Carrot or Stick:
The Evolution of Reciprocal Preferences in a Haystack Model*

Florian Herold†

This version: February 5, 2010

Abstract

This paper studies the evolution and co-evolution of both characteristics of reciprocity - the willingness to reward friendly behavior and the willingness to punish hostile behavior. Firstly, both preferences for rewarding and preferences for punishing can survive in evolution provided individuals interact within separate groups. This holds even with randomly formed groups and even when individual preferences are unobservable, provided players adjust their play to the environment in their group. Secondly, preferences for rewarding survive only in coexistence with self-interested preferences, but preferences for punishing either vanish or dominate the population entirely. Thirdly, the evolution of preferences for rewarding and the evolution of preferences for punishing influence each other decisively. Rewarders can invade a population of self-interested players. The existence of rewarders enhances the evolutionary success of punishers, who then crowd out all other preference types.

JEL-Classification: C72, D63, D64, D83

Keywords: Evolution of Preferences, Reciprocity, Group Selection

1 Introduction

This paper addresses three questions concerning the evolution and co-evolution of the two characteristics of reciprocity - the willingness to reward friendly behavior and the willingness to punish hostile behavior. 1) How can preferences for rewarding, and preferences for punishing, survive the evolutionary competition with purely self-interested preferences? 2) What structural differences distinguish the evolution of the willingness to reward from the evolution of the willingness to punish? 3) How is the evolution of one side of reciprocity influenced by the evolution of the other side?

*I am very grateful to Theodore C. Bergstrom, Eddie Dekel, Florian Englmaier, Robert Evans, Ernst Fehr, Georg Gebhardt, Herbert Gintis, Steffen Huck, Christoph Kuzmics, George J. Mailath, Sten Nyberg, Pedro Rey Biel, Arthur Robson, Larry Samuelson, Klaus M. Schmidt, Ferdinand von Siemens, and participants at several conferences and seminars for very helpful comments and discussions. Funding by the EDGE-Program and the Marie-Curie-Fellowship HPMT-CT-2000-00056 are gratefully acknowledged.

†Contacts: Kellogg School of Management, Northwestern University, 2001 Sheridan Rd., Evanston, IL 60208, Tel.: ++1 857 491-5305, email: f-herold@northwestern.edu

Self-interested preferences are a standard assumption in economic theory. Yet, several experimental studies offer substantial evidence that at least some people are not exclusively driven by self-interest. A significant number of subjects are willing to reward friendly and/or to punish hostile behavior of an opponent even if this is costly and does not maximize their own material payoffs.[1] From an evolutionary standpoint this seems surprising since purely self-interested preferences induce a player to maximize his fitness, which seems optimal for survival. In this paper, we identify conditions under which preferences for rewarding friendly behavior and for punishing unfriendly behavior can survive the evolutionary process. We consider a model in which groups of a relative small size, also called "haystacks", are randomly formed from the infinite player population.[2] Players interact within these groups and adjust their behavior to the distribution of preferences within their group. We show that in such a model rewarders or punishers can survive the evolutionary competition with purely self-interested players. This holds even if individual preferences are unobservable, groups are formed randomly, and players interact anonymously in random pairings within their groups.

We find important structural differences between the evolution of preferences for rewarding and the evolution of preferences for punishing. Rewarders successfully invade a population of self-interested players, but they cannot drive them out completely. Rewarders survive only in coexistence with self-interested types. Punishers on the other hand either drive out other preferences and dominate the population, or they must disappear completely. The option to punish hostile behavior results either in a "culture of punishment" - where all players are willing to punish hostile behavior - or in a "culture of laissez faire" - where nobody is willing to incur the costs of punishing.

If the option to reward friendly behavior and the option to punish hostile actions are both available the evolution of rewarders and the evolution of punisher influence each other in a very interesting way. Rewarders can serve as a catalyst for the evolution of punishers. Rewarders invade a population of self-interested types. Their presence can then enable punishers to invade successfully, and finally to crowd out, both self-interested and rewarding types. Hence the option to reward friendly actions can crucially influence the equilibrium outcome even if in equilibrium nobody uses this option.

To see the gist of the argument consider pairwise interactions of the following structure: a first moving player, the proposer, may either cooperate or defect.[3] Cooperation is costly for the proposer but profitable for a second moving player, the responder. This responder observes the proposer's action and can then, at a personal cost, reward and/or punish the proposer or

[1] For a survey of the experimental literature see Fehr and Gächter [10] or Fehr and Schmidt[11].

[2] The expression "Haystack Model" for settings where players interact only within randomly formed groups which are reshuffled after reproduction goes back to Maynard Smith's [28] example of mice living and replicating over the summer within separate haystacks. At harvest time when the haystacks are cleared mice scramble out into the meadow and mix up completely before colonizing new haystacks in the next summer.

[3] We use the terms "cooperate" and "defect" to indicate an friendly or unfriendly act. The game differs from the prisoners' dilemma.

remain inactive. In Lemma 1 we show that in the role of the proposer it is indeed optimal to have self-interested preference which induce fitness maximizing behavior. This holds for the haystack model as well as for the model without group structure. It is the role of the responder in which the group structure changes the results of the evolutionary process.

In the model without group structure rewarding behavior can not survive evolution and cooperation disappears. While rewarders would be good for the overall population, they have to bear the fitness costs of rewarding whenever cooperation happens, while the fruits of the induced cooperation go equally to rewarders and self-interested players that do not reward. Thus, the proportion of rewarders must eventually fall below the level which is necessary to induce cooperation. There might still be a small fraction of responders with preferences to reward cooperation, but we see no rewarding since there is no cooperation that could be rewarded. In addition, if proposers tremble with any arbitrarily small probability and cooperate occasionally when it is not optimal then rewarders must at times bear the costs of rewarding and disappear completely. Without any group structure punishers have similar difficulties to survive the evolutionary process. When the proportion of punishers is sufficiently large there is an equilibrium in which proposers cooperate to avoid punishment and punishers do equally well as self-interested responders since in this equilibrium they do not have to punish. Yet, this equilibrium is not asymptotically stable. If a small shift towards a lower proportion of punishers occurs, then there is no evolutionary force bringing the state back. Once such small changes happen to lead to a state with such a small proportion of punishers that proposers switch to defection, costly punishments occur, and punishers die out. Note that the state without punishers (and no cooperation) is asymptotically stable and strong evolutionary forces reverse any small random shift in the population state. Furthermore, if proposers tremble with any arbitrarily small probability punishers perform always strictly worse than self-interested responders and evolution leads to a population without punishers and without cooperative behavior.[4]

In the haystack model, in contrast, rewarders and punishers can survive and successfully induce cooperation. The key difference is that in the haystack model proposers adjust their behavior to the composition of the group they happen to play in. Within groups rewarders and punishers do still perform (weakly) worse than self-interested responders, but across groups this is no longer true. In fact, cooperation occurs in groups with a sufficiently high number of rewarders or punishers and everybody in such a group obtains higher fitness than players in uncooperative groups. Rewarders and punishers are more likely to play in these cooperative groups because cooperative groups are precisely those who happen to have a high number of rewarders or punishers. Thus rewarders and punishers are more likely to profit from being in

[4]One caveat may be in place, though. If there is a force of evolutionary drift towards states with equal proportions of punishers and non-punishers and if a relative small proportion of punishers is sufficient to induce cooperation, than this force of drift may suffice to stabilize a state with enough punishers to establish cooperation. See Samuelson [34], chapter 5 and 6 for details. Note though that when the drift is small, the forces stabilizing such a cooperative state with punishers are small, while the forces stabilizing the equilibrium with defecting proposers and no punishers remain strong for small drift.

a cooperative group. Put differently, responders have a marginal effect on the distribution of preferences in their group. Having reciprocal preferences leads to a fitness advantage for the responder when he is pivotal in his group, i.e. his type is decisive for whether the number of reciprocal preferences in his group is just above or just below the threshold for cooperation. This marginal effect is advantageous for a reciprocator and can outweigh the costs of rewarding or punishing.

The structural difference between the evolutionary dynamics for rewarders and the one for punishers is driven by the following. The hope for reward as well as the fear of punishment can induce proposers to cooperate. Yet, when most proposers cooperate it is relatively expensive for responders to reward cooperation whereas the willingness to punish is almost for free. On the other hand, when most proposers defect, the willingness to reward is almost for free, whereas it is expensive to punish defection. A higher fraction of rewarders or punishers leads to a higher fraction of groups in which cooperation occurs. Therefore, rewarders are relatively successful when most responders are self-interested, whereas punishers become more successful when more responders have preferences for rewarding or punishing.

In this paper we considerably relax the observability assumptions typically made in the related literature (compare Section 4). It is crucial for our results, however, that players adapt their behavior to the type composition of their group. In the main section we simply assume that players know the distribution of preferences in their respective groups and react optimally. Would players really learn to do so if they have to learn only from past play? An important feature of this paper is that in Section 3 we consider several key robustness checks of our model, and in particular in Subsection 3.2 we show that we can replace the assumption of an observable group distribution by a simple process of learning by valuation. Our results require only that players learn to play 'as if' they would know the distribution of preferences in their group. We think of equilibrium play as the result of a fast operating learning process, fast compared to the evolution of preferences and fast compared to the frequency of group reshuffling.[5] The learning process does not necessarily need to be fully rational, but should eventually lead to beliefs close to the true distribution of play. Here, as usual in the literature using the indirect evolutionary approach, we assume that the game G is played recurrently, i.e. players act rationally according to their subjective preferences only in a myopic sense. We see the evolution of preferences as a short cut description for a situation where players use somewhat sophisticated, but not too complex heuristics to adapt their behavior. More specifically, in Subsection 3.2 we drop Assumption 2 and consider instead a simple process of learning by valuation. Subjective utilities serve to evaluate outcomes and players choose with high probability the action which lead to the highest average subjective payoff in the past. We show that players' behavior and payoffs are arbitrarily close to the equilibrium play derived under Assumption 2. On the other hand we consider the evolution of preferences to be a very slow process, relative to the fast learning

[5]For a similar interpretation of modeling the evolution of preferences see Ely and Yilankaya [8], p.5. See also Sandholm [36] for an explicit model of such a two speed dynamics.

process and relative to the reshuffling of groups.[6] In Subsection 3.3 we show that for a sufficiently slow speed of the evolution of preferences within-group evolutionary effects are small and do not significantly change our results.

The remainder of the paper is organized as follows: Section 2 presents the model and analyzes three naturally arising settings: 1) Player 2 might have only the costly option of rewarding cooperation. 2) Player 2 might have only the costly option of punishing defection. 3) Player 2 might have both the options - rewarding cooperation and punishing defection. Section 3 discusses the robustness of our results with respect to small trembles, replacing the assumption of an observable group composition by a learning process, and with respect to within-group evolutionary effects. Section 4 discusses the related literature. Section 5 concludes.

2 The Model

We describe the evolution of preferences by adopting the indirect evolutionary approach, pioneered by Güth and Yaari [16]. A game G is played recurrently. Players can differ in their (subjective) preferences over the outcomes of G. Preferences determine players' behavior, behavior in turn determines the (objective) material payoffs, or fitness. Fitness regulates the evolutionary success, i.e. the future occurrence of each preference type.

Let G be a two-player extensive form game of perfect information structure. Both players move exactly once. The first moving player 1, the proposer, can either cooperate (C) or defect (D). We use these terms to indicate that cooperation enhances the fitness of the second moving player 2, the responder, but requires the proposer to incur some fitness costs. This feature is analog to the (sequential) prisoners dilemma, yet the responder's options differ. Depending on the setting he can reward, punish or abstain from doing so after observing the proposer's move (see below). Let X denote the set of decision nodes and O the set of terminal nodes. We call a terminal node $o \in O$ an outcome of the game G. The material payoff function $\pi : O \to \mathbb{R}^2$ represents the fitness players obtain after a certain path of play.

Players may not maximize their material payoff and can differ in their subjective preferences. We describe preferences by (subjective) von Neumann-Morgenstern utilities over outcomes of G. These utilities can depend on the role a player has in the game. Thus each player assigns to each outcome of the game two subjective utilities, one is his utility if he happens to play in the role of the proposer and the second, potentially different, utility is relevant if he is in the role of the responder. Let there be an infinite population of players.[7]

[6] Players could for instance have fixed preferences during their entire live and with high probability pass on their preferences to their offsprings. To avoid the intricacies of sexual reproduction we can e.g. consider a cultural process in which daughters learn to evaluate outcomes from their mother and sons learn their values from their father. Such a very slow evolutionary process might even suggest that our preferences for reciprocity were mostly formed by the conditions of a hunter-gatherer societies which lived in relatively small bands. This is consistent with our haystack model. Reshuffling of groups might perhaps occur a few times during a player's lifetime, and learning to adjust behavior to a new group could be a matter of weeks.

[7] All results can be derived equally in a model of two infinite populations, one proposer population and one

$\Theta_i \subset [0,1]^{|O|}$ (modulus affine transformations) denotes the set of all possible preference types a player can have in player position $i \in \{1,2\}$. For $\theta_i \in \Theta_i$, $\theta_i(o)$ represents the subjective utility that a player in position i of type θ_i assigns to an outcome $o \in O$. Let $\Theta \equiv \Theta_1 \times \Theta_2$ be the set of all preference types. We restrict our analysis to finite sets of preference types and we assume that each preference type in each role has strict preferences over outcomes. Let the simplex Γ denote the set of all probability distributions over Θ and $\gamma \in \Gamma$ represents a population state. Then γ_θ represents the proportion of type $\theta \in \Theta$ in the total population. Furthermore, we assume that $\frac{d_1-c_1}{r}$ and $\frac{d_1-c_1}{p}$ are irrational numbers, which is generically true.[8]

A random process recurrently matches individuals to play G. We will contrast the most common setting in the literature of a single infinite population without any group structure in which players, randomly drawn from the entire population, are matched to play G, with our haystack model, in which matching takes place only within randomly drawn groups of finite size. Each of the infinitely many groups has $2N$ members, N players are randomly drawn from the total population to play in position 1 and N players are randomly drawn from the population to play in position 2. Once in a group players keep their role until the group is resolved.[9]

For a specific group let k_{θ_i} be the number of players of type $\theta_i \in \Theta_i$ in player position i. Thus for any $i \in \{1,2\}$ holds $\sum_{\theta_i \in \Theta_i} k_{\theta_i} = N$. Let $\mathbf{k_i}$ be the vector of the k_{θ_i} and thus $\mathbf{k} \equiv (\mathbf{k_1}, \mathbf{k_2})$ is a vector representing the entire preference type composition of a group.

Assumption 1 Distribution of Groups in the Haystack Model *Each group member in each player position $i \in \{1,2\}$ is drawn independently from the infinite player population with the probability of drawing type $\theta_i \in \Theta_i$ equal to the proportion γ_{θ_i} of that type in the population. Thus the probability that a group has composition $\mathbf{k}$ if the population state is γ is given by*

$$B_{\mathbf{k}}(\gamma) = \prod_{i \in \{1,2\}} \left(\frac{N!}{\prod_{\theta_i \in \Theta_i} k_{\theta_i}!} \prod_{\theta_i \in \Theta_i} (\gamma_{\theta_i})^{k_{\theta_i}} \right). \tag{1}$$

Players act optimally according to their preferences and according to their beliefs about the opponent's play. Furthermore, players have correct beliefs about the distribution of actions they will face. Thus they play perfect Bayesian equilibria of the corresponding two player game of incomplete information Γ.

In the standard model without subgroups a player in position i beliefs to face type θ_j with probability γ_{θ_j}. In our haystack model we assume also that individual preferences are unobservable, relaxing the observability assumption typically made in order to explain the survival of social preferences. Yet, we do want to capture that players adapt to their local environment,

responder population.

[8] This assumption is not essential for our results but simplifies the analysis considerably by guaranteing that proposers have a unique optimal choice of action in each group.

[9] This assumption we make only to keep notation and model description simple. In fact, we obtain exactly the same results if we assume that groups are of size $N+1$ and in each period each player is randomly assigned to his role in the interaction.

and eventually act optimally given the composition of their group.

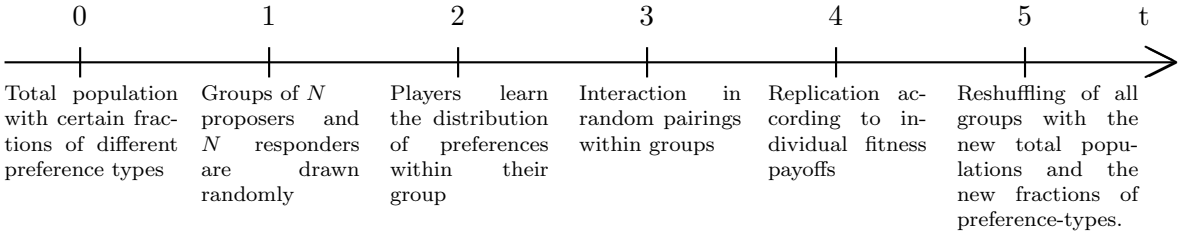
Assumption 2 *In the haystack model players' beliefs about the opponent's play corresponds to the true distribution of actions in their group.*

In Subsection 3.2 we drop Assumption 2 and consider instead a simple process of learning by valuation. Subjective utilities serve to evaluate outcomes and players choose with high probability the action which lead to the highest average subjective payoff in the past. We show that players' behavior and payoffs are arbitrarily close to the equilibrium play derived under Assumption 2.

