von Wangenheim, Georg; Müller, Stephan

Conference Paper

Evolution of cooperation in social dilemmas: signaling internalized norms

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Norms and Culture, No. F10-V1

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: von Wangenheim, Georg; Müller, Stephan (2014) : Evolution of cooperation in social dilemmas: signaling internalized norms, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Norms and Culture, No. F10-V1, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/100340

Evolution of cooperation in social dilemmas: signaling internalized norms.

Stephan Müller, Georg von Wangenheim*

Abstract

Economists have a long tradition in identifying the evolution of cooperation in large, unstructured societies as a puzzle. We suggest a new explanation for cooperation which avoids restrictions of most previous attempts. Our explanation deals with the role of internalized norms for cooperation in large unstructured populations. Even internalized norms, i.e. norms which alter the perceived utility from acting in a cooperative or in an 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. Only when internalization of the norm may be communicated in a reliable way, the picture may change. We derive necessary and sufficient condition for cooperation to be part of an evolutionary stable equilibrium. These conditions relate signaling cost of norm-adopters and non-adopters, the strength of the social norm and parameter measuring the cost of cooperation.

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 obvious advantages of exploiting the good will of others, human beings often cooperate even in large, unstructured societies. However, neither is cooperation universal without exceptions nor is it easy to explain. Economists have a long tradition in identifying the evolution of cooperation in large, unstructured societies as a puzzle (e.g. Axelrod and Hamilton 1981, Fudenberg et al. 2012) and in finding explanations of cooperation based on some structure within the population.

Attempts to solve the puzzle are abundant but have so far most often relied on either or both of two restrictions: On the one hand, explanations have concentrated on structured populations, that is populations in which interaction is not completely anonymous but allows individuals to collect and process information about past behavior of others and about their identity. On the other hand, explanations have been depending on an unexplained ability of social norms to restrict the individuals' action or strategy spaces, in particular with respect to the abuse of punishment.

Among the first group, some strands of the literature deserve special mention.[1] The theory of kin selection focuses on cooperation among individuals who are closely related genetically (Hamilton 1964a; Hamilton 1964b), whereas theories of direct reciprocity focus on the selfish incentives for cooperation in repeated interactions (Trivers 1971; Axelrod 1984). In the case of infinite repetition within one group see Taylor (1976) or Mordecaï (1977) and the Folk-Theorem-type of results (Rubinstein 1979, 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 the cooperators can build a reputation (Nowak and Sigmund 1998; Wedekind and Milinski 2000; Gintis et al. 2001)[2].

Within the second group, we point to the early papers of Hirshleifer and Rasmusen (1989) and Witt (1986), who allow for punishment only after a norm has been violated. Sethi (1996) allows for all possible strategies which condition punishment on the violation of or the compliance with a norm, but adds structure to the society by introducing some exogenous division of the population into some individuals who behaves rationally and the rest whose behavior is determined by routines which are slowly adapted to their environment.

We suggest a new explanation for cooperation which avoids both restrictions. Our explanation deals with 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 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 disregard any information about past behavior of other individuals.

---

[1] A complete review of the literature would of course transcend the limits of an introductory section of a journal article – it would rather require an entire book or even more.

[2] There are other mechanisms that do not rely on informational aspects, but are based on restrictions in strategies: In finitely repeated games cooperation can e.g. 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 the environment which is most unfavorable to cooperation, a completely unstructured society where every interaction occurs among strangers for two reasons. One 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 are 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 which alter the perceived utility from acting in a cooperative or in an 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]. Should internalized norms simply exist, but lack the possibility of being signaled or screened for, they would induce norm bearers to cooperate and to 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 may be communicated in a reliable way, the picture may change, because then behavior may be conditioned on expected behavior of others.

Within this environment, we borrow from the indirect evolutionary approach Güth and Yaari (1992) and Güth (1995) the idea that internalized norms are nothing else than an internal payoff conditional on the behavior of the individual and its partners and 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. Obviously, the two games are 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.

Next to these differences with respect to the environment of interaction, we depart in a fundamental way concerning the behavioral assumptions. We assume that agents play inherited strategies defining both whether an individual signals its norm internalization and whether it cooperates or not. We thus take the stand of Behavioral Economics (as it is often reflected in evolutionary game theory) whereas Güth et al. (2000) applies a rational choice approach with

---

[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 by 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 fixed population size Kandori (1992) and Ellison (1994) show that cooperation can be sustained in a sequential equilibrium if individuals are patient enough. However, such contagious strategies may only uphold complete cooperation of all individuals, require nearly infinite patience in large societies and 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).

agents using Bayesian updating and rationally taking investment decisions with respect to screening. Our model is thus evolutionary both with respect to norm internalization and with respect to behavior, although the speed of the norm internalization dynamics is clearly less than the speed of behavioral adaptation.

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 has been argued before in the field of evolutionary biology (Wright 1999; Smith 2000; Leimar 2001). Only a few 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 if it wouldn't have the implication to induce a far richer set of equilibria concerning the distribution of an internalized norm which can stabilize cooperation. Notably the interplay of those multiple behavioral equilibria may stabilize partial cooperation and dissolves the necessity to introduce evolutionary forces into the dynamics of norm adoption beyond payoff monotonicity that are frequency based as in Gintis et al. (2001).

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 programs are highly complementary Sethi (1996, p. 117). Our results show that the complementarity between these different approaches is limited. We show that there is a substantial difference between assuming that norms simply fix a certain behavior and assuming that norms only give internal incentives to follow this behavior. In the latter case, which is ours', the parameter measuring how strong this incentive is 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 of 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. Finally, Section 6 concludes.

2. The model

The classical Prisoner's 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 of outcomes of their own past interactions. In particular they process no information about identity of opponents or about 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 the 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: Prisoner's Dilemma, where $\alpha > 0, \beta > 0$ and $1+\beta > \alpha$.

