The adaptive rationality of interpersonal commitment

Back, Istvan; Flache, Andreas

Published in:
Rationality and Society

DOI:
10.1177/1043463107085437

Rationality and Society

http://rss.sagepub.com

The Adaptive Rationality of Interpersonal Commitment

István Back and Andreas Flache

Rationality and Society 2008; 20; 65
DOI: 10.1177/1043463107085437

THE ADAPTIVE RATIONALITY OF INTERPERSONAL COMMITMENT

István Back and Andreas Flache

ABSTRACT

Why are people inclined to build friendships and maintain durable, non-reproductive relationships? Previous computational modeling work showed that it can be an efficient survival strategy to choose interaction partners based on relationship length, even if, as a consequence, individuals become unconditionally cooperative in long-term relationships (interpersonal commitment). Such committed individuals can outperform conditional cooperators who play in a fair, reciprocal manner (e.g. tit for tat). However, previous studies did not conduct a sufficiently strict test of the viability of commitment because they did not account for exploiters who specifically take advantage of the tolerance of commitment players. We allow for this by extending previous studies with the possibility of randomly mutating strategies under evolutionary pressures, and thus give a much larger coverage of an infinite strategy space. Our results point to the lack of stable strategies: we find that emerging populations alternate between temporarily stable states. We also show that the viability of strategies increases with increasing levels of interpersonal commitment, and that the effect of interpersonal commitment on viability is larger than the effect of fairness.

KEY WORDS • interpersonal commitment • fairness • reciprocity • agent-based model • evolution

1. Introduction

Of the species of the earth, humans exhibit the highest level of cooperation between genetically unrelated individuals (Gintis 2003). Arguably, cooperation is the de facto key to our evolutionary success. At the same time, cooperation is problematic to explain from a rational actor perspective. Self-interested actors often face a 'social dilemma' (Dawes 1980), where the rational pursuit of individual interests may lead them toward defection, while this in turn entails collectively undesirable outcomes. Game theory

Rationality and Society Copyright © 2008 Sage Publications. Vol. 20(1): 65–83.
http://rss.sagepub.com DOI: 10.1177/1043463107085437

has identified repeated interaction as an important solution for the problem. In a world of harsh competition, repeated encounters reduce uncertainty about the trustworthiness of interaction partners (shadow of the past), while at the same time they create a strategic incentive for cooperation (shadow of the future) (Axelrod 1984; Buskens and Raub 2002; Friedman 1971). Thus durable relationships are expected to be a hotbed of cooperation even in the absence of central enforcement because they provide incentives both to trust others and to honor others' trust.

From this perspective, it is hardly surprising that rational incentives to become committed to long-term cooperative exchange partners are particularly strong in uncertain environments (Kollock 1994; Schüssler 1989). A reduction in uncertainty is often more valuable than a probabilistic increase in payoff from a potential new partner, especially if switching itself is risky or costly or alternatives are scarce. This can explain why in situations where uncertainty may otherwise preclude the desirable outcome of mutual cooperation, social actors often restrict their own freedom of action by using commitment devices such as posting a hostage (Raub 2004). However, this rational explanation of interpersonal commitment behavior is hard to reconcile with the empirical evidence that people tend to stay committed to long-term interaction partners even when (1) alternatives are available, (2) switching costs are low, and (3) uncertainty is of less concern.

A growing body of empirical findings from both interpersonal relationships research (Karreman et al. 2003; Wieselquist et al. 1999) and exchange experiments (Kollock 1994; Lawler and Yoon 1993, 1996) shows that people have a tendency to remain cooperative to interaction partners who are occasionally uncooperative. Moreover, people tend to keep exchanging with the same partner even if more valuable (or less costly) alternatives are available. Such commitments also imply forgiveness and gift-giving without any explicit demand for reciprocation (Lawler 2001; Lawler and Yoon 1993). People help friends and acquaintances in trouble, apparently without calculating present costs and future benefits. Another, extreme example is the case of battered women who stay with their abusive husbands (Rusbult and Martz 1995; Rusbult et al. 1998).

In this article we seek an explanation by following the general lead of the 'indirect evolutionary approach' (Güth and Kliemt 1998), which posits that individuals act rationally in the light of their preferences but also assumes that in the course of biological and cultural evolution individuals with emotions and social preferences – e.g. for fair distributions (cf. Bolton and Ockenfels 2000), or for altruistic punishment (cf. Fehr and Gächter 2002) – may have had a selective advantage, because their preferences produce more viable outcomes than those of pure egoists.

This approach aims to integrate the endogenous explanation of such nonselfish preferences with the classical rational choice assumption that humans act (boundedly) rationally, given their preferences. The core idea is that preferences are embodied in the genotype and guide individual actions. *Subjective* preferences may be harmful to the individual or to the population but genotypes are selected on the basis of *objective* consequences of the actions that preferences produce. As preferences undergo selection and mutation, infeasible and harmful preferences gradually become less widespread in the population, giving way to more 'rational' sets of preferences. However, while the indirect evolutionary approach has proven to be a fruitful way of explaining phenomena such as emotional commitment to a certain course of action, altruistic punishment, or cooperation in the production of collective goods (Frank 1998; Gintis 2003; Güth and Kliemt 1998; Güth and Ockenfels 2002) the phenomenon of interpersonal commitment has received less attention.

