some form of commitment behavior are highly successful under a wide range of conditions. Broadly, commitment is modelled as the extent to which cooperativeness with a particular partner becomes unconditional after some initial cooperative actions of the partner. Counterintuitively, the faster an agent is inclined to solidify its relationships (see Strong Commitment), the less prone it is to exploitation. The reason is that a relationship between two Strong Commitment agents is build up - probabilistically - at least twice as fast as a relationship between a Strong Commitment and a Defector agent.

4.2

Our approach shows that the success of Commitment remains stable even when a much larger range of strategy variation is allowed than in the previous computational experiments of de Vos et al. (2001). We find that strategies which base their behavior on fairness principles generally perform much worse than commitment strategies. A truly fair strategy suffers from its lack of tolerance when interacting with its own kind in an unpredictably ``unfair'' environment, where imbalances in the exchange cannot be avoided due to uncertain hazards. A more tolerant strategy that is nonetheless based on preferences for

fair outcomes proves to be more viable. An interesting result of our model is that strategies that take their own past behavior into account - not just that of their interaction partners - make more successful decisions in general.

4.3

To explicate the reasons for the success of commitment, we also studied the spontaneous formation of exchange network structures in the simulated populations. It turned out that commitment strategies derive a large part of their success from efficient networking: they avoid overloading few designated individuals with interaction requests and instead spontaneously create a structure that ensures an efficient coordination of help requests and help provisions. The problem with fair strategies is that they are predisposed to keeping their relationships in balance so that agents tend to spread interaction requests randomly across the population. During times of great need this structure is inefficient because fair strategies in small groups generate overlapping personal networks so that often too many people try to interact with the same agent at the same time.

4.4

While we tentatively conclude that our results support the hypothesis that commitment strategies are evolutionary viable, we are also aware of a range of potential limitations of our analysis, some of which point to directions for future research.

A first objection to our study might be that we excluded influences of reputation mechanisms on the relative success of strategies.

4.5

There are a number of game theoretical studies and agent based simulations that show how reputation mechanisms can sustain cooperation, because they help cooperators to effectively identify and punish cheaters (see e.g. Takahashi, 2000; Raub and Weesie, 1990). In our analysis, reputation effects were explicitly excluded with the assumption that agents rely only on their own experience about others when making decisions.

4.6

We argue that taking reputation into account may be an unnecessary complication that would not lead to a qualitative change in the outcome of our comparison of Commitment with other versions of conditional cooperation. There is no particular reason to believe that Commitment would benefit less from reputation than other cooperative strategies. On the contrary, as Commitment players build up ties more readily and never abandon them afterwards, it may be particularly useful for Commitment players to use third party information to identify in advance who is a reliable helping partner and who is not.

4.7

A further possible limitation of the present analysis is that we have not yet explored more sophisticated cheating strategies. It is possible that more sophisticated cheating may indeed undermine the viability of Commitment. The Defection strategy derived from AllD that we used in our experiments is not capable of taking advantage of what may be the most decisive weakness of Commitment, its inability to strike back once it is exploited by a partner to whom it has become commited. The only way Commitment punishes opportunists seeking help occassionally, is by giving higher priority to friends with whom more numerous successful interactions have taken place. Similarly, Commitment will not even try to get help from its occassional interaction partners, if it has long standing partners. In other words, instead of detecting cheaters by the number of times they defected, Commitment *detects cooperators* using the number of successful interactions.

4.8

Nevertheless, there may exist more viable strategies outside the range that our analysis has covered. In future research we plan to extend our analysis in this direction. In particular, the effect of more sophisticated cheating strategies can be tested by allowing mutations to randomly generate strategies from a large set of possibilities. Clearly, this requires the modification of the evolutionary dynamic and significantly larger computational power than we used for the present study.

4.9

A final line of future work may follow from resolving a simplifying assumption we made, to ignore possible differences between group members in terms of their attractiveness as exchange partners beyond their strategy. Such differences may for example come from variation in physical strength or more or less favorable local living conditions. Studies by Hegselmann (1996) (cf. Flache and Hegselmann, 1999) suggested that variation in attractiveness may give rise to core periphery network structures in which the strongest population members exchange help with each other, driving weaker actors to the margin of

help exchange networks. However, this work relied on conditionally cooperative strategies that resemble the strategy of Fairness. Accordingly, it is unclear whether variation in individual attractiveness may affect the viability of commitment strategies and also whether commitment strategies would give rise to the exclusion of weak members from exchange networks in the same way as it has been found for Fairness-like strategies.