Groups stay together for a finite number T of periods. In each period players are matched randomly with an opponent within their group and play the game G anonymously. The T periods add up to a duration of $\mathfrak{T}$, a time period assumed to be brief relative to the speed of the evolution of preferences. In particular, while a group stays together, the composition of preferences in the group does not change. Yet, we think of $\mathfrak{T}$ as a long time period relative to the speed of learning to play optimal for a given preference distribution in a given group. This is only relevant when we drop Assumption 2 in Subsection 3.2. Here, under Assumption 2, players play identically in all T periods anyway.

After T periods of interactions preferences are replicated according to received average fitness and the groups are reshuffled. A new cycle starts with the new proportions γ_θ of the different preference types in the total population. Timing of events in our model is illustrated graphically in Figure 1.

Figure 1: Timing of events



For any given population state $\gamma \in \Gamma$ and a given matching scenario we derive the equilibrium play and calculate the average fitness of a player with preference type θ, denoted $\overline{\pi}_\theta(\gamma)$.

We follow a standard approach in evolutionary game theory and describe the evolutionary dynamics by a system of differential equations.[10] A growth rate function g assigns to each population state γ and preference type θ the growth rate $g_\theta(\gamma)$ of the associated population share γ_θ. The evolution of preferences is described by the system of differential equations

$$\dot{\gamma}_\theta = g_\theta(\gamma)\gamma_\theta \qquad \forall\ \theta \in \Theta, \gamma \in \Gamma. \tag{2}$$

[10]The following presentation adapts the definitions in Weibull [42] to our setting.

Assumption 3 *The evolutionary dynamics solves the system(2) of differential equations where g is a regular, payoff monotonic growth rate function.*

Regularity ensures that the dynamics is well defined. Payoff monotonicity means that the proportion of a type with higher average fitness increases relative to types with lower fitness. This captures the process of evolutionary selection. The formal definitions are in the appendix.

2.1 Preferences in player-position 1

What preference types are relevant for our analysis? We are mainly interested in the evolution of preferences for the position of player 2 - reciprocal behavior is only possible in that position. However, the evolutionary success of preferences for player position 2 depends on the behavior of players in position 1. The following shows that we should expect fitness maximizing behavior of players in position 1.

Lemma 1 *Let $\Theta \equiv \Theta_1 \times \Theta_2$ where Θ_1 contains the purely self-interested, fitness maximizing preference type which has (in the role of the proposer) for each outcome a subjective utility that corresponds to his fitness payoff. Under Assumption 1 and Assumption 3*

(a) The fitness of any preference type is weakly dominated by the corresponding preference type which has identical preferences in player-position 2, but self-interested preference in player-position 1.

(b) A preference type who's behavior in the role of player 1 is inconsistent with fitness maximization in any group obtains in every interior population state $\gamma \in int(\Gamma)$ a strictly lower expected fitness than the corresponding preference type which has identical preferences in player-position 2, but self-interested preference in player-position 1.

(c) Any preference type (θ_1, θ_2) in the support of any stationary state $\gamma \in \Gamma$ that contains also a positive proportion of the corresponding type (θ_s, θ_2) must act consistently with fitness maximization in player-position 1 in any group occurring with positive probability in state γ.

(d) In any asymptotically-stable state $\gamma \in \Gamma$ all preference types in the support of γ must act consistently with fitness maximization in player-position 1 in all possible groups.

All proofs are in the Appendix. Lemma 1 refers to our haystack model. Analog statements are true in the model of a single infinite population without group structure. The argument in the setting without group structure is completely analogous to the proof of Lemma 1.

In terms of fitness it is therefore always optimal for proposers to have preferences which simply maximize their own expected fitness. Lemma 1 thus strongly suggests that it is sufficient

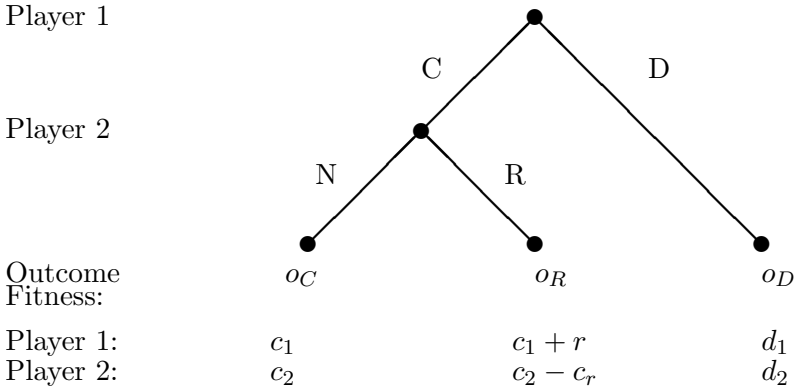
to consider fitness maximizing proposers, which simplifies our further analysis significantly.[11] In this sense Assumption 4 is rather a shortcut than an assumption which allows us to focus on the evolution of preferences for the responder position.

Assumption 4 *All proposers maximize their expected fitness.*

2.2 Setting 1: Costly Rewarding

Setting 1 concentrates on the possibility that player 2 can reward cooperation of player 1.[12] The game G is given by Figure 2. It is also known as the 'trust game'. The fitness payoffs

Figure 2: Interaction in Setting 1



with $c_1 + r > d_1 > c_1$ and $c_2 > c_2 - c_r > d_2$.

capture a social dilemma situation. The fitness obtained after cooperation and rewarding Pareto dominates the fitness after defection, since $c_1 + r > d_1$ and $c_2 - c_r > d_2$. Yet, if all players are known to maximize their fitness, the Pareto dominated outcome is played in the subgame perfect equilibrium. Self-interested responders will not reward because doing so is costly, $c_r > 0$. Anticipating such a reaction, proposers defect as $d_1 > c_1$.

In the evolutionary model we need to distinguish only between two classes of responder types. *Rewarders* have a subjective utilities $\theta_2(o_R) > \theta_2(o_C)$ and therefore reward cooperation. In contrast, *self-interested responders* have preferences with $\theta_2(o_R) < \theta_2(o_C)$ and do not reward cooperation.

In the standard model without any group structure the result of the evolutionary process is essentially the subgame perfect equilibrium of the game with fitness payoffs. The fitness

[11] One caveat may be in place. Some equilibria that are asymptotically stable under Assumption 4 may be only Lyaponov stable when allowing for more preference types in position 1.

[12] We could give player 2 an additional option to reward player 1 also after defection. It is straightforward to show that preferences for rewarding defection cannot be part of any stable equilibrium. Hence, we abstract from this possibility. So our evolutionary argument works only for rewarding conditional on cooperative behavior by the first player. Unconditional rewards (i.e. unconditional altruism) can not survive in our setting of non-assortatively formed groups.

maximizing proposers cooperate only if they believe that the probability of being rewarded is sufficiently high, i.e. if $\gamma_+ \geq \frac{d_1-c_1}{r}$, where γ_+ denotes the proportion of rewarders. Then rewarders obtain fitness of c_2-c_r, strictly less than the payoff c_2 that all self-interested responders obtain. In the payoff monotonic evolutionary dynamics the proportion of rewarders decreases until the proportion γ_+ of rewarders is so small that proposers stop cooperating and both types earn the same payoff d_2. Only population states in which no positive mass of proposers cooperates can be Lyapunov stable. Furthermore, self-interested responders do strictly better than rewarders if proposers tremble occasionally. Then only $\gamma_+ = 0$ is stable.

In our haystack model the proportion of rewarders varies across groups. The probability of a group with k_+ rewarders if the overall proportion of rewarders is γ_+ reduces to a Binomial distribution $B_{N,k_+}(\gamma_+) = \binom{N}{k_+}\gamma_+^{k_+}(1-\gamma_+)^{N-k_+}$. Proposers cooperate only in groups in which the number k_+ of rewarders is larger than $N\frac{d_1-c_1}{r}$. In these groups rewarders earn a high fitness of $(c_2 - c_r)$ and self-interested responders earn an even higher fitness of c_2. In groups with a number of rewarders $k_+ < N\frac{d_1-c_1}{r}$ proposers defect and both responder types earn only a low fitness of d_2. Let $k_+^* \equiv \left\lfloor N\frac{d_1-c_1}{r} \right\rfloor$ denote the greatest integer not greater than $(N\frac{d_1-c_1}{r})$, i.e. proposers cooperate in groups with $k_+ > k_+^*$. In every group self-interested players earn (weakly) greater payoffs than rewarders. Yet rewarders are slightly more likely to be in a group where proposers cooperate. If exactly k^* of the other $(N-1)$ responders are rewarders, a rewarder ends up in a group of cooperation, whereas a self-interested player 2 is in a defective group. The average fitness difference between rewarders and self-interested responders across all groups is given by

$$\bar{\pi}_+(\gamma) - \bar{\pi}_s(\gamma) = (c_2 - c_r - d_2)B_{N-1,k_+^*}(\gamma_+) - c_r \sum_{k_+=k_+^*+1}^{N-1} B_{N-1,k_+}(\gamma_+). \tag{3}$$

Proposition 1 (Coexistence in the unique mixed equilibrium)

Let $\frac{d_1-c_1}{r} < \frac{N-1}{N}$. Then there exists a unique stable equilibrium in which rewarders and self-interested responders coexist.[13]

Remark 1 *If $\frac{d_1-c_1}{r} > \frac{N-1}{N}$ then a monomorphic responder population of rewarders forms the unique stable equilibrium.*

For an intuitive understanding consider first a population consisting almost entirely of self-interested players. Then the vast majority of groups contain few rewarders and thus proposers defect. In these groups rewarders and self-interested responders obtain the same low payoff d_2. Yet, in a small number of groups the number of rewarders is above the threshold k_+^* and proposers

[13] We could slightly generalize this result: Take any trait that (a) when the fraction of a group possessing the trait is less than $1 < k^* < N$, those with and without the trait do equally well; (b) when the fraction is above $k*$, all agents in the groups do better, but those with the trait do worse than those without; (c) agents are randomly assigned to groups. Then there is a positive fraction of agents with the trait in equilibrium. I would like to thank Bob Evans and Herb Gintis for pointing this out.

cooperate. Every responder in these cooperative groups earns higher fitness than responders in groups without cooperation. The share of rewarders in these successful groups is at least $\frac{k_+^*+1}{N}$ and hence far above the share of rewarders in the total population (which is close to zero). Rewarders profit relatively more from these successful groups and can successfully invade a self-interested population. A small proportion of self-interested responders can also invade a population of rewarders. In most groups the number of rewarders is then above the threshold k_+^* and proposers cooperate. Rewarders gain a fitness of only (c_2-c_r) in these groups, whereas a self-interested responders avoid the costs of rewarding and earns the higher payoff of c_2. The share of self-interested responders will grow. Only in the case that $k_+^* = N-1$ self-interested responders cannot invade a population of rewarders. Even if only one self-interested responder invades a group he destroys the cooperative behavior of proposers and free-riding becomes impossible.

Proposition 1 demonstrates that in the haystack model some players cooperate and some responders reward this cooperation in the unique stable equilibrium. The group structure of the haystack model together with the option to reward cooperation results in a Pareto improvement in terms of expected fitness compared to the traditional model without subgroups. Yet, for $k_+^* < N-1$ also non-rewarding self-interested responders survive and defection will occur in some groups. The outcome is still inefficient.

The comparative statics of γ_+^{eq}, the equilibrium proportion of rewarders, follows in case of a mixed equilibrium directly from the condition $\bar{\pi}_+(\gamma_+^{eq}) - \bar{\pi}_s(\gamma_+^{eq}) = 0$, or equivalently

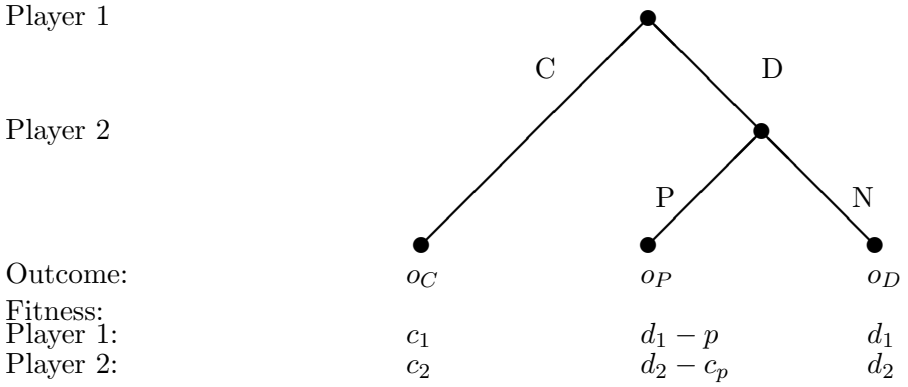
$$\left(\frac{c_2-d_2}{c_r} - 1\right) = \sum_{k=k_+^*+1}^{N-1} \frac{B_{N-1,k}(\gamma_+^{eq})}{B_{N-1,k_+^*}(\gamma_+^{eq})}. \tag{4}$$

The right-hand side is strictly increasing in γ_+ by Assumption 1. For larger values of $\frac{c_2-d_2}{c_r}$, γ_+^{eq} needs to be larger, too. So γ_+^{eq} is strictly increasing in (c_2-d_2), the gains of cooperation for the responder, and strictly decreasing in c_r, the cost of rewarding. Intuitively, higher costs of rewarding reduce the fitness of rewarders, and higher gains from cooperation for the responder increase the gain from inducing it. In addition, if $\frac{c_2-d_2}{c_r}$ goes to one, then γ_+^{eq} goes to zero and if $\frac{c_2-d_2}{c_r}$ goes to infinity, then γ_+^{eq} goes to one. This implies that the equilibrium proportion can be anywhere between zero and one, if we select the right parameters. Furthermore, the right-hand side is decreasing in k_+^*, which represents $\left\lfloor N_2 \frac{d_1-c_1}{r} \right\rfloor$. So γ_+^{eq} is weakly increasing in (d_1-c_1), the cost of cooperation for the proposer and weakly decreasing in the reward r. The larger k_+^* the more rewarders are necessary to establish cooperation. The number of self-interested responders who can free-ride without putting cooperation into danger decreases. Finally we show in the appendix that γ_+^{eq} is strictly decreasing in the group size N, if $\frac{k_+^*}{N}$ is kept constant. The last qualifier is necessary because $k_+^* \equiv \left\lfloor N_2 \frac{d_1-c_1}{r} \right\rfloor$ is only close to being proportional to N. Intuitively, larger groups reduce the probability of being pivotal. Hence, larger groups lead to a lower level of cooperation.

2.3 Setting 2: Costly Punishment

In Setting 2 responders have only the option to punish hostile behavior (i.e. defection) of the proposer.[14] This punishment is costly. The interaction is given by Figure 3.

Figure 3: Interaction in setting 2



with $c_1 + p > d_1 > c_1$, $c_2 > d_2$, and $c_p > 0$.

The payoffs $c_2 > d_2$ capture a situation in which the responder would profit from cooperation of the proposer. Yet, a fitness maximizing proposer cooperates only if he expects to be punished since $c_1 + p > d_1 > c_1$. A self-interested responder, however, cannot credibly commit to incur the costs $c_p > 0$ to punish defection.

In the evolutionary models we distinguish between punishers with subjective utilities $\theta_2(o_D) < \theta_2(o_P)$ and self-interested responders with utilities $\theta_2(o_D) > \theta_2(o_P)$. Proposers maximize their expected fitness and cooperate only if the probability of being punished for defection is sufficiently high.

In the standard evolutionary model without any group structure the only asymptotically stable state is defection, the subgame perfect equilibrium of the game with fitness payoffs. The probability of being punished corresponds to the proportion γ_- of punishers in the responder population. Proposers defect only if $\gamma_- \leq \frac{d_1 - c_1}{p}$. Then punishers gain fitness of $(d_2 - c_p)$, strictly less than the fitness d_2 of a self-interested responder. The proportion of punishers will shrink to zero. Population states $\gamma_- > \frac{d_1 - c_1}{p}$ are stationary as punishers and self-interested responders obtain the same payoff c_2. These states are Lyapunov stable, yet not asymptotically stable. In particular, a sequence of arbitrary small random mutations can lead to the state $\gamma_- = \frac{d_1 - c_1}{p}$ which is instable as arbitrary small mutations to a smaller state end in the basin of attraction of $\gamma_- = 0$. In fact, $\gamma_- = 0$ is the only asymptotically stable state. In addition, if proposers

[14] Again, we could allow for this punishment after cooperation as well as after defection. Yet, similar to Setting 1, preferences which lead to punishment after cooperation (e.g. unconditional spitefulness) can not survive in our setting of randomly composed groups. Again, we simplify the analysis by abstracting from the possibility of punishment after cooperation.

tremble with any small probability $\epsilon > 0$, then punishers do strictly worse and the trajectory from any interior state converges to $\gamma_- = 0$.