A usual assumption in evolutionary models explaining the presence of cooperative behavior is that individuals play inherited strategies that may depart from payoff maximizing behavior. The play of non-maximizing strategies in this line of research is then interpreted as norm-guided (e.g. Sethi 1996). To us this line of argument appears unsatisfactory since apart from showing that such strategies can be sustained in equilibrium it lacks any motivated for why an individual should adhere to that particular norm. We believe that individuals will not stick to behavior which is suboptimal in the current environment. We do not claim, that individuals always do what is best for them from an objective perspective (e.g. maximizes fitness), but they will not stick to suboptimal strategies forever. Hence, in our view any long-lasting departure from the behavior which 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 that determine behavior and behavior 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 $m \in \{\underline{m}, \overline{m}\}$ 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 $\underline{m} = 0, \overline{m} > \alpha$.[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. As in Güth and Ockenfels (2005) we make the natural assumption that beliefs correspond to actual frequencies of types. Without communication the impossibility result

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

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 that there exists a social technology which enables individuals to signal their positive attitude towards cooperation by incurring some costs. Furthermore, agents who actually adopted the norm are supposed to bear lower cost for sending the signal. Let $\overline{k}, \underline{k}$ denote the signaling cost for high types and low types respectively, so that $\overline{k} < \underline{k}$. In the current setup strategies are now given by signal-dependent behavior and an own signal, e.g. "cooperate if high-type signal is received, deviate if low-type signal is received and send high-type signal", denoted $CD\overline{m}$. In general terms, a strategy is denoted by a triple of which the first entry corresponds to behavior in the case of receiving a high-type signal, the second to behavior in the case of receiving a low-type signal, and the third to the signal sent.

What might such a signal be? To give an illustrative example consider the situation where individuals elbow their way through a bargaining sale. There is a rummage table with one good offered as two variants goods A and B. One of the two individuals considered prefers A, the other prefers B. However, for both getting both variants is the first best outcome. They can behave cooperatively leaving each other place to select their preferred variant or try to queue-jump and grab both in which case the other gets none. If both individuals chose not to cooperate, they would grab one of the variants 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 then an act of rational choice.

Evaluation of material payoffs is given

	C	D
C	$1+m$	$-\beta$
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 $CC\overline{m}$, $CD\overline{m}$, $DC\overline{m}$, $DD\overline{m}$ and $CC\underline{m}$, $CD\underline{m}$, $DC\underline{m}$, $DD\underline{m}$. For low types, since defection is dominant behavior, there are only two strategies that reflect their signals, denoted by $D\overline{m}$, $D\underline{m}$. To distinguish the signaling from the non-signal part of the strategy, we will call the former signal and the latter behavior. We will denote the share in the subpopulation of high types playing the strategy $CC\overline{m}$ by $p_{CC\overline{m}}$ and accordingly for any other strategy. Since low types always defect we denote their respective shares by $p_{\overline{m}}$ and $p_{\underline{m}}$.

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 around 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 a population following the former will increase faster than the share of 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 a 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 but – in terms of the composition of mutation-strategies – arbitrary perturbation.

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 a certain strategy 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 as adequate since behavior will adapt faster to differences in payoffs than socially and culturally transmitted norms. We therefore can 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 on the one hand describe the fast dynamics of various signal-behavior strategies and on the other hand the slow dynamics with respect to 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 have to 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 are p-stable and λ-stable, respectively. p-stable equilibria are presented in section 3, λ-stable equilibria are derived in section 4.

3. Equilibria with Exogenous Proportions of Norm Bearers

[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.

For the ease of reading, we only present the equilibria and their stability properties here and leave the derivation to Appendix A (existence) and B (stability). As in many other cases as well, we have – depending on the parameters including λ – separating and pooling equilibria. 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 signal, so called semi-pooling equilibria. Table 3 reports these equilibria.

In the following 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.

Type	Involved strategies	Equilibrium	Support	Conditions for existence	Payoff-Differentials (superscript "f" indicates difference in fitness payoffs)
Separating	High types cooperate against signal and defect else, norm holders send signal, others don't send signal				
	$CD\bar{m}$ $\underline{m}$	$p_{CD\bar{m}}=1,\ p_{\underline{m}}=1$	$\frac{\bar{k}}{1+\bar{m}}<\lambda<\frac{\underline{k}}{1+\alpha}$	$\bar{k}<1+\bar{m}$	$\Pi_{\bar{m}}(CD,\bar{m})-\Pi_{\underline{m}}(\underline{m})=\lambda(1+\bar{m})-\bar{k}$ $\left(\Pi_{\bar{m}}(CD,\bar{m})-\Pi_{\underline{m}}(\underline{m})\right)^{f}=\lambda-\bar{k}$
Low Pooling	High types cooperate, no signal				
	$CC\underline{m}$ $DC\underline{m}$ $\underline{m}$	$p_{CC\underline{m}}+p_{DC\underline{m}}=1$	$\frac{\beta}{(\beta+\bar{m}-\alpha)}\le\lambda$		$\Pi_{\bar{m}}(CC,\underline{m})-\Pi_{\underline{m}}(\underline{m})=\lambda(1+\bar{m})-\beta(1-\lambda)-\lambda(1+\alpha)$ $=\lambda(\bar{m}-\alpha)-\beta(1-\lambda)$ $\left(\Pi_{\bar{m}}(CC,\underline{m})-\Pi_{\underline{m}}(\underline{m})\right)^{f}=-\alpha\lambda-\beta(1-\lambda)<0$
	Complete defection, no signal				
	$CD\underline{m}$ $DD\underline{m}$ $\underline{m}$	$p_{CD\underline{m}}+p_{DD\underline{m}}=1$	$0<\lambda<1$	$p_{CD\underline{m}}\le\frac{1}{\lambda}\min\left\{\frac{\bar{k}+\beta}{1+\bar{m}+\beta},\frac{\bar{k}}{1+\alpha}\right\}$ $=\frac{1}{\lambda}\begin{cases}\frac{\bar{k}+\beta}{1+\bar{m}+\beta},\ (1+\alpha)\beta<(\bar{m}-\alpha+\beta)\bar{k}\\ \frac{\bar{k}}{1+\alpha},\ (1+\alpha)\beta>(\bar{m}-\alpha+\beta)\bar{k}\end{cases}$	$\Pi_{\bar{m}}(CD,\underline{m})-\Pi_{\underline{m}}(\underline{m})=0$ $\left(\Pi_{\bar{m}}(CD,\underline{m})-\Pi_{\underline{m}}(\underline{m})\right)^{f}=0$
High Pooling	High types cooperate, all signal				
	$CC\bar{m}$ $CD\bar{m}$ $\bar{m}$	$p_{CC\bar{m}}+p_{CD\bar{m}}=1$	$\lambda\ge\max\left\{\frac{\underline{k}}{1+\alpha},\frac{\beta}{(\beta+\bar{m}-\alpha)}\right\}$	$\underline{k}<1+\alpha$ $\lambda\ge\frac{\underline{k}}{p_{CD\bar{m}}(1+\alpha)}\Rightarrow p_{CD\bar{m}}>\frac{\underline{k}}{(1+\alpha)}$	$\Pi_{\bar{m}}(CC,\bar{m})-\Pi_{\underline{m}}(\bar{m})$ $=\lambda(\bar{m}-\alpha+\beta)-\beta+\underline{k}-\bar{k}$ $\left(\Pi_{\bar{m}}(CC,\bar{m})-\Pi_{\underline{m}}(\bar{m})\right)^{f}=\lambda(\beta-\alpha)-\beta+\underline{k}-\bar{k}$