Appendix A: Analytical solution of the simplified dilemma

A.1

First we calculate the expected number of times that a trigger player will help a trigger player, n_h , and multiply this by the costs of giving help, f_d . This yields the total loss, $n_h \cdot f_h$, of playing trigger against trigger (which is not incured by an ALLD player playing against trigger).

A.2

Let n_{TT} be the expected number of times before both players revert to eternal defection, i.e. before both trigger players get into distress simultaneously:

$$n_{TT} = \frac{1}{P_d^2} - 1$$

Let n_h be the number of those rounds within these initial n_{TT} rounds that the trigger alter is in distress, while ego is not. Then

$$n_h = P_h \cdot n_{TT}$$

where

$$P_h = \frac{P_d(1 - P_d)}{1 - P_d^2}$$

The conditional probability P_h is calculated as follows. The unconditional probability of the event "alter in distress while ego not" is $P_d(1 - P_d)$ but what we need to know is the probability that this event occurs under the condition that the event "both are in distress" has not yet occurred. To obtain this, we divide the unconditional probability by the probability that the condition occurs.

A.3

Next, we calculate the expected number of times n_d that an ALLD player does not get help from a trigger opponent, while a trigger player would get help at the same time. Multiplied with the costs of not getting help, f_d , this amounts to the expected loss, $n_d \cdot f_d$, that an ALLD player incurs which is not incurred by a trigger player.

A.4

Let n'_{DT} be the length of the period in which a trigger player has not yet lost the help of his opponent, while an ALLD player gets no more help. Let P_n be the probability that in any single round of this part of the game, the ALLD ego is in distress,

while alter is not. Then

$$n_d = n'_{DT} \cdot P_n$$

where

$$P_n = \frac{P_d(1-P_d)}{1-P_d^2}$$

Here P_n is obtained in the same way than P_h above, only this time the roles of ego and alter are reverted (ego in need but alter not), which does not affect the result in this case.

A.5

To obtain n'_{DT} , we need to find the difference between n_{TT} , the expected duration until a trigger player would stop getting help from his trigger opponent, and n_{DT} , the number of times until an ALLD player would stop getting help from this opponent.

A.6

Furthermore, we need to take into account that, if there is a second phase of the game, the first round of that second phase cannot be a round in which an ALLD ego is refused by alter because it is the round in which ego reveals itself as a defector, i. e. ego cannot be in distress. Hence, to obtain n'_{DT} we need to subtract from $n_{TT} - n_{DT}$ exactly that probability.

A.7

To obtain that probability, we need to consider that there may never be a second phase of the game in which an ALLD player would not get help, because the first phase may already end with a round in which both players get into distress simultaneously, in which case also a trigger player would get no more help. The probability for that latter event is again obtained as a conditional probability as follows: the unconditional probability that alter gets into distress but ego not is $P_d(1-P_d)$. There are only two possible events that can occur in the first round after the first phase of the game. Either it is the first round of the second phase, with an unconditional probability of $P_d(1-P_d)$, or there is no second phase and both players are in distress simultaneously, with probability P_d^2 . All other events are excluded by virtue of the precondition that this is the round following the first phase of the game. So the probability we need to subtract from $n_{TT} - n_{DT}$ is $P_d(1-P_d)$ divided by the probability that the condition occurs, $P_d(1-P_d) + P_d^2$. All in all this amounts to

$$n'_{DT} = n_{TT} - n_{DT} - \frac{P_d(1-P_d)}{P_d(1-P_d) + P_d^2}$$

where

$$n_{DT} = \frac{1}{P_d} - 1$$

A.8

Finally, we need to find the condition under which the loss of a trigger player against trigger is equal to that of an ALLD player against trigger. That is, we need to solve

$$n_d \cdot f_d = n_h \cdot f_h$$

This condition yields after some rearrangement the following result:

$$f_h = f_d(1 - P_d)$$

Appendix B: Parameter values used in simulation runs

B.1

We acquired and tested our results using the following parameters:

Model parameters

B.2

probability of distress (P_d) = 0.05, 0.2, 0.5

probability of decision making error (P_e) = 0.0, 0.05, 0.1, 0.5

cost of helping (f_h , measured in fitness) = 1

cost of not getting help (f_d , measured in fitness) = 5, 10, 20

initial fitness (f_i , measured in fitness) = 50, 100, 200

critical fitness (f_c , measured in fitness) = 0

group size (N) = 10, 25, 50
number of sub-rounds in a round (m) = 1, 2, 3

Simulation parameters

B.3

number of rounds in a run = *variable, depending on how long it took the winning strategy to push its opponent(s) out*
number of runs in an experiment = 2000

Appendix C: Pseudocode of the core program

```
/**
 * Main cycle of simulation
 */
begin simulation
    for each experiment
        initialize result variables and charts
        for each run
            initialize parameters, run-level result variables and charts
            initialize society
            for each round
```

```
                initialize round-level result variables
                for each agent A
                    generate a random event R with probability P_d  /* probability
of distress */
                    if R occurs distress A
                end for
                for each subround
                    for each agent A
                        if A is distressed call decideWhomToAskForHelp of A
                    end for
                    for each agent A
                        call decideWhomToGiveHelp of A
                    end for
                end for
                for each agent A
                    update fitness
                    if fitness < f_c remove A from society  /* critical fitness
threshold */
                end for
                call replicator_dynamics()
                update round-level result variables
                update charts if necessary  /* if this is a sample round */
                if society is empty end run
                if there is only one group left in society end run
            end for    /* end of rounds */
            update run-level variables
            update charts
        end for /* end of runs */
        update experiment-level variables
        update charts
    end for /* end of experiments */
    close charts and show results
end simulation


/**
 * Agent deciding whether/whom to give help.
 */
begin decideWhomToGiveHelp
    if nobody asked for help return nobody
    if agent is distressed return nobody
    if agent already gave help return nobody
    generate a random event R with probability P_err  /* decision making error */
    if R does not occur
        determine group G of agents for whom U_d is maximal  /* utility of donation
*/
        if U_d < U_t return nobody              /* threshold utility */
        else return random agent from G
    else
        return random agent from {helpseekers+nobody}
    end if
end decideWhomToGiveHelp


/**
```

```
     * Agent deciding whom to ask for help.
     */
    begin decideWhomToAskForHelp
        remove itself from possible helpers
        remove those who refused before in this round
        if possible helpers is empty return nobody
        generate a random event R with probability P_err  /* decision making error */
        if R does not occur
            determine group G of agents for whom U_s is maximal  /* utility of seeking */
            return random agent from G
        else
            return random agent from possible helpers
        end if
    end decideWhomToAskForHelp


    /**
     * Evolutionary selection process (based on the replicator dynamics)
     */
    begin replicator_dynamics
        for each dead agent
            calculate society_fitness_sum
            for each strategy S in society
                calculate strategy_fitness_sum
                generate a random event R with probability strategy_fitness_sum /
    society_fitness_sum
                if R occurs create and add agent A with strategy S to new_generation
                /* in order to condition successive probability calculations on previous
    ones */
                else decrease society_fitness_sum with strategy_fitness_sum
            end for
        end for
        add new_generation to society
    end replicator_dynamics
```

Acknowledgements

We wish to thank Henk de Vos, Tom Snijders, Károly Takács and three anonymous reviewers for their inspiring and helpful comments. István Back's research was financially supported by the Ubbo Emmius fund of the University of Groningen.

The research of Andreas Flache was made possible by the Netherlands Organization for Scientific Research (NWO), under the Innovational Research Incentives Scheme.

Notes

... fitness[3]

In certain cases these objective consequences may actually be impossible to foresee for the agents or even for the modeller.

... general[4]

See more about the problems of involuntary defection in Fishman (2003) and agents getting stuck in mutual defection in noisy environments in Monterosso et al. (2002)

... threshold.[5]

Since we are interested in the *perception of the strength of a relationship* between agents rather than the *perception of objective costs*, we assume that U_m^D and U_m^S are constant in all interactions.

