Center for European, Governance and Economic Development Research

cege

Discussion Papers

Number 221 – November 2014

EVOLUTION OF COOPERATION IN SOCIAL DILEMMAS: SIGNALING INTERNALIZED NORMS

Stephan Müller, Georg von Wangenheim

Georg-August-Universität Göttingen

ISSN: 1439-2305

Evolution of cooperation in social dilemmas: signaling internalized norms.

Stephan Müller, Georg von Wangenheim[*]

Abstract

The article suggests a new explanation for cooperation in large, unstructured societies that avoids the restrictions required in most previous attempts. Our explanation deals with the role of internalized norms. Even internalized norms, i.e. norms that alter the perceived utility from acting in a cooperative or uncooperative way, will not help to overcome a dilemma in an unstructured society, unless individuals are able to signal their property of being a norm bearer. Only when having the norm may be communicated in a reliable way, can the picture change. We derive necessary and sufficient conditions for cooperation to be part of an asymptotically stable equilibrium of an evolutionary dynamics of signaling norm internalization, behavior and norm adoption. These conditions put the signaling costs of norm-adopters and non-adopters, the strength of the social norm and two parameters measuring the cost of cooperation into relation with each other.

Keywords: Evolution - Cooperation – Signaling

JEL Classifications: A13, D02, D21

[*] Stephan Müller (corresponding author): Göttingen University, Platz der Göttinger Sieben, 3, 37073 Göttingen, Germany and Georg v. Wangenheim: Kassel University, Nora-Platiel-Straße 4, 34109 Kassel, Germany (email stephan.mueller@wiwi.uni-goettingen.de and g.wangenheim@uni-kassel.de)

1. Introduction

Despite the obvious advantages of exploiting the good will of others, human beings often cooperate, even in large, unstructured societies. However, cooperation is neither universal nor is it easy to explain. Economists have a long tradition of finding that the evolution of cooperation in large, unstructured societies is a puzzle (e.g. Axelrod and Hamilton, 1981, Fudenberg et al., 2012); and in explaining cooperation based on some structure within the population.

Attempts to solve the puzzle are abundant but have thus far commonly relied on one or both of two restrictions. The first restriction is that explanations have focused on structured populations, in which interactions are not completely anonymous but allows individuals to collect and process information about past behavior of others and about their identity. The second restriction is that explanations have depended on an unexplained ability of social norms to restrict the individuals' action or strategy spaces, in particular, with respect to the abuse of punishment.

With respect to the first group of restrictions, some strands of the literature deserve special mention.[1] The theory of kin selection focuses on cooperation among individuals who are genetically closely related (Hamilton, 1964a, Hamilton, 1964b), whereas theories of direct reciprocity focus on incentives to cooperate in repeated interactions of self-interested individuals (Trivers, 1971, Axelrod, 1984). For infinite repetition within one group, see Taylor (1976) or Mordecaï (1977) and for Folk-Theorem-type of results Rubinstein (1979) or Fudenberg and Maskin, (1986). For indefinite repetition, see Kreps et al. (1982). The theories of indirect reciprocity and costly signaling show how cooperation in larger groups can emerge when those cooperating can build a reputation (Nowak and Sigmund, 1998; Wedekind and Milinski, 2000; Gintis et al., 2001)[2].

In terms of the second set of exclusions, we point to early papers of Hirshleifer and Rasmusen (1989) and Witt (1986) that allow for punishment only after a norm has been violated. Sethi (1996) allows for all possible strategies which condition punishment on either the violation of or compliance with a norm. However, he then adds structure to the society by introducing some exogenous division of the population – the behavior of some individuals is rational, and for the rest it is determined by routines that are slowly adapted to their environment.

We present a new explanation for cooperation that avoids both restrictions. Our explanation focuses on cooperation in large unstructured populations of individuals whose incentives to use or abuse actions or strategies evolve endogenously from the model. We assume that their behavioral routines adapt to the sum of both objective and subjective payoffs and that their subjective payoffs – which express internalized norms – slowly evolve according to the objective payoffs. This allows us to explain all variation among individuals endogenously and to assume absence of any information on the past behavior of other individuals.

[1] A complete literature review lies outside the scope of an introductory section of a journal article, as it would merit a scholarly work in its own right.

[2] There are other mechanisms that do not rely on informational aspects. Instead, they are based on restrictions in rationality or on extended strategy spaces. In finitely repeated games, cooperation can, for example, result from bounded complexity of strategies (Neyman, 1985), history-dependent payoffs (Janssen et al., 1997) or bounded complexity of beliefs (Harrington, 1987).

We place our model in an environment that is most unfavorable to cooperation, a completely unstructured society where every interaction occurs among strangers. We do this for two reasons. The first reason is methodological: we want to isolate the impact of internalized norms from other factors that might stabilize cooperation. The other is empirical: we believe that in modern societies a non-negligible part of everyday interactions is characterized by cooperation in dilemma situations although they actually do take place in an unstructured environment (for a survey on experimental evidence see Roth, 1995, Cooper et al., 1996).

In such an environment, cooperation cannot be induced by any form of repeated interaction[3] nor by social norms based on sanctions to be inflicted in later interactions. Even internalized norms, i.e. norms that alter the perceived utility from acting in a cooperative or uncooperative way, will not help to overcome a dilemma in an unstructured society, unless – and this is the thrust of the current paper – individuals are able to signal their property of being a norm bearer[4]. If internalized norms simply exist while lacking the possibility of being signaled or screened for, they would induce norm bearers to cooperate and be exploited by others. Hence, norm bearers would have a clear evolutionary disadvantage so that norm adoption would vanish. Only when internalization of the norm can be communicated in a reliable way, may the scenario change, because behavior may then be conditioned on the expected behavior of others.

Within this environment, we borrow two elements from the indirect evolutionary approach (Güth and Yaari, 1992 and Güth, 1995): first, the idea that internalized norms are nothing more than an internal payoff conditional on the behavior of the individual and its partners, and second, the assumption that the adoption of an internalized norm evolves slowly depending on its effects on material, external payoffs. Our approach is thus closely related to Güth et al. (2000), who analyzes the Game of Trust rather than the Prisoners' Dilemma. The two games are clearly similar since in the Game of Trust, the outcome of the first mover trusting and the second mover reciprocating is Pareto superior to the unique Nash equilibrium. In Güth's model, evolution allows for heterogeneity with respect to the evaluation of the material outcome such that some agents will reciprocate and some will exploit trust as second movers. By adding the opportunity of partially informative but costly screening of this evaluation to the standard Game of Trust, Güth opens the path to equilibria in which the first mover trusts and the second reciprocates. We carry this approach over to the Prisoners' Dilemma and concentrate on signaling, instead of screening.

In addition to these differences with respect to the interaction environment, we depart from the standard indirect evolutionary approach in a fundamental way concerning the behavioral assumptions. We assume that agents play inherited strategies defining both whether the agents signal their norm internalization and whether they cooperate or not. We thus take the stand of

[3] Kandori (1992) and Ellison (1994) show that in an environment with similar informational restrictions as in our model, contagious strategies may support cooperation in a social dilemma in an extremely indirect way of repeated interaction. In such strategies, when one player defects in one period, his opponent of that interaction will start to defect from this period onwards, infecting other player who will defect in the future, infecting others and so forth. For any given fixed population size, Kandori (1992) and Ellison (1994) show that cooperation can be sustained in a sequential equilibrium if individuals exhibit enough patience. However, such contagious strategies may only uphold complete cooperation by all individuals in large societies, if patience is nearly infinite. In addition, they are not tolerant with respect to behavioral errors. We therefore do not discuss this approach in detail.

[4] For an empirical paper on the role of costly signaling for the promotion of intragroup cooperation, see Soler (2012).

behavioral economics (as it is often reflected in evolutionary game theory) whereas Güth et al. (2000) apply a rational choice approach with agents using Bayesian updating and making rational investment decisions with respect to screening. Our model is thus evolutionary with respect to both norm internalization and behavior, although the speed of the norm internalization dynamics is clearly less than the speed of behavioral adaptation.

In the field of evolutionary biology it has been argued before that signaling may provide way out of social dilemmas where mechanisms such as reputation, reciprocity or assortative matching are absent or fail to work sufficiently (e.g. Wright, 1999, Smith and Bliege Bird, 2000, Leimar and Hammerstein, 2001). Yet only a few of these approaches incorporate a formal model (Gintis et al., 2001). The novelty of our approach is the derivation of the full set of behavioral equilibria, i.e. all separating, pooling and semi-pooling equilibria of the signaling-extended Prisoners' Dilemma. This would be rather a technical note were it not for the implication of a far richer set of rate-of-norm-adoption equilibria that can stabilize cooperation. Notably, the interplay of those multiple behavioral equilibria may stabilize partial cooperation and dissolves the necessity to introduce specific frequency-based evolutionary forces into the dynamics of norm adoption beyond payoff monotonicity (e.g. Gintis et al., 2001 rely on the replicator dynamics).

Sethi (1996) suggests a linkage between his own approach, i.e. mixing optimizing and non-optimizing behavior in an evolutionary game; and the approach taken by Güth and Yaari (1992) and Güth and Kliemt (1994) in which all agents are assumed to optimize given heterogeneous preferences. Both authors establish the existence of games in which preferences for cooperation or fairness are evolutionary stable. Similarity in results despite differences in methodology suggest that the two research approaches are highly complementary Sethi (1996, p. 117). Our results show that the complementarity between these different approaches is limited. We show that there is substantial difference between assuming that norms simply fix a certain behavior, and assuming that norms only create internal incentives to adhere to this behavior. In our case, the parameter measuring the strength of this incentive affects the range of the other parameters for which cooperation may emerge.

The remainder of the paper proceeds as follows. The model is presented in Section 2. Since we consider a heterogeneous population composed of norm adopters and non-adopters, we first derive equilibria in each sub-population for which the stable equilibria are presented in Section 3. Thereafter, we endogenize heterogeneity and consider equilibria of the two subpopulations in Section 4. Section 5 collects and presents the requirements for partial or full cooperation being part of a stable evolutionary equilibrium. Section 6 concludes.

2. The model

The classical Prisoners' Dilemma (PD) is the most prominent and best-studied example of a social dilemma and serves as the basis for our analysis. The PD is played recurrently in an unstructured population. An *unstructured population* is defined by the anonymity of the interaction, i.e. agents process only information on outcomes of their own past interactions. In particular, they process no information on the opponent's identity or on outcomes in games in which they were not involved. To save space, payoff matrices are given from the row player's perspective. The strategy domain is finite, consisting of two strategies, C – "cooperation" and D –

"defection". In conformity with the standard evolutionary model, we assume that individuals are randomly matched into pairs with each pair having the same probability in each short time period.[5] Any pair will engage in a one-shot PD game. Table 1 below presents the material payoffs of the PD that will be decisive with respect to evolutionary success.