Recently, some authors have begun to use the indirect evolutionary approach to explain interpersonal commitment (Back and Flache 2006; de Vos et al. 2001; de Vos and Zeggelink 1997; Zeggelink et al. 2000). While these analyses suggested that commitment may have been evolutionarily viable, we argue that the tests they used were not strict enough. De Vos and his collaborators argued in a series of papers that in a stylized 'ancestral environment' a strategy based on commitment behavior can outperform a strategy based on calculative reciprocity when both strategies are in competition with one defecting strategy. Researchers modeled commitment as unconditional cooperativeness with a particular partner after some initial cooperative actions of the partner. By contrast, calculative reciprocity (based on fairness principles) continuously keeps track of its interaction balance with alters (partners) and adjusts its cooperativeness accordingly. Using an ecological simulation model, Back and Flache (2006) extended the de Vos model by introducing variation in the extent to which a strategy follows commitment or calculative reciprocity behavior. This study showed that 'strong' commitment strategies outperform 'weaker' forms of commitment and various versions of calculative reciprocators under a wide range of conditions. However, a remaining major limitation of these analyses is that the spontaneous emergence of more sophisticated strategies was not considered. In particular, the possibility was precluded that sophisticated cheaters could emerge who optimally take advantage of the cooperativeness of commitment. Whether and to what extent this may be possible is crucial for the validity of an explanation of interpersonal commitment behavior in terms of its evolutionary advantages in the human ancestral environment.

Accordingly, in the present article we provide a better test of evolutionary explanations for commitment by extending previous analyses with random mutation of strategies. In Section 2, we present our computational model and formulate conjectures. Section 3 contains the results of simulation experiments, followed by a discussion and conclusions in Section 4.

2. Model

We use an abstract decision situation that we call the Delayed Exchange Dilemma (see Back and Flache 2006; de Vos et al. 2001), or DED for short. The DED builds on the well-known repeated Prisoner's Dilemma but contains two major extensions. First, it puts the problem of cooperation into a sequential exchange perspective, which is essentially a generalization of simultaneous exchange. Second, and more important, is that it presents agents with the dilemma of choosing between interaction partners (see also, e.g., Hayashi and Yamagishi 1998). With these extensions the DED becomes ideal for studying commitment-related behavior in uncertain environments.

The DED is played by n agents in successive rounds. Initially, all agents are endowed with f_i points. At the beginning of each round Nature strikes a number of agents, each with a given individually independent probability P_d, who become in need of help from other agents. Agents who are struck by Nature are the initiators of interactions. Each of them asks another agent for help, which is either provided or not. Providing help costs f_h points. Moreover, help giving is time-consuming. Each agent can 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 not more than m agents altogether within the same round, due to time restrictions. If an agent does not get help before the end of the round, it experiences f_d loss in points. If the points of an agent fall below a critical threshold f_c, the agent dies.

To explicitly study the evolutionary viability of commitment and fair reciprocity, we model preferences as a combination of commitment-related traits, fairness-related traits and a general cooperativeness trait. These traits determine the extent to which agents base their decisions on commitment- or fairness-related aspects of a decision situation. Equipped with these preferences agents decide about cooperation and also about choosing interaction partners.

In particular, agents may face two different types of decision situations repeatedly in the DED. When they are hit by distress, they have to select an interaction partner to ask help from. On the other hand, when they themselves are asked to provide help, they need to decide whether to provide it and in the case of multiple requests whom to provide it to. In both cases, agents order possible interaction partners according to the attractiveness of interacting with them. Attractiveness is based on the individual preferences agents have with regard to past interaction histories.

The attractiveness of agent a_j for *giving help to*, calculated by agent a_i, is formalized as:

$$U_{ij}^{G} = comm_i^G \cdot INTFREQ_{ij} + fair_i^G \cdot INTBAL_{ij} + coop_i, \tag{1}$$

where $comm_i^G$ is the preference for commitment in giving, $fair_i^G$ is the preference for fairness in giving, and $coop_i$ is the preference for general cooperativeness. $INTFREQ_{ij}$ is the proportion of cooperative interactions[1] a_i had with a_j compared to the total number of cooperative interactions a_i had with all agents in the population. A cooperative interaction is defined as an interaction in which either a_i helped a_j or a_i received help from a_j. $INTBAL_{ij}$ is the standardized interaction balance between agents a_i and a_j. To obtain this measure, we took into account both the balance of helps and the balance of refusals. The reason is that neither help balance nor refusal balance alone is sufficient to guarantee an overall balance in the exchange relationship. For example, suppose a_i helped a_j equally often as a_j helped a_i but a_i refused to help a_j 10 times more often than a_j refused to help a_i. Despite the equal amount of help given, this exchange relationship clearly cannot be considered perfectly in balance. Technically, we calculated the measure as follows. We subtracted from 1.0 a measure of the overall standardized imbalance. The overall standardized imbalance is obtained by adding the difference between the number of times a_i received help from a_j and a_i gave help to a_j, and the difference between the number of times a_i refused to give help to a_j and a_j refused to give help to a_i, and dividing this by the total number of interactions they had.