We contrast these results with our haystack model. Proposers cooperate only in groups with a number of punishers $k_- > N\frac{d_1-c_1}{p}$. In these groups both types of responders earn the same highest fitness of c_2. In groups below this threshold, proposers defect and the fitness of all responders is lower. Self-interested responders earn the fitness d_2 while punishers get the even lower fitness of $(d_2 - c_p)$. Let $k_-^* \equiv \left\lfloor N\frac{d_1-c_1}{p} \right\rfloor$. Then proposers cooperate in groups with $k_- > k_-^*$. For a given population state γ_- the difference in the average fitness between a punisher and self-interested responder is

$$\bar{\pi}_-(\gamma_-) - \bar{\pi}_s(\gamma_-) = (c_2 - d_2)B_{N-1,k_-^*}(\gamma_-) - c_p \sum_{k_-=0}^{k_-^*-1} B_{N-1,k_-}(\gamma_-). \tag{5}$$

Proposition 2 *Let $\frac{d_1-c_1}{p} > \frac{1}{N}$. The only asymptotically stable states are the two monomorphic populations $\gamma_- = 0$ and $\gamma_- = 1$. The unique mixed equilibrium is not stable.*[15]

Remark 2 *If $\frac{d_1-c_1}{p} < \frac{1}{N}$ then $k_-^* = 0$ and a single punisher in a group induces cooperation. Then only a monomorphic punisher population can be stable.*

In contrast to Setting 1, the option to punish defection drives the population to a monomorphic state. Either a "culture of punishment" develops where all responders are willing to punish, or a "culture of laissez faire" where no responder punishes defectors.

Consider first a population dominated by self-interested responders. A punisher is then typically the only punisher in his group and is unable to enforce cooperation by the proposers. Proposers defect. The punisher incurs the fitness cost of punishing and obtains the minimal fitness $(d_2 - c_p)$ while the self-interested responders earn the higher fitness d_2. Punishers cannot invade a self-interested responder population. Consider now a responder population dominated by punishers. Most groups have enough punishers to induce cooperation by the proposers. Yet then even punishers do not have to execute the punishment and in these groups all responder types earn the maximal fitness of c_2. Only a few groups are below the threshold k_-^* and in these groups punishers earn lower fitness. Yet, if the overall proportion of punishers γ_- is high enough then there is more weight on groups in which a punisher is pivotal to induce cooperation then on groups in which he actually has to punish. Self-interested responders cannot invade a population of punishers. The key difference between Setting 1 and Setting 2 is that rewaders incur costs only if they are successful in inducing cooperation while punisher incur costs only if they fail to induce cooperation. The option to punish may be good or bad for total welfare (measured

[15] Again, we could generalize this result slightly. Take any trait such that (a) when the fraction of a group possessing the trait is above s^* (with $\frac{1}{N} < s^* < \frac{N-1}{N}$) then all agents do equally well; (b) when the fraction is less then s^* then all agents in the group do worse, but those without the trait do better; (c) agents are randomly assigned to groups. Then the two monomorphic equilibria in which either all players do have the trait or all players do not have the trait are stable.

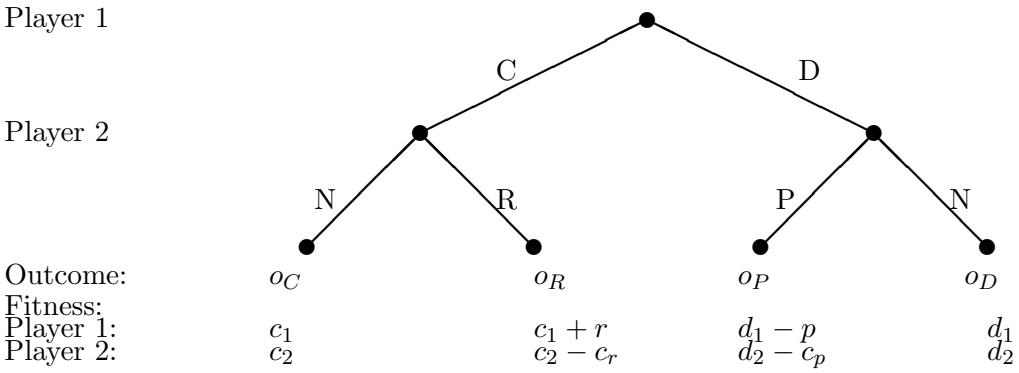
by the sum of the proposer's and the responder's fitness). In the punisher equilibrium we get full cooperation. Yet, the cooperative outcome maximizes welfare only if $d_1 - c_1 \leq c_2 - d_2$ and reduces welfare otherwise. The option to reward cooperation in Setting 1, in contrast, always increases welfare but will typically not maximize it since some defection occurs in equilibrium.

The unstable mixed equilibrium γ_-^{cut} separates the basins of attraction of both stable equilibria and is characterized by $\bar{\pi}_-(\gamma^{cut}) = \bar{\pi}_s(\gamma^{cut})$. The comparative statics of γ_-^{cut} is similar to the comparative statics of γ_+^{eq} in the previous subsection and can be found in the working paper version. In Setting 2 we found two stable equilibria and may end up in an equilibrium of punishers or in an equilibrium of self-interested responders. Setting 3 allows for rewarders and punishers. We then find that, under the right conditions, only a monomorphic responder population of punishers forms a stable equilibrium.

2.4 Setting 3: Costly Rewarding or Costly Punishment

In Setting 3 the responder has both options: costly punishing after defection and costly rewarding after cooperation. This allows us to analyze how the evolution of one side of reciprocity influences the evolution of the other side. The interaction is illustrated in Figure 4. All pro-

Figure 4: Interaction in Setting 3



with $c_1 + r > d_1 > c_1$ and $c_2 > c_2 - c_r > d_2 > d_2 - c_p$.

posers maximize their expected material payoffs. All possible preferences of responders fall into one of four categories: **Self-interested** players neither reward cooperation nor punish defection. **Punishers** do not reward cooperation, but do punish defection. **Rewarders** reward cooperation, but do not punish defection. **Reciprocal** players both reward cooperation and punish defection.[16] Thus we need to distinguish only these four responder types. In fact, it turns out that only the first three types are really important for the dynamics, since typically either

[16] Again, we neglect non-generic cases of preferences which associate the same subjective utility with different outcomes.

rewarding or punishing is the cheaper way to induce cooperation. Accordingly the reciprocal type who rewards and punishes is outperformed either by rewarders or by punishers.

Now the only monomorphic population that is stable is a population consisting entirely of punishers. A monomorphic population of self-interested players is destabilized by rewarders, a monomorphic rewarder population is destabilized by punishers (and also by self-interested players), and a monomorphic population of reciprocal players is not stable against punishers.

Proposition 3 *In Setting 3 under Assumptions 1-4 a monomorphic punisher population is always a stable equilibrium. The remaining monomorphic states are unstable.*

Are there other stable equilibria consisting of several preference types? The answer depends on the parameters of the model. For the right range of parameters this is the only stable equilibrium and the population converges to a pure punisher population from any initial interior state. Rewarders as well as punisher can induce cooperation by the proposers. We know from the analysis of Setting 2 that punisher become more successful when their own fraction is growing. Hence, punishers also profit from a growing proportion of rewarders. Any kind of reciprocity helps to induce cooperation of proposers and reduces the costs of being a punisher. Thus, higher fractions of rewarders and higher fractions punishers enhance the evolutionary success of preferences for punishing. Conversely, we know from Setting 1 that the evolutionary success of rewarders relative to self-interested players decreases if their own proportion becomes too large. Hence, the same holds for too large a fraction of punishers. Furthermore, relative to preferences for punishing, the success of preferences for rewarding is reduced by an increase of the proportion of any type of reciprocity. The higher the fraction of rewarders or punishers, the more groups are above the threshold for cooperation. Thus, costs of rewarding grow, whereas the costs of punishing fall. Higher fractions of rewarders and higher fractions of punishers reduce the evolutionary success of preferences for rewarding relative to the success of preferences for punishing.

This interdependence between the evolution of both types of reciprocity has interesting consequences. Consider an entirely self-interested population. Preferences for punishing cannot invade such a population directly, as shown in Setting 2. Yet preferences for rewarding can invade (see Setting 1). They can serve as a "catalyst" and enable the invasion of punishers. The more punishers invade, the more successful they become, and eventually they drive out self-interested players as well as rewarders.

In fact, we obtain the very strong results of convergence to the pure punisher equilibrium from any interior state whenever the costs of punishing and rewarding are not too high and under an additional technical requirement. We provide the formal statement of this condition in the appendix together with the proof of the next proposition. This condition excludes parameter constellations in which e.g. a punisher cannot successfully enter a population consisting only of reciprocal and self-interested players merely because a single punisher never induces cooperation in such groups. Note that in an extended model, in which $N(d_1 - c_1)$, i.e. either the group size,

or the proposers' payoffs, is a random variable with sufficiently broad support this condition holds automatically without any restrictions on the parameters.

Proposition 4 *In Setting 3 if Condition 1 holds in addition to Assumptions 1-4 and if the costs of punishing $c_p > 0$ are sufficiently small, then the monomorphic punisher population is the only stable equilibrium and the population converges to this stable equilibrium from any interior state.*

Here is an example of an assumption that is sufficient (but certainly not necessary) to ensure that Condition 1 holds.

Assumption 5 *The proposers' material loss p after being punished equals his material gain r after being rewarded, i.e. $p = r$, and $\frac{N(d_1-c_1)}{2r} - \lfloor \frac{N(d_1-c_1)}{2r} \rfloor < \frac{1}{2}$.*

This assumption simplifies the analysis and allows to characterize for which paramter values we have one or two stable equilibria. Under Assumption 5 punishers and rewarders have exactly the same influence on the behavior of proposers in their group. Hence, material payoffs of all other responders are not affected if we replace a punisher by a rewarder or vice versa. It turns out that reciprocal players do hardly influence the result and we can ignore them for the purpose of this discussion. Consider a population consisting only of rewarders and self-interested players. According to Setting 1 this population evolves towards a unique equilibrium containing both preference types. Can a small fraction of punishers invade this equilibrium? The answer depends on how large the proportion γ_+^{eq} of rewarders in the equilibrium of Setting 1 is. Since we assumed $p = r$, the effect of a rewarder on any other responder in his group is precisely the same as the effect of a punisher at the same place. The average fitness difference between rewarders and self-interested responders is still given by Equation 3 if we replace γ_+ in Setting 1 with $\gamma \equiv \gamma_+ + \gamma_-$ in Setting 3. Similarly, the average fitness difference between punishers and self-interested responders is still given by Equation 3 if we replace γ_- in Setting 2 with $\gamma \equiv \gamma_+ + \gamma_-$ in Setting 3. Hence, punishers can invade this equilibrium if and only if the fraction γ_+^{eq} of rewarders in this equilibrium (determined by Equation 4) is higher than the threshold γ^{cut} (determined by the equation $\bar{\pi}_-(\gamma^{cut}) = \bar{\pi}_s(\gamma^{cut})$) at which punishers become more successful than self-interested players. Once preferences for punishing can invade, they drive out all other preferences and the dynamics leads to a monomorphic punisher population.

Corollary 1 *Let γ^{eq} be defined by equation 4 and γ^{cut} by equation 5. In Setting 3 under Assumptions 1-5 the following holds*

a) *if $\gamma_+^{eq} > \gamma_-^{cut}$, then the only stable equilibrium is a monomorphic population, where all players have preferences for punishing. The population converges to this equilibrium from any interior state.*

b) *If $\gamma_+^{eq} < \gamma_-^{cut}$, then there are precisely two stable equilibria. One stable equilibrium is the monomorphic population of preferences for punishing. In the other stable equilibrium preferences for rewarding and self-interested preferences coexist. In this equilibrium the fraction of preferences for rewarding is γ_+^{eq}.*

3 Robustness Checks

In this section we consider several robustness checks of our results. In Subsection 3.1 we consider the robustness with respect to small mistakes in players' choices. In Subsection 3.2 we consider what happens when we drop the assumption that players observe the distribution of preferences in their group and players have to use a simple learning heuristic to adapt their behavior to the composition of their group. In Subsection 3.3 we check robustness of our results when allowing for some evolution of preferences within groups.

All three robustness checks can be seen as a perturbation of our main model. We show that our results change only slightly. The results we obtain with this perturbed processes have a similar structure and the definition below is useful for all three robustness checks. For each setting we consider "perturbed" models with the same type space, players are still interacting randomly and anonymously in groups with N proposers and N responders. Also the interaction remains the same with the same fitness payoffs. We consider situations where players play optimally only in most periods and only in most groups. This helps to check robustness of our results when players occasionally fail to play optimally, either because players tremble occasionally, or because players have to learn to play optimally in their respective group. We do also allow that a small number of groups may change their type composition in subperiods between the reshuffling of groups and the probability distribution changes correspondingly. This is useful for the analysis of within-group evolutionary effects.

Let $P_{\theta,\mathbf{k}}^t(\gamma)$ denote the probability in subperiod t that a player of type θ is in a group that has the constant composition $\mathbf{k}$ in all T subperiods. Similarly, let $B_{\theta,\mathbf{k}}(\gamma)$ denote the probability that a player of type θ ends up in a group of composition $\mathbf{k}$ where the distribution over groups is Multinomial as in Assumption 1.[17]

Fix a setting and the corresponding stage game G. For each group with N players in either positions and of composition $\mathbf{k}$ let $a_{\mathbf{k}}^t$ denote the vector of actions that each member in this group takes at time t given the perturbed process. Let $a_{\mathbf{k}}^*$ denote the vector of actions played in each subperiod in the main model in Section 2 (including Assumption 2). Without loss of generality we assume that all such vectors are sorted in the same sequence of types. Let $W_{\mathbf{k}}^t$ be the probability distribution over action-profiles in subperiod t which is generated by the perturbed process given a group which has the constant type composition $\mathbf{k}$ in all T subperiods.

[17] This probability corresponds to the probability from the Multinomial distribution for groups of size $N-1$ for a group of composition that differs from $\mathbf{k}$ only by reducing the number of players of type θ by one.

Definition 1 (Δ-bounded perturbation of the evolutionary process) *Let $\Delta > 0$. For each of the considered settings we call an evolutionary process a Δ-bounded perturbation of the main model in Section 2 if for all subperiods $t \in \{1, \ldots, T\}$, all types θ, all population states $\gamma \in \Gamma$ and each possible group composition $\mathbf{k}$ it holds that $\left|P^t_{\theta,\mathbf{k}}(\gamma) - B_{\theta,\mathbf{k}}(\gamma)\right| < \Delta$ and if, conditional on being in a group which has composition $\mathbf{k}$ in all T periods, for all $t > \Delta T$ it holds that $W^t_{\mathbf{k}}(a^t_{\mathbf{k}} = a^*_{\mathbf{k}}) \geq (1 - \Delta)$.*

Our haystack model turns out to be robust to perturbations that are bounded by a sufficiently small $\Delta > 0$ in the following sense. A small interval containing any state that was asymptotically stable in the unperturbed model is an asymptotically stable set in the perturbed model. Furthermore, while new asymptotically stable states or sets can potentially arise in the perturbed model, the basin of attraction of this newly emerged stable states vanishes as Δ becomes arbitrarily small. The basin of attraction of the asymptotically stable sets around the stable equilibria from the unperturbed model do not vanish with small Δ. This motivates the following definition.

Definition 2 (δ-stable set) *Let $\delta > 0$. We call an asymptotically stable set S* **δ-stable** *if all γ with $d(\gamma, S) < \delta$ are in the basin of attraction of S.*

Note that on the one hand for sufficiently small $\delta > 0$ all stable equilibria derived in the previous sections are δ-stable.[18] On the other hand, for any given $\delta > 0$ the newly arising stable states of the perturbed model will not remain δ-stable if the perturbation becomes sufficiently small.

Proposition 5 *For each setting there exists a $\overline{\delta} > 0$ such that for all $\delta \in (0, \overline{\delta})$ there is a $\Delta > 0$ such that for all Δ-bounded perturbations of the main model holds:*

in Setting 1 *$[\gamma^{eq}_+ - \delta, \gamma^{eq}_+ + \delta]$ is a δ-stable set and there exists no δ-stable set which is disjoint from $[\gamma^{eq}_+ - \delta, \gamma^{eq}_+ + \delta]$.*

in Setting 2 *the set $[0, \delta]$ and the set $[1 - \delta, 1]$ are δ-stable. There exists no δ-stable set which is disjoint from both of these sets.*

in Setting 3 *the set $\{\gamma : \gamma_- \geq 1 - \delta\}$ is δ-stable. Furthermore, for $\theta \in \{+, \pm, s\}$ the sets $\{\gamma : \gamma_\theta \geq 1 - \delta\}$ are not δ-stable.*

If, in addition, Condition 1 holds, and if the cost of punishing $c_p > 0$ are sufficiently small, then there exists no δ-stable set which is disjoint from $\{\gamma : \gamma_- \geq 1 - \delta\}$. The population converges to this δ-stable set from any interior state γ with $\gamma_- < 1 - \delta$.