Material payoffs are given by:

	C	D
C	1	$-\beta$
D	$1+\alpha$	0

Table 1: Prisoners' Dilemma, where $\alpha > 0, \beta > 0$ and $1+\beta > \alpha$.

A common assumption in evolutionary models that explain the presence of cooperative behavior is that individuals play inherited strategies that may depart from payoff maximizing behavior. Playing non-maximizing strategies in this line of research is then interpreted as norm-guided (e.g. Sethi 1996). This line of argument, however, appears incomplete because while showing that such strategies can be sustained in equilibrium, it lacks motivation behind why an individual would adhere to that particular norm. We believe that individuals will not stick to any behavior that is suboptimal in the current environment. We do not claim that individuals will always do what is best for them from an objective perspective (e.g. maximizes fitness), but we argue that they will not commit to suboptimal strategies forever. Hence, in our view, any long-lasting departure from the behavior that maximizes material payoffs needs to be motivated by a valuation of the outcome of behavior that differs from the material payoffs in a substantial way. In other words, norm-guided behavior is not equivalent to an unmotivated commitment to a certain behavior, but it reflects the subjective valuation of the (physical) outcome of the game. Following this reasoning, we rely on (a variant of) the indirect evolutionary approach, pioneered by Güth and Yaari (1992)[6], i.e. we explicitly model cooperative preferences, which determine behavior, and behavior, which in turn determines fitness.

As a particular internalized norm, we focus on the case of a cooperative norm. Players carrying such an internalized preference gain an additional internal payoff if the behavioral outcome of the stage game is mutual cooperation, i.e. (C, C). We assume that there are two types in the population (high and low types). Let λ denote the share of high types in the population and let $\mathrm{m} \in \{\underline{\mathrm{m}}, \overline{\mathrm{m}}\}$ be their preference parameter measuring the attitude towards cooperation, resulting in the internal payoff matrix depicted in Table 2 below. As Güth et al. (2000) noted in a different setting, the precise level of *m* is behaviorally irrelevant. All *m*-types for whom the same inequality with respect to α holds, form an equivalence class concerning the implied behavior. We therefore normalize $\underline{\mathrm{m}} = 0, \overline{\mathrm{m}} > \alpha$.[7] The value of *m* is assumed to be private information of the agent. In the tradition of Harsanyi (1967, 1968a, 1968b), beliefs about the opponent's type are common knowledge. Like Güth and Ockenfels (2005), we adopt the natural assumption that

[5] An unstructured population need not necessarily engage in uniform or random matches, but departures from those assumptions significantly complicates analysis without changing the qualitative results since we assume that population is unstructured and remains unstructured. Non-random or non-uniform matching might however increase the chance that structure is introduced into the population.

[6] The indirect evolutionary approach has also been applied in different strategic settings (ultimatum game, Huck and Oechssler, 1999) or to analyze the evolutionary stability of altruistic preferences (Bester and Güth, 1998) or of altruistic and spiteful preferences (Possajennikov, 2000).

[7] Assuming $\overline{\mathrm{m}} > \alpha$ is necessary, since otherwise, defection would still be the dominant strategy for norm-adopters.

beliefs correspond to actual frequencies of types. Without communication, the impossibility result of Kandori (1992, Proposition 3) applies, which states that the unique equilibrium is characterized by full defection, i.e. everybody always defects.

Communication is modeled as an additional stage prior to the play of the adjusted PD. In that stage, agents can simultaneously send one message concerning their inner motive. Without loss of generality, we assume the message space to be the same as the type space. The message to be a low type corresponds to sending no message and is costless. As in the standard signaling model (Spence 1973) we assume the existence of a social technology which enables individuals to signal their positive attitude towards cooperation by incurring some costs. Furthermore, agents who adopted the norm are supposed to bear lower costs for sending the signal. Let $\bar{k}, \underline{k}$ denote the signaling cost for high types and low types respectively, so that $\bar{k} < \underline{k}$. In the current setup, strategies are given by signal-dependent behavior and the choice of sending the signal or not, e.g. "cooperate if signal is received, deviate if no signal is received and send signal", denoted $\mathrm{CD}\bar{\mathrm{m}}$. In general, a strategy is denoted by a triple XYm, where the first entry denotes behavior in case of receiving the signal (*C* or *D*), the second denotes behavior in the case of not receiving the signal (*C* or *D*), and the third signifies whether the signal is sent or not ($\bar{\mathrm{m}}$ or $\underline{\mathrm{m}}$, respectively).

What might such a signal be? To give an illustrative example, consider a situation where individuals elbow their way through a rummage sale. There is a table with one good offered as two variants, goods A and B. There are also two individuals, one preferring good A, the other preferring good B. However getting both goods is the first best outcome for both individuals. They can behave cooperatively, allowing the other to select their preferred good; or they can try to queue-jump and grab both goods, in which case, the other gets none. If both individuals chose not to cooperate, they will grab one of the goods by chance, leaving them in expectation with a lower utility then in the cooperative state. Hence, this example is structurally equivalent to a PD. In this scenario, the signal often used is to make room for the other person. Such a signal is costly in terms of time, which usually has some monetary equivalent. If this gesture is received by both individuals, this might lead to mutual cooperation. This example is also instructive in demonstrating that signaling in our context is rather part of the behavioral strategy than an act of rational choice. In the light of this example indeed many acts of courtesy may be understood as a signal for a cooperative attitude. The signals are not limited to this aspect though. Consider a one shot interaction where two agents can realize a project if both

Evaluation of material payoffs is given

	C	D
C	$1+\mathrm{m}$	$-\beta$
D	$1+\alpha$	0

Table 2: PD with preference for cooperation.

Based on the basic behavioral actions C and D, for the high types, there are eight signal-dependent strategies $\mathrm{CC}\bar{\mathrm{m}}, \mathrm{CD}\bar{\mathrm{m}}, \mathrm{DC}\bar{\mathrm{m}}, \mathrm{DD}\bar{\mathrm{m}}$ and $\mathrm{CC}\underline{\mathrm{m}}, \mathrm{CD}\underline{\mathrm{m}}, \mathrm{DC}\underline{\mathrm{m}}, \mathrm{DD}\underline{\mathrm{m}}$. For the low types, since defection is the dominant behavior, there are only two strategies that reflect their signals, denoted by $\mathrm{D}\bar{\mathrm{m}}, \mathrm{D}\underline{\mathrm{m}}$. We will denote the share in the subpopulation of high types playing the strategy $\mathrm{CC}\bar{\mathrm{m}}$ by $\mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}$ and accordingly, for any other strategy. Since low types always defect, we denote their respective shares by $\mathrm{p}_{\bar{\mathrm{m}}}$ and $\mathrm{p}_{\underline{\mathrm{m}}}$.

In evolutionary game theory, there are two approaches with respect to capturing the dynamical aspect of evolution. The first one, due to the work of Smith and Price (1973), centers on the concept of an evolutionary stable strategy and is considered as a "static" approach since typically no reference is given to the underlying process by which behavior changes in the population. The second approach does not attempt to define a particular notion of stability. By explicitly modeling the underlying dynamics, all standard stability concepts used in the analysis of dynamical systems can be applied. We will follow the second approach by modeling the dynamics of the according population shares via payoff-monotone dynamics (see e.g. Bendor and Swistak, 1998 for definitions), i.e. if the fitness payoff of a certain strategy is larger than the one of another, the share of the population following the former will increase faster than the share following the latter, or decrease slower. An equilibrium is defined by the dynamics introduced above. An equilibrium is a distribution in the shares of the population playing certain strategies, such that the dynamical process induces no further adjustments, i.e. an equilibrium is a fixed point of the adjustment process. As a stability concept, we will apply the notion of asymptotic stability (see. e.g. Samuelson, 1997 for definitions). An equilibrium of that type must be reconstituted after a small perturbation, which is arbitrary in terms of the composition of mutation-strategies.

As mentioned above, there are eight strategies for high types and two for low types. We assume that the dynamic accommodation of the population shares playing the various strategies is relatively fast compared to the dynamics of the population share of $\bar{m}$-types, i.e. λ.[8] This assumption will simplify analysis of the dynamics and is considered adequate since behavior will adapt faster to differences in payoffs than socially and culturally transmitted norms. We can therefore analyze these processes separately as long as the faster process is stable. More precisely, we apply the mathematical tool of quasi-stationary approximation, or 'adiabatic elimination' (Haken, 1977, Weidlich and Haag, 1983, used in economics by Samuelson, 1947: 320, already) of fast variables to solve the coupled differential equations which describe our system. The system consists, on the one hand, of the differential equations that describe the fast dynamics of various signal-behavior strategies and, on the other hand, of the differential equations that describe the slow dynamics of norm-adoption. The eight strategies for high types and the two for low types amount to ten differential equations, one per share per strategy, yielding nine independent equations since the size of the total population is fixed. Fixing the size of each subpopulation while analyzing the dynamics of behavioral strategies within each subpopulation reduces the number of independent differential equation by one more, seven for the high types and one for low types. We recall that p_{XYm} and p_m denote the shares of strategies *within* the subpopulations so that $\sum_{X,Y,m} p_{XYm} = 1$ with $X, Y \in \{C, D\}$ and $m \in \{\underline{m}, \bar{m}\}$ and $p_{\underline{m}} + p_{\bar{m}} = 1$.

Given our assumption on the speed of the dynamic processes, we first derive all the behavioral equilibria for a given proportion λ of individuals with a high internal motivation for (mutual) cooperation, and then analyze whether the implied λ-dynamics can support a fully or partially cooperative state. We call the former equilibria 'p-equilibria' and the latter, 'λ-equilibria'. If they are asymptotically stable with respect to the corresponding p- or λ-dynamics, we say that they

[8] This assumption implies that payoff monotonicity is restricted to the fast and to the slow dynamics, but does not comprise the combination of the two.

are p-stable and λ-stable, respectively. The p-stable equilibria are presented in section 3, and λ-stable equilibria are derived in section 4.

3. Equilibria with Exogenous Proportions of Norm Bearers

For ease of reading, we present only the equilibria and their stability properties and leave the derivation in Appendix A (existence) and B (stability). As in many other cases, we have separating and pooling equilibria, depending on the parameters including λ. There are one p-stable separating and three p-stable pooling equilibria. In the separating equilibrium, the subpopulations of the two types of individuals (high and low internal motivation for cooperation) exhibit homomorphic behavior, whereas behavior of types in the pooling equilibria is heteromorphic. However, there is a third type of equilibria where at least one subpopulation applies both types of signals, so called semi-pooling equilibria. Table 3 reports these equilibria.