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

The exception is the p-stable semi-pooling equilibrium at $\lambda = \frac{\underline{k}}{1+\alpha}$ that will be of relevance for one of the inner λ-stable equilibria. In this semi-pooling equilibrium high types always play $CD\bar{m}$ and low types are indifferent between sending the signal or not and therefore $p_{\bar{m}}$ is undefined. The minor importance of all other p-stable semi-pooling equilibria is partly due to the fact that they are characterized by strictly negative fitness differentials between high and low types and partly due 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. That there are both a lower and an upper bound in the support for this equilibrium has the following intuition. If there are too few high types then the cooperative outcome among them cannot compensate for signaling cost. The higher the signaling cost 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 then signaling becomes sufficiently profitable for low types. In other words if there are enough high types that cooperate when receiving the cooperative signal then it becomes profitable for low types to incur the signaling cost. The higher the signal 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 ($\frac{\bar{k}}{1+\bar{m}}$) 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 ($\frac{\underline{k}}{1+\alpha}$), the likelihood to gain the benefit.

In the low pooling cooperative equilibrium nobody signals and high types cooperate. This equilibrium exists if there are sufficiently many high types. Only then 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 below a certain threshold then they will start to prefer playing defective when receiving the low signal. Note that this equilibrium is indeed an equilibrium set, since the strategies $CC\underline{m}$ and $DC\underline{m}$ are equivalent in equilibrium. The share of high types required for this to be an equilibrium increases in the sucker's payoff, since with increasing (absolute) sucker's payoffs cooperative behavior becomes more disadvantageous. This threshold, too, has an intuitive meaning. Note that $\bar{m}-\alpha$ (β) measures the incentive to reciprocate cooperative (defective) behavior, i.e. the condition $\frac{\beta}{\beta+\bar{m}-\alpha} < \lambda$, which can be rewritten as $\lambda(\bar{m}-\alpha) > (1-\lambda)\beta$, 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 earing a payoff of zero. Again, due to lack of distinguishability in equilibrium, equilibrium is indeed a set where $CD\underline{m}$ and $DD\underline{m}$ 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. Given that 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 then they will 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 arises reflecting the incentive compatibility for low types to signal. Note that this equilibrium again is an equilibrium set, since the strategies $CC\bar{m}$ and $CD\bar{m}$ 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 ($\frac{\underline{k}}{1+\alpha}$) 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. In analogy to the derivation of p-equilibria, the differentials in these fitness payoffs among high and low types is the driving force for the evolution of their respective shares. To ease understanding the 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. They then 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 $\bar{k}$, their expected payoff in the cooperative separating equilibrium is $\lambda - \bar{k}$, which is also the expected difference of fitness payoffs:
$\left(\Pi_{\bar{m}}(CD,\bar{m}) - \Pi_{\underline{m}}(\underline{m})\right)^{f} = \lambda - \bar{k}$.

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] Absent signaling costs, differences in material payoffs solely reflect incentives of the underlying PD. More precisely, with probability λ high types meet their own type and realize the cooperative outcome, i.e. they earn 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 a 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.

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

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:

$$\left(\Pi_{\bar{m}}(CC,\bar{m})-\Pi_{\underline{m}}(\bar{m})\right)^{f}=\underline{k}-\bar{k}-\left(\lambda\alpha+(1-\lambda)\beta\right)$$

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 just like the cooperative separating equilibrium breaks down for too high shares of the high type.

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).

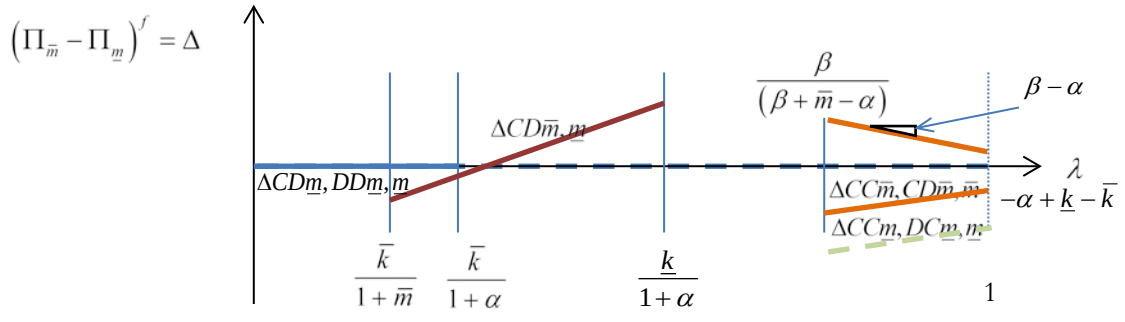

Figure 1: Differences in material payoffs for $\frac{\underline{k}}{1+\alpha}<\frac{\beta}{(\beta+\bar{m}-\alpha)}$

[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.

Payoff differences for semi-pooling equilibria are neglected since their support lies in the interval $\left(\frac{\beta}{(\beta+\bar{m}-\alpha)},1\right]$ 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.