3.1 Small perturbations of play

In this section we consider the robustness of our results when players make occasional mistakes. We assume that players in both positions may, with a small probability, fail to play optimally

[18] E.g. in Setting 1 $\delta < \min\{\gamma^{eq}, 1 - \gamma^{eq}\}$ works, in Setting 2 $\delta < \min\{\gamma^{cut}, 1 - \gamma^{cut}\}$.

and deviate from maximizing their expected subjective utility. A simple example of such a process is one in which every player plays optimally, with probability of at least $(1-\epsilon)$, given the other players' behavior, but trembles with a small probability bounded by ϵ. We allow the probability $\varepsilon_i(\mathbf{k})$ of these mistakes to depend on the player's position $i \in \{1,2\}$, his type θ, the number of periods t he played in the group already, and the composition $\mathbf{k}$ of the respective group he is in. We only require that all this mistakes are bounded by ϵ, i.e. $\varepsilon^t_{i\theta}(\mathbf{k}) < \epsilon$ for all $i \in \{1,2\}$, $\theta \in \Theta$, $t \in \{1,\dots,T\}$, and $\mathbf{k}$.

One issue that arises is whether small probability trembles of players in one position do change the optimal response of the players in the other position - then behavior in a group might change significantly even in instances where no tremble occurs. Trembles by proposers cannot have this effect since responders do observe the proposers action, know at which decision node they are, and choose their optimal action correspondingly. In particular, in instances where the proposer did not tremble, responders' optimal choices match exactly their behavior from the corresponding setting without trembles. The effect of trembles by responders could however have the effect that optimal behavior of the proposers changes in groups that are close to the threshold. Yet it turns out that for sufficiently small trembles this cannot happen and the optimal response for proposers remains unaffected by the trembling of the responders.[19]

Lemma 2 *Let $\Delta > 0$. Fix the setting. Then there is an $\epsilon > 0$ such that if $\varepsilon^t_{i\theta}(\mathbf{k}) < \epsilon$ for all $i \in \{1,2\}$, $\theta \in \Theta$, $t \in \{1,\dots,T\}$, and $\mathbf{k}$, then the model with such small trembles is a Δ-bounded perturbation of the main model in Section 2.*

Proposition 5 together with Lemma 2 confirms that our results are robust to mistakes that proposers and responders commit with small probability. There are δ-stable sets in this perturbed dynamics and if the state is in one of these sets then the state is arbitrarily close to one of the stable equilibria derived in Section 2 without mistakes. Note also, that in Setting 1 and Setting 3 a purely self-interested population forms also a stable equilibrium, but it is not δ-stable for sufficiently small perturbations.

3.2 Learning the distribution of preferences in a group

In this subsection we consider the robustness of our results with respect to dropping Assumption 2. What happens if players, and in particular responders, do not simply know the distribution of preferences in their group but have to learn how to play optimally in their group? We envision this learning process to operate fast within a given group, while the evolution of preferences is a very slow process. After a group is randomly formed players' preferences remain constant until the group is reshuffled after a period of time $\mathfrak{T}$. Given their subjective preferences, players' behavior adjusts fast to optimal behavior and an equilibrium emerges. This equilibrium

[19] Furthermore, while for intermediate trembling probabilities behavior in groups close to the threshold may change, our analysis essentially goes still through with the modification that the threshold k^* needs to be adapted accordingly.

is played for a while before the group dissolves. We divide this time period $\mathfrak{T}$ into T subperiods of length $\frac{\mathfrak{T}}{T}$. In each subperiod the basic game is played once. We are interested in the limit when T becomes large. Then there is enough time for players to learn to play an equilibrium in the early periods and to play this equilibrium afterwards for most of the time period $\mathfrak{T}$. For a given distribution of preferences in a group we consider a simple learning by valuation process, adapted from Jehiel-Samet [22] to our setting.

Each player keeps track of the average subjective utility he obtained whenever he played a certain action available to him at a certain information set. At any information set a player chooses with probability $(1-\epsilon)$ the action which earned him the highest average utility in the past. If both actions did equally well he chooses each action with probability $\frac{1}{2}$. The other action which resulted in a lower average utility in the past is played with a small probability $\epsilon > 0$. We interpret this as trembling or as occasional mistakes. Players cannot avoid to make these mistakes, yet they can learn from them. A different interpretation is to consider this as a primitive form of experimentation. To start the process we assume that each player at any information set plays each action at least once before he repeats an action available at that information set.

Lemma 3 *Fix a setting. For every $\Delta > 0$ and for sufficiently small $\epsilon > 0$ there exists a $T_{\epsilon,\Delta} \in \mathbb{N}$ such that for all $T > T_{\epsilon,\Delta}$ the ϵ-reinforcement learning process (which replaces Assumption 2 of an observable group composition) is a Δ-bounded perturbation of the main model in Section 2.*

Lemma 3, together with Proposition 5, shows that replacing Assumption 2 by a process of learning by valuation does change our results only slightly, provided players play in sufficiently many subperiods before reshuffling occurs. Then, only a small proportion of the time period $\mathfrak{T}$ is needed for learning, after which players act almost "as if" they knew the distribution of preferences in their group.

3.3 Within group evolution

In this section we analyze the robustness of our results with respect to the effect of within-group evolution. So far we assumed that the evolution of preferences only takes place when the groups are reshuffled, and thus the distribution of preferences in each group remained always completely random. This is a good approximation when the speed of the evolution of preferences is slow relative to the frequency of group reshuffling, a point we make precise in this section. One way to think of the assumption that the evolution of preferences occurs only during reshuffling is that players produce offspring in every period which are first in a separate pool, of not yet reproducing potential future players. After T periods before groups are reshuffled a (small) fraction f of the players dies and is replaced randomly by players from the pool of offspring. From this new population new random groups are formed. Then the group composition is always multinomially distributed.

What happens now if we assume instead that evolution can take place in every subperiod. We want to allow for some within-group evolution, while preserving the effect that groups with higher average fitness have more offsprings. A natural way to capture this would be to allow the cooperative groups to grow in size relative to the non-cooperative groups. However, given this is a robustness check of our previous model, we want to keep the group size constant at $2N$ interacting players. We therefore allow groups with higher average fitness in their group to have more offsprings, but these offsprings remain inactive until they either replace an active player or until the next reshuffling of groups occurs.

For any of the three models considered so far, the main model of Section 2, the model with trembles in Subsection 3.1, and the model with learning by valuation in Subsection 3.2, we consider the following extension which allows for within-group evolutionary forces. In any given subperiod every player passes away with probability $w > 0$, independent across players and subperiods, and is replaced by one of the offspring. Then the expected number of subperiods a player lives is given by $\frac{1}{w}$. Thus smaller w correspond to large expected times before a player dies and is replaced. We consider situations where the total population does not grow or shrink. This implies that with a lower probability of dying, the overall number of offspring must be smaller, which corresponds to a rescaling of the fitness payoffs by a common factor. Then a small w corresponds to a slow speed of evolution. The precise way how replacement occurs turns out to be irrelevant for our robustness check. For concreteness, assume that dead players are replaced with offspring from the same group, if there are any, and with random players from the remaining groups otherwise.

In addition, we assume that for any type θ the proportion this type has in the offspring population in the next period is bounded by $C\gamma_\theta$, where $C > 0$ is a potentially large, but finite, constant and γ_θ the proportion of type θ players in the original total population. This assumption is very natural, in particular if we think of situations, where every player reproduces at some rate that is independent from the interaction we investigate plus a growth rate that is derived from the performance in the interaction that we consider.

The key observation for our next result is that if evolution is sufficiently slow relative to the frequency of group reshuffling, than only an arbitrarily small proportion of groups will change their composition and we have a Δ-bounded perturbation of the main model.

Lemma 4 *Fix a setting. Let $\Delta > 0$.*

(a) Consider the main model of Section 2 with the above extension which allows for in-group evolutionary forces. For any T there exists a $\overline{w} > 0$ such that for all probabilities of replacement $w < \overline{w}$ this extended model is a Δ-bounded perturbation of the main model.

(b) Consider the model with trembles of Subsection 3.1 with the above extension which allows for within-group evolutionary forces. For any T there exist an $\epsilon > 0$ and a $\overline{w} > 0$ such that if $\varepsilon^t_{i\theta}(\mathbf{k}) < \epsilon$ for all $i \in \{1,2\}$, $\theta \in \Theta$, $t \in \{1,\ldots,T\}$, and $\mathbf{k}$, then, for all probabilities

of replacement $w < \overline{w}$, this extended model with errors bounded by ϵ is a Δ-bounded perturbation of the main model in Section 2.

(c) Consider the learning by valuation model of Subsection 3.2 with the above extension which allows for within-group evolutionary forces. For sufficiently small $\epsilon > 0$ there exists a $T \in \mathbb{N}$ such that for all $T > T_{\epsilon,\Delta}$ there is a $\overline{w}_T$ such that for all $w < \overline{w}_T$ the ϵ-reinforcement learning process by valuation (which replaces Assumption 2 of an observable group composition) extended by above extension allowing for within-group evolution is a Δ-bounded perturbation of the main model in Section 2.

Thus, as long as the evolution of preferences is slow compared to the frequency of group reshuffling, Lemma 4 together with Proposition 5 shows that in group evolutionary effects are minor and do not fundamentally change our results.

4 Related Literature

The existing evolutionary literature has paid little attention to the structural differences between the evolution of a willingness to reward and a willingness to punish and their mutual interaction and how they can complement each other in undermining the prevalence of purely self-interested preferences and establish cooperation. The question about how reciprocity or social preferences can survive evolution in sporadic interactions has however been tackled by several authors from biology, psychology, economics and other social sciences.For detailed surveys of the literature on reciprocity see Sobel [40] or Sethi and Somanathan [39].[20]

A first branch of literature deviates from random encounters and assumes assortative matching, i.e. players are more likely to play with players of their own type. Then being a cooperative type has the advantage of being more likely to face cooperative players which is an evolutionary advantage. Most of the literature on group selection focuses on this idea to explain the evolutionary survival of social preferences. Price [31] first offered a mathematical description.

[20]Our discussion focuses on sporadic interactions among genetically unrelated individuals. For related individuals kin selection would be crucial. For infinitely repeated interactions purely self-interested individuals can establish cooperation. Guttman [17] points out that already in finitely repeated interactions cooperation is easier to achieve and costs of reciprocal preferences become be very small and a small degree of observability is sufficient to achieve cooperation. There are other interesting evolutionary models which are less directly related to our approach. Eshel, Samuelson and Shaked [9] find that altruists, who live on a circle, can survive in a local imitation process. Huck and Oechssler [21] look at ultimatum games in which costs of punishing unfair behavior are very small compared to the punishment and the inverse group-size. In a finite group punishers, when in the role of the proposer, have a small relative advantage over materialists in their group since they are slightly less likely to be matched with a punisher. This can be sufficient to compensate punishers for the costs of punishing if these costs are very small. There are also models in which, quite different from our approach, players' fitness depends directly on a payoff of the group. In Weibull and Salomonson [43] the groups average payoff can directly enter a players fitness in a positive or negative way, which leads respectively to altruism or spitefulness towards group members. In Kuzmics and Rodriguez-Sickert [26] groups occasionally fight each other and a copy of the winning group replaces the losing group. On the somewhat separate issue of whether evolution selects the Pareto efficient equilibrium in coordination games see Robson [32], Kandori-Mailath-Rob [23], Robson-Vega-Redondo [33], and Kuzmics [25].

Bergstrom [2] investigates the relation between assortative matching and the evolution of cooperation. Notice that initially random groups may become assortative over time by the evolution of preferences inside groups if no reshuffling of groups occurs (compare e.g. Cooper and Wallace [6]).[21] Bergstrom [3] and Sober and Wilson [41] survey this branch of the literature.

A second branch of this literature assumes that preferences are observable. This can give social preferences an advantage over self-interested preferences due to a commitment effect. A reciprocal player, for instance, is credibly committed to rewarding friendly or punishing unfriendly behavior. Therefore, he may induce friendly behavior of a first-moving player, thus enhancing his evolutionary success. The results of Güth and Yaari [16], Güth [15], Bester and Güth [4], and partly Sethi [37] are based on this argument. These results depend crucially on the assumption of observable preferences and on the set of allowed preferences, as Dekel et al. [7], Ely and Yilankaya [8], and Ok and Vega-Redondo [30] point out. They show for symmetric, normal-form games that evolution favors self-interested preferences, or at least leads to equilibrium play 'as if' players were purely self-interested, provided preferences are unobservable, and provided players are randomly matched within one single, infinite population.

A third branch of related literature considers models with types who exhibit behavior called "parochialism". Players interact in symmetric public good games in which, for any given strategy profile of the opponents, it is individually optimal to not contribute, but it would be optimal for all players if everybody would contribute. The number of players in the public good game may be two, in which case the game is the prisoners' dilemma and preferences are assumed to be observable. The number of players can also be larger in which case it is enough to assume that each player knows the distribution of preferences in his group, somewhat similar to Assumption 2 which we make in Section 2 and which we later replace by an explicit learning process in Subsection 3.2. These "parochial" types act to enhance efficiency if they play the public good game with mainly their own type and act to reduce efficiency if they play mainly with self-interested players. They may survive evolution even if the matching is non-assortative. Sethi and Somanathan [38] show in such a setting that conditional altruists who behave in a friendly way towards other altruists but spitefully towards materialists may be more successful than pure materialists. Similarly, and somewhat related to our setting 2, Gintis [14] looks at the evolution of strong reciprocity. The model considers conditional punishers who are willing to punish only if there are enough punishers in their group to induce cooperation and thus there is no need to ever actually carry out the punishments. But if conditional punishers happen to be in such a group where they would be willing to punish, they have to bear some surveillance costs. This additional assumption avoids that conditional punishers do always better than self-interested players. Conditional punishers can survive the evolutionary process in this model, as they can in our model, but the evolutionary dynamics is very different. In Gintis' model conditional punishers can successfully invade a population of self-interested players since punishment never

[21] A different endogenous justification of assortative matching arises if preferences are partly observable, see e.g. Frank [12].

really have to be carried out and surveillance costs are assumed to accrue only in groups where cooperation is induced. The dynamic in our model is almost the opposite. When there are few punishers in the total population, punishers perform poorly compared to self-interested players since there is little cooperation and they have to punish a lot. When there are many punishers in the total population punishers perform better than self-interested players since there is a lot of cooperation and few punishments are necessary.[22] Höffler [19] provides a model which is related to our Setting 1. He considers bounded-rational workers who work in payoff-maximizing firms. Firms can either enforce a certain effort by keeping workers under surveillance or they can save the surveillance costs, pass on some part of the saved costs in form of a higher wage instead, and hope that workers do not cheat. Without surveillance workers can shirk or play fair and they adapt their strategy according to a proportional imitation rule. The results are similar to our findings in Setting 1 in the sense that if most workers cheat, playing fair gives on average across firms higher payoffs, while if most workers play fair, cheating is on average the more successful strategy. In equilibrium some workers play fair, others don't. Our model differs substantially from Höffler's. First, we are really interested in comparing the evolution of rewarders and punishers and in understanding how they influence each other. Second, even if we focus on our Setting 1 where we have only rewarders, we have a quite different model. In Höffler, firms are payoff maximizing by assumption and only workers' strategies evolve. We start with a setting were players in both positions can have any strict preferences over outcomes and derive in Lemma 1 that payoff maximizing behavior will prevail. Thus the asymmetry between proposers (who always maximize their fitness) and responders (who can have reciprocal preferences) arises endogenously. Furthermore, we conduct a number of robustness checks. In particular, we show that our results are robust to dropping the assumption of an observable distribution of preferences and replacing it by a simple learning process.

5 Discussion and Conclusion

Lemma 3 together with Proposition 5 shows that replacing the Assumption of an observable group composition by a simple reinforcement-learning process by valuation changes our results only slightly and in this sense Assumption 2 can be seen as a shortcut for a fast learning process in which players with given preferences adopt fast to the strategic environment of their respective group environment. In addition, Lemma 4 shows that the key results survive in-group evolutionary effects when these are small which is true when the evolution of preferences is a slow process compared to the frequency of reshuffling.

Learning by valuation requires some sophistication but falls short of full rationality. Can we

[22] Bowles and Gintis [5] and Friedman and Singh [13] consider also the case when individual preferences are unobservable for the evolution of types who punish non-cooperative behavior. Friedman and Singh implicitly need the assumption that second-order punishment (i.e. the punishment of non-punishers) is costless. Bowles and Gintis consider a model where punishment takes the form of ostracism. Their model is too complex for an analytical solution but in simulations they also find that punishers can survive.

obtain similar results under even more sophisticated learning?

If proposers choose rationally whether and how long they experiment, while responders play myopically according to their preferences the results will still closely resemble what we found so far. Consider e.g. Setting 1 with the option to reward cooperation (the other settings are treated analogously). Then the proposer essentially faces a two armed bandit problem, where defection corresponds to the certain payoff d_1 while cooperation corresponds to an arm with uncertain distribution over the to potential payoffs $c_1 + r$ and c_1. Suppose a proposer of type θ starts with beliefs $B_{\theta,\mathbf{k}}(\gamma)$ that he is in a group of composition $\mathbf{k}$ and than updates this distribution. Now he maximizes his expected payoff averaged over the T subperiods for which the group stays together. Analogous to Proposition 5, for a given small $\delta > 0$ for a sufficiently large number of subperiods T the interval $[\gamma_+^{eq} - \delta, \gamma_+^{eq} + \delta]$ is a δ stable set and there exists no δ-stable set which is disjoint from $[\gamma_+^{eq} - \delta, \gamma_+^{eq} + \delta]$.[23] When we allow again for in-group evolution, our results remain valid if the speed of the evolution of preferences is sufficiently slow compared to the frequency of reshuffling $\frac{1}{T}$, while T is large enough to give incentives to experiment and learn to adapt the the group players are in.

