Center for European, Governance and Economic Development Research
cege

Discussion Papers

Number 221 – November 2014

COEVOLUTION OF COOPERATION, PREFERENCES, AND COOPERATIVE SIGNALS IN SOCIAL DILEMMAS

Revised Version June 2019

Stephan Müller
Georg von Wangenheim

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

ISSN: 1439-2305

Coevolution of Cooperation, Preferences, and Cooperative Signals in Social Dilemmas

Stephan Müller* Georg von Wangenheim†

February 2019

Abstract

We study the coevolution of cooperation, preferences, and cooperative signals in an environment where individuals engage in a signaling-extended prisoner's dilemma. We prove the existence of a cooperative equilibrium constituted by a (set of) limit cycle(s) and stabilized by the dynamic interaction of multiple Bayesian equilibria. This equilibrium: (1) exists under mild conditions, and (2) can stabilize a population that is characterized by the heterogeneity of behavior, preferences, and signaling. We thereby offer an explanation for the persistent regularities observed in laboratory and field data on cooperative behavior. The cyclicity of the equilibrium offers an alternative account for observed historical changes in (social norms of) cooperation in societies which are not driven by social or environmental shocks.

JEL Classification numbers: C73, D64, D82.

Keywords: Evolutionary Game Theory, Cooperation, Preferences, Signaling.

*Corresponding author, University of Göttingen, Platz der Göttinger Sieben 3, 37073 Göttingen (Germany), *E-mail: stephan.mueller@wiwi.uni-goettingen.de*

†Kassel University, Nora-Platiel-Straße 4, 34109 Kassel (Germany), *E-mail: g.wangenheim@uni-kassel.de*

1 Introduction

Several theories have been proposed to explain the evolution of cooperation among humans when cooperation generates a public benefit at a private cost. In this research, the prisoner's dilemma game (henceforth PD) commonly serves as a metaphor for the problem of cooperation. Since natural selection favors defection in this game, any extension that allows for the emergence of cooperation represents a mechanism to promote cooperation. It has been argued that the essential feature of any mechanism to foster cooperation is that cooperative acts must occur more often between cooperators than expected, based on population averages. In other words, the mechanism must induce a positive assortment between cooperative types (Queller, 1985; Fletcher and Zwick, 2004).[1]

The mechanisms proffered in the literature may vary substantially in how they induce this assortment. Positive assortment can, for instance, arise because of direct reciprocity in repeated interactions (Trivers, 1971; Axelrod, 1984; Fudenberg and Maskin, 1986), indirect reciprocity based on image scores (Alexander, 1987; Nowak and Sigmund, 1998; Wedekind and Milinski, 2000; Panchanathan and Boyd, 2004), or network reciprocity where players interact only with their neighbors (Nowak and May, 1992; Hubermann and Glance, 1993; Nowak et al., 1994; Killingback et al., 1999).

In solving the puzzle of cooperation in social dilemmas the literature so far has primarily focused on providing mechanisms that support the existence of a cooperative equilibrium. We extend this literature by providing an explanation for the following conspicuous regularities of this puzzle. First, there is a persistent pattern showing that cooperation is only partial, i.e., only a fraction of the population plays cooperatively when individual rationality calls for defective behavior.[2] Second, the elicitation of preferences in the laboratory and in the field, as well as studies on revealed preferences, show that individuals substantially differ in their cooperative attitudes (e.g., Andreoni and Miller, 1993; Cooper et al., 1996; Ockenfels and Weimann, 1999, Fischbacher and Gächter, 2010). Thus, the heterogeneity in behavior does not seem to result from mixed strategy play, but appears to be a consequence of differences in preferences. Third, it is the rule rather than the exception that human interactions are accompanied by communication, particularly if the interaction is of a strategic nature. Humans also differ in this respect and show different

[1] Indeed, many models of the evolution of altruism share an underlying mathematical structure – that of Hamilton's Price equation formulation of inclusive fitness theory (Hamilton, 1964a,b). Hamilton's relatedness coefficient can be interpreted as the degree of positive assortment of types and need make no reference to common descent (McElreath and Boyd, 2007).

[2] See Rapaport and Chammah, 1965, and Dawes, 1980, for reviews of these experiments in sociology and psychology. For a survey of some of the studies by economists, see Roth, 1988.

ways and intensity of preplay-communication in laboratory and field studies. Importantly, it has been shown that communication influences cooperative behavior (Dawes et al., 1977; Ostrom and Walker, 1991; Brosig, 2002).[3] In this paper we study the coevolution of these three behavioral dimensions and offer an explanation for the presence of the aforementioned regularities.

All three phenomena take place at a population level, we therefore take an evolutionary perspective to study these related dimensions of heterogeneity. As a stylized social dilemma, the action set of the PD incorporates the two diametrically opposed behaviors of defection and cooperation. To account for the potential heterogeneity of preferences in equilibrium, we consider an evolutionary model with two types of individuals: 'opportunists,' who maximize individual fitness, and 'conditional cooperators,' who have a preference for joint cooperation.[4] To emphasize the necessity to communicate about preferences, we study the evolution of cooperation in social dilemmas in one-shot interactions without social information, such as reputation, which puts other mechanisms like direct or indirect reciprocity out of operation. Any mode of communication hardly comes without any cost, be it material cost because of effort exerted, resources spent or forgone opportunities. On the other hand, compliance to some code of conduct as a signal of cooperativeness may cause internal costs if it contradicts an individual's preferences. To account for these aspects, we incorporate the communication of types via costly signaling.[5] Taken together, we study a population game with a PD preceded by a round of communication as a stage game. In our evolutionary approach we analyze whether the potential heterogeneity in behavior, preferences and communication can indeed be present in an evolutionary stable equilibrium. If this turns out to be the case this would provide a rationale for the observed regularities.

The standard criticism of many preference evolution models is that the evolutionary stability of preferences which do not implement a Nash equilibrium in the fitness game hinges on the assumption of (partial) observability of preferences (Dekel, Ely, and Yilankaya, 2007). In our signaling framework preferences are not assumed to be observable but, of course, might be revealed in (partially) separating equilibria. Importantly, we do not assume that either type has a cost advantage in signaling cooperativeness neither

[3]It is a stylized fact in experimental research that the opportunity of communication has a robust and strong positive impact on cooperation, for an overview see Sally, 1995.

[4]There is evidence from laboratory and field experiments that the majority of individuals can be assigned to one of these two classes: Keser and van Winden, 2000; Fischbacher et al., 2001; Frey and Meier, 2004; Fischbacher and Gächter, 2010.

[5]Costly signaling is present in many species, including humans (Zahavi, 1977; Grafen, 1990; Maynard Smith, 1991; Johnstone, 1995; Wright, 1999).

in fitness nor in utility terms. We depart from standard applications of the 'indirect' evolutionary approach pioneered by Güth and Yaari (1992) in one important manner. Instead of applying the static notion of evolutionary stable strategies (Maynard Smith and Price, 1973), we explicitly study the dynamic stability of the Bayesian equilibria of the signaling-extended PD. Importantly, considering (locally) the full set of Bayesian equilibria and their dynamic stability puts us in the position to study the transition across different Bayesian equilibria.

As our main result we prove the existences of a cooperative equilibrium which is stabilized by the dynamic interplay of separating, semi-pooling, and pooling equilibria of the signaling-extended PD. This equilibrium is constituted by a (set of) limit cycle(s) and is characterized by full heterogeneity with respect to behavior, preferences, and signaling. In this equilibrium the share of cooperators and the share of signalers oscillates. In our model such an equilibrium exists under mild conditions on signaling cost and other model parameters. In particular, this equilibrium is the only cooperative equilibrium which exists if signaling is costless in terms of fitness. Quite surprisingly this equilibrium is also consistent with cooperators bearing higher signaling than opportunists both in terms of fitness and utility. Thus, contrary to previous results in the literature, sustaining cooperation does not hinge on the assumption of some material cost advantage for cooperative types. Furthermore, the oscillatory property of the (set of) limit cycle(s) offers interesting insights into the economics of social change in (norms of) cooperation.

The remainder of the paper is organized as follows. The following section discusses the related theoretical literature in more detail. Our model is presented in section 3. Section 4 presents the set of stable perfect Bayesian equilibria (PBE) for a given composition of preferences. This share of conditional cooperators is endogenized in section 5. Before we conclude in section 7, we discuss our findings in the penultimate section 6.

2 Related Theoretical Literature

In this section we focus on literature which considers the problem of cooperation in social dilemmas under incomplete information regarding the opponent's type. Most closely related to our approach are the papers of Guttman (2003, 2013), Gintis et al. (2001), and Panchanathan and Boyd (2004). Guttman (2003) is motivated by the seminal paper of Kreps et al. (1982). Therein the authors show that if one of two players assigns a small probability that the opponent will play the 'tit-for-tat' strategy then, in a finitely repeated prisoner's dilemma (PD), cooperation can be an equilibrium outcome for at least some of

the stages. In an 'indirect' evolutionary framework Guttman endogenizes the uncertainty assumed by Kreps et al. (1982) regarding the opponent's preferences. More precisely, the model considers a community consisting of 'opportunists' and 'reciprocators' who have a preference for mutual cooperation. Furthermore, agents send a costless, random signal that has some informational value for the receiver with respect to the recognition of the opponent's type. Players are randomly matched to play a finitely-repeated PD. In the unique evolutionary equilibrium, both reciprocators and opportunists coexist.[6] Although the evolutionary equilibrium is characterized by a heteromorphic population, the equilibrium behavior of reciprocators and opportunists differs only in the last round, i.e., both types show almost identical behavior in equilibrium. Furthermore, if the likelihood of emitting the cooperative signal is independent of the taste parameter measuring the preference for mutual cooperation then the endogenization of the taste parameter leads to an all-reciprocator equilibrium, and thereby to full cooperation. Thus, the model is less suitable for explaining the regularities of heterogeneous preferences and behavior, particularly in environments with very few repetitions.

Guttman (2013) studies the evolution of an inherited preference to match other agents' contribution to the provision of public goods. Under complete information and randomly matched groups, the unique evolutionary stable matching rate equals one. The model provides a potential explanation for the existence of conditional cooperation, which does not rely on reputation or group selection. However, as the informational assumptions are rather strict, we circumvent this by considering a signaling environment where types are only revealed by equilibrium play. Furthermore, the model predicts a unique preference value and therefore cannot account for the heterogeneity in preferences and behavior, which is the focus of our paper.