In the following paragraphs, we will take a closer look at the separating and pooling equilibria. We will refer to the first of these equilibria as the '*cooperative separating equilibrium*', to the second as the '*low pooling cooperative equilibrium*', to the third as the '*low pooling defective equilibrium*' and to the fourth as the '*high pooling cooperative equilibrium*'. It turns out that the semi-pooling equilibria with one exception are less important for the implied λ-dynamics and are therefore not further discussed. The exception is the p-stable semi-pooling equilibrium at $\lambda = \frac{\mathrm{k}}{1+\alpha}$ that will be of relevance for one of the inner λ-stable equilibria. In this semi-pooling equilibrium, high types always play $\mathrm{CD}\bar{\mathrm{m}}$ and low types are indifferent between sending the signal or not, and therefore $\mathrm{p}_{\bar{\mathrm{m}}}$ is undefined. The minor importance of all other p-stable semi-pooling equilibria is partly due to their being characterized by strictly negative fitness differentials between high and low types and partly to their limited λ-support (see Figure 1-Figure 2).

In the cooperative separating equilibrium, the high types recognize each other and cooperate only among themselves. The intuition behind the fact that the support of this equilibrium has both a lower and an upper is as follows: If there are too few high types, then the cooperative outcome among them cannot compensate for the signaling costs. The higher the signaling costs relative to the (non-material) reward for a cooperative outcome, the higher the required share of high types in the population. If on the other hand, there are too many high types, signaling becomes sufficiently profitable for low types. In other words, if there are enough high types that cooperate when receiving the cooperative signal, it becomes profitable for low types to incur the signaling costs. The higher the signaling cost for low types relative to what can be gained from defection against a cooperative opponent, the higher is the share of high types needed for signaling to become a profitable strategy for low types. The thresholds for the share of high types have a precise economic interpretation. For high types, the cost-benefit ratio from signaling ($\frac{\bar{\mathrm{k}}}{1+\bar{\mathrm{m}}}$) must be smaller than the probability to gain the benefit (λ). The reverse holds true for low types, i.e. their cost-benefit ratio from signaling must exceed ($\frac{\mathrm{k}}{1+\alpha}$), the likelihood of gaining the benefit.

Type	Involved strategies	Equilibrium	Support	Conditions for existence	Payoff differentials (superscript "f" indicates the difference in fitness payoffs)
Separating	High types cooperate against signal and defect against no signal. Norm holders signal, while others do not signal.				
	$\mathrm{CD}\bar{\mathrm{m}}$ $\underline{\mathrm{m}}$	$p_{\mathrm{CD}\bar{\mathrm{m}}}=1,\ p_{\underline{\mathrm{m}}}=1$	$\frac{\bar{k}}{1+\bar{m}}<\lambda<\frac{\underline{k}}{1+\alpha}$	$\bar{k}<1+\bar{m}$	$\Pi_{\bar{m}}\left(\mathrm{CD}\bar{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\underline{\mathrm{m}}\right)=\lambda\left(1+\bar{m}\right)-\bar{k}$ $\left(\Pi_{\bar{m}}\left(\mathrm{CD}\bar{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\underline{\mathrm{m}}\right)\right)^{f}=\lambda-\bar{k}$
Low Pooling	High types cooperate, no signal				
	$\mathrm{CC}\underline{\mathrm{m}}$ $\mathrm{DC}\underline{\mathrm{m}}$ $\underline{\mathrm{m}}$	$p_{\mathrm{CC}\underline{\mathrm{m}}}+p_{\mathrm{DC}\underline{\mathrm{m}}}=1$ $p_{\underline{\mathrm{m}}}=1$	$\lambda\geq\frac{\beta}{\bar{m}-\alpha+\beta}$		$\Pi_{\bar{m}}\left(\mathrm{CC}\underline{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\underline{\mathrm{m}}\right)=-\left(\lambda\left(\alpha-\bar{m}\right)+\left(1-\lambda\right)\beta\right)$ $\left(\Pi_{\bar{m}}\left(\mathrm{CC}\underline{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\underline{\mathrm{m}}\right)\right)^{f}=-\left(\lambda\alpha+\left(1-\lambda\right)\beta\right)<0$
	Complete defection, no signal				
	$\mathrm{CD}\underline{\mathrm{m}}$ $\mathrm{DD}\underline{\mathrm{m}}$ $\underline{\mathrm{m}}$	$p_{\mathrm{CD}\underline{\mathrm{m}}}+p_{\mathrm{DD}\underline{\mathrm{m}}}=1$ $p_{\underline{\mathrm{m}}}=1$	$0<\lambda<1$	$p_{\mathrm{CD}\underline{\mathrm{m}}}\leq\frac{1}{\lambda}\min\left\{\frac{\bar{k}+\beta}{1+\bar{m}+\beta},\frac{\bar{k}}{1+\alpha}\right\}$	$\Pi_{\bar{m}}\left(\mathrm{CD}\underline{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\underline{\mathrm{m}}\right)=0$ $\left(\Pi_{\bar{m}}\left(\mathrm{CD}\underline{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\underline{\mathrm{m}}\right)\right)^{f}=0$
High Pooling	High types cooperate, all signal				
	$\mathrm{CC}\bar{\mathrm{m}}$ $\mathrm{CD}\bar{\mathrm{m}}$ $\bar{\mathrm{m}}$	$p_{\mathrm{CC}\bar{\mathrm{m}}}+p_{\mathrm{CD}\bar{\mathrm{m}}}=1$ $p_{\bar{\mathrm{m}}}=1$	$\lambda\geq\max\left\{\frac{\bar{k}}{1+\alpha},\frac{\beta}{\bar{m}-\alpha+\beta},\right\}$	$\underline{k}<1+\alpha$ $\lambda\geq\frac{\underline{k}}{p_{\mathrm{CD}\bar{\mathrm{m}}}\left(1+\alpha\right)}\Rightarrow p_{\mathrm{CD}\bar{\mathrm{m}}}\geq\frac{\underline{k}}{\left(1+\alpha\right)}$	$\Pi_{\bar{m}}\left(\mathrm{CC}\bar{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\bar{\mathrm{m}}\right)=\underline{k}-\bar{k}-\left(\lambda\left(\alpha-\bar{m}\right)+\left(1-\lambda\right)\beta\right)$ $\left(\Pi_{\bar{m}}\left(\mathrm{CC}\bar{\mathrm{m}}\right)-\Pi_{\underline{m}}\left(\bar{\mathrm{m}}\right)\right)^{f}=\underline{k}-\bar{k}-\left(\lambda\alpha+\left(1-\lambda\right)\beta\right)$

Table 3: p-stable equilibria (p-stable semi-pooling equilibria are referred to Appendix C)

In the low pooling cooperative equilibrium, nobody signals and high types cooperate. This equilibrium exists if there are sufficiently many high types. Only then can they compensate for the loss from being cooperative against low types by the cooperative outcome among each other. In other words, if the share of high types falls below a certain threshold, then they will start to prefer defecting when receiving the low signal. Note that this equilibrium is indeed an equilibrium set, since the strategies $\mathrm{CC}\underline{\mathrm{m}}$ and $\mathrm{DC}\underline{\mathrm{m}}$ are equivalent in equilibrium. The share of high types required for this to be an equilibrium increases in the sucker's payoff, since cooperative behavior becomes more disadvantageous with increasing (absolute) sucker's payoffs. This threshold, too, has an intuitive meaning. Note that $\bar{\mathrm{m}}-\alpha$ (β) measures the incentive to reciprocate cooperative (defective) behavior. In essence, the condition $\frac{\beta}{\beta+\bar{\mathrm{m}}-\alpha}<\lambda$, which can be rewritten as $\lambda\left(\bar{\mathrm{m}}-\alpha\right)>\left(1-\lambda\right)\beta$, states that the expected gain from reciprocating cooperative behavior must exceed the expected gain from reciprocating defective behavior.

In the low pooling defective equilibrium, nobody sends the cooperative signal and everybody defects earning a payoff of zero. Again, due to lack of distinguishability in equilibrium, equilibrium is indeed a set where $\mathrm{CD}\underline{\mathrm{m}}$ and $\mathrm{DD}\underline{\mathrm{m}}$ might be played by high types. This set of equilibrium reflects the benchmark solution in the underlying game and exists for all population compositions between high types and low types.

In the high pooling cooperative equilibrium, everybody signals and high types cooperate. This equilibrium exists if there are sufficiently many high types. If the latter's proportion is large enough, they can compensate for the loss from being cooperative against low types by the cooperative outcome among each other. In other words, if the share of high types falls beneath a certain threshold, they will then start to prefer to play defective while receiving the low signal. Contrary to the low pooling equilibrium, an additional restriction with respect to the share of high types will arise, reflecting the incentive compatibility for low types to signal. Note that this equilibrium is again an equilibrium set, since the strategies $\mathrm{CC}\bar{\mathrm{m}}$ and $\mathrm{CD}\bar{\mathrm{m}}$ are equivalent in equilibrium. The share of high types required for this to be an equilibrium weakly increases in the sucker's payoff and the signaling cost for low types. Since with increasing (absolute) sucker's payoffs, cooperative behavior and sending the signal for low types respectively become more disadvantageous. Here, for low types, the reverse logic applies in comparison to the separating cooperative equilibrium, i.e. for low types to find it worthwhile to signal, their cost-benefit ratio ($\frac{\underline{\mathrm{k}}}{1+\alpha}$) must be smaller than the likelihood to profit from signaling (λ). The lower bound stemming from incentive constraint for high types bears the same logic as in the low pooling cooperative equilibrium.

4. Endogenous Proportion of Norm Bearers

We now analyze the dynamics of the share of high types in the population for which we assume that the p-dynamic has reached a stable p-equilibrium, as we assumed that inner motives evolve far more slowly than behavioral frequencies. The evolution of the proportion of norm bearers is determined by its relative fitness. Fitness is measured by the material payoffs as presented in

Table 1. Thus, any preference parameter measuring the evaluation of material payoffs will be neglected when calculating fitness payoffs. Analogous to the derivation of p-equilibria, the differentials in these fitness payoffs among high and low types are the driving force for the evolution of their respective shares. To ease the understanding of the differentials of fitness payoff differentials, we provide some intuition for their size in the relevant p-stable equilibria.

In the cooperative separating equilibrium, both types defect in all interactions, except when two individuals of the high type meet. In this case, they cooperate. The low type will thus always earn a fitness payoff of zero, and the high type will earn a fitness payoff of one with probability λ, i.e. the probability that he interacts with another individual of the high type. Since high types unconditionally bear the signaling cost $\overline{k}$, their expected payoff in the cooperative separating equilibrium is $\lambda-\overline{k}$, which is also the expected difference of fitness payoffs: $\left(\Pi_{\overline{m}}\left(CD\overline{m}\right)-\Pi_{\underline{m}}\left(\underline{m}\right)\right)^{f}=\lambda-\overline{k}$.

Obviously, this fitness advantage of the high type grows in the share of high types in the population.