So far all robustness checks did not change the results significantly. Dropping Assumption 2 may however change results significantly in a setting in which all players, and in particular the responders, are fully rational, the structure of the haystack model is common knowledge, and players are very patient. A complete analysis of such a setting is very complex and beyond the scope of this paper. There is reason though to conjecture that results might change significantly. Self-interested responders now understand that their behavior in early periods may influence the behavior of proposers in later periods and may therefore be willing to reward or punish in the hope of inducing cooperation of this player in future periods, in a sense trying to build a reputation for the group they are in.

The result that rewarders serve as a catalyst for the evolution of punishers does not only highlight an important interaction between rewards and punishment. It also reinforces the point made in Dekel et al. [7] that it is important for an evolutionary analysis to consider all possible preference types. The effect that rewarders enhance the success of punishers extends beyond the context of the haystack model. The insight that rewards and punishments have a similar effect on incentives to a proposer generalizes, and so does the fact that rewards are costly once you achieve cooperation, while punishments are costly when you don't. In fact, Kuzmics and Rodriguez-Sickert [26] find a such an effect in their model considering the evolution of moral codes which can be enforced by punishments or rewards. In their model rewards are not part of any equilibrium, but they do facilitate the evolution of beneficial codes which are enforced by punishments. This interplay of rewards and punishments seems to extend even beyond the

[23] The proof is similar to the proof of Proposition 5. Sketch: First realize, that it is enough to show that payoffs on $[\delta, \gamma_+^{eq} - \delta]$ and $[\gamma_+^{eq} + \delta, 1 - \delta]$ do not change their sign. Then we can show for all $\gamma \in [\delta, 1 - \delta]$ and any Δ_1 and Δ_2 there is a $\tilde{t}$ such that after t periods with a probability of at least $1 - \Delta_1$ the proposer assigns a belief of at least $(1 - \Delta_2)$ to the composition of the group he actually happens to be in. Furthermore, for every such $\tilde{t}$ there is a T such that for all $\gamma \in [\delta, 1 - \delta]$ it is optimal for a proposer of any type to experiment for at least t periods.

evolutionary context. Suppose a policy maker wants to give cost efficient incentive that change a populations behavior to a better norm. Suppose people react to such incentives but this reaction is slowly due to momentum. The cost efficient way seems to start by offering rewards until a significant proportion follows the new norm. Then punishments become the cheaper way to give incentives and the policy maker changes from offering rewards for following the new norm to threatening punishment for violating it.

In summary, this non-assortative group selection model offers an explanation for the evolutionary survival of both sides of reciprocal preferences. Despite the fact that individual behavior and preferences are unobservable, individuals continue to have a marginal effect on the distribution of preferences in their group, and this influences the behavior of the other players in their group. This effect is sufficient to enable preferences for rewarding and preferences for punishing to survive in the evolutionary competition with self-interested preferences. Both preferences for rewarding and preferences for punishing can induce cooperative behavior. But there is an intrinsic difference between the two preference types: preferences for rewarding tend to coexist with self-interested preferences, whereas preferences for punishing tend either to dominate the population completely or to vanish entirely. Furthermore, rewarders enhance the evolution of preferences for punishing. Preferences for rewarding are able to invade a self-interested population and may then, as a "catalyst", enable the invasion of preferences for punishing. Punishers, on the other hand, crowd out rewarders and may even drive them out completely.

A Proofs

A.1 Proof of Lemma 1

Lemma 1 follows from two observations. First, the distribution of types of the responders a proposer faces in his group is independent of his own type by Assumption 1. Second, responders condition their responses only on the proposer's action but not on his type or the composition of types in their group. This holds since players have preferences over outcomes only and for a responder choosing an action is equivalent to choosing an outcome of the game. In other words, the response profile which a proposer faces conditional on his own action is independent of his own type. Note also that the preference type a player has in position 1 has no influence on his fitness when he acts in role 2 since the type distribution of the other players in his group is independent of his type and since proposers have preferences over outcomes only. Since self-interested preferences of a proposer induce fitness maximizing behavior for any given response profile, a preference type $\theta = (\theta_s, \theta_2)$ with self-interested preferences when in role 1 must weakly dominate any preference type $\theta' = (\theta_{1s}, \theta_2)$ which has the identical preferences θ_2 when in role 2 and differs only in his type when in position 1, which shows part (a). Together with the observation that under Assumption 1 for every interior state every group composition has a strictly positive probability Part (b) follows also immediately. Part (c) follows since a

type (θ_1, θ_2) which deviates in position 1 from fitness maximization in some groups that occur with positive probability will earn less fitness than type θ_{1s}, θ_2 would by replacing type (θ_1, θ_2) while in all other groups type θ_{1s}, θ_2 earns at least the same expected fitness. Thus, type θ_{1s}, θ_2 earns a strictly higher expected fitness and must have a higher growth rate by payoff monotonicity, which contradicts stationarity. Part (d) we show by contradiction. Suppose there is an asymptotically stable state γ with a strictly positive proportion γ_θ of a type $\theta = (\theta_1, \theta_2)$ where proposers of type θ_1 deviate for some group composition from fitness maximizing behavior. Let $\theta' \equiv (\theta_s, \theta_2)$ denote the corresponding type with self-interested preferences when in the role of the proposer and identical preferences θ_2 when in the role of the responder. Consider any neighborhood of γ. It must contain a a state $\tilde{\gamma}$ which is an interior state with $\tilde{\gamma}_\theta < \gamma_\theta$ and $\tilde{\gamma}_{\theta'} > \gamma_{\theta'}$. It is well known that the solution trajectory starting from any interior state will always stay in the interior of Γ which implies by part (b) and payoff monotonicity that for any time $\tau > 0$ holds $\frac{\xi_{\theta'}(\tau,\tilde{\gamma})}{\xi_\theta(\tau,\tilde{\gamma})} > \frac{\tilde{\gamma}_{\theta'}}{\tilde{\gamma}_\theta} > \frac{\gamma_{\theta'}}{\gamma_\theta}$, which contradicts convergence to state γ.

A.2 Proof of Proposition 1

For $0 < \gamma_+ < 1$ holds $\bar{\pi}_+(\gamma_+) - \bar{\pi}_s(\gamma_+) = c_r B_{N-1,k_+^*}(\gamma_+) \left(\left(\frac{c_2 - d_2}{c_r} - 1 \right) - \sum_{k=k_+^*+1}^{N-1} \frac{B_{N-1,k}(\gamma_+)}{B_{N-1,k_+^*}(\gamma_+)} \right)$. Since, for $k > k_+^*$, $\frac{B_{N-1,k}(\gamma_+)}{B_{N-1,k_+^*}(\gamma_+)} = \left(\prod_{i=k_+^*+1}^{k} \frac{N-i}{i} \right) \left(\frac{\gamma_+}{1-\gamma_+} \right)^{(k-k_+^*)}$ it follows that $\frac{B_{N-1,k}(\gamma_+)}{B_{N-1,k_+^*}(\gamma_+)}$ is strictly increasing in γ_+, $\lim_{\gamma_+ \to 0} \frac{B_{N-1,k}(\gamma_+)}{B_{N-1,k_+^*}(\gamma_+)} = 0$, and $\lim_{\gamma_+ \to 1} \frac{B_{N-1,k}(\gamma_+)}{B_{N-1,k_+^*}(\gamma_+)} = \infty$. Since $0 < \frac{c_2 - d_2}{c_r} - 1 < \infty$ the difference $\bar{\pi}_+(\gamma_+) - \bar{\pi}_s(\gamma_+)$ is positive for γ_+ sufficiently close to 0 and negative for γ_+ sufficiently close to 1. The states $\gamma_+ = 0$ and $\gamma_+ = 1$ are therefore not stable. Furthermore, the term in the large brackets is strictly increasing in γ_+. Due to continuity this implies a unique asymptotically stable equilibrium characterized by $\bar{\pi}_+(\gamma_+^{eq}) - \bar{\pi}_s(\gamma_+^{eq}) = 0$.

A.3 Proof of the comparative statics of γ^{eq} with N

We assumed $\frac{k_+^*}{N} \equiv q$ constant, i.e. $k_+^* = qN$ with $0 < q < 1$. We can rearrange the equilibrium condition 4 into

$$c_2 - c_r - d_2 = c_r \sum_{k=1}^{N-1-k_+^*} \left(\left(\prod_{l=1}^{k} \frac{N - k_+^* - l}{k_+^* + l} \right) \left(\frac{\gamma}{1-\gamma} \right)^k \right) \tag{6}$$

$$= c_r \sum_{k=1}^{N(1-q)-1} \left(\left(\prod_{l=1}^{k} \frac{(1-q)N - l}{qN + l} \right) \left(\frac{\gamma}{1-\gamma} \right)^k \right). \tag{7}$$

Now we prove that for constant γ the right hand side is strictly increasing in N. Since the left hand side is constant, γ^{eq} has to fall in order to equilibrate the two sides again.
The number of terms increases with N. Since all terms in equation 6 are positive it is sufficient

to prove that each term increases in N. By extending N to real numbers we find

$$\frac{\partial}{\partial N}\left(\frac{(1-q)N-l}{qN+l}\right)=\frac{l}{(qN+l)^2}>0. \tag{8}$$

A.4 Proof of Proposition 2

For $0<\gamma_-<1$ the difference in the average fitness between a punisher and self-interested responder can be rewritten $\bar{\pi}_-(\gamma_-)-\bar{\pi}_s(\gamma_-)=c_pB_{N-1,k_-^*}(\gamma_-)\left(\left(\frac{c_2-d_2}{c_p}\right)-\sum_{k=0}^{k_-^*-1}\frac{B_{N-1,k}(\gamma_-)}{B_{N-1,k_-^*}(\gamma_-)}\right)$. For $k<k_-^*$ holds $\frac{B_{N-1,k}(\gamma_-)}{B_{N-1,k_-^*}(\gamma_-)}=\left(\prod_{i=k+1}^{k_-^*}\frac{i}{N-i}\right)\left(\frac{1-\gamma_-}{\gamma_-}\right)^{\left(k_-^*-k\right)}$ and thus $\frac{B_{N-1,k}(\gamma_-)}{B_{N-1,k_-^*}(\gamma_-)}$ is strictly decreasing in γ_-, $\lim_{\gamma_-\to 0}\frac{B_{N-1,k}(\gamma_-)}{B_{N-1,k_-^*}(\gamma_-)}=\infty$, and $\lim_{\gamma_-\to 1}\frac{B_{N-1,k}(\gamma_-)}{B_{N-1,k_-^*}(\gamma_-)}=0$. This implies that the difference $\bar{\pi}_-(\gamma_-)-\bar{\pi}_s(\gamma_-)$ is negative for sufficiently small γ_- and positive for γ_- sufficiently close to one (note that in all terms of the sum we have $k_-<k_-^*$). Thus $\gamma_-=0$ and $\gamma_-=1$ are asymptotically stable. Furthermore, the term in the large brackets is strictly increasing in γ_-. Due to continuity this implies a unique interior rest point γ_-^{cut}, which is unstable and characterized by $\bar{\pi}_-(\gamma_-^{cut})-\bar{\pi}_s(\gamma_-^{cut})=0$.

A.5 Technical Definitions and Condition used in Proofs of Setting 3

Let $k_\pm^*\equiv\lfloor\frac{N(d_1-c_1)}{r+p}\rfloor$. This is the number of reciprocal players just not sufficient to induce cooperation.

For $k_\pm\le k_\pm^*$ let $k_+^*(k_\pm)\equiv\lfloor\frac{N(d_1-c_1)-k_\pm(r+p)}{r}\rfloor$. This is the number of rewarders which, for a given number of reciprocal players, is just not sufficient to induce cooperation.

For $k_\pm\le k_\pm^*$ let $k_-^*(k_\pm)\equiv\lfloor\frac{N(d_1-c_1)-k_\pm(r+p)}{p}\rfloor$. This is the number of punisher which, for a given number of reciprocal players, is just not sufficient to induce cooperation.

We define $\mathcal{U}_{+,\pm}\equiv\{(k_+,k_\pm):k_+r+k_\pm(r+p)<N(d_1-c_1)\}$, and for any $(k_+,k_\pm)\in\mathcal{U}_{+,\pm}$ we define $k_-^*(k_+,k_\pm)\equiv\lfloor\frac{N(d_1-c_1)-k_+r-k_\pm(r+p)}{p}\rfloor$. For a given number k_+ of rewarders and a given number $k_\pm$ of reciprocal players in a group $k_-^*(k_+,k_\pm)$ corresponds to the highest number of punishers which is still not sufficient to induce cooperation in that group.

Let $\mathcal{U}_{-,\pm}\equiv\{(k_-,k_\pm):k_-p+k_\pm(r+p)<N(d_1-c_1)\}$. For $(k_-,k_\pm)\in\mathcal{U}_{-,\pm}$ let $k_+^*(k_-,k_\pm)\equiv\lfloor\frac{N(d_1-c_1)-k_-p-k_\pm(r+p)}{r}\rfloor$. This is the number of rewarders which, for a given number of punishers k_- and reciprocal players $k_\pm$ in a group, is just not sufficient to induce cooperation.

Condition 1 *For $i,j\in\{+,-\}$ with $i\neq j$ let $k_i^*(k_\pm^*)=0$ hold and let $k_j^*(k_i^*(k_\pm),k_\pm)=0$ for all $k_\pm<k_\pm^*$.*

Condition 1 ensures that for any number of reciprocal players that is not yet sufficient to induce cooperation a single rewarder (jointly with a number of punishers) can potentially be pivotal in inducing cooperation and that a single punisher (jointly with a number of rewarders) can be pivotal in inducing cooperation. Note also that Assumption 5 is sufficient to guarantee Condition 1.

A.6 Proof of Proposition 3

Step 1: First we show that a monomorphic population of punishers is asymptotically stable. It is sufficient to show that punishers earn strictly higher fitness than any other type when the proportion of non-punishers in a population state is sufficiently small.[24] First consider a type who rewards after cooperation, i.e. a rewarder or a reciprocal responder. Almost all groups cooperate and he has to bear the costs c_r of rewarding with a probability arbitrarily close to one. One the other hand he can outperform a punisher only in non-cooperative groups which have a proportion arbitrarily close to zero. Hence punishers strictly outperform rewarders if the proportion of punishers is close enough to one. Now consider the expected fitness difference between a punisher and a self-interested responder. Being a punisher increases fitness by (c_2-d_2) in groups where this has a pivotal effect, but costs c_p in groups where cooperation is not achieved.

$$\begin{aligned}
&\bar{\pi}_-(\gamma) - \bar{\pi}_s(\gamma) \\
&= \sum\nolimits_{(k_+,k_\pm)\in\mathcal{U}_{+,\pm}} \gamma_+^{k_+}\gamma_\pm^{k_\pm}\gamma_s^{N-1-k_+-k_\pm-k_-^*(k_+,k_\pm)}\gamma_-^{k_-^*(k_+,k_\pm)}\left[(c_2-d_2)\binom{N-1}{k_+,k_\pm,k_-^*(k_+,k_\pm)}\right. \\
&\qquad\left. -c_p\sum_{k_-<k_-^*(k_+,k_\pm)}\binom{N-1}{k_+,k_\pm,k_-}\left(\frac{\gamma_s}{\gamma_-}\right)^{k_-^*(k_+,k_\pm)-k_-}\right] \qquad (9)
\end{aligned}$$

See Subsection A.5 for the definition of $k_-^*(\cdot,\cdot)$ and related expressions. As γ_- is sufficiently close to one, γ_s becomes arbitrarily small and the term in the square bracket is strictly positive for all k_+. Hence punishers strictly outperform self-interested players (and all other types) in a sufficiently small neighborhood of the monomorphic punisher population.

Step 2: The claim that a monomorphic population of rewarders or a monomorphic population of self-interested players are both unstable follows directly from Proposition 1. A monomorphic population of reciprocal responders is not stable since any invading proportion of punishers will be more successful since they save the costs of rewarding but still obtain full cooperation.

Proof of Proposition 4

The following lemma is useful in proving Proposition 4.