Contrary to Guttman (2003), Gintis et al. (2001) consider an environment with no repeated or assortative matching. Furthermore, the signaling in their model is costly. In a multi-player public good game individuals can signal their type by providing a group benefit at a personal cost. These signals may in turn influence a partner's acceptance or rejection of potentially profitable allies. They show that an honest signaling of underlying quality can be evolutionary stable. Necessary conditions for the existence are that signaling is more costly to so-called high-quality types and that partners prefer to ally with high-quality types. They show that the payoff difference between high and low types is positive. As a consequence, the frequency of high types would increase over time.

[6]The survival of reciprocators hinges on the assumption that the costless signal emitted by all subjects has some small but positive correlation with the actual type.

This eventually undermines the separating equilibrium, since it has only limited support.[7] More precisely, once the share of high types exceeds a certain threshold, high types no longer find it a best response to signal their quality. As a consequence, cooperation could break down. Without an exhaustive search for Nash equilibria and an analysis of their dynamic stability – which is part of our approach – we just do not know. In the model of Gintis et al. (2001), the monotonic increase in the share of high-quality types is stabilized by the *ad hoc* introduction (see p.112, eq.(12)) of other forces on the population dynamics. Indeed, without introducing the exogenous frequency dependency no heteromorphic population could be stabilized.

The general theme of this strand of literature (see also Lotem et al., 2003; Panchanathan and Boyd, 2004; Macfarlan et al., 2013) is that costly forms of generosity (like contributions to public goods) serves as a signal of trustworthiness, facilitating the formation of cooperative partnerships in the future. A general problem with the signaling hypothesis is that it does not explain why quality is signaled by doing good (noted by Gintis et al., 2001, themselves).[8] Indeed, quality could be signaled by other costly activity like conspicuous consumption.[9] In contrast, in our approach the nature of the signaling technology is not limited to forms of generosity. Moreover, to the best of our knowledge there is no paper which offers an explanation of heterogeneity in preferences, behavior, and communication and does not hinge on some *ad hoc* advantage for non-opportunists.

Another paper related to our approach is Janssen (2008) which studies the evolution of cooperation in a one-shot PD environment based on the recognition of the opponent's trustworthiness. Agents costlessly display symbols and they learn which symbols are important to estimate an opponent's trustworthiness. The simulation-based results show both cooperative and defective behavior. In contrast, the evolution of agents' taste parameters shows the tendency toward homogeneity, since almost all agents in the long run value cooperation over defection (see the statistics of parameter α in Table 4). Since the result hinges on the assumption that agents can withdraw from playing the game, it does not apply to our idea of random interaction in an unstructured population that cannot be circumvented.

[7] That is, the range of the share of high-quality types, such that the conditions for the existence of the honest signaling equilibrium are met, is an open interval with a measure of less than one.

[8] Hopkins (2014) provides a potential solution to the problem. If the ability to reason about others' mental state, i.e., having a "theory of mind," is associated with empathy, then humans that possess these attributes can signal their capability by pro-social acts.

[9] After all, the prominent example for costly signaling in the context of sexual selection is the peacock's tail.

3 Model

In this section we will first describe the primitives of the Bayesian game played by the randomly paired individuals. We thereafter present the evolutionary framework in which we will study the induced population dynamics.

3.1 Underlying Bayesian Game

The game played by the paired individuals consists of two stages. In the first stage, the communication stage, the two subjects simultaneously exchange messages. In the second stage the two individuals play a one-shot prisoner's dilemma. We describe the details of the second stage first.

The classical prisoner's dilemma is the most prominent and best-studied example of a social dilemma. In this game players can either cooperate (C) or defect (D), i.e., the set of actions available to player i is $A_i = \{C, D\}$. The resulting payoffs are given in the left table of Figure 1. These payoffs reflect the material payoffs or fitness which are decisive for the evolutionary success of different strategies. We consider two types of preferences. These preferences represent a subject's evaluation of these outcomes in terms of utility. Opportunists are assumed to show preferences which are in line with material payoffs, i.e., defection is the dominant strategy for those subjects. Without loss of generality we assume that for those subjects utility and material payoffs coincide. Conditional cooperators, however, have preferences over material outcomes which make cooperation a best reply to an opponents' cooperative behavior. In other words, such preferences make mutual cooperation a Nash equilibrium. Note that any such preference if represented by a utility function $u_i : A_i \times A_{-i} \to \mathbb{R}$ has the property that $u_i(C, C) > u_i(D, C)$. If we neglect the case where a cooperative preference makes cooperation the dominant strategy of the game, then such preferences can be represented by a single parameter m which adds sufficient utility to the mutual cooperation outcome. To be of any behavioral significance, we assume $m > \alpha$, such that mutual cooperation becomes a Nash equilibrium if two cooperators interact.[10] The resulting payoffs for this type of subjects are given in the right table of Figure 1. Given this parameterization of individuals' cooperative preferences, we will refer to opportunists as *low types* and to conditional cooperators as *high types*.

Given the two types of individuals, the type space of player i is $\Theta_i = \{H, L\}$, H for high types, and L for low types. We therefore set as the state space of the Bayesian

[10] 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.

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

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

Figure 1: Material (left table) and utility (right table) payoff in the PD, $\alpha, \beta > 0$ and $1 + \beta > \alpha$.

game $\Omega = \Theta_1 \times \Theta_2$. The preference for joint cooperation is assumed to be the private information of the agent. Thus, the signaling function $\tau_i : \Omega \to \Theta_i$ of player i is given by $\tau_i(\omega) = \omega_i$, i.e., the individual signal contains information about the own type, but no information about the opponent's type. In the tradition of Harsanyi (1967, 1968a, 1968b), beliefs about the opponent's type are common knowledge. Like Guttman (2003) and Güth and Ockenfels (2005), we adopt the natural assumption that beliefs correspond to actual frequencies of types in the population.

In the first stage, after receiving their signals, the two players can simultaneously exchange messages about their preferences.[11] As in the standard signaling model (Spence, 1973) we assume the existence of a social technology which enables individuals to signal their positive attitude toward cooperation by incurring some costs. Indeed, note that in no evolutionary stable equilibrium would individuals bear a cost to actually signal to be an opportunist. Research on many species including humans (Zahavi, 1977; Grafen, 1990; Maynard Smith, 1991; Johnstone, 1995; Wright 1999) supports the assumption of the existence of such a signaling technology. Thus, without loss of generality, we assume that the message space of player i is given by $M_i = \{m, 0\}$, where the costless message 0 corresponds to not sending a message. The message m corresponds to the costly signal to be a high type. Sending a message can be costly in terms of material or utility payoff. Let k_H, k_L denote the utility signaling cost for high types and low types, respectively. To distinguish utility and material signaling cost we denote material signaling cost by k_H^f and k_L^f, respectively. Importantly, *a priori* we impose no restriction on the relation of signaling cost in terms of utility or fitness across types. In particular, we allow for type-independent signaling cost, i.e., $k_H = k_L$ and $k_H^f = k_L^f$. Higher lying cost for opportunists, for example, would suggest that $k_H \leq k_L$. There is evidence that lying cost are large and widespread (Abeler et al. 2015). The relation of material singling cost across types which could reflect opportunity cost of a time-consuming signaling stage is more ambiguous. We refer the reader to our discussion of the relation and the nature of signaling costs in

[11] Note that without communication, the impossibility result of Kandori (1992, Proposition 3) applies to such an environment which states that the unique equilibrium is characterized by full defection, i.e., everybody always defects.

section 6.

For a given type θ_i a strategy $\mathfrak{s}_i$ of player i specifies an action depending on the message received and a message to be sent, i.e., $\mathfrak{S}_i = \{f : M_{-i} \to A_i\} \times M_i$. Utility payoffs are given by $u_i(\mathfrak{s}_i, \mathfrak{s}_{-i}, \theta_i) = u_i(a_i, a_{-i}, \theta_i) - \mathbf{1}_m \cdot k_{\theta_i}$, where $a_{-i} = \mathfrak{s}_{-i}(m_i), \forall i$ and $\mathbf{1}_m$ equals one if $m_i = m$ and zero otherwise. Material payoffs are given by $u_i^f(\mathfrak{s}_i, \mathfrak{s}_{-i}, \theta_i) = u_i^f(a_i, a_{-i}, \theta_i) - \mathbf{1}_m \cdot k_{\theta_i}^f$. Payoffs $u_i(a_i, a_{-i}, \theta_i)$ and $u_i^f(a_i, a_{-i}, \theta_i)$ are given in Figure 1. Taken together, this constitutes the Bayesian game $\Gamma = \langle N, (\mathfrak{S}_i), (u_i), \Omega, (\Theta_i), (\tau_i), (p_i) \rangle$, with prior beliefs p_i on Ω. The adoption of the assumption that the prior beliefs correspond to actual frequencies in the population gives us $p_i(\tau_i^{-1}(H)) = \lambda$ and $p_i(\tau_i^{-1}(L)) = 1 - \lambda$, where λ denotes the share of high types in the population. To distinguish the type-specific strategies of a player we denote the set of strategies for high (low) types by $\mathfrak{S}_H$ ($\mathfrak{S}_L$).

The (pure strategy) Bayesian Nash equilibria of Γ correspond to the Nash equilibria of the strategic game $\Gamma_N = \langle N, (\mathscr{S}_i), (\tilde{u}_i) \rangle$, where $\tilde{u}_i = E_\omega[u_i]$ denotes the expected utility and $\mathscr{S}_i$ is the set of all pairs of mappings (ψ_i, ϕ_i), where $\psi_i : \Theta_i \times M_{-i} \to A_i$ and $\phi_i : \Theta_i \to M_i$. That is, $\mathscr{S}_i$ is the set of plans of actions contingent on the own signal $\tau_i(\omega)$ and on the signal $m_{-i} \in M_{-i}$ received from the opponent, represented by ψ_i. Moreover, by ϕ_i it specifies which signal $m_i \in M_i$ to be sent contingent on the own type. Note that for any player we can identify $\mathscr{S}$ by $\mathfrak{S}_H \times \mathfrak{S}_L$, where the first entry specifies the type-specific strategy for a high type and the second entry the type-specific strategy for a low type. Let $K = |\mathscr{S}| = |\mathfrak{S}_H||\mathfrak{S}_L|$ denote the number of pure strategies.