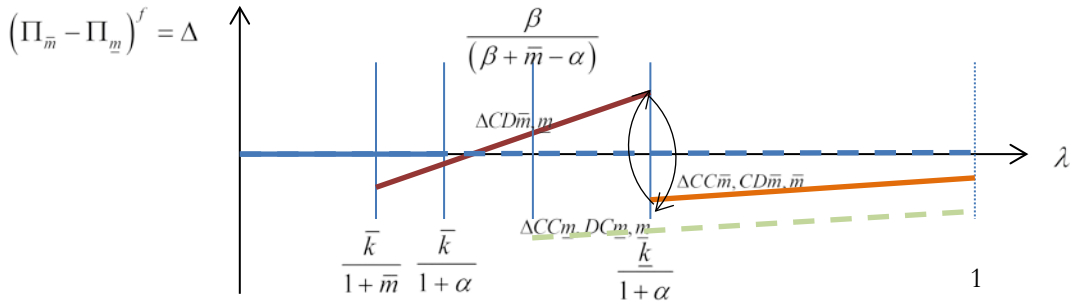

Figure 2: Differences in material payoffs for $\frac{\underline{k}}{1+\alpha}\geq\frac{\beta}{(\beta+\bar{m}-\alpha)}$

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, 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 $\lambda=1$ a strictly positive payoff difference in favor of high types at $\lambda<1$, at the lower boundary a strictly negative payoff difference at $\lambda>0$ is necessary and sufficient.

We first analyze whether there exist λ-stable equilibria with full cooperation. 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 iff $\underline{k}\geq 1+\alpha$ and $\bar{k}<1$
(2) by the high pooling cooperative equilibrium iff: $\underline{k}<1+\alpha$ and either
$\underline{k}-\bar{k}>\alpha$ or $\underline{k}-\bar{k}=\alpha>\beta$.

All proofs are in Appendix D.

Although the existence of fully cooperative equilibria might seem surprising at the first glance, a closer look at the stated conditions reveals how strong they are. 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 $1+\alpha$ form playing the dominant defective strategy at cost $\underline{k}$. The second qualification $\overline{k} < 1$ stems from the incentive compatibility constraint for high types, since they could always earn zero by not-signaling and defective behavior. In the case of the high pooling cooperative equilibrium the difference in the signaling cost must exceed the material reward to defect on a cooperative opponent.

The restrictiveness of Lemma 1 draws our attention to stable inner 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 $1-\frac{\underline{k}-\overline{k}-\alpha}{\beta-\alpha}$. All other equilibria are characterized by either strictly negative or by strictly increasing payoff differentials. The high pooling cooperative equilibrium exists and is λ-stable, if $1-\frac{\underline{k}-\overline{k}-\alpha}{\beta-\alpha}$ 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 a λ-stable inner equilibrium at $1-\frac{\underline{k}-\overline{k}-\alpha}{\beta-\alpha}$ if and only: $\frac{\beta}{\overline{m}-\alpha+\beta}\overline{m} < \underline{k}-\overline{k} < \alpha$ and $\beta < \frac{1+\beta}{1+\alpha}\underline{k}-\overline{k}$.

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 by high types only. Looking at the conditions we observe that the existence of stable inner equilibria requires that norm adopters' costs of signaling 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, (1) the supports of the p-equilibria need to be adjacent, (2) the differences of fitness payoffs have to exhibit the same properties as for the stable λ-equilibrium constituted by only one p-equilibrium (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 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 the argument, we suggest the reader to consider Figure 2 while reading the following argument.

Condition (1) requires that $\frac{\underline{k}}{1+\alpha} \geq \frac{\beta}{\overline{m}-\alpha+\beta}$ (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 we have the fitness difference given by $\Pi^f_{\overline{m}}(CD,\overline{m}) - \Pi^f_{\underline{m}}(\underline{m}) = \lambda - \overline{k}$ for $\lambda \leq \frac{\underline{k}}{1+\alpha}$. This difference must be strictly positive at $\lambda = \frac{\underline{k}}{1+\alpha}$, whence $1+\alpha < \frac{\underline{k}}{\overline{k}}$. 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 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 $\left(\Pi_{\overline{m}}(CC,\overline{m}) - \Pi_{\underline{m}}(\overline{m})\right)^f = \lambda(\beta-\alpha) - \beta + \underline{k} - \overline{k}$, which has to be negative. We Hence get $\frac{\underline{k}}{1+\alpha} < \frac{\overline{k}+\beta}{1+\beta}$.

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

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

where the first inequality is strict if $p_{\overline{m}} < 1$ and the second inequality requires $\lambda^* \equiv \frac{\underline{k}}{1+\alpha} > \frac{\beta}{m-\alpha+\beta} \equiv \tilde{\lambda}$. Hence continuity of the payoffs and Lipschitz-continuity of the dynamics implies that for all λ sufficiently close to λ^* and all sufficiently large $p_{CD\overline{m}} + p_{CC\overline{m}}$ we have $\dot{p}_{CD\overline{m}} + \dot{p}_{CC\overline{m}} > 0$. Hence, as once $p_{CD\overline{m}} + p_{CC\overline{m}}$ has become large enough close to any of the three relevant p-stable equilibria, $p_{CD\overline{m}} + p_{CC\overline{m}}$ will continue to grow for all $p_{\overline{m}}$. Second, we observe that if $p_{CD\overline{m}}$ 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 keep the first argument valid. Third, if $p_{CD\overline{m}} + p_{CC\overline{m}}$ is large enough and thus increases, $\Pi_{\overline{m}}(CD\overline{m}) < \Pi_{\overline{m}}(CC\overline{m})$ only occurs for ever decreasing ranges of large $p_{\overline{m}}$. Hence for every payoff-monotone dynamic $p_{CC\overline{m}}$ will be smaller after every full cycle and will never again reach its previous maximum level. Hence, $p_{CD\overline{m}}$ will eventually be large enough to secure that our second argument is valid.