In the two (partially) cooperative pooling equilibria, individuals of the high type cooperate in reaction to the signal they send, and all individuals of the low type copy this signal but still defect.[10] Leaving aside signaling costs for a moment, differences in material payoffs then reflect payoffs of unconditional cooperators and defectors in the underlying PD. More precisely, with probability λ, high types meet their own type and realize the cooperative outcome, earning 1. With the residual probability, they meet a low type and lose β. Low types always defect and only earn positive payoffs when matched with high types, which happens with probability λ and earns them $1+\alpha$. A fitness differential to the advantage of the high types thus cannot result from playing the game itself, but only form sufficiently large differences in signaling cost (see Table 3). Obviously, if no signal is sent, as is the case in the low pooling cooperative equilibrium, the fitness payoff of the high type can only be smaller than that of the low type,

$$\left(\Pi_{\overline{m}}\left(CC\underline{m}\right)-\Pi_{\underline{m}}\left(\underline{m}\right)\right)^{f}=-\left(\lambda\alpha+\left(1-\lambda\right)\beta\right)<0 .$$

Only in the high pooling cooperative equilibrium, the signaling cost disadvantage of the low type may outweigh the disadvantage of the high type from playing cooperatively in the game, so that the high type earns a higher fitness payoff than the low type, $\left(\Pi_{\overline{m}}\left(CC\overline{m}\right)-\Pi_{\underline{m}}\left(\overline{m}\right)\right)^{f}=\underline{k}-\overline{k}-\left(\lambda\alpha+\left(1-\lambda\right)\beta\right)$.

Obviously, the fitness payoff difference increases (declines) in the share of the high types if defection is more (less) tempting against defection than against cooperation, i.e. if β is larger (smaller) than α. If the proportion of the high type in the population is too small, it is either not worthwhile to mimic the other type, or the chances to meet another high-type individual are so low that cooperation ceases to be the best reaction to the signal sent by all individuals. For these small shares of the high type in the population, the pooling cooperative equilibria break down

[10] This implies that the other signal is never sent, which explains why the high type is indifferent between the two behavioural actions C and D to this never-observed signal.

just like the cooperative separating equilibrium discussed earlier breaks down for shares of the high type that are too large.

In the pooling defective equilibrium, both types always defect without sending signals and thus all earn the same fitness (and behavioral) payoff of zero.

The following two figures depict the differences in material payoffs for the various p-stable equilibria (see Table 3).

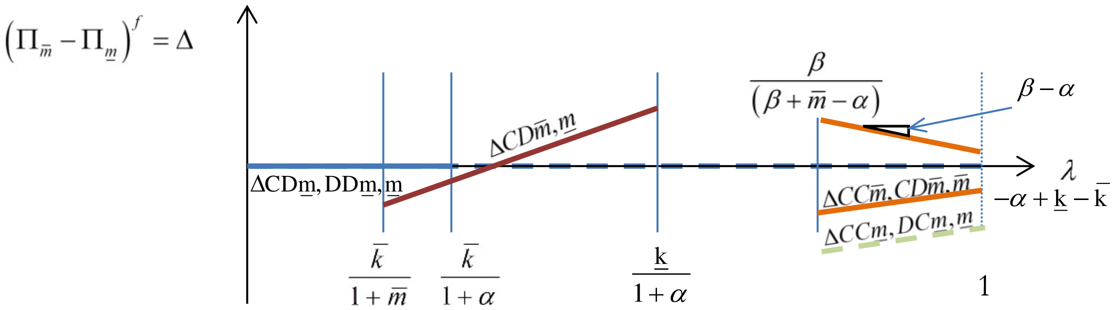


Figure 1: Differences in material payoffs for $\frac{\underline{k}}{1+\alpha}<\frac{\beta}{(\beta+\overline{m}-\alpha)}$

Payoff differences for semi-pooling equilibria are neglected since their support lies in the interval $\left(\frac{\beta}{(\beta+\overline{m}-\alpha)},1\right)$ and the difference is strictly negative for all. Hence, their presence will have no important implications for the dynamics of the share of high types.

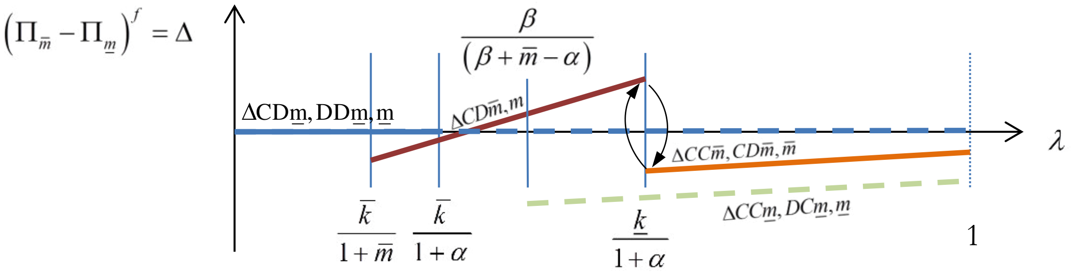


Figure 2: Differences in material payoffs for $\frac{\underline{k}}{1+\alpha}\geq\frac{\beta}{(\beta+\overline{m}-\alpha)}$

A stable λ-equilibrium may be realized around one p-stable equilibrium or by the interplay of several such equilibria. We first concentrate on the first case, which we further differentiate into corner equilibria (Lemma 1) and inner equilibria (Lemma 2) and then turn to the second case (Lemma 3).

In the first case, the difference in fitness payoffs between high and low types must vanish to constitute a stationary point at this particular value of the share of high types λ. For stability, in the neighborhood of an equilibrium λ^*, high types must earn strictly more than low types for $\lambda<\lambda^*$ and strictly less for $\lambda>\lambda^*$. In terms of Figure 1 and Figure 2, the stationary point is a zero of the linear payoff difference for a certain p-stable equilibrium, and stability is equivalent to a negative slope of the payoff difference function. Of course, the requirement with respect to the zero and the slope is only relevant for inner equilibria. At the upper bound of the domain, $\lambda=1$, a strictly positive payoff difference in favor of high types at $\lambda < 1$ is necessary and sufficient for having a corner equilibrium. At the lower bound of the domain, $\lambda = 0$, a strictly negative payoff difference at $\lambda > 0$ is necessary and sufficient for having a corner equilibrium.

We first analyze whether λ-stable equilibria with full cooperation exist. Since only high types may cooperate, this is equivalent to asking whether there is a λ-stable equilibrium at $\lambda = 1$ with cooperating high types. Since high types in the low pooling cooperative equilibrium face an evolutionary disadvantage for all population compositions, this p-stable equilibrium cannot induce a stable cooperative λ-equilibrium (partial or full). Hence, there are two potential candidates left, the separating cooperative equilibrium and the high pooling equilibrium. The following lemma states the conditions such that a locally stable equilibrium with only high types present in the population who cooperate with each other exists.

Lemma 1 The PD can be fully resolved as a locally λ-stable equilibrium only in two ways:

(1) by the separating cooperative equilibrium if and only if $\underline{k} \geq 1+\alpha$ and $\bar{k} < 1$

(2) by the high pooling cooperative equilibrium if and only if $\underline{k} < 1+\alpha$ and either $\underline{k} - \bar{k} > \alpha$ or $\underline{k} - \bar{k} = \alpha > \beta$.

All proofs are in Appendix D.

The existence of fully cooperative equilibria seems surprising at first glance, but a closer look at the stated conditions for their existence reveals how rarely they occur. In the case of the separating cooperative equilibrium, the condition corresponds to a scenario where the signaling cost for low types are so severe that it will never pay for them to signal. More precisely, in a cooperative separating equilibrium with $\lambda = 1$, a single low-type mutant would earn $1+\alpha$ from playing the dominant defective strategy at cost $\underline{k}$. The second qualification $\bar{k} < 1$ stems from the incentive compatibility constraint for high types, since they could always earn zero by not-signaling and exhibiting defective behavior. In the case of the high pooling cooperative equilibrium, the difference in the signaling cost must exceed the material reward of defecting on a cooperative opponent.

The restrictiveness of Lemma 1 draws our attention to inner stable equilibria. The only candidate for such a λ-equilibrium supported by only one p-stable equilibrium is one associated with the high pooling cooperative equilibrium at $1 - \frac{\underline{k} - \bar{k} - \alpha}{\beta - \alpha}$. All other equilibria are characterized by either strictly negative or strictly increasing payoff differentials. The high pooling cooperative equilibrium exists and is λ-stable if $1 - \frac{\underline{k} - \bar{k} - \alpha}{\beta - \alpha}$ is inside the λ-support of this equilibrium and the fitness differential decreases in λ, which is the case if $\beta - \alpha < 0$ (see Figure 1). Taking these conditions together yields:

Lemma 2 The high pooling cooperative equilibrium constitutes an inner λ-stable equilibrium at $1 - \frac{\underline{k} - \bar{k} - \alpha}{\beta - \alpha}$ if and only if: $\frac{\beta}{\bar{m} - \alpha + \beta}\bar{m} < \underline{k} - \bar{k} < \alpha$ and $\beta < \frac{1+\beta}{1+\alpha}\underline{k} - \bar{k}$.

Note that the first condition implies $\beta-\alpha<0$, which guarantees stability. As expected, the conditions presented in Lemma 2 are less restrictive as compared to the requirements for an equilibrium formed only by high types. Looking at the conditions, we observe that the existence of inner stable equilibria requires that the costs of signaling for norm adopters must differ sufficiently from the corresponding costs of non-adopters.

What remains to be studied is whether separating λ-equilibrium constituted by the interplay of several p-equilibria exists. For this to be the case requires: (1) the supports of the p-equilibria need to be adjacent, (2) around the point where the supports are adjacent, the differences of fitness payoffs of the relevant p-equilibria must be positive for less-than-equilibrium shares of high types and negative for more-than-equilibrium shares of high types, and (3) after λ moves from the support of one p-equilibrium to the support of another, the behavioral frequencies have to be within the basin of attraction of the "new" equilibrium if they have been sufficiently close to the "old" equilibrium. In our case, we may have such an equilibrium only at $\lambda=\frac{\underline{k}}{1+\alpha}$ where three equilibria interplay: the separating cooperative equilibrium, a semi-pooling cooperative equilibrium (last row in Appendix C), and the high pooling cooperative equilibrium. To facilitate understanding of this argument, we recommend that the reader views Figure 2 while reading the following argument.