Lemma 5 *In Setting 3 let Condition 1 hold in addition to Assumptions 1-4. Then there exists a $\tilde{c}_p > 0$ such that for all $c_p \leq \tilde{c}_p$ there exist values $\underline{\gamma}_s$ and $\overline{\gamma}_s$ with $0 < \underline{\gamma}_s < \overline{\gamma}_s < 1$ such that $\bar{\pi}_+(\gamma) > \bar{\pi}_s(\gamma)$ whenever $1 > \gamma_s \geq \underline{\gamma}_s$, $\bar{\pi}_-(\gamma) > \bar{\pi}_s(\gamma)$ whenever $0 < \gamma_s \leq \overline{\gamma}_s$, and $\bar{\pi}_\pm(\gamma) > \bar{\pi}_s(\gamma)$ whenever $\underline{\gamma}_s < \gamma_s \leq \overline{\gamma}_s$.*

[24] Then Step 1a) follows e.g. from Lyapunov's direct method (see e.g. [42]) with Lyapunov function $L(\gamma) \equiv (\gamma_- - 1)^2$.

Proof of Lemma 5: (See Subsection A.5 for the definition of $k_-^*(\cdot,\cdot)$ and related expressions.)

$$
\begin{aligned}
&\bar{\pi}_+(\gamma) - \bar{\pi}_s(\gamma) \\
= \;& (c_2 - d_2 - c_r) \sum_{(k_-,k_\pm)\in\mathcal{U}_{-,\pm}} \binom{N-1}{k_-, k_\pm, k_+^*(k_-,k_\pm)} \gamma_+^{k_+^*(k_-,k_\pm)} \gamma_-^{k_-} \gamma_\pm^{k_\pm} \gamma_s^{N-1-k_+^*(k_-)-k_--k_\pm} \\
& - c_r \sum_{(k_-,k_\pm)\in\mathcal{U}_{-,\pm}} \sum_{k_+>k_+^*(k_-,k_\pm)} \binom{N-1}{k_-, k_\pm, k_+} \gamma_+^{k_+} \gamma_-^{k_-} \gamma_\pm^{k_\pm} \gamma_s^{N-1-k_+-k_--k_\pm} \\
& - c_r \sum_{(k_-,k_\pm)\notin\mathcal{U}_{-,\pm}} \sum_{k_+\geq 0} \binom{N-1}{k_-, k_\pm, k_+} \gamma_+^{k_+} \gamma_-^{k_-} \gamma_\pm^{k_\pm} \gamma_s^{N-1-k_+-k_--k_\pm}
\end{aligned}
$$

$$
\begin{aligned}
\geq \;& \sum_{(k_-,k_\pm)\in\mathcal{U}_{-,\pm}} \gamma_-^{k_-} \gamma_\pm^{k_\pm} \gamma_+^{k_+^*(k_-,k_\pm)} \gamma_s^{N-1-k_--k_\pm-k_+^*(k_-,k_\pm)} \left[\frac{c_2-d_2-c_r}{2} \binom{N-1}{k_-, k_\pm, k_+^*(k_-,k_\pm)} \right. \\
& \left. - c_r \sum_{k_+>k_+^*(k_-,k_\pm)} \binom{N-1}{k_-, k_\pm, k_+} \left(\frac{\gamma_+}{\gamma_s}\right)^{k_+-k_+^*(k_-,k_\pm)} \right] \\
+ \;& \sum_{k_\pm<k_\pm^*} \gamma_-^{k_-^*(k_\pm)} \gamma_\pm^{k_\pm} \gamma_s^{N-1-k_-^*(k_\pm)-k_\pm} \left[\frac{c_2-d_2-c_r}{2} \binom{N-1}{k_-^*(k_\pm), k_\pm, 0} \right. \\
& \left. - c_r \sum_{k_->k_-^*(k_\pm)} \sum_{k_+\geq 0} \binom{N-1}{k_-, k_\pm, k_+} \left(\frac{\gamma_-}{\gamma_s}\right)^{k_--k_-^*(k_\pm)} \left(\frac{\gamma_+}{\gamma_s}\right)^{k_+} \right] \\
+ \;& \gamma_\pm^{k_\pm^*} \gamma_s^{N-1-k_\pm^*} \left[\frac{c_2-d_2-c_r}{2} \binom{N-1}{0, k_\pm^*, 0} - c_r \sum_{k_->k_-^*(k_\pm^*)} \sum_{k_+\geq 0} \binom{N-1}{k_-, k_\pm^*, k_+} \left(\frac{\gamma_-}{\gamma_s}\right)^{k_-} \left(\frac{\gamma_+}{\gamma_s}\right)^{k_+} \right. \\
& \left. - c_r \sum_{k_\pm>k_\pm^*} \sum_{k_-\geq 0} \sum_{k_+\geq 0} \binom{N-1}{k_-, k_\pm, k_+} \left(\frac{\gamma_-}{\gamma_s}\right)^{k_-} \left(\frac{\gamma_\pm}{\gamma_s}\right)^{k_\pm-k_\pm^*} \left(\frac{\gamma_+}{\gamma_s}\right)^{k_+} \right] \qquad (10)
\end{aligned}
$$

where we split the first sum of positive terms into two equal parts and then dropped in the second part all positive terms except those for which $k_+^*(\cdot) = 0$. For γ_s sufficiently close to one γ_+, γ_-, and $\gamma_\pm$ both become arbitrarily close to zero, the negative terms after c_r go to zero and the total terms in the square brackets are strictly positive. Hence, there is a $\underline{\gamma}_s < 1$ such that for all γ with $\underline{\gamma}_s < \gamma_s < 1$ holds $\bar{\pi}_+(\gamma) > \bar{\pi}_s(\gamma)$. Note that $\underline{\gamma}_s$ does not depend on the cost of punishing c_p.

Now let $\alpha \equiv \inf_{\gamma_s=\underline{\gamma}_s}\{(\bar{\pi}_+(\gamma) - \bar{\pi}_s(\gamma))\}$. Since $\bar{\pi}_\pm(\gamma) \geq \bar{\pi}_+(\gamma) - c_p$ we have for $c_p < \frac{\alpha}{2}$ that $\bar{\pi}_\pm(\gamma) > \bar{\pi}_s(\gamma) + \frac{\alpha}{2}$ for all γ with $\gamma_s = \underline{\gamma}_s$. Since all payoffs are continuous in γ this extends to a neighborhood and we can select an $\overline{\gamma}_s$ such that for all γ with $\underline{\gamma}_s < \gamma_s < \overline{\gamma}_s$ holds $\bar{\pi}_\pm(\gamma) > \bar{\pi}_s(\gamma)$.

Finally, we show for any value $\overline{\gamma}_s$ such that $\underline{\gamma}_s < \overline{\gamma}_s < 1$ that for sufficiently small costs of punishment $c_p > 0$ holds $\bar{\pi}_-(\gamma) > \bar{\pi}_s(\gamma)$ for all γ with $0 < \gamma_s < \overline{\gamma}_s$. For such γ at least one of the following three conditions must hold. $\frac{\gamma_-}{3} > \overline{\gamma}_s$ or $\frac{\gamma_+}{3} > \overline{\gamma}_s$ or $\frac{\gamma_\pm}{3} > \overline{\gamma}_s$. In the first case the expression after c_p in Equation 9 is bounded and for sufficiently small $c_p > 0$ the results follows.

In the second case an analogous argument applies to the following equation

$$\begin{aligned}
&\bar{\pi}_-(\gamma) - \bar{\pi}_s(\gamma) \\
&= \sum_{(k_-,k_\pm)\in\mathcal{U}_{-,\pm}} \gamma_+^{k_+^*(k_-,k_\pm)} \gamma_\pm^{k_\pm} \gamma_s^{N-1-k_+^*(k_-,k_\pm)-k_\pm-k_-} \gamma_-^{k_-} \left[(c_2 - d_2) \binom{N-1}{k_+^*(k_-,k_\pm), k_\pm, k_-} \right. \\
&\qquad \left. -c_p \sum_{k_+ < k_+^*(k_-,k_\pm)} \binom{N-1}{k_+, k_\pm, k_-} \left(\frac{\gamma_s}{\gamma_+}\right)^{k_+^*(k_-,k_\pm)-k_+} \right]. \qquad (11)
\end{aligned}$$

In the third case the analogous argument applies to

$$\begin{aligned}
&\bar{\pi}_-(\gamma) - \bar{\pi}_s(\gamma) \\
&= \sum_{(k_+,k_-)\in\mathcal{U}_{+,-}} \gamma_+^{k_+} \gamma_\pm^{k_\pm^*(k_+,k_-)} \gamma_s^{N-1-k_+-k_\pm^*(k_+,k_-)-k_-} \gamma_-^{k_-} \left[(c_2 - d_2) \binom{N-1}{k_+, k_\pm^*(k_+,k_-), k_-} \right. \\
&\qquad \left. -c_p \sum_{k_\pm < k_\pm^*(k_+,k_-)} \binom{N-1}{k_+, k_\pm, k_-} \left(\frac{\gamma_s}{\gamma_\pm}\right)^{k_\pm^*(k_+,k_-)-k_\pm} \right]. \qquad (12)
\end{aligned}$$

Jointly these three cases complete the proof of Lemma 5.

We now use it to prove Proposition 4. As a preamble note that from any interior state, and for any regular selection dynamics, the population state does not reach the boundaries in finite time, i.e. no preference-type vanishes completely in finite time.[25]

First, we show that for any initial interior state γ with $\gamma_s > \underline{\gamma}_s$ we will after a finite time reach a state with $\gamma_s \leq \underline{\gamma}_s$, and once $\gamma_s \leq \underline{\gamma}_s$ holds it will continue to hold forever. The latter follows from $\min\{\bar{\pi}_-(\gamma), \bar{\pi}_+(\gamma), \bar{\pi}_\pm(\gamma)\} > \bar{\pi}_s(\gamma)$for all γ with $\underline{\gamma}_s < \gamma_s < \overline{\gamma}_s$. Consider any initial interior state $(\gamma_+^0, \gamma_-^0, \gamma_\pm^0, \gamma_s^0)$. Then, as long as $\underline{\gamma}_s < \gamma_s$, we must have $\frac{\gamma_s^t}{\gamma_+^t} \leq \frac{\gamma_s^0}{\gamma_+^0}$, which, since $\gamma_+^t \leq 1 - \gamma_s^t$, implies in particular $\gamma_s^t \leq \frac{\gamma_s^0}{\gamma_s^0+\gamma_+^0}$. Consider now the set $A \equiv \{\gamma : \underline{\gamma}_s \leq \gamma_s \leq \frac{\gamma_s^0}{\gamma_s^0+\gamma_+^0}\}$. For all $\gamma \in A$ holds $\bar{\pi}_+(\gamma) > \bar{\pi}_s(\gamma)$ and hence by payoff monotonicity $g_+(\gamma) > g_s(\gamma)$.

Since A is compact we have that $\epsilon \equiv \inf_{\gamma\in A}\{g_+(\gamma) - g_s(\gamma)\} > 0$. Hence, for all $t \geq 0$ such that $\gamma^t \in A$

$$\frac{d}{dt}\left(\frac{\gamma_s^t}{\gamma_+^t}\right) = \frac{\dot{\gamma}_s^t}{\gamma_+^t} - \frac{\gamma_s^t \dot{\gamma}_+^t}{\left(\gamma_+^t\right)^2} = \left(g_s\left(\gamma^t\right) - g_+\left(\gamma^t\right)\right)\frac{\gamma_s^t}{\gamma_+^t} \leq -\epsilon\frac{\gamma_s^t}{\gamma_+^t} \qquad (13)$$

and hence $\frac{\gamma_s^t}{\gamma_+^t} \leq \frac{\gamma_s^0}{\gamma_+^0} e^{-\epsilon t}$ and therefore $\gamma_s^t \leq \frac{\gamma_s^0}{\gamma_+^0} e^{-\epsilon t}$ for all $t \geq 0$ such that $\gamma^t \in A$.[26] It follows directly that for all $t \geq -\frac{1}{\epsilon}\ln\left(\frac{\underline{\gamma}_s}{\gamma_s^0}\gamma_+^0\right)$ it must hold that $\gamma_s^t \leq \underline{\gamma}_s$.

Second, for any, arbitrarily small number ε with $\underline{\gamma}_s > \varepsilon > 0$, we can find a time $\tilde{t}$ such that $\gamma_s^t < \varepsilon$ holds for all times $t > \tilde{t}$. This follows from an argument completely analogous to

[25] Compare e.g. Weibull [42] page. 141

[26] See Weibull [42]. p.146 for a related formal argument in the context of showing that the share of a strictly dominated strategy converges to zero for payoff monotonic dynamics.

the previous one, where we consider the compact set $B \equiv \{\gamma : \varepsilon \leq \gamma_s \leq \underline{\gamma}_s\}$ and note that $\inf_{\gamma \in B}\{g_-(\gamma) - g_s(\gamma)\} > 0$.

We complete step 3 by showing that γ_+^t and $\gamma_\pm^t$ also converge to zero, and hence γ_- converges to one. It suffices to show that there is an $\varepsilon > 0$ such that for all γ with $\gamma_s \leq \varepsilon$ (again a compact set) holds $\overline{\pi}_-(\gamma) > \max\{\overline{\pi}_+(\gamma), \overline{\pi}_\pm(\gamma)\}$. We find a lower bound for $\overline{\pi}_-(\gamma) - \max\{\overline{\pi}_+(\gamma), \overline{\pi}_\pm(\gamma)\}$ by distinguishing only between groups in which all $N-1$ other players (excluding the one under consideration) are either rewarders or punisher and groups in which at least one player is self-interested. In the first type of groups punishers earn $c_r > 0$ more than rewarders, while in the other groups the fitness advantage of rewarders is bounded (e.g. by $(c_2 - d_2 + c_p)$). When γ_s goes to 0, the total probability mass on the first type of groups goes to one and the total probability mass on the second type of groups goes to zero. Hence, $\overline{\pi}_-(\gamma) > \max\{\overline{\pi}_+(\gamma), \overline{\pi}_\pm(\gamma)\}$ for sufficiently small γ_- which completes step 3.

A.7 Proof of Proposition 5

Step 1: Fix a setting. Let $\overline{\pi}_{i,\Delta}(\gamma)$ denotes the expected average fitness of type i in state $\gamma \in \Gamma$ in the Δ-bounded perturbation of the model. Then, for every $\epsilon > 0$ there exists a $\Delta > 0$ such that for all $\gamma \in \Gamma$ and all $\theta, \theta' \in \{+, -, s\}$ holds

$$\left|\left(\overline{\pi}_\theta(\gamma) - \overline{\pi}_{\theta'}(\gamma)\right) - \left(\overline{\pi}_{\theta,\Delta}(\gamma) - \overline{\pi}_{\theta',\Delta}(\gamma)\right)\right| < \frac{\epsilon}{2}. \tag{14}$$

To show this, let $b \equiv c_2 - d_2 + c_p$ denote the maximal possible per subperiod difference in material payoffs for responders. Let K denote the set of all possible group compositions $\mathbf{k}$ and thus $|K|$ denotes the number of different $\mathbf{k} \in K$. We show that for all $\theta \in \{+, -, s\}$, $|\overline{\pi}_\theta(\gamma) - \overline{\pi}_{\theta,\Delta}(\gamma)| < (2|K|+1)b\Delta$, and thus equation 14 follows with $\Delta \equiv \frac{\epsilon}{4(2|K|+1)b}$ from the triangular inequality.

Let $\tilde{P}^t_{\theta,\mathbf{k}}(\gamma)$ denote the probability that a player of type θ is in a group that has the constant composition $\mathbf{k}$ in all T subperiods and in subperiod t the action profile played in this group is $a^*_{\mathbf{k}}$. Thus, $\tilde{P}^t_{\theta,\mathbf{k}}(\gamma) = P_{\theta,\mathbf{k}}(\gamma) W^t_{\mathbf{k}}(a^t_{\mathbf{k}} = a^*_{\mathbf{k}})$.

We have

$$\begin{aligned}
\left|\tilde{P}^t_{\theta,\mathbf{k}}(\gamma) - B_{\theta,\mathbf{k}}(\gamma)\right| &= \left|P_{\theta,\mathbf{k}}(\gamma) W^t_{\mathbf{k}}(a^t_{\mathbf{k}} = a^*_{\mathbf{k}}) - B_{\theta,\mathbf{k}}(\gamma)\right| \\
&= \left|\left(P_{\theta,\mathbf{k}}(\gamma) - B_{\theta,\mathbf{k}}(\gamma)\right) W^t_{\mathbf{k}}(a^t_{\mathbf{k}} = a^*_{\mathbf{k}}) - B_{\theta,\mathbf{k}}(\gamma)\left(1 - W^t_{\mathbf{k}}(a^t_{\mathbf{k}} = a^*_{\mathbf{k}})\right)\right| \\
&\leq \left|P_{\theta,\mathbf{k}}(\gamma) - B_{\theta,\mathbf{k}}(\gamma)\right| + B_{\theta,\mathbf{k}}(\gamma)\left|1 - W^t_{\mathbf{k}}(a^t_{\mathbf{k}} = a^*_{\mathbf{k}})\right| < 2\Delta.
\end{aligned} \tag{15}$$

and thus

$$\begin{aligned}
\left|\overline{\pi}_\theta(\gamma) - \overline{\pi}_{\theta,\Delta}(\gamma)\right| &\leq b\Delta + (1-\Delta)\left(\sum_{\mathbf{k}\in K}\left|\tilde{P}^t_{\theta,\mathbf{k}}(\gamma) - B_{\theta,\mathbf{k}}(\gamma)\right| b\right) \\
&< b\Delta + (1-\Delta)\, b2|K|\Delta \leq (2|K|+1)\, b\Delta
\end{aligned} \tag{16}$$

Step 2: Setting 1: Let $0 < \overline{\delta} < \frac{1}{4}\min\{\gamma^{eq}, (1-\gamma^{eq})\}$. Consider any $\delta \in (0, \overline{\delta})$.