When comparing *INTFREQ* and *INTBAL*, notice that while a committed agent a_i will find an interaction partner a_j more attractive the more often it helped a_j, a fair agent will be negatively influenced by the same fact. Note also that, for simplicity, this model treats the impact of helping and refusing to help on the interaction balance as equally large.

In the actual implementation, every time an agent has to make a decision, there is also a probability P_e that the agent will not use the above utility calculation but will choose randomly from the set of available decisions, each being equally likely. 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.

The attractiveness of agent a_j for *asking help from* is defined in a similar way, the difference being that agents may put different weights on the two history-specific decision parameters, and that there is no cooperativeness parameter:

$$U_{ij}^{A} = comm_i^A \cdot INTFREQ_{ij} + fair_i^A \cdot INTBAL_{ij}, \tag{2}$$

Before agents make a decision, be it help giving or help asking, they calculate the corresponding one of these two types of attractiveness respectively for each agent who asked for help (U^G), or for each other agent in the population (U^A). In the case of help giving, they choose a partner with the highest attractiveness, if that attractiveness is above an agent-specific threshold u_i^t. Notice that $INTFREQ_{ij}$ and $INTBAL_{ij}$ are always smaller than or equal to 1. We allow $comm_i$, $fair_i$ and $coop_i$ to take values from [−1; 1]. Thus we allow the attractiveness threshold u_i^t to take values from [−3; 3].

If the attractiveness of all possible agents is below the threshold attractiveness, no help is given to anyone. Otherwise, if there is more than one other agent with highest attractiveness,[2] the agent selects one of the others with equal probability. In the case of help seeking, agents also choose a partner with the highest attractiveness but there is no threshold, i.e. agents in distress always ask someone for help.

Definition 1. *(Strategy) A strategy is a combination of four traits for help-giving behavior ($comm^G$, $fair^G$, $coop$, u^t) and two traits for help-asking behavior ($comm^A$, $fair^A$).*

The heart of our model is an evolutionary dynamic that captures *random mutation* of strategies and *selection* of objectively successful ones. The implementation of this process is based on the replicator dynamics (Taylor and Jonker 1978). Broadly, the replicator dynamics dictate that if a generation of genotypes (strategies) undergoes reproduction, the net reproduction rate of a genotype is proportional to its relative success compared to other genotypes in the current generation. Genotypes which perform below average, in particular, have a negative reproduction rate. In our case, genotypes (strategies) represent subjective preferences.

To prevent a population from growing without bounds, thus modeling resource scarcity in an implicit way, we keep the size of the population constant, in the following way. At the end of each round we count how many agents have died and replace them with new agents in the next round. Each new agent *A* has the same strategy as a randomly selected other agent *B*, present in the population who has reached a minimum age *n* (measured in the number of interactions it had). The probability of choosing this other agent *B* is proportionate to the share of points *B* holds within the group of all agents older than *n*. Before *A* is added to the population, with probability P_{mut}, its strategy may undergo mutation. A mutation occurs in exactly one, randomly chosen trait, with equal probabilities for all traits, thus $P = \frac{1}{9}$ for each trait. The new value of the trait is a uniformly distributed random value from the interval [−3; 3] for the attractiveness threshold, and from [−1; 1]for all other traits.

3. Conjectures

To guide the simulation experiments, in the following we formulate a number of conjectures derived from previous work.

Definition 2. *(Stability) Stability of a strategy s is equal to the number of consecutive rounds it existed in a population in a given simulation run, counting from the first round it appeared until the round in which it became extinct. A strategy s is infinitely stable if it does not become extinct.*

Generalizing from analytical results about the evolutionary stability of strategies in repeated games that are simpler than the DED (cf. Bender and Swistak 2001), we expect that there is no single strategy that is superior to all others in the dilemma we study. In other words, for every incumbent strategy there exists another (mutant) strategy that can take advantage of the incumbent's weakness.

Conjecture 1. *There is no infinitely stable strategy in an infinitely played game of DED.*

Nevertheless, the length of time a strategy exists (stability) carries an important message about its viability. Since mutations constantly arise and threaten to push other strategies out of the population, stability is an indicator for the number of attacks a strategy could withstand. Therefore, stability of a strategy will be one of the indicators of its viability.[3] The other measure is typical longevity within a strategy (variable longevity, the average age at death of agents belonging to a strategy). Note that in our model, there is no upper age limit on reproducibility, in other words, agents keep reproducing until they die, which makes longevity a suitable measure for viability.