Condition (1) requires that $\frac{\underline{k}}{1+\alpha}\geq\frac{\beta}{\bar{m}-\alpha+\beta}$ (cf. Table 3 and Appendix C). Condition (2) has implications for the fitness differences of the p-stable equilibria. For the cooperative separating p-equilibrium, the fitness difference is given by $\left(\Pi_{\bar{m}}\left(CD\bar{m}\right)-\Pi_{\underline{m}}\left(\underline{m}\right)\right)^{f}=\lambda-\bar{k}$ for $\lambda\leq\frac{\underline{k}}{1+\alpha}$. This difference must be strictly positive at $\lambda=\frac{\underline{k}}{1+\alpha}$, whence $1+\alpha<\frac{\underline{k}}{\bar{k}}$. In other words, the relative disadvantage for low types in terms of signal costs must exceed the relative incentive to defect given the opponent cooperates. Given this inequality and a share of high types sufficiently close to, but lower than $\lambda=\frac{\underline{k}}{1+\alpha}$, the share of the high type increases when the p-dynamics has reached the cooperative separating equilibrium. For the high pooling cooperative equilibrium, the fitness difference is given by $\left(\Pi_{\bar{m}}\left(CC\bar{m}\right)-\Pi_{\underline{m}}\left(\bar{m}\right)\right)^{f}=\underline{k}-\bar{k}-\left(\lambda\alpha+\left(1-\lambda\right)\beta\right)$, which has to be negative. Hence, we get $\frac{\underline{k}}{1+\alpha}<\frac{\bar{k}+\beta}{1+\beta}$.

To see that Condition (3) is satisfied under certain conditions we present our argument in three steps. First, we draw the reader's attention to the fact that for all three of the considered equilibria, we have $p_{CD\bar{m}}+p_{CC\bar{m}}=1$. This implies that for $\lambda=\frac{\underline{k}}{1+\alpha}$ we have:

$$\begin{aligned}
&\Pi_{\bar{m}}\left(CD\bar{m}\right)=\lambda\left(1+\bar{m}\right)-\left(1-\lambda\right)p_{\bar{m}}\beta-\bar{k}\\
&\geq\Pi_{\bar{m}}\left(CC\bar{m}\right)=\lambda\left(1+\bar{m}\right)-\left(1-\lambda\right)\beta-\bar{k}\\
&>\max_{X}\left(\Pi_{\bar{m}}\left(X\right)\right)\quad\text{where}\quad X\in\left\{C,D\right\}^{2}\times\left\{\underline{m},\bar{m}\right\}\backslash\{CD\bar{m},CC\bar{m}\}
\end{aligned}$$

where the first inequality is strict if $\mathrm{p}_{\bar{\mathrm{m}}} < 1$ and the second inequality requires $\lambda^* \equiv \frac{\underline{\mathrm{k}}}{1+\alpha} > \frac{\beta}{\bar{\mathrm{m}} - \alpha + \beta} \equiv \tilde{\lambda}$. Hence, continuity of the payoffs and Lipschitz-continuity of the dynamics implies that for all λ sufficiently close to λ^* and all sufficiently large $\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}} + \mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}} = 1$ we have $\dot{\mathrm{p}}_{\mathrm{CD}\bar{\mathrm{m}}} + \dot{\mathrm{p}}_{\mathrm{CC}\bar{\mathrm{m}}} = 1$. Hence, once the system is close enough to any of the three relevant p-stable equilibria, and in particular once $\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}} + \mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}$ has become large enough, $\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}} + \mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}$ will continue to grow for all $\mathrm{p}_{\bar{\mathrm{m}}}$. Second, we observe that if $\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}}$ is large enough and the p-dynamics is sufficiently fast compared to the λ-dynamics, then λ will always stay close enough to λ^* to uphold the validity of the first argument. Third, if $\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}} + \mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}$ is large enough and thus increases, $\Pi_{\bar{\mathrm{m}}}(\mathrm{CD}\bar{\mathrm{m}}) < \Pi_{\bar{\mathrm{m}}}(\mathrm{CC}\bar{\mathrm{m}})$ only occurs for ever decreasing ranges of large $\mathrm{p}_{\bar{\mathrm{m}}}$. Hence, for every payoff-monotone dynamic $\mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}$ will be smaller after every full cycle and will never again reach its previous maximum level. Hence, $\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}}$ will eventually be large enough to ensure the validity of our second argument.

Hence, once our full dynamic system is close enough to λ^* and the λ-dynamic is slow enough, the system will rotate between the separating equilibrium and the high pooling equilibrium in ever smaller cycles (note that this does not necessarily imply that a fixed point is reached because a limit cycle may exist). We summarize all conditions in the following:

Lemma 3 If $\frac{\beta}{\bar{\mathrm{m}} - \alpha + \beta} < \frac{\underline{\mathrm{k}}}{1+\alpha} \underset{\substack{\text{if = then} \\ \alpha>\beta}}{\leq} \frac{\bar{\mathrm{k}} + \beta}{1+\beta}$ and $\bar{\mathrm{k}} < \frac{\underline{\mathrm{k}}}{1+\alpha}$, an inner λ-stable equilibrium exists at $\lambda = \frac{\underline{\mathrm{k}}}{1+\alpha}$, in which (1) high-type individuals cooperate among each other but also with those low-type individuals who signal to be of the high type and (2) the proportion of low-type individuals who signal to be of the high type fluctuates.

Note that the conditions in Lemma 2 and Lemma 3 are mutually exclusive, i.e. there is at most one stable inner equilibrium.

We have so far not considered the case of $\frac{\underline{\mathrm{k}}}{1+\alpha} \leq \frac{\beta}{\bar{\mathrm{m}} - \alpha + \beta}$. If there is equality ($\frac{\underline{\mathrm{k}}}{1+\alpha} = \frac{\beta}{\bar{\mathrm{m}} - \alpha + \beta}$), an equilibrium of the type discussed in Lemma 3 still exists at $\lambda = \frac{\underline{\mathrm{k}}}{1+\alpha}$, but it is unstable (the argument on condition 3 fails). If the inequality is strict ($\frac{\underline{\mathrm{k}}}{1+\alpha} < \frac{\beta}{\bar{\mathrm{m}} - \alpha + \beta}$), there is a gap between the λ-supports of the separating cooperative equilibrium and the high pooling cooperative equilibrium (see Figure 1). In the interval $\left(\frac{\underline{\mathrm{k}}}{1+\alpha}, \frac{\beta}{\bar{\mathrm{m}} - \alpha + \beta}\right)$, the defective pooling equilibrium is the unique equilibrium. Should the population start at the cooperative separating p-equilibrium with a positive fitness differential, then it will eventually drive the share of high-type individuals beyond the λ-support of this equilibrium so that $\mathrm{p}_{\bar{\mathrm{m}}}$ starts to grow. Once it grows too much, the strategy $\mathrm{DD}\bar{\mathrm{m}}$ yields the largest behavioral payoff to high-type individuals while $\mathrm{CD}\bar{\mathrm{m}}$ yields only the second largest. Hence, the share of always defecting high-type individuals $\mathrm{p}_{\mathrm{DD}\bar{\mathrm{m}}}$ must grow and $\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}}$ must decline because the shares of the other strategies (with even lower behavioral payoffs) are already zero. Less cooperation by high-type individuals reduces the advantages that low-type individuals accrue from falsely signaling to be of the high type. Hence, $\mathrm{p}_{\bar{\mathrm{m}}}$ will eventually decline again. A behavioral equilibrium exists in which only some low-type individuals signal the wrong type and only some high-type individuals cooperate after receiving the high signal while the others always defect, but this equilibrium is not stable (see Appendix B). Consequently, $\mathrm{p}_{\mathrm{DD}\bar{\mathrm{m}}}$ will eventually grow large enough to move the population in the attraction region of the defective separating equilibrium, where it will remain. We admit that the evolution may become more complex when $\mathrm{p}_{\bar{\mathrm{m}}}$ and $\mathrm{p}_{\mathrm{DD}\bar{\mathrm{m}}}$ both become so large that $\mathbf{CD\bar{m}}$ becomes less profitable than $\mathbf{DC\bar{m}}$. There may then be payoff monotonic dynamics for which $\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}$ starts to grow, although slower than $\mathrm{p}_{\mathrm{DD}\bar{\mathrm{m}}}$. If this happens, false signaling by high types may eventually become reasonable. However, as the low pooling equilibrium with cooperation only of the high types fails to exist in the interval $\left(\frac{\underline{\mathrm{k}}}{1+\alpha},\frac{\beta}{\bar{\mathrm{m}}-\alpha+\beta}\right)$, we conjecture that the population will eventually end up in the defective pooling equilibrium as the unique behavioral equilibrium.

Conjecture If $\frac{\underline{\mathrm{k}}}{1+\alpha}<\frac{\beta}{\bar{\mathrm{m}}-\alpha+\beta}$, no λ-stable inner equilibrium exists at $\lambda=\frac{\underline{\mathrm{k}}}{1+\alpha}$.

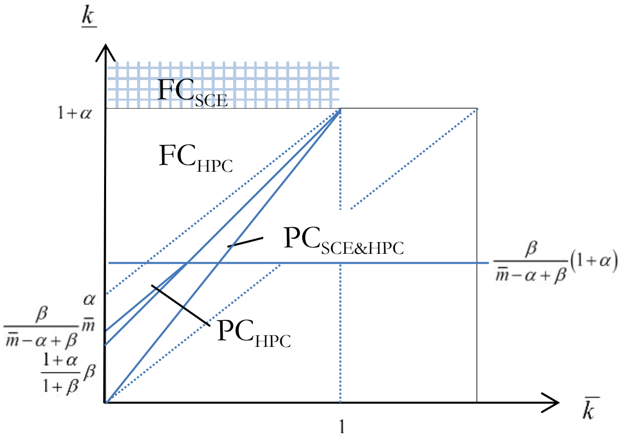


Figure 3: Parameter region for partial or full cooperation

Figure 3 illustrate the conditions of Lemmas 1 through 3 graphically. For illustrative purposes, we assume $\beta-\alpha<0$ and $\frac{\underline{\mathrm{k}}}{1+\alpha}\geq\frac{\beta}{\bar{\mathrm{m}}-\alpha+\beta}$ so that all inner equilibria may exist for some parameter ranges. In the figure, areas marked by FC and PC represent parameter combinations for which full and partial cooperation occur, respectively. More specifically, the indexes mark parameter ranges for which cooperation is induced by the separating cooperative equilibrium (SCE), the high pooling cooperative equilibrium (HPCE), or the interplay of the two and a semi-pooling equilibrium (SCE&HPCE).

It is worth noting that the strength of the cooperative norm measured by $\bar{\mathrm{m}}$ has a direct impact on the parameter set allowing for λ-stable inner equilibria (see Figure 3). As $\bar{\mathrm{m}}$ gets closer to the incentive to defect α, the parameter region supporting a separating cooperative equilibrium ($PC_{SCE\&HPCE}$) becomes smaller and smaller. Although the exact size of $\bar{m}$ is not important for the behavioral consequence for each individual as long as $\bar{m} > \alpha$ holds true, the exact size of $\bar{m}$ does matter for the size of the parameter range for which evolutionary stable equilibria characterized by partial cooperation exist.