3.2 Population Dynamics

We now turn to the description of our evolutionary framework. We assume that the symmetric signaling-extended one-shot PD Γ_N is played recurrently in a large but finite population by randomly matched pairs of individuals. Agents are assumed to only process information on the outcomes of their own past interactions. In particular, they do not process any information on the opponent's identity or on outcomes in games in which they were not involved. By these assumptions we refrain from imposing additional structure on the interactions which may drive the emergence of cooperation and the properties of cooperative equilibria. This allows us to study the interdependencies between preferences, behavior, and communication in isolation.

Given our distinction between material payoffs and utility, we employ the indirect evolutionary approach, pioneered by Güth and Yaari (1992),[12] in which all players are as-

[12]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

sumed to be rational, and the evolutionary forces determine the population's composition of players with different preferences. In other words, preferences determine behavior and behavior in turn determines fitness. Recent criticism of this approach (Dekel et al., 2007) is concerned with the assumption of the observability of agents' preferences. However, in our model preferences are not observable, we only assume that agents have correct beliefs about the distribution of types in the population.

In our model, on the one hand evolutionary forces shape the composition of preferences in the population, i.e., the share of high types λ. On the other hand, selection also shapes the composition of strategies in the population, which individuals apply in the signaling-extended PD. We make the natural assumption that the inherited behavior evolves independently across types. This assumption excludes biases like the case that a certain strategy as a high type it is more likely to be played for some low type strategies than for others. We will refer to $(\mathbf{p}_H, \mathbf{p}_L)$ as the *population state*, where $\mathbf{p}_\theta$ denotes the vector of shares in the population which play strategy $\mathfrak{s}_\theta \in \mathfrak{S}_\theta$. In modeling the dynamics of the population state and λ we build on the general selection dynamics defined on $\Delta_{|\mathfrak{S}_H|} \times \Delta_{|\mathfrak{S}_L|} \times [0,1]$ in terms of growth-rates. That is, $\dot{p}_{\mathfrak{s}_\theta} = g_{\mathfrak{s}_\theta}(\mathbf{p}_H, \mathbf{p}_L, \lambda) \cdot p_{\mathfrak{s}_\theta},\ \mathfrak{s}_\theta \in \mathfrak{S}_\theta$ and $\dot{\lambda} = h(\mathbf{p}_H, \mathbf{p}_L, \lambda) \cdot \lambda \cdot (1-\lambda)$[13], where the functions $g_{\mathfrak{s}_\theta}, h : X_1 \times X_2 \times X_3 \to \mathbb{R}$ with open domains X_1, X_2, and X_3 containing $\Delta_{|\mathfrak{S}_H|}$, $\Delta_{|\mathfrak{S}_L|}$, and $[0,1]$ specify the respective growth rate per time unit. Taken together, this gives rise to the following coupled system of differential equations:

$$\dot{p}_{\mathfrak{s}_\theta} = g_{\mathfrak{s}_\theta}(\mathbf{p}_H, \mathbf{p}_L, \lambda) \cdot p_{\mathfrak{s}_\theta},\ \mathfrak{s}_\theta \in \mathfrak{S}_\theta,\ \theta \in \{H, L\} \tag{1}$$

$$\dot{\lambda} = h(\mathbf{p}_H, \mathbf{p}_L, \lambda) \cdot \lambda \cdot (1-\lambda) \tag{2}$$

Given the non-negativity of a population state and the share λ, in this context the literature on dynamic systems commonly imposes some regularity conditions which guarantee that starting from an interior point the system remains in $\Delta_{|\mathfrak{S}_H|} \times \Delta_{|\mathfrak{S}_L|} \times [0,1]$. The following assumptions are a weaker notion of Samuelson and Zhang's (1992) regularity condition. We make the assumptions $\sum_{\mathfrak{s}_\theta \in \mathfrak{S}_\theta} \dot{p}_{\mathfrak{s}_\theta} = 0,\ \forall (\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \Delta_{|\mathfrak{S}_H|} \times \Delta_{|\mathfrak{S}_L|} \times [0,1],\ \theta \in \{H, L\}$ and $\dot{\lambda} + (1 - \lambda) = 0$. Moreover, we assume that $p_{\mathfrak{s}_\theta} = 0$ implies $\dot{p}_{\mathfrak{s}_\theta} \geq 0$. Analogously, $\lambda = 0$

and Güth, 1998), of altruistic and spiteful preferences (Possajennikov, 2000) or of risk preferences (Wärneryd, 2002).

[13] For the sake of a more convenient representation we refrain from modeling the shares of low types and high types separately since one is the residual of the other as both add up to one. As a technical consequence the additional factor $(1-\lambda)$ arises if we want to place the condition of Lipschitz continuity on the function $h(\mathbf{p}_H, \mathbf{p}_L, \lambda)$. Lipschitz continuity of $h(\mathbf{p}_H, \mathbf{p}_L, \lambda)$ then implies the Lipschitz continuity for the relevant growth rate function for high and low types, i.e., $h(\mathbf{p}_H, \mathbf{p}_L, \lambda) \cdot (1-\lambda)$ and $h(\mathbf{p}_H, \mathbf{p}_L, \lambda) \cdot \lambda$.

implies $\dot{\lambda} \geq 0$ and $\lambda = 1$ implies $\dot{\lambda} \leq 0$. We also make the standard assumption of Lipschitz continuity of $\dot{p}_{\mathfrak{s}_\theta}$ and $\dot{\lambda}$ which guarantees the existence and uniqueness of a solution for the dynamic system (1)–(2) by the Picard-Lindelöf theorem. Let $\xi : \mathbb{R} \times C \to C$ denote the induced solution mapping of the system (1)–(2), where $C = \Delta_{|\mathfrak{S}_H|} \times \Delta_{|\mathfrak{S}_L|} \times [0,1]$. We are ultimately interested in stable sets of the dynamic system $(\mathbb{R}, C, \xi)$. As a classical notion of stability we will apply the concept of asymptotic stability. Intuitively, this concept requires that any small perturbations of the set induce a movement back to the set. Formally, a closed set $A \subset C$ is *asymptotically stable* if it is Lyapunov stable and if there exists a neighborhood B^* of A such that $\xi(t, x^0) \xrightarrow{t \to \infty} A$ for all $x^0 \in B^* \cup C$.[14]

So far no relation between payoffs and changes in the population state or in the share of high types has been imposed. We make the common assumption that the dynamics of the population state and the share of high types satisfy payoff-monotonicity. Payoff-monotonicity captures the idea that a pure strategy with a higher payoff grows at a higher rate. Importantly, in the indirect evolutionary approach payoff-monotonicity for $g_{\mathfrak{s}_\theta}$ refers to utility payoffs, as behavior in the underlying game Γ_N is driven by the evaluation of material payoffs according to subjects' preferences. In contrast, payoff-monotonicity of h refers to material payoffs as the ultimate survival of a preference is determined by the induced fitness of the subject carrying that preference. When material payoffs differ across individuals of the same type we assume for simplicity that payoff-monotonicity refers to the average material payoffs within each type. Formally, $g_{\mathfrak{s}_\theta}(\mathbf{p}_H, \mathbf{p}_L, \lambda) > g_{\mathfrak{s}'_\theta}(\mathbf{p}_H, \mathbf{p}_L, \lambda)$ if and only if $\Pi_\theta(\mathfrak{s}_\theta) > \Pi_\theta(\mathfrak{s}'_\theta)$, where $\Pi_\theta(\mathfrak{s}_\theta) = \sum_{\mathfrak{t}_\theta \in \mathfrak{S}_\theta} p_{\mathfrak{t}_\theta} u(\mathfrak{s}_\theta, \mathfrak{t}_\theta, \theta)$ denotes the expected payoff for the type-specific strategy $\mathfrak{s}_\theta$. Analogously, $\dot{\lambda} > 0$ if and only if $\Pi^f_H > \Pi^f_L$, where $\Pi^f_\theta = \sum_{\mathfrak{s}_\theta \in \mathfrak{S}_\theta} p_{\mathfrak{s}_\theta} \Pi^f_\theta(\mathfrak{s}_\theta)$ with $\Pi^f_\theta(\mathfrak{s}_\theta) = \sum_{\mathfrak{t}_\theta \in \mathfrak{S}_\theta} p_{\mathfrak{t}_\theta} u^f(\mathfrak{s}_\theta, \mathfrak{t}_\theta, \theta)$.

Note that for payoff-monotonicity to also be satisfied at the boundaries of $\Delta_{K_H} \times \Delta_{K_L} \times [0,1]$, Samuelson and Zhang (1992) assume that the limits $\lim\limits_{p_{\mathfrak{s}_\theta} \to 0} \dfrac{\dot{p}_{\mathfrak{s}_\theta}}{p_{\mathfrak{s}_\theta}}$ exist and are finite. In our setup this corresponds to the upper bounds for the growth-rate functions $g_{\mathfrak{s}_\theta}$. However, this is sufficient but is not necessary for payoff-monotonicity. Note further that the assumption of bounded growth-rate functions implies that $\lim_{p_{\mathfrak{s}_\theta} \downarrow 0} \dot{p}_{\mathfrak{s}_\theta} = 0$ irrespective of the payoff associated with this strategy. We make the assumption of bounded growth-rate functions only for those strategies with inferior payoffs, allowing the growth-rate functions for a strategy $\mathfrak{s}_\theta$ with maximal payoff to be unbounded, which is not in contradiction to the Lipschitz continuity and payoff-monotonicity of the dynamics $\dot{p}_{\mathfrak{s}_\theta}$. Given humans'

[14] $\xi(t, x^0) \xrightarrow{t \to \infty} A$ is defined as the distance $d(\xi(t, x^0), A)$ converging to zero as $t \to \infty$. A closed set $A \subset C$ is *Lyapunov stable* if every neighborhood $B \subset A$ contains a neighborhood B^0 of A such that $\gamma^+(B^0 \cap C) \subset B$, where $\gamma^+(Z)$ is defined as the union of all semi-orbits $\gamma^+(z)$ with $z \in Z$.

cognitive abilities to consider counterfactuals and their potential for seizing profitable opportunities via behavioral innovations we think that the following assumption is more appropriate. If at $p_{\mathfrak{s}_\theta} = 0$, $\Pi_\theta(\mathfrak{s}_\theta) > \Pi_\theta(\mathfrak{t}_\theta)$, $\forall \mathfrak{t}_\theta \neq \mathfrak{s}_\theta \in \mathfrak{S}_\theta$ then $\lim_{p_{\mathfrak{s}_\theta} \downarrow 0} \dot{p}_{\mathfrak{s}_\theta} \in (0, \infty)$. In some sense this assumption reflects the common economic wisdom that there are no \$10 bills lying on the street. We will refer to this assumption as the *social innovation assumption*. Note at this point that this assumption is sufficient but not necessary to prove our main results in the realm of the very general class of regular payoff-monotone selection dynamics. We will elaborate on this in section 6.