Back and Flache (2006) found that the most successful strategies in the DED exhibited some level of interpersonal commitment and that committed agents outcompeted fair reciprocators. These results suggest the following two conjectures, which we will test under the new assumption of random mutation:

Conjecture 2. *Individual preferences for interpersonal commitment and fairness have a positive effect on viability.*

Conjecture 3. *The positive effect of commitment preferences on viability is stronger than the effect of fairness preferences.*

According to de Vos et al. (2001) commitment works best under harsh conditions: the more agents are challenged by Nature to survive, the more compelled they are to cooperate with each other in durable relationships. More technically, they found that the larger the probability of distress, the larger the proportion of commitment strategies surviving, relative to the defector strategy. This leads us to test:

Conjecture 4. *Environmental harshness has a positive interaction effect on stability with the level of cooperation and interpersonal commitment of a strategy.*

4. Results

Binmore (1998) argued forcefully that the outcome of computer tournaments and simulations of evolutionary dynamics strongly depends on the set of strategies that are initially present in a population. To avoid our results becoming biased by a restrictive set of starting conditions, we ran a large number of replications of our simulation runs, each time with a population whose initial strategy is randomly chosen from the strategy space defined by the six traits. We did not find any significant effects of features of initial starting strategies on the outcomes of simulation runs. The reason is that soon after the initial rounds of a simulation run, mutation ensures the emergence of a large variety of different strategies in the population.

We allow this population to play the DED game. In the course of the game agents start to lose points, some of them eventually die, while others reproduce. At some point, random mutations occur in the initial strategy, creating a potential invader. The better a mutated strategy performs in the DED compared to agents of the original strategy, the larger is its probability of reproducing and increasing its proportion within the agent population. The simulation run ends with either the extinction of all agents[4] or after an arbitrarily chosen large number of rounds (10 million). We then repeat the simulation run with another, randomly generated initial population.

During each simulation run we record all strategies and their key characteristics that have ever appeared through random mutations. These characteristics include on the individual level the traits of the strategy ($comm^A$, $fair^t$, *coop*, u^t, $comm^G$, $fair^G$); the average longevity measured in rounds of game play on the strategy level; and finally a population-level variable measuring the overall level of cooperation and defection (SOCCOOP).[5]

Initial parameters

To preserve the comparability of our results we started our simulations with the same initial parameters (where applicable) that were used in earlier work. These 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 meaning of each parameter consult the Model section above.) These parameters impose a set of conditions under which for strictly instrumental agents the choice between purposeful defection and (conditional) cooperation is as difficult as possible. The parameters are determined such that in a simplified two-person version of the game, perfectly rational actors would be indifferent between choosing a conditionally cooperative and a fully defecting strategy if they meet a conditionally cooperative partner. In this way, we implement a setting in which the problem of cooperation is particularly hard to solve and thus provide a strict test of the viability of cooperative strategies, including commitment and fairness. If cooperation placed an excessively high burden on agents, or, conversely, if cooperation entailed no real sacrifices, the model would hardly yield any interesting insights. (For the detailed game theoretical derivation using trigger strategies see Back and Flache 2006.) We refer to this parameter setting as the baseline condition. After obtaining results for the baseline condition, we conduct experiments in which we systematically vary the level of environmental harshness (f_d). Furthermore, we run additional experiments with varying parameter combinations to test the sensitivity of the results to variation in model parameters.

Stability

In support of conjecture 1, our simulation results show that strategies keep changing endlessly in all initial parameter settings – we found no infinitely stable strategy in the DED. We simulated 175 runs altogether, each of which started with a different randomly chosen initial strategy and consisted of maximally 10,000,000 rounds. During these runs more than 4.7 million mutations took place altogether, generating as many strategies. However, in none of these runs have we recorded any strategy that existed longer than 220,000 rounds.

We may of course simply have not encountered the infinitely stable strategies during our random walks in this vast strategy space. However, judging by the vast coverage of the strategy space by our method, this seems implausible.

A plausible explanation for the lack of infinitely stable strategies is that for each strategy there exists a better response that takes advantage of the strategy's weakness. Sooner or later mutations generate this better response and the original strategy is gradually pushed out of existence. If a strategy is too cooperative, opportunistic exploiters take advantage of this and flourish. Later, in a harsh world of mainly exploiters, where everybody is suffering, two cooperators who appear randomly at the same time and find each other will survive and reproduce more easily than others, given that they have a sound method of excluding defectors from cooperative interactions.

That cooperativeness eventually declines again may be explained by the gradual loss of the ability to exclude defectors through 'evolutionary drift' (e.g. Bendor and Swistak 2001) or by the emergence of new defecting mutants who have developed the ability to behave such that they are not excluded from exchanges between incumbent cooperators. Figure 1 illustrates these dynamics of average age at death and helping behavior, for a typical simulation run. The upper part of the figure shows how the average age at death (measured in interactions) changes over time within one simulation round. Compare this with the level of cooperation, generated for the same simulation run, in the lower part of the figure: periods of high refusal rates coincide with short lives.

The importance of interpersonal commitment

Conjectures 2 and 3 relate the strength of the commitment preference within a strategy directly to viability, the average length of agents' life

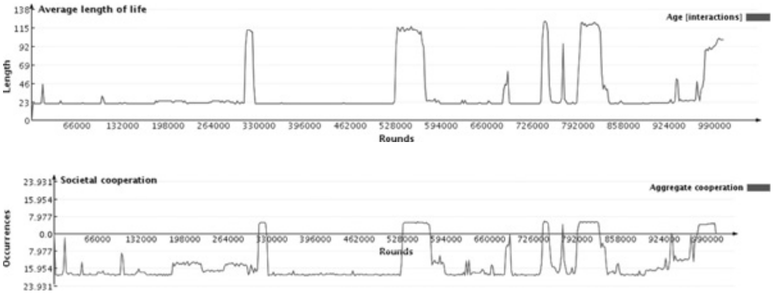

Figure 1. Age at death and cooperation in the baseline condition (single run, initial 1 million rounds)

within a strategy. Commitment is measured by the *comm*^{G} and *comm*^{A} traits, distinguished for giving and asking respectively. The higher these traits are, within the [−1;1] interval, the more an agent is inclined to choose and cooperate with long-term interaction partners. If they are positive, the agent has a preference for commitment; if they are negative, the agent has a preference against being committed; and when the values are close to zero, the agent is indifferent to the concept of commitment.