... to.[6]

Note, that for the sake of simplicity in explaining the behavior of strategy classes, we will assume that the probability of a decision making error $P_e = 0$ throughout this section.

... Fairness.[7]

Social psychologists also refer to this type of behavior as equity (cf. Smaniotto, 2004).

... simulations.[8]

We stop simulation runs when a strategy completely pushed out its opponents from the agent society.

... times[9]

$n = 2000$ independent runs were usually sufficient.

... other.[10]

In other words, we are not covering exhaustively the entire parameter space, only a very large part of it. It may still be possible that there is a specific trait combination that is superior under some conditions to the ones examined.

....[11]

A strategy is described by four trait parameters for donation, four trait parameters for seeking and the utility threshold: $[eh^D, er^D, ah^D, ar^D | eh^S, er^S, ah^S, ar^S | U_t]$. If parameters for donation and seeking are equal, we provide only four trait parameters instead of eight. Find more information above, in section 2.1.

...[12]

Note that in order the maintain the clarity of the graphs, we reduced the number of simulation runs actually plotted on the figures.

... partners.[13]

Therefore, it is important to define AllC as $[0,0,0,0|0]$ and not e.g. as $[1,1,1,1|1]$ or $[2,2,2,2|2]$ since in the latter two cases it would become a variant of commitment (i.e. it will attribute more utility to those he interacted with more times in the past). While the latter two definitions are perfectly identical to each other, the first definition prescribes surprisingly different behavior with regard to partner selection.

... roughly[14]

Depending on the actual number of stochastically arisen distresses.

References

Abreu, D. (1988). On the theory of infinitely repeated games with discounting. *Econometrica*, 56(2):383.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, New York.

Binmore, K. (1998). Review of `R. Axelrod, The complexity of cooperation: Agent-based models of competition and collaboration; Princeton UP 1997'. *Journal of Artificial Societies and Social Simulation*, 1(1). http://jasss.soc.surrey.ac.uk/1/1/review1.html.

Boster, F., Rodriguez, J., Cruz, M., and Marshall, L. (1995). The relative effectiveness of a direct request message and a pregiving message on friends and strangers. *Communication Research*, 22(4):475-484.

Cosmides, L. (1989). The logic of social exchange: Has natural section shaped how humans reason? studies with the Watson selection task. *Cognition*, 31:187-276.

Cosmides, L. and Tooby, J. (1993). Cognitive adaptations for social exchange. In Barkow, J. H., Cosmides, L., and Tooby, J., editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pages 163-228.

de Vos, H., Smaniotto, R., and Elsas, D. A. (2001). Reciprocal altruism under conditions of partner selection. *Rationality and Society*, 13(2):139-183.

de Vos, H. and Zeggelink, E. P. H. (1997). Reciprocal altruism in human social evolution: the viability of reciprocal altruism with a preference for 'old-helping-partners'. *Evolution and Human Behavior*, 18:261-278.

Falk, A., Fehr, E., and Fischbacher, U. (2001). Driving forces of informal sanctions. In *NIAS Conference Social Networks, Norms and Solidarity*.

Fehr, E., Fischbacher, U., and Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature - An Interdisciplinary Biosocial Perspective*, 13(1):1-25.

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

Fehr, E. and Schmidt, K. (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, CXIV:817-851.

Fishman, M. A. (2003). Indirect reciprocity among imperfect individuals. *Journal of Theoretical Biology*, 225(3):285-292.

Flache, A. and Hegselmann, R. (1999). Rationality vs. learning in the evolution of solidarity networks: A theoretical comparison. *Computational and Mathematical Organization Theory*, 5(2):97-127.

Friedman, J. (1971). A non-cooperative equilibrium for supergames. *Review of Economic Studies*, 38:1-12.

Güth, W. and Kliemt, H. (1998). The indirect evolutionary approach. *Rationality and Society*, 10(3):377-399.

Hegselmann, R. (1996). Solidarität unter ungleichen. In Hegselmann, R. and Peitgen, H.-O., editors, *Modelle sozialer Dynamiken -- Ordnung, Chaos und Komplexität*, pages 105-128. Hölder-Pichler-Tempsky, Wien.