Finally, we assume that the dynamics of the population state $(\mathbf{p}_H, \mathbf{p}_L)$ is much faster than the adjustment of the composition of preferences λ in the population. This is a common assumption in applications of the indirect evolutionary approach as the adjustment of behavior is presumably faster than the dynamics of underlying preferences. This assumption allows us to make use of the adiabatic elimination technique (see e.g., Haken, 1977). Under this assumption system (1) is said to be slaved by system (2). However, the slaved system reacts to system (2). Since we are looking for stable points or sets of the system (1)–(2), the adiabatic technique allows us to solve (2) approximately by putting $\dot{p}_{\mathfrak{s}_\theta} = 0, \forall \mathfrak{s}_\theta \in \mathfrak{S}_\theta,\ \theta \in \{H, L\}$. In other words, we study the dynamics of the type composition $\dot{\lambda}$ under the assumption that the fast dynamics of the composition of strategies $(\dot{\mathbf{p}}_L, \dot{\mathbf{p}}_H)$ has reached a stationary point. Intuitively, this assumption guarantees that the dynamic system $(\mathbb{R}, C, \xi)$ will mostly be in the region of attraction of one of the stable (sets of) population states. Formally, let $P(\lambda)$ denote the union of all stable (sets of) population states with $\dot{p}_{\mathfrak{s}_\theta} = 0,\ \forall \mathfrak{s}_\theta \in \mathfrak{S}_\theta,\ \theta \in \{H, L\}$ at any given λ and let $U_\iota \subset X_1 \times X_2 \times X_3$ denote an open ι-neighborhood of $\cup_\lambda P(\lambda) \times [0, 1]$, with $\iota > 0$. We will assume the following:

$$\frac{\|(\dot{p}_H, \dot{p}_L)\|}{\|\dot{\lambda}\|} > \sqrt{2} \cdot \sigma \gg 1,\ \forall (\mathbf{p}_H, \mathbf{p}_L, \lambda) \in X_1 \times X_2 \times X_3 \setminus U_\iota, \tag{3}$$

where σ parameterizes the relative speed of the two dynamic processes. As a next step we first derive the stable (sets of) populations states for any fixed composition of types $\tilde{\lambda}$ in section 4. That is, we study the system (1) for $\lambda \equiv \tilde{\lambda}$. In section 5 we study the induced dynamics of λ.

4 Stable Bayesian Nash Equilibria with Exogenous Proportion of High Types

In this section we present the stable (sets of) population states for each $\lambda \in (0,1)$. Note that these (sets of) states correspond to the symmetric Nash equilibria of the mixed extension of Γ_N. In the game Γ_N, a strategy is a type-continent and signal-contingent plan. Given the two actions C and D, the two types and the two signals, there are 64 pure strategies. Since defection is the dominant strategy for low types, we can eliminate all strategies which specify cooperative behavior for any signal received for the contingency of being a low type. Thus, $\mathfrak{S}_H = \{CCs, CDs, DCs, DDs, CCns, CDns, DCns, DDns\}$ and $\mathfrak{S}_L$ reduces to $\tilde{\mathfrak{S}}_L = \{s, ns\}$, where the first entry in, for instance, CDs specifies the action if the cooperative signal is received, the second entry refers to the action in the event of no signal, and finally s or ns indicate whether the signal is sent or not. CDs, s will denote the corresponding type-contingent strategy. Hence, there are $\tilde{K} = 16$ type-contingent strategies remaining.

It turns out that in our signaling-extended PD Γ_N, there exists one stable separating and three stable pooling equilibria. There are also stable semi-pooling equilibria, however, only one of them is relevant for our subsequent analysis. We will introduce this equilibrium in the next chapter when we endogenize the proportion of high types.[15] The following Proposition 1 reports the stable signaling equilibrium and the stable pooling equilibria.

Proposition 1 *In the signaling extended prisoner's dilemma there exist the following stable Perfect Bayesian Equilibria:*

(i) Cooperative Separating Equilibrium (CSE): $p_{CDs,ns} = 1$

(ii) Cooperative High-Pooling Equilibrium (CHPE): $p_{CDs,s} = 1 - p_{CCs,s} \geq \frac{k_L}{(1+\alpha)\lambda}$

(iii) Cooperative Low-Pooling Equilibrium (CLPE): $p_{CCns,ns} + p_{DCns,ns} = 1$

(iv) Defective Low-Pooling Equilibrium (DLPE): $p_{DDns,ns} = 1 - p_{CDs,ns} \leq \frac{1}{\lambda} \min\left\{\frac{k_H+\beta}{1+m+\beta}, \frac{k_H}{1+\alpha}\right\}$

Table 1 presents the conditions for the existence and the λ-support[16] of these equilibria.

Proof. We leave the derivation and the analysis of stability to Appendix B. □

[15] For a derivation of all semi-pooling equilibria assuming $k_H \leq k_L$ we refer the reader to Appendix D of our working paper version available under `http://wwwuser.gwdg.de/~cege/Diskussionspapiere/DP221_Appendix_D.pdf`.

[16] Here, the λ-support of an equilibrium corresponds to the set of all λ such that the equilibrium under consideration exists.

Strategies	λ-support Condition for Existence	Differences in utility payoffs Differences in average material payoffs
Cooperative Separating Equilibrium		
CDs, ns	$\frac{k_H}{1+m} \leq \lambda \leq \frac{k_L}{1+\alpha}$ $k_H < 1 + m$	$\Delta\Pi(CDs, ns) = \lambda(1 + m) - k_H$ $\Delta\Pi^f(CDs, ns) = \lambda - k_H^f$
Cooperative High-Pooling Equilibrium		
CCs, s CDs, s	$\lambda \geq \max\{\frac{k_L}{1+\alpha}, \frac{\beta}{\beta+m-\alpha}\}$ $k_L < 1 + \alpha$	$\Delta\Pi(CCs, s) = \Delta\Pi(CDs, s)$ $= k_L - k_H - (\lambda(\alpha - m) + (1 - \lambda)\beta)$ $\Delta\Pi^f(CCs, s) = \Delta\Pi^f(CDs, s)$ $= k_L^f - k_H^f - (\lambda\alpha + (1 - \lambda)\beta)$
Cooperative Low-Pooling Equilibrium		
$CCns, ns$ $DCns, ns$	$\lambda \geq \frac{\beta}{\beta+m-\alpha}$	$\Delta\Pi(CCns, ns) = \Delta\Pi(DCns, ns)$ $= -(\lambda(\alpha - m) + (1 - \lambda)\beta)$ $\Delta\Pi^f(CCns, ns) = \Delta\Pi^f(DCns, ns)$ $= -(\lambda\alpha + (1 - \lambda)\beta) < 0$
Defective Low-Pooling Equilibrium		
$CDns, ns$ $DDns, ns$	$0 < \lambda < 1$	$\Delta\Pi(CDns, ns) = \Delta\Pi(DDns, ns) = 0$ $\Delta\Pi^f(CDns, ns) = \Delta\Pi^f(DDns, ns) = 0$

Table 1: Separating and pooling equilibria. Note that the equilibria are only stable in the interior of their support. $\Delta\Pi = \Pi_H - \Pi_L$ and superscript f indicates fitness payoffs.

In the *cooperative separating equilibrium*, players apply the strategy CDs, ns. Thus, 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 bound is as follows: If there are too few high types then the cooperative outcome among them cannot compensate for the signaling costs. If, on the other hand, there are too many high types, signaling becomes sufficiently profitable for low types. In the *cooperative low-pooling equilibrium*, nobody signals and high types cooperate. This equilibrium exists if there are a sufficient number of high types. Only then can high types be compensated for the loss from playing cooperatively against low types by the cooperative outcome among each other. In the *defective low-pooling equilibrium*, nobody sends the cooperative signal and everybody defects and earns a payoff of zero. Again, because of the lack of distinguishability in equilibrium, this equilibrium is indeed a set where $CDns, ns$ and

$DDns, ns$ might be played. This equilibrium set reflects the benchmark solution in the PD without communication and exists for all population compositions between high types and low types. In the *cooperative high-pooling equilibrium*, everybody signals and high types cooperate. This equilibrium exists if there are a sufficient number of 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.

5 Endogenous Proportion of High Types

We now analyze the dynamics of the share of high types (λ) in the population for which we assume that the dynamics have reached a stable equilibrium, as we assumed that inner motives evolve far more slowly than behavioral frequencies. For the sake of a more convenient presentation, we will assume that $k_L, k_H < 1 + \alpha$ which is sufficient for the existences of all equilibria presented in Table 1. In other words, we focus on the more meaningful case of signaling devices which are less costly than the maximum material gain from sending the cooperative signal, i.e., $u_i^f(D, C, \theta) - u_i^f(D, D, \theta)$.

The evolution of the proportion of conditional cooperators is determined by their relative fitness. Fitness is measured by the material payoffs as presented in Figure 1. Analogous to the derivation of the PBE, the differentials in these fitness payoffs among high and low types are the driving force for the evolution of their respective shares. These fitness-payoff differentials are given in Table 1 and depicted as functions of λ in Figure 2.