What we find is that among the most stable 1355 strategies (where stability is at least 50,000 rounds), 776 strategies (57.3%) are positive on both commitment traits. Among the same strategies, only 385 (28.4%) are positive on both *fair*^{G} and *fair*^{A}, and 717 (52.9%) on *coop*. This suggests that if a strategy is highly stable, its decision-making process is likely to be guided by preferences for unconditional cooperation with old interaction partners. These preferences appear to be far more important for success than being fair or simply being cooperative (*coop* trait).

To get a closer insight into the separate contributions of the traits to a strategy's success and to compare in particular the relative importance of the *fair*^{G} and *fair*^{A} traits to the importance of the *comm*^{G} and *comm*^{A} traits, we conducted a linear regression analysis with average longevity within a strategy as the dependent variable (see Table 1). Before performing the analysis we filtered out highly unstable strategies (STAB < 2000 rounds) and strategies with low longevity (longevity < 75 interactions). The reason for filtering out highly unstable strategies is that, due to the stochastic nature of the simulation, it often happens that strategies that would otherwise be stable cannot grow to a critical mass in the population to stabilize. In this case, they distort the association between strategy features and viability. The reduced sample consisted of 34,143 strategies.

We estimate three models (see Table 1), gradually extending the set of independent variables included. With the first model we test the

Table 1. OLS regression with dependent variable longevity

	Model 1		*Model 2*		*Model 3*	
	Unstandardized coefficients		Unstandardized coefficients		Unstandardized coefficients	
	B	*Std. Error*	*B*	*Std. Error*	*B*	*Std. Error*
(Constant)	97.035**	(0.088)	97.325*	(0.233)	97.510**	(0.230)
$comm^G$	1.754**	(0.128)	1.272**	(0.144)	1.136**	(0.144)
$fair^G$	1.393*	(0.134)	0.143	(0.165)	0.043	(0.162)
coop	5.625**	(0.126)	5.752**	(0.134)	5.770**	(0.132)
$comm^A$	–0.754**	(0.129)	–0.651**	(0.129)	–0.560**	(0.128)
$fair^A$	–4.340**	(0.135)	–4.556**	(0.136)	–3.798**	(0.136)
u^t	–6.798**	(0.051)	–6.920**	(0.076)	–6.963**	(0.075)
f_d			0.038	(0.019)	–0.006	(0.018)
$comm^G \times f_d$			–0.456**	(0.068)	–0.507**	(0.068)
$fair^G \times f_d$			–1.106**	(0.076)	–1.284**	(0.075)
$comm^G \times comm^A$					0.639**	(0.074)
$fair^G \times fair^A$					–2.415**	(0.076)
	R	**R^2-adj.**	**R**	**R^2-adj.**	**R**	**R^2-adj.**
	0.670	0.448	0.673	0.452	0.685	0.469

Note:** Significant at the $p < 0.001$ level.

effects of the strength of preferences on viability. The second model adds environmental harshness and its interaction effects; while the third model adds interaction effects between similar preferences. In the first model we see that the effects of having preferences for commitment and for fairness in giving are both positive but the effect of the commitment preference is larger, as expected based on conjectures 2 and 3. We also see that the cooperation preference has a very large effect. Preferences in asking are negative but for fairness the coefficient is much larger in absolute value. This suggests that strategies that restrict their partner search too much either to old partners or to partners with balanced exchange ratios are disadvantaged because their search space is overly reduced. The attractiveness threshold u^t has a very large negative effect. The explanation is that u^t is very important in deciding whether a strategy is initially cooperating or defecting (niceness). This in turn is crucial for the ability to bind future helping partners or establish mutually cooperative balanced exchange relationships.

In Model 2 we include the main and interaction effects of environmental harshness (f_d, cost of not getting help). To test the sensitivity of our result to the choice of environmental harshness, we repeated our simulations for a range of parameters. Namely, we were interested in the effect of variation in the proportion of the cost of cooperation and the cost of being cheated. We reran the simulation with $f_d = 5, 10, 20, 30$ and added these repetitions to the original dataset obtained in the baseline condition. Then we tested for significant effects of f_d on the dependent variable *LONGEVITY*.