5. Collecting requirements for equilibria with cooperation

By combining Lemmas 1 through 3 from the previous section, we deduce a theorem on cooperation in an unstructured population:

Theorem In an unstructured society, cooperation in a PD may exist and be stable due to the possibility of signaling the existence of inner payoffs for (mutual) cooperation, which do not affect fitness, if the costs of falsely signaling to have such inner payoffs are sufficiently large. These costs must be larger to reach full cooperation than to reach partial cooperation.

In our model, 'sufficiently large' translates to $\underline{k} - \bar{k} > \alpha$ or $\underline{k} - \bar{k} = \alpha > \beta$ for full cooperation (Lemma 1). For partial or full cooperation (Lemmas 1 through 3), 'sufficiently large' translates to:

$$\underline{k} \underset{\substack{\geq \\ \text{for last} \\ \text{term}}}{>} (1+\alpha)\max\left\{\min\left\{\frac{\bar{k}+\alpha}{1+\alpha},\frac{\beta}{\bar{m}-\alpha+\beta}\right\},\bar{k}\right\} \text{ if } \alpha \leq \beta \text{ and to}$$

$$\underline{k} > (1+\alpha)\max\left\{\min\left\{\frac{\beta\bar{m}}{\bar{m}-\alpha+\beta}+\bar{k},\frac{\beta}{\bar{m}-\alpha+\beta}\right\},\bar{k}\right\} \text{ if } \alpha > \beta .$$

Figure 4 and Figure 5 illustrate the interrelation between the costs for low types to signal falsely and the extent of the inner motive for mutual cooperation. This relation is determined by the various inequality conditions for existence of partial or full cooperation stated in the theorem above. Figure 4 reveals the negative relation between these two parameters, i.e. in order to sustain some level of cooperation, lower signalling costs for low-types must be compensated by a higher inner motive for mutual cooperation of the high-types. Here, the aforementioned interdependence of $\bar{m}$ and the presence of cooperative equilibria is directly observable. Although the precise level of $\bar{m}$ is not decisive with respect to its behavioural consequence, its level plays a crucial role with respect to the size of the set of parameters such that partial or full cooperation could be sustained as an equilibrium outcome. Furthermore, we observe that this set of parameters is strictly decreasing in the signalling cost for the high type. Finally, Figures 4 and Figure 5 show that the chances for cooperation diminish with increasing β. In essence, the riskier or more painful cooperation occurs when matched with defective behaviour, the higher requirements have to be met with respect to signalling costs for low types and the inner motive for mutual cooperation. A mirror argument applies with respect to parameter α, measuring the incentive to defect on cooperation in the underlying game. The following corollary summarizes these insights.

Corollary (1) The range of signalling cost for the low type allowing for partial or full cooperation is weakly increasing in the social norm for mutual cooperation $\bar{m}$.

(2) The set of $(\underline{k},\bar{m})$-pairs allowing for partial or full cooperation is strictly increasing in signalling cost for the high type $\bar{k}$ and strictly decreasing in the Sucker's payoff β and the incentive to defect on cooperation α.

The theorem reveals that in case of full cooperation, almost always it is only the incentive to defect on a cooperative player α relative to the difference in signalling costs that matters. Whereas for stable partial cooperation, the relation of α and β is relevant. The loss from playing cooperatively on a defective opponent β must be less than what a player could gain from defecting on a cooperative player. Intuitively, this explains the edge of defective players over cooperative players for shares of the latter that exceed the equilibrium level and vice versa. Reflecting on both incentives in case of a partially cooperative equilibrium is also plausible since both behaviors are present in equilibrium, whereas fully cooperative equilibria are characterized by solely cooperative actions. In that case, only the price for cooperation given the monomorphic cooperative behavior α is relevant.

Interdependence between the size of the inner motive and the cost to send a false signal

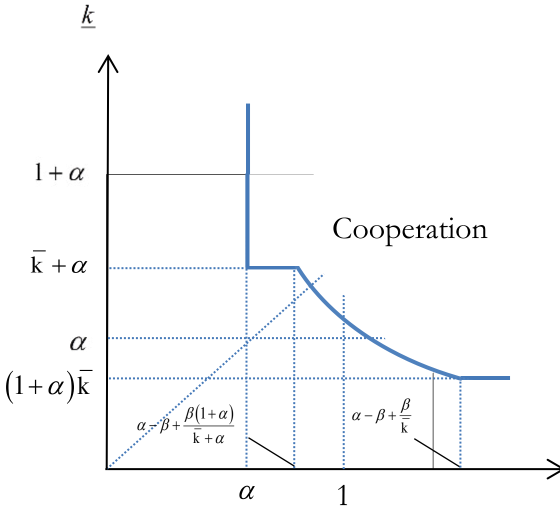
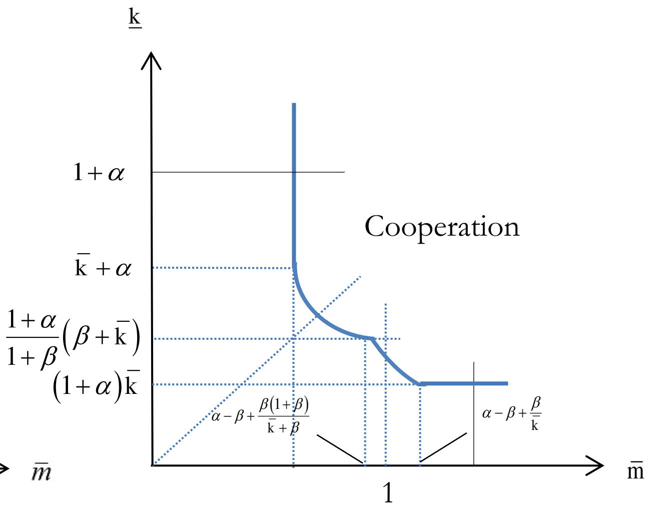


Figure 4: $\alpha \leq \beta$



$\underline{k}$
$1+\alpha$
Cooperation
$\bar{k}+\alpha$
$\frac{1+\alpha}{1+\beta}(\beta+\bar{k})$
$(1+\alpha)\bar{k}$
$\alpha-\beta+\frac{\beta(1+\beta)}{\bar{k}+\beta}$
$\alpha-\beta+\frac{\beta}{\bar{k}}$
1
$\bar{m}$

Figure 5: $\alpha > \beta$

6. Conclusion

In this paper, we analyze an evolutionary model where individuals are able to signal that they internalized a particular social norm, namely a norm for mutual cooperation. This preference was embedded in a Prisoners' Dilemma. In section 5, we present a theorem that states necessary and sufficient conditions for full or partial cooperation to be prevalent in a stable equilibrium. These conditions refer to the difference in signaling cost between the cooperative and the opportunistic type, the extent of the cooperative norm and the model parameters of the PD, i.e. the temptation to defect and the sucker's payoff. We obtain several interesting results. First, it is true that the exact size of the behavioral parameter measuring the internal bias in favor of mutual cooperation is not important for the behavioral consequence for each individual. However, when it comes to the presence of stable equilibria characterized by partial cooperation its size and its relation to the incentive to defect do become relevant. More precisely, the stronger the inner motive to

cooperate is, the less restrictive are the conditions on the spread in signaling cost. Second, for cooperative agents to coexist with defecting agents in a stable equilibrium, it is not necessary that the signaling technology fully cancels the incentive to defect. Since this would be necessary for many corresponding results that are based on some sort of involuntary redistribution (e.g. punishment), our approach may explain cooperation in more cases than the latter approaches. Furthermore, the range of signalling cost for the low-type individuals allowing for partial or full cooperation is weakly increasing in the strength of the social norm for mutual cooperation. Finally, the set of pairs of signalling cost for the defective type and level of cooperative norm allowing for partial or full cooperation is strictly increasing in signalling cost for the high type and strictly decreasing in the sucker's payoff and the incentive to defect on cooperation.
We achieved these results by analyzing the evolution of norms concerning cooperation in the PD with one of the most general class of dynamics considered in evolutionary game theory, namely the class of payoff-monotone dynamics. Existing literature has already demonstrated that signaling may point a way out of a social dilemma where mechanisms as reputation, reciprocity or assortative matching are absent or fail to work sufficiently well. Yet only a few approaches incorporate a formal model. The novelty of our approach is the derivation of the full set of behavioral equilibria, i.e. all separating, pooling and semi-pooling equilibria of the signaling-extended PD. This would be only a technical note if it did not induce a richer set of equilibria concerning the distribution of an internalized norm that can stabilize cooperation. In particular, it is worthwhile to observe the existence of an inner equilibrium, i.e. an equilibrium where norm bearers and non-bearers coexist, that is stabilized by the interplay of a separating, a semi-pooling and a pooling equilibrium of the evolutionary signaling game. It is exactly this interplay that stabilizes the share of norm bearers and dissolves the necessity to introduce evolutionary forces into the dynamics of norm adoption beyond payoff monotonicity that are frequency based[11].

Since cooperative equilibria exist when agents may signal their cooperative attitude, large societies aiming for more cooperation are not completely limited to the reduction of anonymity in social interaction (and hence, giving up some of the advantages of large societies) or the use of formal institutions. Politicians may also try to provide hard-to-falsify signals of internal motives to cooperate in areas where interaction is rather anonymous. Then, informal institutions may spontaneously and easily evolve even in large unstructured interaction environments. Even if politics cannot alter the underlying incentives of the social dilemma to the extent that the dilemma aspect would indeed vanish, partial reduction of the incentive to defect or partial insurance for the suckers' payoff may be sufficient to allow for cooperation to evolve. The share of norm bearers in our model is driven by evolutionary forces that are beyond the scope of any policy measure. However, politics might have some leverage on how strong the internal sanctions are that support the norm once it is internalized. Hence, strengthening the internalized norms will also increase the chance for cooperation.

If we argue that it is foremost the spontaneous institutions that repel defection in large unstructured societies, then these insights lead us to argue that concepts of institutions should

[11] Gintis et al. (2001) show in one of the few formal evolutionary signaling models that a stable separating equilibrium may exist. However, under general payoff monotonicity, this equilibrium would cease to exist since their type that corresponds to our high-types face an evolutionary advantage. As a consequence, their share of the population would increase and eventually exceed the threshold beyond which the separating equilibrium breaks down.

not require that all individuals adhere to the behavior prescribed by the spontaneous institution. Instead, a definition of institutions should allow for a substantial share of the population to deviate from its rule. We add a theoretical basis to this insight, which seems obvious from an empirical point of view.

We have not modeled the interplay of different PD situations in a society. Without going into any detail here, we conjecture from our signaling model that cooperation in one PD may serve as a signal to have the internal cooperation in order to fare better in another PD. The temptation to defect in the first game would be the cost to falsely signal having the internal motivation to cooperate. Hence, the interplay between different PD situations does not allow for scaling up: temptation in the first game cannot be larger than in the second game, or cooperation there cannot be complete. Further research is needed on the details of the interplay between different PD games in an unstructured society.