A stable inner equilibrium, i.e., an equilibrium where both high types and low types coexist, may be realized around one stable PBE or by the interplay of several PBEs. We first concentrate on the first case (Proposition 2) before turning to the second case (Proposition 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 $\lambda > \lambda^*$. The only candidate where a stable heteromorphic population is supported by a single PBE is one associated with the cooperative high-pooling equilibrium (CHPE) at $1 - \frac{\alpha-(k_L^f-k_H^f)}{\alpha-\beta}$. This is illustrated in Figure 2. All other equilibria are characterized by either strictly negative or strictly increasing payoff differentials. The CHPE exists and is stable if $1 - \frac{\alpha-(k_L^f-k_H^f)}{\alpha-\beta}$ is inside the λ-support of this equilibrium and the fitness differential decreases in λ, which is the case if $\alpha - \beta > 0$ (see Table 1). Taking these conditions together yields:

Proposition 2 *The set* $\left\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) | \lambda = \frac{k_L^f - k_H^f - \beta}{\alpha - \beta}, p_{CCs,s} + p_{CDs,s} = 1, p_{CDs} \geq \frac{k_L}{(1+\alpha)\lambda}\right\}$ *is a stable equilibrium set if and only if* $\max\{\beta + \frac{k_L}{1+\alpha}(\alpha - \beta), \frac{\beta}{\beta+m-\alpha}m\} < k_L^f - k_H^f < \alpha$. *In this equilibrium (CHPE) both types send the signal and high types cooperate.*

Proof. Stability requires a negative slope of the fitness difference function, i.e., $\alpha - \beta > 0$ (see Table 1). Let us first consider $\frac{k_L}{1+\alpha} \leq \frac{\beta}{\beta+m-\alpha}$. In this case, the within-support condition amounts to $\frac{\beta}{\beta+m-\alpha} < 1 - \frac{\alpha-(k_L^f-k_H^f)}{\alpha-\beta} < 1$, rearranging yields $\frac{\beta}{\beta+m-\alpha}m < k_L^f - k_H^f < \alpha$.

If on the other hand $\frac{k_L}{1+\alpha} > \frac{\beta}{\beta+m-\alpha}$, the within-support condition amounts to $\frac{k_L}{1+\alpha} < 1 - \frac{\alpha-(k_L^f-k_H^f)}{\alpha-\beta} < 1$, rearranging yields $\beta + \frac{k_L}{1+\alpha}(\alpha - \beta) < k_L^f - k_H^f < \alpha$.

Note that the first pair of inequalities implies that $\alpha - \beta > 0$, because $\frac{\beta}{\beta+m-\alpha}m < \alpha \iff m(\beta - \alpha) < \alpha(\beta - \alpha) \iff \beta - \alpha < 0$. Thus, the two pairs of inequalities are necessary and sufficient. □

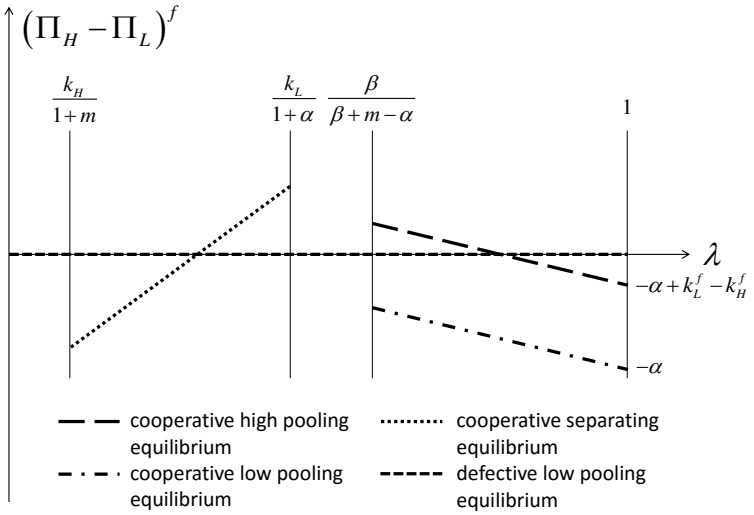

Figure 2: Differences in material payoffs under the conditions of Proposition 2.

The conditions in Proposition 2 reveal that the existence of inner stable equilibria requires that the material signaling costs for high types must exceed the corresponding costs for low types. The spread in signaling costs, however, does not have to compensate for the entire incentive to defect on cooperative behavior (α) for partial cooperation to be supported by the CHPE. Note that the necessary difference in material signaling costs increases in α and β. In other words, the higher the temptation to defect and the higher the suckers' payoff in absolute terms, the higher the required disadvantage in terms of

material signaling cost for low types will be. Interestingly, although the precise level of m is not decisive with respect to its behavioral consequence, its level plays a role for partial cooperation induced by the CHPE. The needed spread in signaling costs weakly decreases in the strength of the preference for conditional cooperation m. That is, if high types are more inclined to conditionally cooperate, the signaling device needs to be materially less disadvantageous for low types.

The equilibrium supported by the CHPE is characterized by partial cooperation and the heterogeneity of preferences. However, the equilibrium only exists if we low types bear substantial material signaling costs exceeding those of high types by more than the suckers' payoff in absolute terms. The equilibrium also requires $\alpha > \beta$. Moreover, the CHPE, like any pooling equilibrium, cannot account for heterogeneity regarding communication. All limitations will be overcome by the second case, i.e., when there is an equilibrium constituted by the interplay of several stable PBEs.

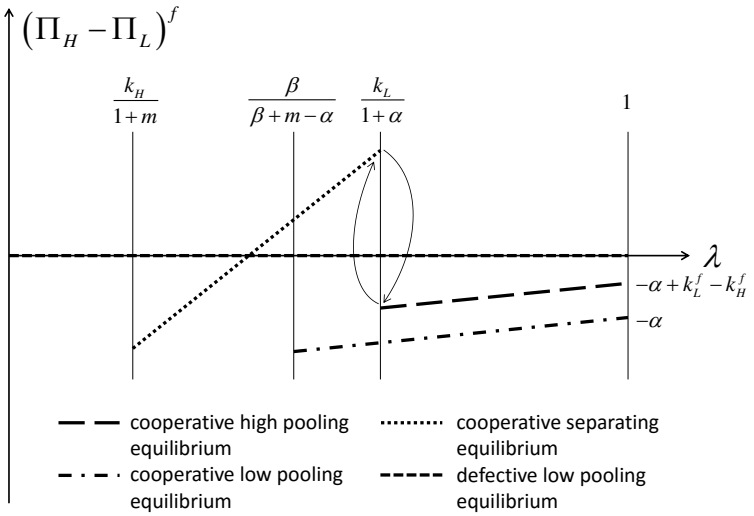

Figure 3: Differences in material payoffs under the conditions of Proposition 3 for $k_L^f > k_H^f$.

In our case, we may have such an equilibrium only at $\lambda^* \equiv \frac{k_L}{1+\alpha}$ where two equilibria interplay: the cooperative separating equilibrium and the cooperative high-pooling equilibrium (see Figure 3). In more detail, high types may earn higher material payoffs in the cooperative separating equilibrium by realizing the gains from mutual cooperation. This induces a growth of their share in the population. At λ^* this equilibrium ceases to exist since low types start to find it profitable to send the cooperative signal. Hence, the system $(\mathbf{p}_H(t), \mathbf{p}_H(t), \lambda(t))$ could be attracted by the cooperative high-pooling equilib-

rium. If high types face a material disadvantage in this equilibrium caused by the signal's loss of its discriminatory power, then their share will decrease. This cyclic process may eventually converge, kept alive only by random drift or may lead to a limit cycle where even without random forces a cyclic process is established. In what follows, we derive the conditions under which either of these two cyclic processes exists.

At this point we introduce the aforementioned semi-pooling equilibrium which exists at λ^*. Its existence and its potential stability make it relevant for the dynamic analysis around λ^*. In this equilibrium all high types play CDs and there is pooling among low types in the sense that some send the signal others don't.[17]

For the interaction of the CSE and CHPE it is necessary that the supports of these two equilibria are adjacent, which corresponds to the following condition:

$$\frac{k_L}{1+\alpha} > \frac{\beta}{\beta + m - \alpha} \tag{4}$$

Moreover, note that the material payoff differences reported in Table 1 imply that

$$\text{(a)} \;\; k_H^f < \frac{k_L}{1+\alpha} \qquad \text{and} \qquad \text{(b)} \;\; k_L^f - k_H^f < \beta + \frac{k_L}{1+\alpha}(\alpha - \beta) \tag{5}$$

are necessary and sufficient conditions for the difference between material payoffs of the high types and of the low types being (a) positive in the CSE and (b) negative in the CHPE, respectively, in the relevant neighborhoods of λ^*. In addition, at λ^* the difference in the average material payoffs of the semi-pooling equilibrium strictly decreases in the share of signal senders among the low types, p_s, and coincides with the corresponding difference of the cooperative separating equilibrium and the cooperative high-pooling equilibrium if $p_s = 0$ and $p_s = 1$, respectively. For λ^* a fixed point thus exists and is unique at $p_s^* \equiv \frac{\lambda^* - k_H^f}{\lambda^*(1+\alpha-\beta)+\beta-k_L^f}$ for any payoff-monotone dynamic. This fixed point can be either stable or unstable. For the latter we can show that there is a limit cycle around this point as long as the dynamics of $(\mathbf{p}_H, \mathbf{p}_L)$ are sufficiently fast.

Proposition 3 *Let the cyclicity conditions* (4) *and* (5) *hold. There exists a* $\bar{k}_H > k_L$ *such that if* $k_H < \bar{k}_H$*, then*

(i) for $k_H^f(1+\alpha-\beta) < k_L^f - \beta$, $(\mathbf{p}_H^*, \mathbf{p}_L^*, \lambda^*)$ *is a stable equilibrium point,*

(ii) for $k_H^f(1+\alpha-\beta) > k_L^f - \beta$*, there exists a* $\underline{\sigma} > 0$ *such that there is a stable (set of) limit cycle(s) around* $(\mathbf{p}_H^*, \mathbf{p}_L^*, \lambda^*)$ *for all* $\sigma > \underline{\sigma}$*. The limit cycles are characterized by* $p_{CDs,s} + p_{CDs,ns} = 1$*,*

[17] For further details see Appendix B

where $\lambda^* = \frac{k_L}{1+\alpha}$, *and* $(\mathbf{p}_H^*, \mathbf{p}_L^*) \in \Delta_{|\mathfrak{S}_H|} \times \Delta_{|\mathfrak{S}_L|}$ *with* $p_{CDs,s} = 1 - p_{CDs,ns} = p_s^*$.