What we see is that although its main effect is not significant, there are negative interaction effects with $fair^G$ and $comm^G$ but with $fair^G$ the effect is much larger. The negative interaction with $comm^G$ is clearly inconsistent with conjecture 4 and thus with the results previous studies (e.g. de Vos et al. 2001) reported for models without mutation. At the same time harshness is less of a problem for commitment players than for fairness players. While the net effect of $comm^G$ remains positive (including the main effect), $fair^G$ no longer has a significant main effect. One interpretation is that whatever beneficial effect a preference for fairness may have, the effect is strongly mitigated when the environment is harsh. This is consistent with the explanation that Back and Flache (2006) gave for the weaker performance of fairness strategies in their experiments. Fairness players tend to avoid unbalanced exchange accounts by spreading their help requests across a large number of potential partners. However, the harsher the environment, the more likely it is that help requests are directed by multiple help seekers at the same target. Accordingly, in harsh environments fairness players are likely to lose more points than commitment players, who coordinate their requests in a more efficient way.

Finally, in Model 3 we control for the interaction effects between being committed in both giving and asking and being fair in both giving and asking. This helps us understand whether having consistent preferences for both giving and asking has an impact on the dependent variable. While for commitment the effect is relatively small but positive, the effect for fairness is much larger and is in a negative direction. In addition, the interaction effect between the commitment traits slightly outweighs the negative effect of $comm^A$. This suggests that, in terms of viability, it is disadvantageous to be committed in giving but not in asking (or vice versa) but such a mismatch in fairness preferences is nonetheless beneficial. This intuitively makes sense: if you have friends you usually ask for help, you also want to give them help when they ask, and vice versa.

5. Discussion and Conclusions

In this article we examined the arms race between two tactics for cooperation under evolutionary pressures. One of them is conditional cooperation or reciprocity (cf. Axelrod 1984), the other one is commitment. The simple idea behind conditional cooperation is this: 'Be cooperative but retaliate against those who cheated on you before.' Since Axelrod (1984), conditions that may trigger retaliation (defection) have been refined and sophisticated, adapting conditional cooperation to various challenges, such as asymmetric uncertainty and random noise.

In contrast with conditional cooperation, commitment is based on a very different idea: 'Be generally cooperative but always favor long-term exchange partners.' Thus, the main question for commitment is not to decide whether to cooperate or defect but to select exchange partners. At first this seems to make commitment excessively cooperative, and vulnerable to exploitation. In a large enough interdependent population, however, partner selection substitutes the need for explicit punishment.

Previous computational modeling work (Back and Flache 2006; de Vos et al. 2001; de Vos and Zeggelink 1997; Zeggelink et al. 2000) has pointed to the evolutionary advantages of commitment under conditions resembling the human ancestral environment. In this article, we reported a stricter test of the underlying evolutionary explanation. Unlike previous work, our computational study allowed for the random mutation of competing strategies and thus implemented a much tougher evolutionary selection to which both fairness and commitment strategies were exposed. Under this stricter test, our results still point to certain evolutionary advantages of interpersonal commitment but the findings also highlight weaknesses of commitment and put the results of earlier research into perspective. We found that, under the postulated conditions of the ancestral environment, traits both for commitment and for fairness in giving increase viability, and, as expected, helping old interaction partners (commitment) was more important than helping in a fair, reciprocal way. At the same time, we found that tendencies for both commitment and fairness have negative effects when it comes to seeking help. This suggests that previous studies may have overemphasized the evolutionary advantages of commitment. We found that it is beneficial for agents to bind potential partners (commitment in helping) but it is disadvantageous to restrict the search for help too much to these partners (commitment in seeking). In a similar vein, our results are also inconsistent with the argument of de Vos et al. (2001) that environmental harshness strengthens the effects of commitment. Instead, we found that the positive effects of commitment on survival weaken when the environment becomes harsher. However, we could show that commitment players are less affected by harshness than are fairness players.

Our study has both supported and refined evolutionary accounts of interpersonal commitment. At the same time, this previous work has its limitations, some of which we believe do not affect the central conclusions, while others point to a need for future research. Previous theoretical work may suggest, in particular, that the viability of commitment is seriously hampered by the 'dyadic' nature of this strategy. Bendor and Swistak (2001) have shown that dyadic strategies (strategies that only sanction defections that cause harm to the sanctioner) can never be evolutionarily stable, while 'social strategies' that also sanction non-cooperation between third parties are stable. In a nutshell, the reason is that social strategies leave no room for benefiting from second-order free riding because a second-order free-rider would be punished by every compliant group member. However, we believe that this is not a serious problem for our theory of (dyadic) commitment. While dyadic strategies are not eternally stable, it has also been shown that reciprocal dyadic strategies (including commitment) can be relatively more stable (but not perfectly stable) compared to non-reciprocal strategies. At the same time, Bendor and Swistak (2001) do not deny that social strategies impose a higher burden of cognitive complexity and information gathering on agents than do dyadic strategies. To the extent that this creates fitness costs, the advantage of social strategies may turn into a disadvantage. Moreover, social strategies may be relatively more vulnerable to environmental uncertainty and noise, because 'erroneous' defections may disrupt more relationships than just the dyad in which they occurred. In sum, while – consistently with our results – Bendor and Swistak's argument implies that the dyadic strategy of commitment is not eternally stable, it is plausible to assume that at least under uncertainty conditions it also has some fitness advantages as compared to social strategies.