Hence, once our full dynamic system is close enough to λ^* and the λ-dynamic is slow enough, the system will cycle 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; a limit cycle may exist.) We summarize all conditions in the following

Lemma 3 If $\frac{\beta}{\bar{m}-\alpha+\beta} < \frac{\underline{k}}{1+\alpha} \underset{\alpha>\beta}{\overset{\text{if = then}}{\leq}} \frac{\bar{k}+\beta}{1+\beta}$ and $\bar{k} < \frac{\underline{k}}{1+\alpha}$ an inner λ-stable equilibrium exists at $\lambda = \frac{\underline{k}}{1+\alpha}$, in which high-type individuals cooperate among each other but also with those low-type individuals who signal to be of the high type and 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 $\frac{\underline{k}}{1+\alpha} \leq \frac{\beta}{\bar{m}-\alpha+\beta}$. Given that we have $\tilde{p}_{\bar{m}}\big|_{\lambda=\frac{\bar{k}}{1+\alpha}} < 1$ and there is a gap between the λ-supports of the separating cooperative equilibrium and of the high pooling cooperative equilibrium in case of a strict inequality (see Figure 1) and instability of the equilibrium around $\lambda = \frac{\underline{k}}{1+\alpha}$ in case of an equality. In the interval $\left(\frac{\underline{k}}{1+\alpha}, \frac{\beta}{\bar{m}-\alpha+\beta}\right)$ the defective separating equilibrium is the unique equilibrium. Should the population start at the cooperative separating p-equilibrium with positive fitness differential then it will eventually drive the share of high-type individuals beyond the λ-support of this equilibrium so that $p_{\bar{m}}$ starts to grow. Once it grows too much, the strategy $DD\bar{m}$ yields the largest behavioral payoff to high-type individuals and $CD\bar{m}$ only the second-largest. Hence the share of always defecting high-type individuals $p_{DD\bar{m}}$ must grow and $p_{CD\bar{m}}$ 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 advantageousness of low type's signaling the false type so that $p_{\bar{m}}$ will eventually decline again. A behavioral equilibrium 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 exists, but is not stable (see Appendix B). As a consequence, $p_{DD\bar{m}}$ will eventually grow large enough to bring 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 $p_{\bar{m}}$ and $p_{DD\bar{m}}$ both become so large that $CD\bar{m}$ becomes less profitable than $DC\bar{m}$. Then there may be payoff monotonic dynamics for which $p_{DC\bar{m}}$ starts to grow, although slower than $p_{DD\bar{m}}$. If this happens, eventually false signaling by high types may become reasonable. However, as the low pooling equilibrium fails to exist in the interval $\left(\frac{\underline{k}}{1+\alpha}, \frac{\beta}{\bar{m}-\alpha+\beta}\right)$, we conjecture that the population will eventually end up in the defective separating equilibrium as the unique behavioral equilibrium:

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

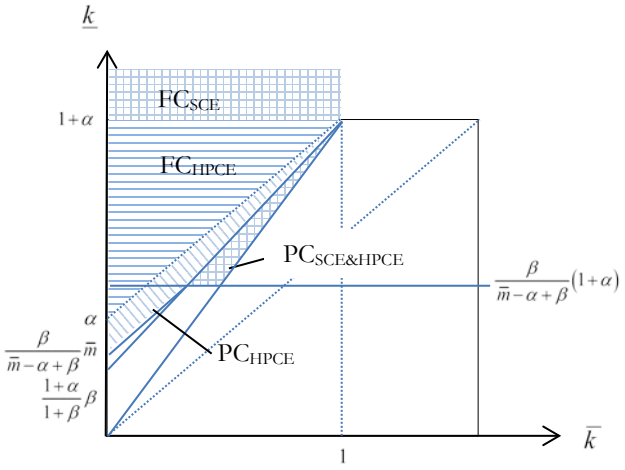

Figure 3: parameter region for partial or full cooperation

Figure 3 illustrate the conditions of Lemma 1-Lemma 3 graphically. For all inner equilibria potentially to exist we assume $\beta-\alpha<0$ and $\frac{\underline{k}}{1+\alpha}\geq\frac{\beta}{\bar{m}-\alpha+\beta}$. SCE denotes the separating cooperative equilibrium, HPCE the high pooling cooperative equilibrium and SCE&HPCE the inner equilibrium that is generated by the interplay of SCE, HPCE and a semi-pooling equilibrium. Finally FC and PC indicate full or partial cooperation respectively.

It is worth noting that the strength of the cooperative norm measured by $\bar{m}$ has a direct impact on the parameter set allowing for λ-stable inner equilibria (see Figure 3). As $\bar{m}$ gets closer the incentive to defect α then the parameter region supporting a separating cooperative equilibrium becomes smaller and smaller. Although the size of $\bar{m}$ is not important for the behavioral consequence for each individual, but only its relation to α, its size matters with respect to the presence of evolutionary stable equilibria characterized by partial cooperation.

5. Collecting requirements for equilibria with cooperation

Combining Lemma 1-3 of the previous Section, we can state 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 specific model, 'sufficiently large' translates to

$\underline{k}-\bar{k}>\alpha$ or $\underline{k}-\bar{k}=\alpha>\beta$ for full cooperation (Lemma 1)

and

either

$$\underset{\substack{\geq \\ \text{for 2.} \\ \text{term}}}{\underline{k}} \geq (1+\alpha)\max\left\{\frac{\beta}{\bar{m}-\alpha+\beta}, \bar{k}\right\}$$

or

$$\alpha > \beta \quad \text{and} \quad \underline{k} > \begin{cases} \dfrac{\beta}{\bar{m}-\alpha+\beta}\bar{m}+\bar{k}, & 0 \leq \bar{k} < \dfrac{\beta}{\bar{m}-\alpha+\beta}(1+\alpha-\bar{m}) \\ (1+\alpha)\max\left\{\dfrac{\beta}{\bar{m}-\alpha+\beta}, \bar{k}\right\}, & \dfrac{\beta}{\bar{m}-\alpha+\beta}(1+\alpha-\bar{m}) < \bar{k} \leq 1 \end{cases}$$

for partial or full cooperation (Lemma 1-Lemma 3)

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 cost 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 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 Figure 5 and Figure 5 show that chances for cooperation diminish with increasing β, i.e. the riskier cooperation or the more painful cooperation is when matched with defective behaviour the higher the requirements with respect to signalling costs for low types and the inner motive for mutual cooperation. A mirror argument applies with respect to the 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 only the incentive to defect on a cooperative player α relative to the difference in signalling costs 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