Proof. The strategy of the proof is as follows. Our ultimate goal is to establish the conditions for asymptotic stability of $(\mathbf{p}_H^*, \mathbf{p}_L^*, \lambda^*)$, where $\mathbf{p}_H^* = \{(p_{\mathfrak{s}_H}) \in \Delta_8 | p_{CDs} = 1\}$, $\mathbf{p}_L^* = (p_s^*, 1 - p_s^*)$, with $p_s^* = \frac{\lambda^* - k_H^f}{\lambda^*(1+\alpha-\beta)+\beta-k_L^f}$, and $\lambda^* = \frac{k_L}{1+\alpha}$ and the asymptotic stability of a set $\mathcal{A}$ which contains $(\mathbf{p}_H^*, \mathbf{p}_L^*, \lambda^*)$ if this fixed point is itself unstable according to the dynamics given by (1)–(2). We will prove this by first concentrating on the two-dimensional dynamics resulting from restricting the dynamic system (1)–(2) to p_{CDs}=1, i.e., we consider the following two-dimensional dynamic system:[18]

$$\dot{p}_s = g_s(p_s, \lambda) \cdot p_s \cdot (1 - p_s) \tag{6}$$

$$\dot{\lambda} = h(p_s, \lambda) \cdot \lambda \cdot (1 - \lambda), \tag{7}$$

We will then extend the argument in several steps to the full-dimensional case.

The proof proceeds in six steps. In the first step we show that under certain conditions the dynamic system (6)–(7) has an asymptotically stable fixed point at (p_s^*, λ^*) (Lemma 1). In step two we prove for the dynamic system (6)–(7) the existence of an asymptotically stable set $\mathcal{A}_2 \subset [0,1]^2$ which contains (λ^*, p_s^*) (Lemma 2) under certain conditions. Moreover, we show that the forward orbit of any point $(\lambda, p_s) \in \mathcal{A}_2$ circles around the fixed point (p_s^*, λ^*). By step three we show that under condition (ii) $\mathcal{A}_2$ contains either a stable limit cycle or a neutrally-stable center (Lemma 3). Step four extends the result to the dynamic system (1)–(2) restricted to $p_{CCs} + p_{CDs} = 1$ (Lemma 4). In Lemma 5 we show the payoff superiority of the strategies CCs and CDs in a full-dimensional set containing $\mathcal{A}_2$. Finally, we prove the existence of a full-dimensional Lyapunov stable set $\mathcal{A}$ with the property that for any state $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}$ the strategies CCs and CDs earn the highest payoffs (Lemma 6). As a consequence, any perturbation from $p_{CDs} = 1$ other than CCs will vanish. By Lemma 4 this finishes the proof. We will assume for the remainder of the proof that the cyclicity conditions (4) and (5) hold.

Step One

Lemma 1 $p_s^* = \frac{\lambda^* - k_H^f}{\lambda^*(1+\alpha-\beta)+\beta-k_L^f}$, $\lambda^* = \frac{k_L}{1+\alpha}$ *is a fixed point of the dynamic system* (6)–(7). *The fixed point is asymptotically stable if* $k_H^f(1+\alpha-\beta) < k_L^f - \beta$. *It is unstable if* $k_H^f(1+\alpha-\beta) > k_L^f - \beta$.

[18] The additional factor $(1 - p_s)$ results from the fact that $p_{ns} = 1 - p_s$, see footnote 13.

Proof. Given our assumption that payoff-monotonicity in the dynamics across types, i.e., $\dot{\lambda}$, refers to average material payoffs for each type, we have $\dot{\lambda} = 0$ if and only if $\Pi_H^f(CDs) = p_s\Pi_L^f(s) + (1-p_s)\Pi_L^f(ns)$. Moreover, $\dot{p}_s = 0$ if and only if $\Pi_L(s) = \Pi_L(ns)$. This gives us the following system of equations:

$$\dot{p}_s = 0 \quad \Leftrightarrow \quad \lambda(1+\alpha) - k_L = 0 \tag{8}$$

$$\dot{\lambda} = 0 \quad \Leftrightarrow \quad \lambda - k_H^f - (1-\lambda)\beta p_s = p_s\big(\lambda(1+\alpha) - k_L^f\big) \tag{9}$$

$$\Leftrightarrow \quad \lambda(1 - p_s(1+\alpha-\beta)) = k_H^f + \beta p_s - p_s k_L^f \tag{10}$$

The equivalences (8)–(9) hold for $p_s, \lambda \in (0,1)$. Solving $\dot{p}_s = \dot{\lambda} = 0$ yields p_s^* and λ^* as the unique fixed point of the dynamic system (6)–(7). Since we assumed that $k_L < 1+\alpha$, we have $\lambda^* \in (0,1)$. Cyclicity conditions (5a) and (5b) guarantee that $p_s^* \in (0,1)$.

In order to study the stability of this solution, we linearize the system (6)–(7) at λ^* and p_s^* and study the eigenvalues of the corresponding characteristic matrix. This gives us

$$\chi_{1,2} = \frac{1-p_s^*(1+\alpha-\beta)}{2}\lambda^*(1-\lambda^*) \pm \tag{11}$$

$$\sqrt{\left(\frac{1-p_s^*(1+\alpha-\beta)}{2}\lambda^*(1-\lambda^*)\right)^2 - (1+\alpha)(\lambda^*(1+\alpha-\beta)+\beta-k_L^f)\lambda^*(1-\lambda^*)p^*(1-p^*)} \tag{12}$$

as eigenvalues of the characteristic matrix. Since the last term of the radicand is positive by cyclicity conditions (5), the fixed point (p_s^*, λ^*) is stable if $1 - p_s^*(1+\alpha-\beta) < 0$ and unstable if $1 - p_s^*(1+\alpha-\beta) > 0$. Inserting p_s^* into the condition for stability and reformulating gives us: $\frac{\beta - k_L^f + k_H^f(1+\alpha-\beta)}{\lambda^*(1+\alpha-\beta)+\beta-k_L^f} < 0$. Again, since $\lambda^*(1+\alpha-\beta)+\beta-k_L^f > 0$ this reduces to $k_H^f(1+\alpha-\beta)+\beta-k_L^f < 0$. □

In the next step we focus on the case where the fixed point (p_s^*, λ^*) is unstable.

Step Two

We introduce the following definition, where σ corresponds to the parameter in (3).

Definition $\mathcal{A}_2(\mu,\sigma) = \left\{(p_s,\lambda) \in [0,1]^2 \middle| \lambda \in (\lambda^-(p_s), \lambda^+(p_s))\right\}$ *with* $\lambda^-(p_s)$ *and* $\lambda^+(p_s)$ *being characterized by the differential equations and boundary conditions:*

$$\frac{\partial \lambda^+(p_s)}{\partial p_s} = \max\left\{\frac{\mu}{\sigma}, (1+\mu)\frac{\dot{\lambda}}{\dot{p}_s}\right\} \text{ and } \lambda^+(0) = \lambda^* + \mu \tag{13}$$

$$\frac{\partial \lambda^-(p_s)}{\partial p_s} = \max\left\{\frac{\mu}{\sigma}, (1+\mu)\frac{\dot{\lambda}}{\dot{p}_s}\right\} \text{ and } \lambda^-(1) = \lambda^* - \mu \tag{14}$$

where $\dot{p}_s$ *and* $\dot{\lambda}$ *are given by equations* (6) *and* (7)*, and* $\mu > 0, \sigma > 1$.

First, as long as $\dot{\lambda}/\dot{p}_s \geq \frac{\mu}{\sigma(1+\mu)}$ the boundary functions $\lambda^-(p_s)$ and $\lambda^+(p_s)$ solve the differential equation $\partial\lambda(p_s)/\partial p_s = (1+\mu)\dot{\lambda}/\dot{p}_s$. Note that payoff-monontonicity implies $\dot{p}_s(p_s, \lambda) \neq 0$ for all $\lambda \neq \lambda^*$ as long as $p_s \in (0,1)$. Importantly, given that $p_{CDs} = 1$, we will show below that signaling earns low types the highest payoff for population states $p_s = 0$ and $\lambda \in (\lambda^*, 1)$. Hence, $\dot{p}_s(0,\lambda) > 0, \forall\lambda \in (\lambda^*, 1)$ by the social innovation assumption. Analogously, given that $p_{CDs} = 1$, strategy *ns* earns low types the highest payoff for population states with $p_s = 1$ and $\lambda \in (0, \lambda^*)$. Hence, $\dot{p}_s(1,\lambda) < 0, \forall\lambda \in (0, \lambda^*)$. Thus, $1+\mu)\dot{\lambda}/\dot{p}_s$ is well defined. Existence and uniqueness of a solution to the differential equation are guaranteed by the Picard-Lindelöf theorem. Second, for $\dot{\lambda}/\dot{p}_s < \frac{\mu}{\sigma(1+\mu)}$ the boundary functions are linear functions with slope μ/σ. Consequently, $\lambda^-(p_s)$ and $\lambda^+(p_s)$ are well defined. See Figure 4 for an illustration (see Appendix A for details).

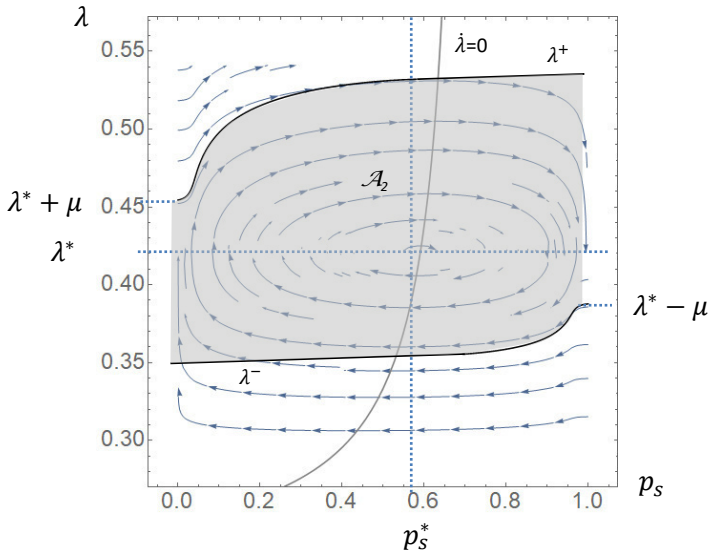

Figure 4: Vector field for modified replicator dynamics. Parameters: $m = 2, \alpha = 0.9, \beta = 0.5, k_H = k_H^f = 0.25, k_L = k_L^f = 0.8, \sigma = 19, nn = 20, np = 1, \mu = 1/30$.

Lemma 2 *Let* $k_H^f(1+\alpha-\beta) > k_L^f - \beta$*. There exists an* $\overline{\mu}$ *such that for all* $\mu \in (0, \overline{\mu})$ *there exists a* $\sigma(\mu) > 0$ *such that* $\mathcal{A}_2(\mu, \sigma)$ *is an asymptotically stable set of* (6)–(7) *for all* $\sigma > \sigma(\mu)$.