The analysis for a more general norm than the one we considered is left open to future research. We believe that the size of the parameter measuring the strength of the internalized norm is not driven by evolutionary forces, since no fitness payoff differences depend on it. However, the size of the parameter does determine the range in which cooperative equilibria exist. Hence, if two separate populations with different levels of the internalized norms are considered, the one with the higher value is more likely to evolve towards a cooperative state. If in the course of time, both populations start interacting with each other, a cooperative population might induce cooperation in a defective population and vice versa. To analyze such an environment may be relevant for studying migrational effects on cooperation.

References

Axelrod, R. 1984. The Evolution of Cooperation. New York, NY.: Basic Book. Inc.

Axelrod, R., W. D. Hamilton. 1981. The evolution of cooperation. *Science* **211**(4489) 1390–1396.

Bendor, J., P. Swistak. 1998. Evolutionary equilibria: Characterization theorems and their implications. *Theory and decision* **45**(2) 99–159.

Bester, H., W. Güth. 1998. Is altruism evolutionarily stable? *Journal of Economic Behavior & Organization* **34**(2) 193–209.

Cooper, R., D. V. DeJong, R. Forsythe, T. W. Ross. 1996. Cooperation without reputation: experimental evidence from prisoner's dilemma games. *Games and Economic Behavior* **12**(2) 187–218.

Ellison, G. 1994. Cooperation in the prisoner's dilemma with anonymous random matching. *The Review of Economic Studies* **61**(3) 567–588.

Fudenberg, D., D. G. Rand, A. Dreber. 2012. Slow to anger and fast to forgive: cooperation in an uncertain world. *The American Economic Review* **102**(2) 720–749.

Fudenberg, D., E. Maskin. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54** 533–554.

Gintis, H., E. A. Smith, S. Bowles. 2001. Costly signaling and cooperation. *Journal of Theoretical Biology* **213**(1) 103–119.

Güth, W. 1995. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory* **24**(4) 323–344.

Güth, W., H. Kliemt. 1994. Competition or Co-operation: on the evolutionary economics of trust, exploration and moral attitutes. *Metroeconomics* **45**(2) 155–187.

Güth, W., A. Ockenfels. 2005. The coevolution of morality and legal institutions: an indirect evolutionary approach. *Journal of Institutional Economics* **1**(2) 155–174.

Güth, W., M. Yaari. 1992. An evolutionary approach to explain reciprocal behavior in a simple strategic game. *U. Witt. Explaining Process and Change–Approaches to Evolutionary Economics. Ann Arbor* 23–34.

Güth, W., H. Kliemt, B. Peleg. 2000. Co-evolution of Preferences and Information in Simple Games of Trust. *German Economic Review* **1**(1) 83–110.

Haken, H. 1977. Synergetics. An Introduction. Nonequilibrium Phase Trasitions and Self-organization in Physics, Chemistry, and Biology. Berlin: Springer.

Hamilton, W. D. 1964a. The genetical evolution of social behavior. I. *Journal of Theoretical Biology* **7**(1) 1–16.

Hamilton, W. D. 1964b. The genetical evolution of social behaviour. II. *Journal of Theoretical Biology* **7**(1) 17–52.

Harrington, J. E. 1987. Finite rationalizability and cooperation in the finitely repeated Prisoners' Dilemma. *Economics Letters* **23**(3) 233–237.

Harsanyi, J. C. 1967. Games with Incomplete Information Played by" Bayesian" Players, I-III. Part I. The Basic Model. *Management Science* **14**(3) 159–182.

Harsanyi, J. C. 1968a. Games with incomplete information played by 'Bayesian'players, part III. The basic probability distribution of the game. *Management Science* **14**(7) 486–502.

Harsanyi, J. C. 1968b. Games with Incomplete Information Played by "Bayesian" Players Part II. Bayesian Equilibrium Points. *Management Science* **14**(5) 320–334.

Hirshleifer, D., E. Rasmusen. 1989. Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior & Organization* **12**(1) 87–106.

Huck, S., J. Oechssler. 1999. The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior* **28**(1) 13–24.

Janssen, M. C. W., J. Gorter, van de Meerendonk, Sjoerd. 1997. Cooperation in a modified version of the finitely repeated prisoners' dilemma game. *Journal of Economic Behavior & Organization* **32**(4) 613–619.

Kandori, M. 1992. Social norms and community enforcement. *The Review of Economic Studies* **59**(1) 63–80.

Kurz, M. 1977. Altruistic equilibrium. Bela Balassa and Richard Nelson, Economic Progress, Private Values and Public Policy 177–200.

Kreps, D. M., P. Milgrom, J. Roberts, R. Wilson. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of economic theory* **27**(2) 245–252.

Leimar, O., P. Hammerstein. 2001. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**(1468) 745–753.

Maynard Smith, J., G. R. Price. 1973. The Logic of Animal Conflict. *Nature* **246** 15

Mordecaï, K. 1977. Altruistic equilibrium. Bela Balassa and Richard Nelson, Economic Progress, Private Values and Public Policy 177–200.

Neyman, A. 1985. Bounded complexity justifies cooperation in the finitely repeated prisoners' dilemma. *Economics Letters* **19**(3) 227–229.

Nowak, M. A., K. Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* **393**(6685) 573–577.

Possajennikov, A. 2000. On the evolutionary stability of altruistic and spiteful preferences. *Journal of Economic Behavior & Organization* **42**(1) 125–129.

Roth, A. Bargaining experiments, in: Kagel J., Roth A., The Handbook of Experimental Economics, 1995. Princeton: Princeton University Press.

Rubinstein, A. 1979. Equilibrium in supergames with the overtaking criterion. *Journal of economic theory* **21**(1) 1–9.

Samuelson, L. 1997. Evolutionary games and equilibrium selection. Cambridge: MIT Press.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge: Harvard Univ. Press.

Sethi, R. 1996. Evolutionary stability and social norms. *Journal of Economic Behavior & Organization* **29**(1) 113–140.

Smith, E. A., R. Bliege Bird. 2000. Turtle hunting and tombstone opening: public generosity as costly signaling. *Evolution and Human Behavior* **21**(4) 245–261.

Soler, M. 2012. Costly signaling, ritual and cooperation: evidence from Candomblé, an Afro-Brazilian religion. *Evolution and Human Behavior* **33**(4) 346–356.

Spence, M. 1973. Job market signaling. *The Quarterly Journal of Economics* **87**(3) 355–374.

Taylor, M. 1976. *Anarchy and cooperation*. London: Wiley.

Trivers, R. L. 1971. The evolution of reciprocal altruism. *Quarterly review of biology* **46**(1) 35–57.

Wedekind, C., M. Milinski. 2000. Cooperation through image scoring in humans. *Science* **288**(5467) 850–852.

Weidlich, W., G. Haag. 1983. Concepts and models of a quantitative sociology: The dynamics of interacting populations. Springer-Verlag (Berlin and New York).

Witt, U. 1986. Evolution and stability of cooperation without enforceable contracts. *Kyklos* **39**(2) 245–266.

Wright, J. 1999. Altruism as a signal: Zahavi's alternative to kin selection and reciprocity. *Journal of Avian Biology* **30**(1) 108–115.

Appendices

Appendix A and B are large documents and are provided separately. They can be downloaded via the following link: http://wwwuser.gwdg.de/~cege/Diskussionspapiere/DP221_Appendix