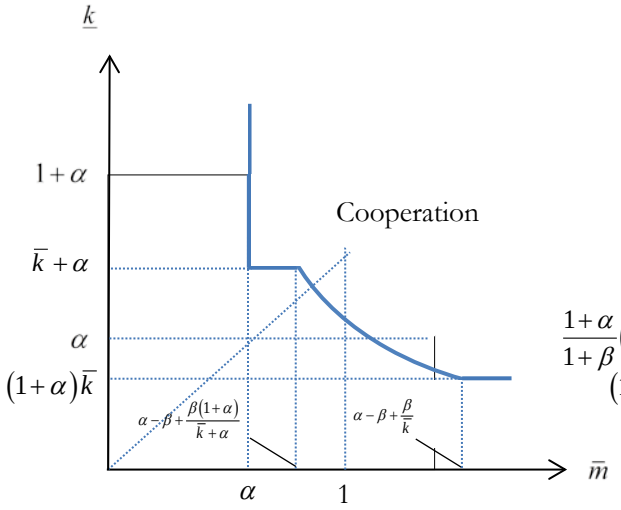

Figure 4: $\alpha \leq \beta$

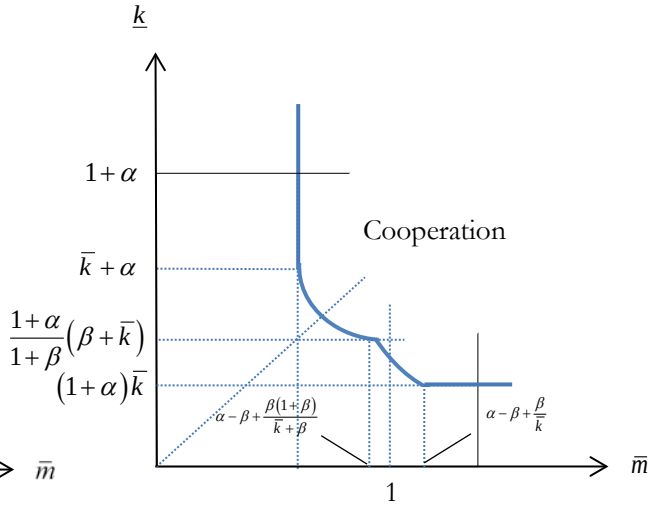

Figure 5: $\alpha > \beta$

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 provide a theorem that states necessary and sufficient condition for full or partial cooperation to be prevalent in a stable equilibrium. These conditions reflect on the difference in signaling cost between the cooperative and the opportunistic type, the extent of the cooperative norm and the model parameters of the Prisoner's dilemma, i.e. the temptation to defect and the sucker's payoff. We obtain several interesting results. First of all, although it is true that the size of the behavioral parameter measuring the evaluation of certain material outcomes is not important for the behavioral consequence for each individual, but only its relation to the incentive to defect, its size matters with respect to the presence of evolutionary stable equilibria characterized by partial cooperation. More precisely, the stronger the inner motive to cooperate the less restrictive 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 allowing for partial or full cooperation is weakly increasing in 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 cooperative 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 Prisoners' Dilemma with one of the most general class of dynamics considered in evolutionary game theory, namely the class of payoff-monotone dynamics. 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 has been argued before in the literature. 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 Prisoners' Dilemma. This would be rather a technical note if it wouldn't have the implication to induce a far richer set of equilibria concerning the distribution of an internalized norm which can stabilize cooperation. In particular notably is 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 given that 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 for areas where interaction is rather anonymous. Then informal institutions may spontaneously evolve more easily even in large unstructured interaction environments. Even if politics cannot alter the underlying incentives of the social dilemma to the extent such 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 the strength of the norm once incorporated. Hence, strengthening the internalized norms will also increase the chance for cooperation.

If we argue that it is spontaneous institutions which repel defection in large unstructured societies, then these insights lead us to argue that concepts of institutions should not require that all individuals adhere to the behavior prescribed by the spontaneous institution. Rather 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 Prisoners' Dilemma situations in a society. Without going into any detail here, we conjecture from our signaling model that cooperation in one Prisoners' Dilemma may serve as a signal to have the internal cooperation in order to better fare in another Prisoners' Dilemma. The temptation to defect in first game would be the costs to falsely signal having the internal motivation to cooperate. Hence the interplay between different Prisoners' Dilemma situations does not allow for scaling up: temptation in first game cannot be larger than in second game, or cooperation there cannot be complete. Further research needs to be done on the details of the interplay between different Prisoners' Dilemma games in an unstructured society.

---

[11] Although Gintis et al. (2001), as one of the few formal evolutionary signaling models, show the existence of an stable separating equilibrium, it would under payoff monotonicity only cease to exist as the type that correspond to our high types face an evolutionary advantage. Thereby their share in the population would increase and eventually exceed the threshold beyond which the separating equilibrium breaks down.

The analysis for more general norm than the one we considered is left for future research. So far, we think that the size of the parameter measuring 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 population start interacting with each other a cooperative population might induce cooperation in an defective population and vice versa. To analyze such an environment might 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., E. Maskin. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica: Journal of the Econometric Society* 533–554.

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.

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

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.

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

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.

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.

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, Kagel J., Roth A., The Handbook of Experimental Economics, 1995. Princeton University Press, Princeton, NJ.

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. MIT Press (Cambridge, Mass.).

Samuelson, P. A. 1947. Foundations of Economic Analysis (Atheneum, New York).

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

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

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

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.

Trivers, R. L. 1971. The evolution of reciprocal altruism. *Quarterly review of biology* 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* 108–115.

Appendices

Appendix A and B are large documents and are available upon request.