Proof. As a first step we partition the unit square $[0,1]^2$ according to the dynamics of p_s and λ. Note that the function $\lambda(p_s) = \frac{p_s(\beta - k_L^f) + k_H^f}{1 - p_s(1+\alpha-\beta)}$ which solves (9) has a singularity at $p_s^{crit} = \frac{1}{1+\alpha-\beta}$. Moreover, $\lambda(p_s) \to \frac{k_L^f - \beta}{1+\alpha-\beta}$ as $p_s \to \pm\infty$ and $\lambda(0) = k_H^f$. Also note that

$\lambda(0) = k_H^f < \frac{k_L}{1+\alpha} = \lambda^*$, because of the cyclicity condition (5a). Additionally, we have $\lambda(1) = \frac{k_H^f+\beta-k_L^f}{\beta-\alpha}$. We distinguish two cases. First, $p_s^{crit} \in (0,1)$ implies that $\beta < \alpha$ and thereby that $\lambda(1) < \lambda^*$ because $k_L^f - \beta < k_H^f(1+\alpha-\beta) < k_H^f + (\alpha-\beta)\lambda^*$ where the first inequality follows from the assumption of this Lemma and the second from the cyclicity condition (5a). Second, if $p_s^{crit} \notin [0,1]$ then $\beta > \alpha$. This implies that $\lambda(1) > \lambda^*$ because of $k_L^f - \beta < k_H^f + (\alpha-\beta)\lambda^* < k_H^f(1+\alpha-\beta)$. Given equation (10) these insights imply:

$$\dot{\lambda} > 0 \Leftrightarrow \begin{cases} \lambda > \frac{k_H^f + p_s(\beta - k_L^f)}{1 - p_s(1+\alpha-\beta)} \ , & p_s < p_s^{crit} \\ \lambda < \frac{k_H^f + p_s(\beta - k_L^f)}{1 - p_s(1+\alpha-\beta)} \ , & p_s > p_s^{crit} \end{cases} \tag{15}$$

Finally, $\dot{p}_s = 0$ if and only if $\lambda = \lambda^*$. See Figure 5 for an illustration.

In a second step we show that we can choose μ and σ such that $\mathcal{A}_2(\mu,\sigma) \subset [0,1] \times [\lambda^* - 2\mu, \lambda^* + 2\mu]$. Let us first focus on $\lambda^+(p_s)$. According to equations (6)–(7), We may have $\dot{p}_s = 0$ only if $p_s \in \{0,1\}$ or $\lambda = \lambda^*$. Our adiabatic-elimination assumption guarantees that $\frac{\|\dot{p}_s\|}{\|\dot{\lambda}\|} > \sigma$ outside an ι-neighborhood of $0 \times [0,1]$, $1 \times [0,1]$, and $(0,1) \times \lambda^*$ in $[0,1]^2$. This is because $\sqrt{2}\frac{\|\dot{p}_s\|}{\|\dot{\lambda}\|} = \frac{\|(\dot{p}_s, \dot{p}_{ns})\|}{\|\dot{\lambda}\|} = \frac{\|(\dot{p}_H, \dot{p}_L)\|}{\|\dot{\lambda}\|} > \sqrt{2}\sigma$. Thus, we have:

$$\frac{\partial \lambda^+(p_s)}{\partial p_s} \leq \max\left\{\frac{\mu}{\sigma}, \frac{1+\mu}{\sigma}\right\} = \frac{1+\mu}{\sigma}, \forall p_s \in [\iota, 1-\iota] \tag{16}$$

Moreover, $\Pi_L(s) > \Pi_L(ns)$ for $p_s = 0$ and $\lambda > \lambda^*$. Thus, our social innovation assumption implies for $\lambda > \lambda^*$ and $p_s = 0$ we have $\dot{p}_s > 0$. Thus, for any $\lambda' > \lambda^*$, $\frac{\|\dot{\lambda}\|}{\|\dot{p}_s\|}$ is bounded above for all $(p_s, \lambda) \in [0,\iota] \times [\lambda', 1]$. As a consequence, for $p_s \in [0,\iota]$, $\lambda^+(p_s)$ is bounded by a linear function in ι. For a sufficiently small ι, if $p_s > 1 - \iota$ then $\lambda^+(p_s)$ is given by a linear function with slope μ/σ because of $\dot{\lambda} < 0$. Taken together, for any $\mu > 0$ we can set ι sufficiently small and $\sigma(\mu)$ sufficiently large such that $\lambda^+(1) < \lambda^* + 2\mu$. By an analogous argument we can guarantee that $\lambda^-(0) > \lambda^* - 2\mu$ because $\Pi_L(s) < \Pi_L(ns)$ for $p_s = 1$ and $\lambda < \lambda^*$.

The upper bound $\overline{\mu}$ must satisfy the following conditions:

(i) $\lambda^+(1)$ in the support of CHPE, $\lambda^-(0)$ in the support of CSE

$$\Leftarrow \lambda^* - 2\overline{\mu} > \frac{k_H}{1+m} \text{ and } \lambda^* + 2\overline{\mu} < 1 \tag{17}$$

(ii) $\dot{\lambda}(0,\lambda) > 0, \forall \lambda \in (\lambda^* - 2\overline{\mu}, \lambda^*)$, $\dot{\lambda}(1,\lambda) < 0, \forall \lambda \in (\lambda^*, \lambda^* + 2\overline{\mu})$

$$\Leftarrow \lambda^* - 2\overline{\mu} > k_H^f \text{ and } \lambda^* + 2\overline{\mu} < \frac{k_H^f + \beta - k_L^f}{\beta - \alpha} \tag{18}$$

Condition (i) guarantees that the boundary functions of $\mathcal{A}_2(\mu,\sigma)$ will limit the trajectories to end up in the support of the CSE while approaching $p_s = 0$, and in the support of the CHPE while approaching $p_s = 1$, respectively. The sufficiency of equation (17) follows from the definitions of $\lambda^+(p_s)$, and $\lambda^-(p_s)$. Condition (ii) implies $\dot{\lambda} > 0 (< 0)$ for trajectories in $\mathcal{A}_2(\mu,\sigma)$ near the CSE (CHPE). The sufficiency of (18) results from (15). The existence of a $\overline{\mu} > 0$ which satisfies inequalities (17) and (18) is guaranteed by the non-empty support for the CSE and by the cyclicity conditions (5a), and (5b). In summary, for $\mu < \overline{\mu}$ and $\sigma > \sigma(\mu)$, $\mathcal{A}_2(\mu,\sigma)$ is partitioned into four segments: (1) $\dot{p}_s, \dot{\lambda} > 0$, (2) $\dot{p}_s > 0, \dot{\lambda} < 0$, (3) $\dot{p}_s, \dot{\lambda} < 0$, and (4) $\dot{p}_s < 0, \dot{\lambda} > 0$ (see Figure 5).

In the last step we show that for $\mu < \overline{\mu}$ and $\sigma > \sigma(\mu)$ any trajectory at the boundaries $\lambda^+(p_s)$ and $\lambda^-(p_s)$ points into $\mathcal{A}_2(\mu,\sigma)$. Note that by definition $\lambda^+(p_s) > \lambda^*$. For $\lambda > \lambda^*$, we have $\dot{p}_s > 0$ for all $p_s < 1$ including zero due to the social innovation assumption. Hence, as long as $\dot{\lambda} > 0$ the definition of $\lambda^+(p_s)$ implies $\frac{\partial \lambda^+(p_s)}{\partial p_s} \geq (1+\mu)\frac{\dot{\lambda}}{\dot{p}_s} > \frac{\dot{\lambda}}{\dot{p}_s}$. That is, the slope of the boundary function $\lambda^+(p_s)$ is strictly greater than the slope of any trajectory at the boundary $\lambda^+(p_s)$. Thus, at the boundary $\lambda^+(p_s)$ the trajectories point into $\mathcal{A}_2(\mu,\sigma)$ for $\dot{\lambda} > 0$. If $\dot{\lambda} \leq 0$ the definition of $\lambda^+(p_s)$ implies $\frac{\partial \lambda^+(p_s)}{\partial p_s} = \frac{\mu}{\sigma} > 0 > \frac{\dot{\lambda}}{\dot{p}_s}$. Hence, any trajectory intersecting $\lambda^+(p_s)$ points into $\mathcal{A}_2(\mu,\sigma)$ for $\dot{\lambda} \leq 0$. Note that $\mu < \overline{\mu}$ and $\sigma > \sigma(\mu)$ guarantee that the trajectory at $(1, \lambda^+(1))$ points into $\mathcal{A}_2(\mu,\sigma)$. A similar argument establishes the analogous results for $\lambda^-(p_s)$.

Hence, $\mathcal{A}_2(\mu,\sigma)$ is asymptotically stable with respect to the dynamics (6)–(7). See Figure 4 for an example. □

Step Three

In what follows, we need the concept of strict circulation. let $\xi_2(\cdot, p_s, \lambda)$ denote the induced solution mapping of the dynamic system (6)–(7). We say that a forward orbit of system (6)–(7) *strictly circulates* around $(p_s^*, \lambda^*) \in \mathcal{A}_2(\mu,\sigma)$ if starting from any point in $(p_s, \lambda) \in \mathcal{A}_2(\mu,\sigma)$ the trajectory $\xi_2(\cdot, p_s, \lambda)$ in the phase plane moves either clockwise or counterclockwise relative to (p_s^*, λ^*) for all $t > 0$ but never changes direction. Formally, let $\mathbf{x}(t) \equiv \xi_2(t, p_s, \lambda) - (p_s^*, \lambda^*)$ then $\xi_2(\cdot, p_s, \lambda)$ *strictly circulates* around $(p_s^*, \lambda^*) \in \mathcal{A}_2(\mu,\sigma)$ if $x_1(t)x_2(t+dt) - x_2(t)x_1(t+dt) < 0, \forall t > 0$, or $x_1(t)x_2(t+dt) - x_2(t)x_1(t+dt) > 0, \forall t > 0$, where $dt > 0$ is sufficiently small.