Let $\epsilon \equiv \min_{\gamma_+ \in [\frac{\delta}{2}, \gamma^{eq}-\frac{\delta}{2}] \cup [\gamma^{eq}+\frac{\delta}{2}, 1]} (|\overline{\pi}_+(\gamma) - \overline{\pi}_s(\gamma)|) > 0$. If we choose $\Delta > 0$ sufficiently small, then, by step 1, $(\overline{\pi}_{+,\Delta}(\gamma) - \overline{\pi}_{s,\Delta}(\gamma)) > 0$ on $[\frac{\delta}{2}, \gamma^{eq} - \frac{\delta}{2}]$ and $(\overline{\pi}_{+,\Delta}(\gamma) - \overline{\pi}_{s,\Delta}(\gamma)) < 0$ on $[\gamma^{eq} + \frac{\delta}{2}, 1]$. Hence, $[\gamma^{eq} - \delta, \gamma^{eq} - \delta]$ is a δ-stable set and no disjoint set can be δ-stable. (Note that $\{\gamma : \gamma_+ = 0\}$ is asymptotically stable, but not δ-stable since the basin of attraction is smaller than $\frac{\delta}{2}$.

Setting 2: Let $0 < \overline{\delta} < \frac{1}{4}\min\{\gamma^{cut}, (1-\gamma^{cut})\}$. Consider any $\delta \in (0, \overline{\delta})$. Let $\epsilon \equiv \min_{\gamma_- \in [\frac{\delta}{2}, \gamma^{cut}-\frac{\delta}{2}] \cup [\gamma^{cut}+\frac{\delta}{2}, 1]} (|\overline{\pi}_-(\gamma) - \overline{\pi}_s(\gamma)|) > 0$. If we choose $\Delta > 0$ sufficiently small, according to step 1, then $(\overline{\pi}_{-,\Delta}(\gamma) - \overline{\pi}_{s,\Delta}(\gamma)) < 0$. on $[0, \gamma^{cut} - \frac{\delta}{2}]$ and $(\overline{\pi}_{-,\Delta}(\gamma) - \overline{\pi}_{s,\Delta}(\gamma)) > 0$ on $[\gamma^{cut} + \frac{\delta}{2}, 1 - \frac{\delta}{2}]$. Hence, $[0, \delta]$ is a δ-stable set and so is $[1-\delta, 1]$. There cannot exist any further δ-stable set. Note in particular, that any small set around the equilibrium state(s) in $[\gamma^{cut} - \frac{\delta}{2}, \gamma^{cut} + \frac{\delta}{2}]$ is not δ stable since, for instance at the upper value of the boundary plus δ, γ_- must be increasing.

Setting 3: (the steps are analogous to the proof of Proposition 3 and 4.) Step 1: From step 1 of proof of Proposition 3 we know that there is a $\hat{\delta} > 0$ such that for all γ with $1 > \gamma_- \geq \hat{\delta}$ holds $\overline{\pi}_-(\gamma) - \max\{\overline{\pi}_s(\gamma), \overline{\pi}_+(\gamma)\} > 0$. Let $\delta < \frac{\hat{\delta}}{4}$. Thus,
$\epsilon \equiv \min_{\gamma \in \{\gamma : 1-3\delta \leq \gamma_- \leq 1-\delta\}} (\overline{\pi}_-(\gamma) - \max\{\overline{\pi}_s(\gamma), \overline{\pi}_+(\gamma)\}) > 0$ and hence, again by step 1, we can find a $\Delta > 0$ such that $(\overline{\pi}_{-,\Delta}(\gamma) - \max\{\overline{\pi}_{s,\Delta}(\gamma), \overline{\pi}_{+,\Delta}(\gamma)\}) > 0$ for all γ in $\{\gamma : 1 - 3\delta \leq \gamma_- \leq 1 - \delta\}$ and thus $\{\gamma : \gamma_- \geq 1 - \delta\}$ is δ-stable.

Step 2: From setting 1 and setting 2 it follows immediately that for $\theta \in \{+, s\}$ the sets $\{\gamma : \gamma_\theta \geq 1 - \delta\}$ are not δ-stable.

Step 3: Let c_p sufficiently small such that Lemma 5 in the proof of Proposition 4 applies. Then we can find $0 < \underline{\gamma}_s < \overline{\gamma}_s < 1$ such that $\bar{\pi}_+(\gamma) > \bar{\pi}_s(\gamma)$ whenever $1 > \gamma_s \geq \underline{\gamma}_s$, and $\bar{\pi}_-(\gamma) > \bar{\pi}_s(\gamma)$ whenever $0 < \gamma_s \leq \overline{\gamma}_s$. Furthermore we know from the last paragraph of step 3 of the proof of Proposition 4 that there is an $\tilde{\gamma}_s > 0$ such that for all γ with $\gamma_s < \tilde{\gamma}_s$ holds $\overline{\pi}_-(\gamma) - \overline{\pi}_+(\gamma) > 0$.

For $\delta > 0$, let $\epsilon_1 \equiv \min_{\gamma_s \in [\underline{\gamma}_s, 1-\delta]} \{(\bar{\pi}_+(\gamma) - \bar{\pi}_s(\gamma))\}$, $\epsilon_2 \equiv \min_{\gamma_s \in [\frac{\tilde{\gamma}_s}{2}, \overline{\gamma}_s]} \{(\bar{\pi}_-(\gamma) - \bar{\pi}_s(\gamma))\}$, $\epsilon_3 \equiv \min_{\gamma_s \in [0, \tilde{\gamma}_s]} \{(\bar{\pi}_-(\gamma) - \bar{\pi}_+(\gamma))\}$, and $\epsilon \equiv \min\{\epsilon_1, \epsilon_2, \epsilon_3\}$. Step 1 guaranties that we can choose a $\Delta > 0$ such that the payoff structure that we need for step 3 of the proof of Proposition 4 is preserved for Δ-bounded perturbations and the argument there still applies.

A.8 Proof of Lemma 2

$|P_{\theta,\mathbf{k}}(\gamma) = B_{\theta,\mathbf{k}}(\gamma)| < \Delta$ holds automatically since the composition of types in a group is not changed in this robustness check and therefore $P_{\theta,\mathbf{k}}(\gamma) = B_{\theta,\mathbf{k}}(\gamma)$. Furthermore, we assumed the generic case where parameters in each of the settings without mistakes are such that the expected payoff for player 1 after one action is always strictly different from the other actions. Let g denote the minimum of the absolute value of this payoff difference for all group compositions. Payoffs are bounded. Let b denote the maximal payoff difference. If the mistake probabilities

of responders are bounded by ϵ, then, for a given action, the change in expected payoff for proposers is bounded by ϵb and if we choose $\epsilon < \frac{g}{2b}$ the proposers' optimal choice of action is not influenced by these small enough mistake probabilities. Then we can find for any $\Delta > 0$ a sufficiently small ϵ with $0 < \epsilon < \frac{g}{2b}$ such that $W((a_1, \ldots, a_T) \neq (e_1, \ldots, e_T)) < \Delta$.[27] Hence, $W(E_\Delta(\mathbf{a}) \geq W(\mathbf{a} = \mathbf{e}) > 1 - \Delta$, and thus W is a Δ-perturbation of $\mathbf{a}$.

A.9 Proof of Lemma 3

$|P_{\theta,\mathbf{k}}(\gamma) = B_{\theta,\mathbf{k}}(\gamma)| < \Delta$ holds automatically since the composition of types in a group is not changed in this robustness check, thus $P_{\theta,\mathbf{k}}(\gamma) = B_{\theta,\mathbf{k}}(\gamma)$.

At a time t we call the action that scored the highest average subjective utility up to that time the "intended" action of a player. The other action with a lower score is called the "unintended action". Let $V_t(C)$ denote the proposer's valuation of action C at time t, i.e. the average payoff player 1 received until time t in periods where he player C. Correspondingly, $V_t(D)$ denotes his valuation of the action D. Let $\Delta > 0$. Define $\tilde{\Delta} \equiv min_{k_+,k_-} \left| c_1 - d_1 + \frac{k_+}{N_2} r + \frac{k_-}{N_2} p \right| > 0$, with $k_- = 0$ in Setting 1 and with $k_+ = 0$ in Setting 2. Consider an ϵ-reinforcement learning process with $0 < \epsilon < \min\{\frac{\Delta}{8}, 1 - \left(1 - \frac{\Delta}{8}\right)^{\frac{1}{2N}}, \frac{\tilde{\Delta}}{8(r+p)}\}$.

Step 1: After $\tilde{t}_{\epsilon,\Delta}$ periods in a group the intended action of all responders corresponds with probability of at least $1 - \frac{\Delta}{8}$ to their respective preference type and will not change afterwards.

To see this consider a responder. Once he has tried all actions available at a certain node his valuation of each action will not change further. At each node he will afterwards choose the action which gives him the highest subjective utility with probability $(1 - \epsilon)$), i.e. a rewarder will reward after cooperation, a punisher will punish after defection and a self-interested type will neither reward nor punish. Note that after three periods all (of the maximally two) decision nodes of player 2 have been reached at least once and there is maximally one decision node left that has not yet been reached twice. Let $\tilde{t}_\delta$ be a time with $\tilde{t}_\delta > \frac{ln\left(1-(1-\delta/8)^{\frac{1}{N}}\right)}{ln(1-\epsilon)} + 3$ For all times $t > \tilde{t}_\delta$ the probability that for any given player all final nodes have been reached (and therefore every action has been tried by player 2 and his intended action will now always correspond to his preference type) is greater than $1 - (1 - \epsilon)^{t-3} > \left(1 - \frac{\delta}{8}\right)^{\frac{1}{N}}$. Hence, the probability that all players in position 2 in a given group play according to their respective preference type after time $\tilde{t}_\delta$ is greater than $\left(1 - \frac{\delta}{8}\right)$.

Step 2: There exists a time $\tilde{\tilde{t}}_{\epsilon,\Delta}$ such that for all $T > \tilde{t}_{\epsilon,\Delta} + \tilde{\tilde{t}}_{\epsilon,\Delta}$ with probability greater than $1 - \frac{\Delta}{8}$ the proportion of periods in which, for a given group, the intended action of any proposer in that group deviates from the one derived in Section 2[28] is below $\frac{\Delta}{8}$.

Conditional on being in a situation in which the intended action of player 2 corresponds to

[27] The proof, e.g. by induction, is straightforward.

[28] I.e. cooperation for $c_1 - d_1 + \frac{k_+ r + k_- p}{N_2} > 0$ and defection for $c_1 - d_1 + \frac{k_+ r + k_- p}{N_2} < 0$

his preference type the expected payoff of cooperation and of defection for player 1 is respectively

$$\begin{aligned} E_\epsilon(C) &= c_1 + \frac{k_+}{N_2} r + \epsilon \left(1 - \frac{2k_+}{N_2}\right) r \\ E_\epsilon(D) &= d_1 - \frac{k_-}{N_2} p - \epsilon \left(1 - \frac{2k_-}{N_2}\right) p \end{aligned} \tag{17}$$

Since $\epsilon < \frac{\tilde{\Delta}}{8(r+p)}$, we have $|E(C) - E_\epsilon(C)| < \frac{\tilde{\Delta}}{8}$ and $|E(D) - E_\epsilon(D)| < \frac{\tilde{\Delta}}{8}$.

We consider now without loss of generality the case when $E(C) > E(D)$. (The argument for the case $E(C) < E(D)$ is completely analogous.) Then we know from the definition of $\tilde{\Delta}$ that $E(C) - E(D) \geq \tilde{\Delta}$. We can thus find a time $\tilde{\tilde{t}}_{\epsilon,\Delta}$ such that for all $t > \left(\tilde{t}_{\epsilon,\Delta} + \tilde{\tilde{t}}_{\epsilon,\Delta}\right)$ with a probability of at least $\left(1 - \frac{\Delta}{4}\right)$ holds for at least a proportion $\left(1 - \frac{\Delta}{8}\right)$ that $|V_t(C) - E_\epsilon(C)| \leq \frac{\tilde{\Delta}}{4}$ and $|V_t(D) - E_\epsilon(D)| \leq \frac{\tilde{\Delta}}{4}$. This follows e.g. by the Law of the Iterated Logarithm: For any sum $S_n = \sum_{i=1}^n X_i$ of i.i.d. random variables X_i with expected value μ and standard deviation σ holds almost surely $\limsup_{n\to\infty} \left(\frac{S_n - \mu n}{\sqrt{2n \ln \ln n}}\right) = \sigma$, which implies here in particular that almost surely $\limsup_{t\to\infty} |V_t(C) - E_\epsilon(C)| = 0$ and $\limsup_{t\to\infty} |V_t(D) - E_\epsilon(D)| = 0$, which implies the statement above. Conditional on being in such a period, we have

$$V_t(C) \geq E_\epsilon(C) - \frac{\tilde{\Delta}}{4} > E(C) - \frac{\tilde{\Delta}}{4} - \frac{\tilde{\Delta}}{4} \geq E(D) + \tilde{\Delta} - \frac{\tilde{\Delta}}{2} > E_\epsilon(D) - \frac{\tilde{\Delta}}{4} + \frac{\tilde{\Delta}}{2} \geq V_t(D). \tag{18}$$

Thus for any $T > T_1 \equiv \frac{8\left(\tilde{t}_{\epsilon,\Delta} + \tilde{\tilde{t}}_{\epsilon,\Delta}\right)}{\Delta}$ holds for any group that with probability greater than $\left(1 - \frac{\Delta}{4}\right)$ that at least in a proportion $\left(1 - \frac{\Delta}{4}\right)$ of the periods the intended actions of all players in the group corresponds to the action profile derived under Assumption 2. Considering only those periods, we have i.i.d trembles with probability ϵ, and in any such period we have with a probability of at least $(1-\epsilon)^{2N} > 1 - \frac{\Delta}{8}$ that $a_t^\epsilon = a_t$. By the law of large numbers we can find a sufficiently large $T_2 > T_1$ such that for all $T > T_2$ with probability greater than $1 - \frac{\Delta}{4}$ from these periods a proportion greater or equal $(1 - \frac{\Delta}{4})$ has a realization $a_t^\epsilon = a_t$.

Together, we have with probability greater than $1 - \frac{\Delta}{2}$ that for a proportion greater than $1 - \frac{\Delta}{2}$ holds $a_t^\epsilon = a_t$. This establishes that the distribution generated by this ϵ-reinforcement learning is a Δ bounded perturbation of the action profile $\mathbf{a}$ derived under Assumption 2.

A.10 Proof of Lemma 4

We only have to show that for sufficiently small $w > 0$ holds that for all $\gamma \in \Gamma$

$$\left| P_{\theta,\mathbf{k}}^t(\gamma) - B_{\theta,\mathbf{k}}(\gamma) \right| < \Delta, \tag{19}$$

since conditionally on on type θ being in a group that has composition $\mathbf{k}$ in all T periods the property that $W_{\mathbf{k}}^t(a_{\mathbf{k}}^t = a_{\mathbf{k}}^*) \geq (1 - \Delta)$ for all $t > T\Delta$ follows immediately from the previous analysis (in case (a) we have $a_{\mathbf{k}}^t = a_{\mathbf{k}}^*$ in all periods anyway, in case (b) the proof of Lemma 2

needs only to be conditioned on type θ being in a group that has composition $\mathbf{k}$ in all T periods, and in case (c) the same is true for the proof of Lemma 3).

Let $f^t_{\theta,w}(\gamma)$ denote the fraction of of type θ players in subperiod t which are born after the last reshuffling of groups. Then, with $P^T_{\mathbf{k}}(w)$ denoting the probability that a given group of composition $\mathbf{k}$ has never a change in its composition in all T periods, for sufficiently small w holds

$$\begin{aligned} B_{\theta,\mathbf{k}}(\gamma) \geq P^t_{\theta,\mathbf{k}}(\gamma) &\geq (1 - f^t_{\theta,w}(\gamma))P^T_{\mathbf{k},w}(\gamma)B_{\theta,\mathbf{k}}(\gamma) \geq (1 - \frac{\Delta}{2})(1 - \frac{\Delta}{2})B_{\theta,\mathbf{k}}(\gamma) \qquad (20) \\ &\geq B_{\theta,\mathbf{k}}(\gamma) - \Delta. \end{aligned}$$

This holds since, when w goes to zero, $P^T_{\mathbf{k},\gamma}(w) = (1-w)^{NT}$ and $(1 - f^t_{\theta,\gamma}(w))$ both converge uniformly to one. The last statement follows since the proportion of any type θ in the offspring population is bounded from above by $\gamma_\theta C$, where $C > 0$ is a positive constant which does not depend on γ. Thus, $f^t_{\theta,w}(\gamma) \geq \frac{\gamma_\theta C\left(1-(1-w)^t\right)}{\gamma_{\theta(1-w)^t}} = C\frac{1-(1-w)^t}{(1-w)^t}$, where $(1-w)^t$ is the proportion of players (of any type in the total population) in period t who where already born before period 1, and $\left(1-(1-w)^t\right)$ is the proportion of offsprings who were born after the last group reshuffling.