Appendix C – Stable Semi-Pooling Equilibria

Involved strategies	Equilibrium	Support	Conditions for existence	Payoff-Differentials (superscript "f" indicates difference in fitness payoffs)
$CC\bar{m}$ $CD\underline{m}$ $\underline{m}$	$\left\{\begin{array}{l} p_{CD\underline{m}} = \dfrac{\bar{k}+(1-\lambda)\beta}{\lambda(1+\bar{m})} \\ p_{CC\bar{m}} = 1-\dfrac{\bar{k}+(1-\lambda)\beta}{\lambda(1+\bar{m})} \\ p_{\underline{m}} = 1 \end{array}\right\}$	1. $\dfrac{\beta}{(\bar{m}-\alpha+\beta)} < \dfrac{\bar{k}}{(1+\alpha)} < \dfrac{1+\bar{m}}{(1+\alpha)}$: $\dfrac{\bar{k}+\beta}{(1+\bar{m}+\beta)} < \lambda < 1$ 2. $\dfrac{\bar{k}}{(1+\alpha)} \le \dfrac{\beta}{(\bar{m}-\alpha+\beta)}$: $1-\dfrac{(\bar{m}-\alpha)}{\beta}\dfrac{\bar{k}}{(1+\alpha)} < \lambda < 1$		$\Pi_{\bar{m}}(CD,\underline{m}) - \Pi_{\underline{m}}(\underline{m}) = \lambda(p_{CC\bar{m}})(1+\bar{m}) - \lambda p_{CC\bar{m}}(1+\alpha)$ $= \lambda(\bar{m}-\alpha)p_{CC\bar{m}}$ $= \lambda(\bar{m}-\alpha)\left(1-\dfrac{\bar{k}+(1-\lambda)\beta}{\lambda(1+\bar{m})}\right) > 0$ $\left(\Pi_{\bar{m}}(CD,\underline{m}) - \Pi_{\underline{m}}(\underline{m})\right)^f = -\alpha\lambda\left(1-\dfrac{\bar{k}+(1-\lambda)\beta}{\lambda(1+\bar{m})}\right) < 0$
$DC\bar{m}$ $CD\underline{m}$ $\underline{m}$ $\bar{m}$	$\left\{\begin{array}{l} p_{CD\underline{m}} = \dfrac{1}{2}\left[1+\dfrac{\underline{k}}{\lambda(1+\alpha)}\right] \\ p_{DC\bar{m}} = \dfrac{1}{2}\left[1-\dfrac{\underline{k}}{\lambda(1+\alpha)}\right] \\ p_{\underline{m}} = \dfrac{1}{2}\left[1+\dfrac{1}{(1-\lambda)\beta}\left[\dfrac{(1+\bar{m})}{(1+\alpha)}\underline{k}-\bar{k}\right]\right] \\ p_{\bar{m}} = \dfrac{1}{2}\left[1-\dfrac{1}{(1-\lambda)\beta}\left[\dfrac{(1+\bar{m})}{(1+\alpha)}\underline{k}-\bar{k}\right]\right] \end{array}\right\}$	$0 < \dfrac{\beta+(\underline{k}-\bar{k})}{(\bar{m}-\alpha+\beta)} < \lambda$ $< 1-\dfrac{1}{\beta}\left[\dfrac{(1+\bar{m})}{(1+a)}\underline{k}-\bar{k}\right] < 1$	$\beta(1+\alpha) >$ $\dfrac{(\bar{m}-\alpha+\beta)}{(\bar{m}-\alpha)}\left[(1+\bar{m})\underline{k}-(1+\alpha)\bar{k}\right]$ $+\dfrac{\underline{k}-\bar{k}}{\bar{m}-\alpha}\beta(1+\alpha)$	$\Pi_{\bar{m}}(CD,\underline{m}) - \Pi_{\underline{m}}(\underline{m})$ $= \lambda p_{DC\bar{m}}(1+\bar{m}) + (1-\lambda)\left[p_{\bar{m}}(-\beta)\right] - \lambda p_{DC\bar{m}}(1+\alpha)$ $= \lambda(\bar{m}-\alpha)p_{DC\bar{m}} - \beta(1-\lambda)p_{\bar{m}}$ $\left(\Pi_{\bar{m}}(CD,\underline{m}) - \Pi_{\underline{m}}(\underline{m})\right)^f = -\alpha\lambda - \beta(1-\lambda)p_{\bar{m}} < 0$

$DC\bar{m}$ $CD\underline{m}$ $\underline{m}$	$\left\{\begin{array}{l} p_{CD\underline{m}} = \frac{1}{2}\left[1+\frac{\bar{k}+(1-\lambda)\beta}{\lambda(1+\bar{m})}\right] \\ p_{DC\bar{m}} = \frac{1}{2}\left[1-\frac{\bar{k}+(1-\lambda)\beta}{\lambda(1+\bar{m})}\right] \\ p_{\bar{m}} = 0 \end{array}\right\}$	$\lambda > \max\left\{\frac{\bar{k}+\beta}{(1+\bar{m}+\beta)}, 1-\frac{1}{\beta(1+\alpha)}\left((1+\bar{m})\underline{k}-(1+\alpha)\bar{k}\right), 1-\frac{\bar{m}-\alpha}{(\bar{m}-\alpha+\beta)(1+\bar{m})+(1+\alpha)\beta}(1+\bar{m}+\bar{k})\right\}$ $= \begin{cases} \frac{\bar{k}+\beta}{(1+\bar{m}+\beta)}, & (1+\alpha)\beta < (\bar{m}-\alpha+\beta)\bar{k} \\ 1-\frac{(\bar{m}-\alpha)}{(\bar{m}-\alpha+\beta)(1+\bar{m})+(1+\alpha)\beta}(1+\bar{m}+\bar{k}), & (\bar{m}-\alpha+\beta)\bar{k} < (1+\alpha)\beta < \frac{(\bar{m}-\alpha+\beta)(1+\bar{m})+(1+\alpha)\beta}{(\bar{m}-\alpha)}\underline{k}-\frac{(\bar{m}-\alpha+\beta)(1+\alpha)+(1+\alpha)\beta}{(\bar{m}-\alpha)}\bar{k} \\ 1-\frac{1}{\beta(1+\alpha)}\left((1+\bar{m})\underline{k}-(1+\alpha)\bar{k}\right), & \frac{(\bar{m}-\alpha+\beta)(1+\bar{m})+(1+\alpha)\beta}{(\bar{m}-\alpha)}\underline{k}-\frac{(\bar{m}-\alpha+\beta)(1+\alpha)+(1+\alpha)\beta}{(\bar{m}-\alpha)}\bar{k} < (1+\alpha)\beta \end{cases}$	$\Pi_{\bar{m}}(CD,\underline{m})-\Pi_{\underline{m}}(\underline{m}) = \lambda p_{DC\bar{m}}(1+\bar{m})-\lambda p_{DC\bar{m}}(1+\alpha)$ $= \lambda(\bar{m}-\alpha)p_{DC\bar{m}} > 0$ $\left(\Pi_{\bar{m}}(CD,\underline{m})-\Pi_{\underline{m}}(\underline{m})\right)^f = -\alpha\lambda < 0$
$CD\bar{m}$ $\bar{m}$ $\underline{m}$	$p_{CD\bar{m}} = 1$	$\lambda = \frac{\underline{k}}{(1+\alpha)}$	1. $\underline{k} < (1+\alpha)$ 2. $p_{\bar{m}} < \frac{\lambda(\bar{m}-\alpha)}{(1-\lambda)\beta}$ note that 3. is only binding if: $\frac{\lambda(\bar{m}-\alpha)}{(1-\lambda)\beta} < 1 \Leftrightarrow \lambda < \frac{\beta}{(\bar{m}-\alpha+\beta)}$ $\Leftrightarrow \frac{\underline{k}}{(1+\alpha)} < \frac{\beta}{(\bar{m}-\alpha+\beta)}$	$\Pi_{\bar{m}}(CD,\bar{m})-\Pi_{\underline{m}}(\underline{m}) = \frac{\underline{k}}{(1+\alpha)}(1+\bar{m})-\bar{k}-\beta(1-\lambda)p_{\bar{m}}$ $\left(\Pi_{\bar{m}}(CD,\bar{m})-\Pi_{\underline{m}}(\underline{m})\right)^f = \frac{\underline{k}}{(1+\alpha)}-\bar{k}-\beta(1-\lambda)p_{\bar{m}}$