Appendix C – Stable Semi-Pooling Equilibria

	Equilibrium	Support	Conditions for existence	Payoff differentials (superscript “f” indicates difference in fitness payoffs)
CC$\bar{\mathrm{m}}$ CD$\underline{\mathrm{m}}$ $\underline{\mathrm{m}}$	$\mathrm{p}_{\mathrm{CD}\underline{\mathrm{m}}}=\frac{\bar{\mathrm{k}}+(1-\lambda)\beta}{\lambda(1+\bar{\mathrm{m}})}$ $\mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}=1-\frac{\bar{\mathrm{k}}+(1-\lambda)\beta}{\lambda(1+\bar{\mathrm{m}})}$ $\mathrm{p}_{\underline{\mathrm{m}}}=1$	1. $\frac{\beta}{(\bar{\mathrm{m}}-\alpha+\beta)}<\frac{\bar{\mathrm{k}}}{(1+\alpha)}<\frac{1+\bar{\mathrm{m}}}{(1+\alpha)}$: $\frac{\bar{\mathrm{k}}+\beta}{(1+\bar{\mathrm{m}}+\beta)}<\lambda<1$	2. $\frac{\bar{\mathrm{k}}}{(1+\alpha)}\leq\frac{\beta}{(\bar{\mathrm{m}}-\alpha+\beta)}$: $1-\frac{(\bar{\mathrm{m}}-\alpha)}{\beta}\frac{\bar{\mathrm{k}}}{(1+\alpha)}<\lambda<1$	$\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\underline{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})=\lambda(\mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}})(1+\bar{\mathrm{m}})-\lambda\mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}(1+\alpha)$ $=\lambda(\bar{\mathrm{m}}-\alpha)\mathrm{p}_{\mathrm{CC}\bar{\mathrm{m}}}$ $=\lambda(\bar{\mathrm{m}}-\alpha)\left(1-\frac{\bar{\mathrm{k}}+(1-\lambda)\beta}{\lambda(1+\bar{\mathrm{m}})}\right)>0$ $\left(\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\underline{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})\right)^{\mathrm{f}}=-\alpha\lambda\left(1-\frac{\bar{\mathrm{k}}+(1-\lambda)\beta}{\lambda(1+\bar{\mathrm{m}})}\right)<0$
DC$\bar{\mathrm{m}}$ CD$\underline{\mathrm{m}}$ $\underline{\mathrm{m}}$ $\bar{\mathrm{m}}$	$\mathrm{p}_{\mathrm{CD}\underline{\mathrm{m}}}=\frac{1}{2}\left[1+\frac{\underline{\mathrm{k}}}{\lambda(1+\alpha)}\right]$ $\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}=\frac{1}{2}\left[1-\frac{\underline{\mathrm{k}}}{\lambda(1+\alpha)}\right]$ $\mathrm{p}_{\underline{\mathrm{m}}}=\frac{1}{2}\left[1+\frac{1}{(1-\lambda)\beta}\left[\frac{(1+\bar{\mathrm{m}})}{(1+\alpha)}\underline{\mathrm{k}}-\bar{\mathrm{k}}\right]\right]$ $\mathrm{p}_{\bar{\mathrm{m}}}=\frac{1}{2}\left[1-\frac{1}{(1-\lambda)\beta}\left[\frac{(1+\bar{\mathrm{m}})}{(1+\alpha)}\underline{\mathrm{k}}-\bar{\mathrm{k}}\right]\right]$	$0<\frac{\beta+(\underline{\mathrm{k}}-\bar{\mathrm{k}})}{(\bar{\mathrm{m}}-\alpha+\beta)}<\lambda$ $<1-\frac{1}{\beta}\left[\frac{(1+\bar{\mathrm{m}})}{(1+\mathrm{a})}\underline{\mathrm{k}}-\bar{\mathrm{k}}\right]<1$	$\beta(1+\alpha)>$ $\frac{(\bar{\mathrm{m}}-\alpha+\beta)}{(\bar{\mathrm{m}}-\alpha)}\left[(1+\bar{\mathrm{m}})\underline{\mathrm{k}}-(1+\alpha)\bar{\mathrm{k}}\right]$ $+\frac{\underline{\mathrm{k}}-\bar{\mathrm{k}}}{\bar{\mathrm{m}}-\alpha}\beta(1+\alpha)$	$\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\underline{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})$ $=\lambda\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}(1+\bar{\mathrm{m}})+(1-\lambda)\left[\mathrm{p}_{\bar{\mathrm{m}}}(-\beta)\right]-\lambda\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}(1+\alpha)$ $=\lambda(\bar{\mathrm{m}}-\alpha)\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}-\beta(1-\lambda)\mathrm{p}_{\bar{\mathrm{m}}}$ $\left(\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\underline{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})\right)^{\mathrm{f}}=-\alpha\lambda-\beta(1-\lambda)\mathrm{p}_{\bar{\mathrm{m}}}<0$
DC$\bar{\mathrm{m}}$ CD$\underline{\mathrm{m}}$ $\underline{\mathrm{m}}$	$\mathrm{p}_{\mathrm{CD}\underline{\mathrm{m}}}=\frac{1}{2}\left[1+\frac{\bar{\mathrm{k}}+(1-\lambda)\beta}{\lambda(1+\bar{\mathrm{m}})}\right]$ $\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}=\frac{1}{2}\left[1-\frac{\bar{\mathrm{k}}+(1-\lambda)\beta}{\lambda(1+\bar{\mathrm{m}})}\right]$ $\mathrm{p}_{\bar{\mathrm{m}}}=0$	$\lambda>\max\left\{\begin{array}{l}\frac{\bar{\mathrm{k}}+\beta}{(1+\bar{\mathrm{m}}+\beta)},1-\frac{1}{\beta(1+\alpha)}\left((1+\bar{\mathrm{m}})\underline{\mathrm{k}}-(1+\alpha)\bar{\mathrm{k}}\right),\\1-\frac{\bar{\mathrm{m}}-\alpha}{(\bar{\mathrm{m}}-\alpha+\beta)(1+\bar{\mathrm{m}})+(1+\alpha)\beta}(1+\bar{\mathrm{m}}+\bar{\mathrm{k}})\end{array}\right\}$		$\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\underline{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})=\lambda\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}(1+\bar{\mathrm{m}})-\lambda\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}(1+\alpha)$ $=\lambda(\bar{\mathrm{m}}-\alpha)\mathrm{p}_{\mathrm{DC}\bar{\mathrm{m}}}>0$ $\left(\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\underline{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})\right)^{\mathrm{f}}=-\alpha\lambda<0$
CD$\bar{\mathrm{m}}$ $\bar{\mathrm{m}}$ $\underline{\mathrm{m}}$	$\mathrm{p}_{\mathrm{CD}\bar{\mathrm{m}}}=1$	$\lambda=\frac{\underline{\mathrm{k}}}{(1+\alpha)}$	1. $\underline{\mathrm{k}}<(1+\alpha)$ 2. $\mathrm{p}_{\bar{\mathrm{m}}}<\frac{\lambda(\bar{\mathrm{m}}-\alpha)}{(1-\lambda)\beta}$	$\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\bar{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})=\frac{\underline{\mathrm{k}}}{(1+\alpha)}(1+\bar{\mathrm{m}})-\bar{\mathrm{k}}-\beta(1-\lambda)\mathrm{p}_{\bar{\mathrm{m}}}$ $\left(\Pi_{\bar{\mathrm{m}}}(\mathrm{CD},\bar{\mathrm{m}})-\Pi_{\underline{\mathrm{m}}}(\underline{\mathrm{m}})\right)^{\mathrm{f}}=\frac{\underline{\mathrm{k}}}{(1+\alpha)}-\bar{\mathrm{k}}-\beta(1-\lambda)\mathrm{p}_{\bar{\mathrm{m}}}$

Appendix D - Proofs

Proof (*Lemma 1*) Full cooperation can only be achieved with only high-types present in the population, i.e. $\lambda = 1$. There are only two equilibria which support cooperation among high-types that are supported at $\lambda = 1$ under certain conditions and potentially exhibit a fitness advantage for high-types (necessary for local stability), the separating cooperative equilibrium and the high pooling cooperative equilibrium. With respect to the former, the support condition amounts to $\frac{\underline{k}}{1+\alpha} \geq 1 \Leftrightarrow \underline{k} \geq 1+\alpha$, and the fitness condition to $\bar{k} < 1$ (see Table 5). With respect to the latter, the support condition amounts to $\underline{k} < 1+\alpha$, and the fitness condition to $\left(\beta-\alpha\right)-\beta+\underline{k}-\bar{k} > 0 \Leftrightarrow \underline{k}-\bar{k} > \alpha$. If $\underline{k}-\bar{k} = \alpha$ stability requires a strict positive difference in fitness payoffs for high-types for λ close to 1, i.e. $\beta-\alpha < 0$. QED

Proof (*Lemma 2*) The first pair of inequalities $\frac{\beta}{\bar{m}-\alpha+\beta}\bar{m} < \underline{k}-\bar{k} < \alpha$ arises from the condition of the root ($1-\frac{\underline{k}-\bar{k}-\alpha}{\beta-\alpha}$) of the fitness difference for the high pooling cooperative equilibrium to lie in the support of this equilibrium, i.e. $\max\left\{\frac{\underline{k}}{1+\alpha}, \frac{\beta}{\left(\beta+\bar{m}-\alpha\right)}\right\} < 1-\frac{\underline{k}-\bar{k}-\alpha}{\beta-\alpha} < 1$. Stability requires a negative slope of the fitness difference function, i.e. $\beta-\alpha$. Let us first consider $\frac{\underline{k}}{1+\alpha} \leq \frac{\beta}{\left(\beta+\bar{m}-\alpha\right)}$. In this case, the within-support condition amounts to $\frac{\beta}{\bar{m}-\alpha+\beta} < 1-\frac{\underline{k}-\bar{k}-\alpha}{\beta-\alpha} < 1$, rearranging yields $\frac{\beta}{\bar{m}-\alpha+\beta}\bar{m} < \underline{k}-\bar{k} < \alpha$. If on the other hand $\frac{\underline{k}}{1+\alpha} > \frac{\beta}{\left(\beta+\bar{m}-\alpha\right)}$, the within-support condition amounts to $\frac{\underline{k}}{1+\alpha} < 1-\frac{\underline{k}-\bar{k}-\alpha}{\beta-\alpha} < 1$, rearranging yields $\beta-\frac{\underline{k}}{1+\alpha}\left(\beta-\alpha\right) < \underline{k}-\bar{k} < \alpha$. Summarizing gives us $\frac{\beta}{\bar{m}-\alpha+\beta}\bar{m} < \underline{k}-\bar{k} < \alpha$ and $\beta-\frac{\underline{k}}{1+\alpha}\left(\beta-\alpha\right) < \underline{k}-\bar{k} \Leftrightarrow \beta < \frac{1+\beta}{1+\alpha}\underline{k}-\bar{k}$. Note that $\frac{\beta}{\bar{m}-\alpha+\beta}\bar{m} < \underline{k}-\bar{k} < \alpha$ implies that $\beta-\alpha < 0$, because $\frac{\beta}{\bar{m}-\alpha+\beta}\bar{m} < \alpha \Leftrightarrow \bar{m}\left(\beta-\alpha\right) < \alpha\left(\beta-\alpha\right) \overset{\bar{m}>\alpha}{\Leftrightarrow} \beta-\alpha < 0$ QED

Proof (*Lemma 3*) For an inner λ-stable equilibrium to exist at $\lambda = \frac{\underline{k}}{1+\alpha}$, we need (1) the connectedness of the supports of the involved equilibria, (2) a fitness advantage for high-types to the left of $\lambda = \frac{\underline{k}}{1+\alpha}$, (3) a fitness disadvantage for high types to the right of $\lambda = \frac{\underline{k}}{1+\alpha}$ and finally (4) for being an inner equilibrium $\lambda = \frac{\underline{k}}{1+\alpha} \in \left(0,1\right)$. (1) gives us $\frac{\beta}{\bar{m}-\alpha+\beta} \leq \frac{\underline{k}}{1+\alpha}$, (2) yields

$\frac{\underline{\mathrm{k}}}{1+\alpha}-\bar{\mathrm{k}}>0$, (3) amounts to $\frac{\underline{\mathrm{k}}}{1+\alpha}(\beta-\alpha)-\beta+\underline{\mathrm{k}}-\bar{\mathrm{k}}<0 \Leftrightarrow \frac{\underline{\mathrm{k}}}{1+\alpha}<\frac{\bar{\mathrm{k}}+\beta}{1+\beta}$, if high-types and low-types fare equally well at $\lambda=\frac{\underline{\mathrm{k}}}{1+\alpha}$, then for stability, high-types need to earn strictly less to the right of $\lambda=\frac{\underline{\mathrm{k}}}{1+\alpha}$. In essence, if $\frac{\underline{\mathrm{k}}}{1+\alpha}=\frac{\bar{\mathrm{k}}+\beta}{1+\beta}$, then $\beta-\alpha<0$, (4) is equivalent to $\underline{\mathrm{k}}<1+\alpha$. (1) and (3) are equivalent to:

$$\frac{\beta}{\bar{\mathrm{m}}-\alpha+\beta}\le\frac{\underline{\mathrm{k}}}{1+\alpha}\underset{\substack{\text{if = then}\\ \alpha>\beta}}{\le}\frac{\bar{\mathrm{k}}+\beta}{1+\beta} \tag{*}$$

(2) and (4) are equivalent to:

$$\bar{\mathrm{k}}<\frac{\underline{\mathrm{k}}}{1+\alpha}<1 \tag{**}$$

Note that (2) and (3) imply (4), hence what remains is: $\frac{\beta}{\bar{\mathrm{m}}-\alpha+\beta}\le\frac{\underline{\mathrm{k}}}{1+\alpha}\underset{\substack{\text{if = then}\\ \alpha>\beta}}{\le}\frac{\bar{\mathrm{k}}+\beta}{1+\beta}$ and $\bar{\mathrm{k}}<\frac{\underline{\mathrm{k}}}{1+\alpha}$

QED