References

[1] Benaim, M. and Weibull, J.W. (2003),"Deterministic Approximation of Stochastic Evolution in Games" *Econometrica* Vol. 71, 873-903

[2] Bergstrom, Theodore C. (2001), "The Algebra of Assortative Encounters and the Evolution of Cooperation", to appear in: *International Game Theory Review*

[3] Bergstrom, Theodore C. (2002), "Evolution of Social Behavior: Individual and Group Selection", *Journal of Economic Perspectives* Vol. 16, 2, pp. 67-88

[4] Bester, Helmut and Güth, Werner (1998),"Is altruism evolutionary stable?", *Journal of Economic Behavior and Organization* 34, 193-209

[5] Bowles, Samuel and Gintis, Herbert (2004),"The evolution of strong reciprocity: cooperation in heterogeneous populations", *Theoretical Population Biology*, 65, 17-28

[6] Cooper, Ben and Wallace, Chris (2001), "Group Selection and the Evolution of Altruism", Oxford Discussion Paper Series, Nr. 67

[7] Dekel, Eddie, Jeffrey C. Ely and Okan Yilankaya, (2007), "Evolution of Preferences", *Review of Economic Studies*, 74, 685-704

[8] Ely, Jeffrey C. and Okan Yilankaya, (2001) "Nash Equilibrium and the Evolution of Preferences", *Journal of Economic Theory* 97, 255-272

[9] Eshel, Ilan, Larry Samuelson and Avner Shaked (1998) "Altruists, Egoists, and Hooligans in a Local Interaction Model",*American Economic Review*, 88,1, 157-179

[10] Fehr, Ernst and Gächter, Simon (2000),"Fairness and retaliation: The economics of reciprocity", *Journal of Economic Perspectives*, 14, 159-181

[11] Fehr, Ernst and Schmidt, Klaus (2000),"Theories of Fairness and Reciprocity - Evidence and Economic Applications", (paper prepared for the invited session of the 8th World Congress of the Econometric Society)

[12] Frank, R.H. (1987),"If homo economicus could choose his own utility function, would he want one with a conscience?", *American Economic Review* 77, 593-604

[13] Friedman, D. and Singh, N. (1999), "On the viability of vengance", UC Santa Cruz, Mimeo

[14] Gintis, Herbert (2000), "Strong Reciprocity and Human Sociality", Journal of Theoretical Biology 206, 169 - 179

[15] Güth, Werner (1995), "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives", *International Journal of Game Theory* 24, 323-44

[16] Güth, Werner and Yaari (1992), Menahem, "An evolutionary approach to explain reciprocal behavior in a simple strategic game," in: U. Witt(Editor), *Explaining Process and Change: Approaches in Evolutionary Economics*, Ann Arbor: The University of Michigan Press, 23-34

[17] Guttman, Joel M. (2003), "Repeated Interaction and the Evolution of Preferences for Reciprocity", *The Economic Journal*, 113, 631-656

[18] Hart, Sergiu (2000), "Evolutionary dynamics and backward induction", *Games and Economic Behavior* 41, 227-264

[19] Höffler, Felix (1999),"Some play fair, some don't. Reciprocal fairness in a stylized principal-agent problem", *Journal of Economic Behavior & Organization*, Vol.38, 113-131

[20] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis and R. McElreath (2001), "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies", *American Economic Review*

[21] Huck, S. and Oechssler, J. (1999),"The Indirect Evolutionary Approach to Explaining Fair Allocations", *Games and Economic Behavior* 28, 13-24

[22] Jehiel, Philippe and Dov Samet (2005) "Learning to play games in extensive form by valuation", *Journal of Economic Theory* 124, 129-148

[23] M. Kandori, G. Mailath and R.Rob (1993) "Learning, Mutation, and Long-Run Equilibria in Games.", *Econometrica*, 61:29-56

[24] Kuzmics, C. (2003), "Stochastic Evolutionary Stability in Generic Extensive Form Games of Perfect Information",*Games and Economic Behavior*, forthcoming

[25] Kuzmics, C. (2003),"Individual and Group Selection in Symmetric 2-Player Games", Mimeo

[26] Kuzmics, C. and Rodriguez-Sickert, C. (2009) "The Evolution of Moral Codes of Behavior", Mimeo

[27] Levine, D.K., (1998),"Modeling altruism and spitefulness in experiments", *Review of Economic Dynamics*, 1, 593-622

[28] Maynard Smith, J. (1964), "Group Selection and Kin Selection", *Nature*, March 14, 201, 1145-147

[29] Nöldeke, G. and Samuelson, L. (1993), "An evolutionary analysis of backward and forward induction", *Games and Economic Behavior*, 5, 425-454

[30] Ok, E.A. and Vega-Redondo,F. (2001),"On the evolution of individualistic preferences: an incomplete information scenario", *Journal od Economic Theory*, 97, 231-254

[31] Price, G.R. (1970),"Selection and Covariance", *Nature* 277, 520-521

[32] Robson, Arthur J. (1990),"Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake", *Journal of theoretical Biology*, 144, 379-396

[33] Robson, Arthur J. and Fernando Vega-Redondo (1996), "Efficient Equilibrium Selection in Evolutionary Games with Random Matching", *Journal of Economic Theory*, 70, 65-92

[34] Samuelson, L. (1997), "Evolutionary Games and Equilibrium Selection", MIT Press

[35] Samuelson,L. (2001),"Analogies, Adaptation, and Anomalies", *Journal of Economic Theory*, 97, 320-366

[36] Sandholm, B. (2001), "Preference Evolution, Two-Speed Dynamics, and Rapid Social Change", *Review of Economic Dynamics*, 4, 637-679

[37] Sethi, R. (1996), "Evolutionary stability and social norms", *Journal of Economic Behavior and Organization*, 29, 113-140

[38] Sethi, R. and Somananthan, E. (2001), "Preference Evolution and Reciprocity", *Journal of Economic Theory*, Vol. 97, 273-297

[39] Sethi, R. and Somananthan, E. (2003), "Understanding Reciprocity", *Journal of Economic Behavior and Organization*, 50, 1-27

[40] Sobel, J. (2005), "Interdependent Preferences and Reciprocity", *Journal of Economic Literature*, XLIII, 392-436

[41] Sober, E. and Wilson, D.S. (1998),"UNTO OTHERS - The Evolution and Psychology of Unselfish Behavior", Cambridge(M.A.): Harvard University Press

[42] Weibull, J. W. (1995) "Evolutionary Game Theory", MIT Press

[43] Weibull, J. W. and Salomonson, M. (2005) "Natural selection and social preference", *Journal of Theoretical Biology*,

B Supplementary Material

B.1 Definitions from Evolutionary Game Theory

The following presentation adapts definitions from Weibull [42], Chapert 6 to our setting.

Definition 3 *A regular growth-rate function is a Lipschitz continuous function $g : X \to \mathbb{R}^{|\Theta|}$ with open domain $X \subset \mathbb{R}^{|\Theta|}$ containing Γ, such that $g(\gamma) \cdot \gamma = 0$ for all $\gamma \in \Gamma$.*

These regularity conditions on the growth rates guarantee a unique solution to system 2 of differential equations. The condition $g(\gamma)\cdot\gamma = 0$ guarantees that the sum of the population shares remains 1, as $\sum_{\theta\in\Theta} \dot{\gamma}_\theta = g(\gamma) \cdot \gamma = 0 \quad \forall\gamma \in \Gamma$. For any initial state $\gamma^0 \in \Gamma$ the dynamic will remain in Γ. Together with the Lipschitz continuity on an open domain $X \supset \Gamma$ this guarantees a unique solution $\xi(\cdot, \gamma^0) : \mathbb{R} \to X$ through every initial state $\gamma^0 \in \Gamma$.Furthermore $\xi(\tau, \gamma^0)$ is continuous in $\tau \in \mathbb{R}$ and $\gamma^0 \in \Gamma$.

Definition 4 relates the evolutionary success of each type to the obtained fitness payoffs. Types that earn greater average fitness have higher growth-rates.

Definition 4 *A regular growth-rate function g is called payoff monotonic if*

$$\bar{\pi}_\theta(\gamma) < \bar{\pi}_{\theta'}(\gamma) \Leftrightarrow g_\theta(\gamma) < g_{\theta'}(\gamma) \qquad \forall\ \theta, \theta' \in \Theta, \gamma \in \Gamma. \tag{21}$$

Payoff monotonicity captures the process of evolutionary selection. Whether a limit point of this dynamic process is robust to mutations depends on the stability of this state.

Definition 5 *A population-state $\gamma \in \Gamma$ is called* **Lyapunov stable** *if every neighborhood B of γ contains a neighborhood B^0 of γ such that $\xi(\tau, \gamma^0) \in B$ for all $\gamma^0 \in B^0 \cap \Gamma$and $\tau \geq 0$.*
A state $\gamma \in \Gamma$ is called **asymptotically stable** *(or simply* **stable***) if it is Lyapunov stable and there exists a neighborhood B^* such that $\lim_{\tau\to\infty} \xi(\tau, \gamma^0) = \gamma$ holds for all $\gamma^0 \in B^* \cap \Gamma$.*

Intuitively, Lyapunov stability guaranties that sufficiently small mutations do not lead far away from the considered population state. Asymptotic stability requires in addition that after small mutations the population is driven back to the stable state by the selection dynamics. These stability concepts can be generalized to sets.

Definition 6 *A* **closed set** *$A \subset \Gamma$ is* **Lyaponov stable** *if every neighborhood B of A contains a neighborhood B^0 of A such that $\xi(\tau, \gamma^0) \in B$ for all $\gamma^0 \in B^0 \cap \Gamma$ and $\tau \geq 0$.*

A closed set $A \subset \Gamma$ is **asymptotically stable** *if it is Lyapunov stable and if there exists a neighborhood B^* of A such that $\xi(\tau, x^0)_{\tau\to\infty} \to A$ for all $x^0 \in B^* \cap \Gamma$.*[29]

B.2 Numerical Examples

B.2.1 Setting 1

Figure 5 illustrates the evolutionary dynamics for a numerical example in Setting 1.

[29]We say $\xi(\tau, x^0)_{\tau\to\infty} \to A$, where A is a closed set, if the distance $d(\xi(\tau, x^0), A)\tau \to \infty \to 0$. The distance $d(y, A)$ between a point $y \in \Gamma$ and a closed set A is defined as the minimal distance between y and any point $a \in A$, i.e. $d(y, A) = \min_{a\in A} d(y, A)$.

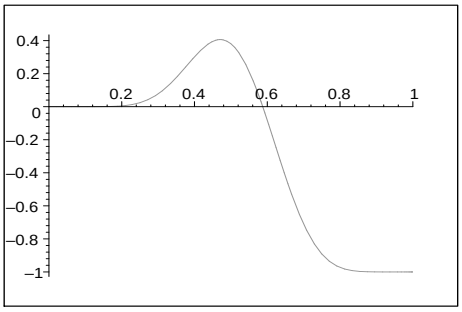


Figure 5: The difference in average fitness between rewarders and self-interested responders $(\bar{\pi}_+ - \bar{\pi}_s)$ plotted as function of γ_+ for $N = 20, d_1 = 1, c_1 = 0, r = 2, d_2 = 5, c_2 = 0, c_r = 1$.The fraction of rewarders in the stable equilibrium of this example is $\gamma^{eq} \approx 0.5876$. If the fraction γ_+ of rewarding individuals is below γ_+^{eq} then they earn a higher average fitness and their fraction γ_+ increases. If $\gamma_+ > \gamma_+^{eq}$ rewarding players earn less and γ_+ decreases. Due to the continuity of the evolutionary dynamics, γ_+ converges to γ_+^{eq}.

B.2.2 Setting 2

Figure 6 illustrates the evolutionary dynamics in Setting 2.

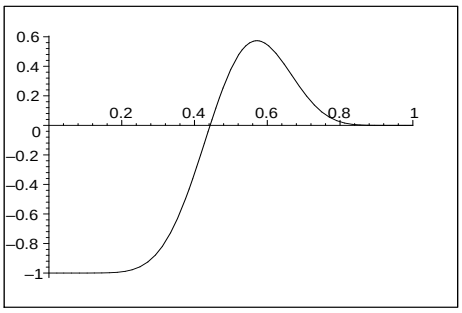


Figure 6: The difference in average fitness between punishers and self-interested individuals $(\bar{\pi}_- - \bar{\pi}_s)$ is plotted as a function of γ_- for $N = 20, d_1 = 1, c_1 = 0, p = 2, d_2 = 5, c_2 = 0, c_p = 1$. The mixed equilibrium at $\gamma^{cut} \approx 0.443$ is unstable and separates the basins of attraction of both stable monomorphic equilibria. If $\gamma_- < \gamma_-^{cut}$ punishers perform worse and γ decreases to 0. If $\gamma > \gamma^{cut}$ punishers perform better and γ_- increases to 1.

B.2.3 Setting 3

Figure 7 shows the dynamics for parameters that fall into Part a of Corollary 1

B.3 Comparative Statics of γ^{cut} in setting 2

Let γ^{cut} be the fraction of punishers in the unstable mixed equilibrium. This fraction separates the basins of attraction of the stable equilibria. If the initial fraction of punishing players is below the cutoff γ^{cut} then this fraction decreases until the entire population has self-interested

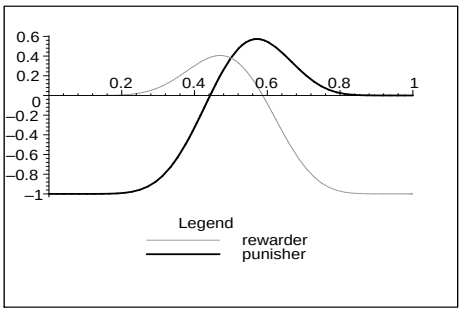


Figure 7: Setting 3 with $\gamma_+^{eq} > \gamma_-^{cut}$: $(\bar{\pi}_+ - \bar{\pi}_s)$ and $(\bar{\pi}_- - \bar{\pi}_s)$ as functions of $\gamma \equiv \gamma_+ + \gamma_-$ for $N = 20, d_1 = 1, c_1 = 0, r = 2, d_2 = 5, c_2 = 0, c_r = 1$.

preferences and nobody punishes defection. If, on the other hand, the initial fraction of punishing players is above the cutoff γ^{cut}, then this fraction increases until the entire population has preferences for punishing. One might therefore interpret the value of γ^{cut} as an indicator for how likely it is to end up in one or the other equilibrium.[30] The comparative statics of γ^{cut} is analogous to setting 1 and can be derived directly from equation **??**.

Higher costs of punishing diminish the basin of attraction of the punisher equilibrium:

Proposition 6 *If the costs c_p - which a player 2 has to bear in order to punish - increase, then γ^{cut} increases, i.e. there have to be initially more punishers in order to end up in the punishing equilibrium. Furthermore,* $\lim_{c_p \to 0} \gamma^{cut} = 0$ *and* $\lim_{c_p \to \infty} \gamma^{cut} = 1$.

The intuition is straightforward: the higher the number of punishing players, the cheaper it is to be a punisher. If the costs of punishing increase, punishers become less fit. Hence punishing players need a higher fraction of punishers in order to be at least as successful as non-punishers.

Higher gains from cooperation for player 2 are good for punishers. Hence the basin of attraction for their equilibrium becomes larger:

Proposition 7 *If the gains of cooperation for the players 2 $(c_2 - d_2)$ increase, then γ^{cut} decreases, i.e. a lower initial fraction of punishing players is necessary in order to end up in the punishing equilibrium. Furthermore,* $\lim_{(c_2 - d_2) \to 0} \gamma^{cut} = 1$ *and* $\lim_{(c_2 - d_2) \to \infty} \gamma^{cut} = 0$.

Again, the intuition is straightforward: the higher the gains of cooperation for a player 2, the higher his profit from being pivotal in inducing cooperation of players 1. Therefore a lower fraction of punishers is necessary in order to make punishing more successful than non-punishing.

Lemma 6 *If the threshold k_-^* of punishing players 2 in a group above which the players 1 start to cooperate increases then γ^{cut} increases, i.e. there are more punishing players necessary in order to end up in the punishing equilibrium.*

Intuitively, a higher threshold k_-^* makes it more probable for an individual to be in a group in which the number of punishers is too low to induce cooperation. In these groups being a punisher is costly. Therefore, fitness of punishers is lower and a higher initial fraction of punishers is necessary to make punishing more successful than non-punishing.

[30] Again, this interpretation is in the spirit of the model by Kandori et. al [23], where the size of the basins of attraction determines the long run equilibrium

Corollary 2 *If player 1's costs for cooperation* $(d_1 - c_1)$ *increase, then* γ^{cut} *increases weakly, i.e. a higher or equal fraction of punishers is necessary in order to end up in the punishing equilibrium.*

Corollary 3 *If player 2's losses due to a punishment* p *increase, then* γ^{cut} *decreases weakly.*