Karremans, J., Lange, P. V., Ouwerkerk, J., and Kluwer, E. (2003). When forgiving enhances psychological well-being: the role of interpersonal commitment. *Journal of Personality and Social Psychology*, 84(5):1011-1026.

Kollock, P. (1993). An eye for an eye leaves everyone blind - cooperation and accounting systems. *American Sociological Review*, 58(6):768-786.

Kollock, P. (1994). The emergence of exchange structures: An experimental study of uncertainty, commitment, and trust. *American Journal of Sociology*, 100(2):313-45.

Lawler, E. and Yoon, J. (1993). Power and the emergence of commitment behavior in negotiated exchange. *American Sociological Review*, 58(4):465-481.

Lawler, E. and Yoon, J. (1996). Commitment in exchange relations: Test of a theory of relational cohesion. *American Sociological Review*, 61(1):89-108.

Lawler, E. J. (2001). An affect theory of social exchange. *American Journal of Sociology*, 107(2):321-52.

Mills, J. R. and Clark, M. S. (1982). Communal and exchange relationships. In Wheeler, L., editor, *Annual Review of Personality and Social Psychology*. Beverly Hills, Calif.: Sage.

Mills, J. R. and Clark, M. S. (1994). Communal and exchange relationships: Controversies and research. In Erber, R. and Gilmour, R., editors, *Theoretical Frameworks for Personal Relationships*. Hillsdale, N.J.: Lawrence Erlbaum.

Monterosso, J., Ainslie, G., Mullen, P. A. C. P. T., and Gault, B. (2002). The fragility of cooperation: A false feedback study of a sequential iterated prisoner's dilemma. *Journal of Economic Psychology*, 23(4):437-448.

Nesse, R. M. (2001). Natual selection and the capacity for subjective commitment. In Nesse, R. M., editor, *Evolution and the Capacity for Commitment*. New York : Russell Sage Foundation.

Nowak, M. A. and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364:56-58.

Raub, W. and Weesie, J. (1990). Reputation and efficiency in social interaction: an example of network effects. *American Journal of Sociology*.

Rusbult, C. and Martz, J. (1995). Remaining in an abusive relationship - an investment model analysis of nonvoluntary dependence. *Personality and Social Psychology Bulletin*, 21(6):558-571.

Rusbult, C., Martz, J., and Agnew, C. (1998). The investment model scale: Measuring commitment level, satisfaction level, quality of alternatives, and investment size. *Personal Relationships*, 5(4):357-391.

Schüssler, R. (1989). Exit threats and cooperation under anonimity. *The Journal of Conflict Resolution*, 33:728-749.

Schüssler, R. and Sandten, U. (2000). Exit, anonymity and the chances of egoistical cooperation. *Analyse & Kritik*, 22.

Smaniotto, R. C. (2004). *`You scratch my back and I scratch yours' versus `love thy neighbour': two proximate mechanisms of reciprocal altruism*. PhD thesis, ICS/University of Groningen.

Takahashi, N. (2000). The emergence of generalized exchange. *American Journal of Sociology*, 105(4):1105-1134.

Taylor, P. D. and Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40:145-156.

Vanberg, V. and Congleton, R. (1992). Rationality, morality and exit. *American Political Science Review*, 86:418-431.

Wieselquist, J., Rusbult, C., Agnew, C., Foster, C., and Agnew, C. (1999). Commitment, pro-relationship behavior, and trust in close relationships. *Journal of Personality and Social Psychology*, 77(5):942-66.

Wu, J. and Axelrod, R. (1995). How to cope with noise in the iterated prisoner's dilemma. *Journal of Conflict Resolution*.

Yamagishi, T., Hayashi, N., and Jin, N. (1994). Prisoner's dilemma networks: selection strategy versus action strategy. In Schulz, U., Albers, W., and Mueller, U., editors, *Social Dilemmas and Cooperation*, pages 311-326. Heidelberg: Springer.

Zeggelink, E., de Vos, H., and Elsas, D. (2000). Reciprocal altruism and group formation: The degree of segmentation of reciprocal altruists who prefer old-helping-partners. *Journal of Artificial Societies and Social Simulation*, 3(3). http://jasss.soc.surrey.ac.uk/3/3/1.html.

Return to Contents of this issue