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 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 $\frac{\underline{k}}{1+\alpha} \geq 1 \Leftrightarrow \underline{k} \geq 1+\alpha$, the fitness condition to $\bar{k} < 1$ (see Table 5). With respect to the latter the support condition amoutns to $\underline{k} < 1+\alpha$, the fitness condition to $(\beta - \alpha) - \beta + \underline{k} - \bar{k} > 0 \Leftrightarrow \underline{k} - \bar{k} > \alpha$. If $\underline{k} - \bar{k} = \alpha$ stability requires a strikt 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 on 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. $\max\left\{\frac{\underline{k}}{1+\alpha}, \frac{\beta}{(\beta + \bar{m} - \alpha)}\right\} < 1 - \frac{\underline{k} - \bar{k} - \alpha}{\beta - \alpha} < 1$. Stability requires a negative slope oft he fitness difference function, i.e. $\beta - \alpha$. Lets first consider $\frac{\underline{k}}{1+\alpha} \leq \frac{\beta}{(\beta + \bar{m} - \alpha)}$. In that 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 $\frac{\underline{k}}{1+\alpha} > \frac{\beta}{(\beta + \bar{m} - \alpha)}$ the within-support condition amounts to $\frac{\underline{k}}{1+\alpha} < 1 - \frac{\underline{k} - \bar{k} - \alpha}{\beta - \alpha} < 1$, rearranging yields $\beta - \frac{\underline{k}}{1+\alpha}(\beta - \alpha) < \underline{k} - \bar{k} < \alpha$. Summarizing gives us $\frac{\beta}{\bar{m} - \alpha + \beta} \bar{m} < \underline{k} - \bar{k} < \alpha$ and $\beta - \frac{\underline{k}}{1+\alpha}(\beta - \alpha) < \underline{k} - \bar{k} \Leftrightarrow \beta < \frac{1+\beta}{1+\alpha} \underline{k} - \bar{k}$. Note that $\frac{\beta}{\bar{m} - \alpha + \beta} \bar{m} < \underline{k} - \bar{k} < \alpha$ implies that $\beta - \alpha < 0$, because

$$\frac{\beta}{\bar{m} - \alpha + \beta} \bar{m} < \alpha \Leftrightarrow \bar{m}(\beta - \alpha) < \alpha(\beta - \alpha) \overset{\bar{m} > \alpha}{\Leftrightarrow} \beta - \alpha < 0$$

QED

Proof (*Lemma 3*) For an inner λ-stable equilibrium to exist at $\lambda = \frac{\underline{k}}{1+\alpha}$ we need (1) the connectivness 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 typest o the right of $\lambda = \frac{\underline{k}}{1+\alpha}$ and finally (4) for being an inner equilibrium $\lambda = \frac{\underline{k}}{1+\alpha} \in (0,1)$. (1) gives us $\frac{\beta}{\bar{m} - \alpha + \beta} \leq \frac{\underline{k}}{1+\alpha}$, (2) yields $\frac{\underline{k}}{1+\alpha} - \bar{k} > 0$, (3) amounts to $\frac{\underline{k}}{1+\alpha}(\beta - \alpha) - \beta + \underline{k} - \bar{k} < 0 \Leftrightarrow \frac{\underline{k}}{1+\alpha} < \frac{\bar{k} + \beta}{1+\beta}$, if high types and low types fare equally well at

$\lambda = \frac{\underline{k}}{1+\alpha}$ then for stability high types need to earn strictly less to the right of $\lambda = \frac{\underline{k}}{1+\alpha}$ i.e. if $\frac{\underline{k}}{1+\alpha} = \frac{\overline{k}+\beta}{1+\beta}$ then $\beta - \alpha < 0$, (4) is equivalent to $\underline{k} < 1+\alpha$. (1) and (3) are equivalent to

$$\frac{\beta}{\overline{m}-\alpha+\beta} \leq \frac{\underline{k}}{1+\alpha} \underset{\alpha>\beta}{\overset{\text{if = then}}{\leq}} \frac{\overline{k}+\beta}{1+\beta} \qquad (*)$$

(2) and (4) are equivalent to

$$\overline{k} < \frac{\underline{k}}{1+\alpha} < 1 \qquad (**)$$

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

QED