A more obvious limitation of our work is the lack of a direct empirical test for the existence of a commitment trait in contemporary societies. To be sure, while we presented a theoretical argument for a preference for building committed relationships, this work was motivated by laboratory research that showed that commitment in exchange is positively related to uncertainty (Kollock 1994). Moreover, it has been demonstrated that people attach positive feelings to the mere existence of long-term exchange relationships, in addition to the material benefits that result from them (Lawler 2001; Lawler and Yoon 1993). In a similar vein, Smaniotto (2004) showed, using scenario experiments, that subjects are more willing to provide help and to report emotions of commitment, if a scenario provides ‘commitment cues’ such as another person being in need, or being a friend. More recently, neurobiology is turning its interest to uncovering traits and mechanisms underlying human sociability and affiliation. Kosfeld et al. (2005) managed to artificially increase the level of trust, a key element in committed relationships, by administering oxytocin, a hormone that acts as a neurotransmitter in the brain, to participants of an experiment. Animal research on mammals suggests that social bonding can be modulated by various hormones, including oxytocin, vasopressin, opioids, corticotropin-releasing hormone, dopamine and adrenal steroids, including corticosterone or cortisol (c.f Carter 2005). Even more to the point, Depue and Morrone-Strupinsky (2005) provide support for the existence of a neurobiological system in humans that regulates reward via opiate functioning when people create and dissolve social bonds.

But all these findings do not give sufficient insight yet into the underlying mechanism or trait for interpersonal commitment. In particular they do not allow us to disentangle conclusively rational commitment and our indirect evolutionary explanation that posits ‘irrational’ emotions as a proximate mechanism driving commitment. It is of course impossible to empirically test an ultimate (evolutionary) explanation for commitment. But future work should devise tests, such as laboratory experiments, that allow us to rule out rival hypotheses derived from competing proximate explanations for commitment. In order to test whether a positive feeling, i.e. a preference for commitment, exists as a possible evolutionary remnant in contemporary populations, it is of key importance to find support for at least two hypotheses. The first is that the preference for commitment is stable across situations with varying materialistic payoffs, i.e. people behave according to the preference even when this is not in their rational self-interest. And the second is that the preference is stable across different cultures.

To conclude, there is reason to believe that humans may have been selected for some form of commitment behavior in their evolutionary past. One possible explanation for the success of commitment that we offer is the following. Both conditional cooperation and commitment have a tendency to cooperate, which is the only recipe for success under conditions of high interdependence and uncertainty. They both have a method to exclude defectors from the bliss of cooperative interactions: conditional cooperation retaliates against defectors, while commitment leaves them for better partners. This means that commitment does not purposefully defect when it is able to help. Or, from another perspective, while conditional cooperation operates by punishment, commitment operates by reward.

Appendix: Pseudocode

The evolutionary process, executed at the end of each run:

```
...
end of round
begin evolutionary process
   for each dead agent A
    choose with a probability equal to its fitness
      share within N-old agents an agent B who is
      alive and is at least N-old
    generate a new agent C with a strategy identical
      to that of B
    mutate the strategy of C with a probability Pmut
   end for
end evolutionary process
start of new round
...
```

NOTES

1. Interactions take place always between exactly two agents. Possible interactions are giving help (cooperation) and refusing to help (defection). Asking for help is always followed by one of these.
2. This is unlikely, as the preference parameters are high-precision real values and interaction histories tend to differ with time.
3. We will not use here stability concepts from the evolutionary game theory literature (e.g., evolutionary stability or asymptotic stability) because they do not allow the expression of the relative stability of strategies.
4. Extinction is possible if all agents die within one round and thus there is no basis for the distribution of strategies in the next generation.
5. SOCCOOP measures the difference between the per-round average number of cooperation (helps) and defection (refusals) in the entire population.

REFERENCES

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

Back, I. and Flache, A. 2006. ‘The Viability of Cooperation Based on Interpersonal Commitment.’ *Journal of Artificial Societies and Social Simulation* 9(1). http://www.soc.surrey.ac.uk./JASS/1/1/review1.html

Bendor, J. and P. Swistak 2001. ‘The Evolution of Norms.’ *American Journal of Sociology* 106: 1493–545.

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://www.soc.surrey.ac.uk./JASSS/1/1/review1.html.

Bolton, G. and A. Ockenfels. 2000. 'ERC – A Theory of Equity, Reciprocity and Competition.' *American Economic Review* 90(1): 166–93.

Buskens, V. and W. Raub. 2002. 'Embedded Trust: Control and Learning.' *Advances in Group Processes* 19: 167–202.

Carter, C. 2005. 'Biological Perspectives on Social Attachment and Bonding.' In *Attachment and Bonding: A New Synthesis*, eds C. Carter, L. Ahnert, K. Grossmann, S. Hrdy, S. Porges and N. Sachser, pp. 85–100. Cambridge, MA: MIT Press.

Dawes, R. 1980. 'Social Dilemmas.' *Annual Review of Psychology* 31: 169–93.

deVos, H., R. Smaniotto and D. A. Elsas. 2001. 'Reciprocal Altruism under Conditions of Partner Selection.' *Rationality and Society* 13(2): 139–83.

de Vos, H. and E. P. H. Zeggelink. 1997. 'Reciprocal Altruism in Human Social Evolution: The Viability of Reciprocal Altruism with a Preference for "old-helping-partners".' *Evolution and Human Behavior* 18: 261–78.

Depue, R. A. and J. V. Morrone-Strupinsky. 2005. 'A Neurobehavioral Model of Affiliative Bonding: Implications for Conceptualizing a Human Trait of Affiliation.' *Behavioral Brain Sciences* 28: 313–95.

Fehr, E. and S. Gächter. 2002. 'Altruistic Punishment in Humans.' *Nature* 415: 137–140.

Frank, R. H. 1988. *Passions within Reason*. New York: W. W. Norton & Company.

Friedman, J. 1971. 'A Non-cooperative Equilibrium for Supergames.' *Review of Economic Studies* 38: 1–12.

Gintis, H. 2003. 'Solving the Puzzle of Prosociality.' *Rationality and Society* 15(2): 155–87.

Güth, W. and H. Kliemt. 1998. 'The Indirect Evolutionary Approach.' *Rationality and Society* 10(3): 377–99.

Güth, W. and A. Ockenfels. 2002. 'The Coevolution of Morality and Legal Institutions – An Indirect Evolutionary Approach.' Discussion Papers on Strategic Interaction 2002–06, Max Planck Institute of Economics, Strategic Interaction Group. Available at http://ideas.repec.org/p/esi/discus/2002-06.html.

Hayashi, N. and T. Yamagishi. 1998. 'Selective Play: Choosing Partners in an Uncertain World.' *Personality and Social Psychology Review* 2: 276–89.

Karremans, J., P. V. Lange, J. Ouwerkerk and E. Kluwer. 2003. 'When Forgiving Enhances Psychological Well-being: The Role of Interpersonal Commitment.' *Journal of Personality and Social Psychology* 84(5): 1011–26.

Kollock, P. 1994. 'The Emergence of Exchange Structures: An Experimental Study of Uncertainty, Commitment, and Trust'. *American Journal of Sociology* 100(2): 313–45.

Kosfeld, M., M. Heinrichs, P. Zak, U. Fischbacher and E. Fehr. 2005. 'Oxytocin Increases Trust in Humans.' *Nature* 435(2): 673–6.

Lawler, E. J. 2001. 'An Affect Theory of Social Exchange.' *American Journal of Sociology* 107(2): 321–52.

Lawler, E. and J. Yoon. 1993. 'Power and the Emergence of Commitment Behavior in Negotiated Exchange.' *American Sociological Review* 58(4): 465–81.

Lawler, E. and J. Yoon. 1996. 'Commitment in Exchange Relations: Test of a Theory of Relational Cohesion.' *American Sociological Review* 61(1): 89–108.

Raub, W. 2004. 'Hostage Posting as a Mechanism of Trust: Binding, Compensation, and Signaling.' *Rationality and Society* 16: 319–65.

Rusbult, C. and J. Martz. 1995. 'Remaining in an Abusive Relationship – An Investment Model Analysis of Nonvoluntary Dependence.' *Personality and Social Psychology Bulletin* 21(6): 558–71.

Rusbult, C., J. Martz and C. Agnew. 1998. 'The Investment Model Scale: Measuring Commitment Level, Satisfaction Level, Quality of Alternatives, and Investment Size.' *Personal Relationships* 5(4): 357–91.

Schüssler, R. 1989. 'Exit Threats and Cooperation under Anonymity.' *The Journal of Conflict Resolution* 33: 728–49.

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. Available online at http://irs.ub.rug.nl/ppn/269506959.

Taylor, P. D. and L. B. Jonker. 1978. 'Evolutionary Stable Strategies and Game Dynamics.' *Mathematical Biosciences* 40: 145–56.

Wieselquist, J., C. Rusbult, C. Agnew, C. Foster and C. Agnew. 1999. 'Commitment, Pro-relationship Behavior, and Trust in Close Relationships.' *Journal of Personality and Social Psychology*, 77(5): 942–66.

Zeggelink, E., H. de Vos and D. Elsas. 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://www.soc.surrey.ac.uk/JASSS/3/3/1.html.

ISTVÁN BACK received his master's degree in computer science in 2003 from the Budapest University of Technology and Economics, Hungary. He finished his PhD at the Interuniversity Center for Social Science Theory and Methodology (ICS) at the University of Groningen, the Netherlands. His main research interest lies in explaining interpersonal commitment behavior using evolutionary models and laboratory experiments.

ADDRESS: Interuniversity Center for Social Science Theory and Methodology (ICS), University of Groningen, Grote Rozenstraat 31, 9712 TC Groningen, The Netherlands [email: istvan@back.hu].

ANDREAS FLACHE is currently an associate professor for methodology in the Sociology Department of the University of Groningen, the Netherlands. He applies computational and game theoretical modeling, laboratory experimentation and survey research. His main research interests are cooperation and social integration, the dynamics of social networks and social control, and the relationships between these phenomena. He has published his work recently in, for example, *JASSS, Journal of Mathematical Sociology, Journal of Conflict Resolution, Rationality and Society,* and *Proceedings of the National Academy of Sciences.*

ADDRESS: Sociology Department, University of Groningen, The Netherlands [email: a.flache@rug.nl].