Lemma 3 *Let $k_H^f(1+\alpha-\beta) > k_L^f - \beta$, $\mu \in (0, \overline{\mu})$, and $\sigma > \sigma(\mu)$. Then for any $(p_s, \lambda) \in \mathcal{A}_2(\mu,\sigma)$ the trajectory $\xi_2(\cdot, p_s, \lambda)$ of the dynamic system (6)–(7) strictly circulates around (p_s^*, λ^*) and converges to a limit cycle or a neutrally-stable center in $\mathcal{A}_2(\mu,\sigma)$.*

Proof. First we show that $\xi_2(\cdot, p_s, \lambda)$ with $(p_s, \lambda) \in \mathcal{A}_2(\mu, \sigma)$ strictly circulates around (p_s^*, λ^*). Remember that $\dot{p}_s > 0$ if and only if $\lambda > \lambda^*$ and that the condition for $\dot{\lambda} > 0$ is given by equation (15). Note that $k_H^f(1+\alpha-\beta) > k_L^f - \beta$ also implies that $\partial\lambda(p_s)/\partial p_s > 0$, where $\lambda(p_s)$ solves $\dot{\lambda} = 0$. Thus, for $\mu \in (0, \overline{\mu})$, and $\sigma > \sigma(\mu)$ the loci of $\dot{\lambda} = 0$ and $\dot{p}_s = 0$ partition $\mathcal{A}_2(\mu, \sigma)$ into four segments as shown in Figure 5 which depicts a generic phase plane for the dynamics (6)–(7). Clearly, for sufficiently small $dt > 0$ in any of the four segments we have $x_1(t)x_2(t+dt) - x_2(t)x_1(t+dt) < 0, \forall t > 0$.

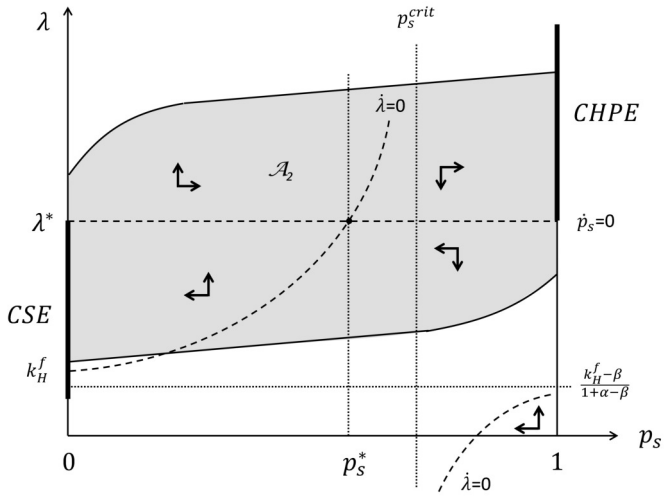

Figure 5: Dynamics in $\mathcal{A}_2(\mu, \sigma)$ for $k_H^f(1 + \alpha - \beta) > k_L^f - \beta$, i.e., for an unstable fixed point at (p_s^*, λ^*).

We now turn to the convergence of all forward orbits in $\mathcal{A}_2(\mu, \sigma)$. As a consequence of the Poincaré-Bendixson Theorem,[19] $\mathcal{A}_2(\mu, \sigma)$ contains at least one limit cycle. This is because the condition $k_H^f(1 + \alpha - \beta) > k_L^f - \beta$ implies that (p_s^*, λ^*) is a repellor because both eigenvalues in equations (11)–(12) have positive real parts. Thus, there exists an open neighborhood V of (p_s^*, λ^*) such that all trajectories that intersect the boundary ∂V point to the interior of the set $\mathcal{A}_2(\mu, \sigma) \setminus V$. Moreover, the asymptotic stability of $\mathcal{A}_2(\mu, \sigma)$ (Lemma 2) implies that any trajectory that intersects the boundary of $\mathcal{A}_2(\mu, \sigma)$ points to the interior of $\mathcal{A}_2(\mu, \sigma)$. Applying the Poincaré-Bendixson Theorem to $\mathcal{A}_2(\mu, \sigma) \setminus V$ implies the existence of at least one limit cycle. Since all trajectories circulate around (p_s^*, λ^*) and $\mathcal{A}_2(\mu, \sigma) \setminus V$ is an attractor they must converge to a limit cycle or a neutrally stable center in $\mathcal{A}_2(\mu, \sigma)$. □

In the next three steps we provide the extension of Lemma 1 and Lemma 3 to the full-dimensional case. In a first step (Lemma 4) under the restriction to $p_{CDs} + p_{CCs} = 1$ we

[19] See Bendixson (1901).

show the existence of a set $\mathcal{A}_3 \supset \mathcal{A}_2(\mu,\sigma)$ which allows for small positive p_{CCs} such that any trajectory in this set converges to $\mathcal{A}_2(\mu,\sigma)$. In Lemma 5, allowing for an arbitrary perturbation from $p_{CDs}=1$, we show that as long as the population state is close to $p_{CDs}+p_{CCs}=1$ and p_{CDs} is not too small the strategies CDs and CCs will earn the highest payoffs. This holds for all $p_s \in [0,1]$ and in a neighborhood of λ^*. Thus, if we set μ sufficiently low and $\sigma > \sigma(\mu)$ then $\mathcal{A}_2(\mu,\sigma)$ will satisfy these conditions, i.e., the strategies CDs and CCs will earn the highest payoffs. In the last step (Lemma 6) we show based on Lemma 5 the existence of a Lyapunov stable set $\mathcal{A} \supset \mathcal{A}_3$ with the property that $\dot{p}_{\neg CCs} < 0$ for all $(\mathbf{p_H},\mathbf{p_L},\lambda) \in \mathcal{A}$, where $p_{\neg CCs} \geq 0$ refers to the sum of perturbations from $p_{CDs}=1$ others than CCs. Thus, Lemma 6 informs us that we can essentially restrict to the case considered in Lemma 4.

Step Four

In order to state the first result, let $\xi_3(\cdot,\mathbf{p}_H,\mathbf{p}_L,\lambda)$ denote the induced solution mapping of the dynamic system (1)–(2) under the restriction to $p_{CDs}+p_{CCs}=1$.

Lemma 4 *Let $\mu<\overline{\mu}$ and $\sigma>\sigma(\mu)$. Then there exist $\nu>0$ such that $\xi_3(t,\mathbf{p}_H,\mathbf{p}_L,\lambda) \xrightarrow{t\to\infty} \mathcal{A}_2(\mu,\sigma)$, $\forall(\mathbf{p}_H,\mathbf{p}_L,\lambda)\in\mathcal{A}_3=\cup_{i=1}^3\mathcal{A}_3(i)$, where $\mathcal{A}_3(1)=A_1\cap B$, $\mathcal{A}_3(2)=D\cap B$, and $\mathcal{A}_3(3)=A_2\cap B$ with*

$$A_1=\left\{(\mathbf{p}_H,\mathbf{p}_L,\lambda)\in C\middle|(p_s,\lambda)\in\mathcal{A}_2(\mu,\sigma),\ p_{CCs}>1-\frac{k_L}{\lambda(1+\alpha)},\ p_{CCs}+p_{CDs}=1\right\}$$
$$A_2=\left\{(\mathbf{p}_H,\mathbf{p}_L,\lambda)\in C\middle|(p_s,\lambda)\in\mathcal{A}_2(\mu,\sigma),\ p_{CCs}<1-\frac{k_L}{\lambda(1+\alpha)},\ p_{CCs}+p_{CDs}=1\right\}$$
$$B=\left\{(\mathbf{p}_H,\mathbf{p}_L,\lambda)\in C\middle|(p_s,\lambda)\in\mathcal{A}_2(\mu,\sigma),\ p_{CDs}\geq 1-\nu,\ p_{CCs}+p_{CDs}=1\right\}$$
$$D=\left\{(\mathbf{p}_H,\mathbf{p}_L,\lambda)\in C\middle|(p_s,\lambda)\in\mathcal{A}_2(\mu,\sigma),\ p_{CCs}=1-\frac{k_L}{\lambda(1+\alpha)},\ p_{CCs}+p_{CDs}=1\right\}.$$

Proof. Figure 6 illustrates the construction of $\mathcal{A}_3$ and the partition $\mathcal{A}(i)$. Note that the specification of the partition is indeed not necessary to prove the convergence result. However, the construction is informative for Lemma 6 where the construction of $\mathcal{A}$ builds on this partition.

Given our assumption that payoff-monotonicity in the dynamics across types, i.e., $\dot{\lambda}$, refers to average material payoffs for each type and focusing on $p_s,\lambda\in(0,1)$, we have $\dot{\lambda}=0$ if and only if $p_{CDs}\Pi_H^f(CDs)+p_{CCs}\Pi_H^f(CCs)=p_s\Pi_L^f(s)+(1-p_s)\Pi_L^f(ns)$. Moreover,

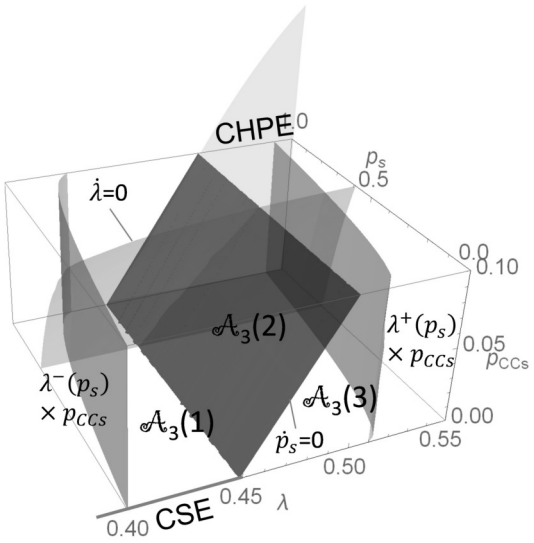

Figure 6: Partition of $\mathcal{A}_3$.

$\dot{p}_s = 0$ if and only if $\Pi_L(s) = \Pi_L(ns)$. This gives us the following system of inequalities:

$$\dot{p}_s \geq 0 \quad \Leftrightarrow \quad p_{CCs} \leq 1 - \frac{k_L}{\lambda(1+\alpha)} \tag{19}$$

$$\dot{\lambda} \leq 0 \Leftrightarrow (1 - p_{CCs})(\lambda - k_H^f - (1-\lambda)\beta p_s) + p_{CCs}(\lambda - k_H^f - (1-\lambda)\beta) \leq \tag{20}$$

$$p_s\big(\lambda(1+\alpha) - k_L^f\big) + (1-p_s)\lambda(1+\alpha)p_{CCs} \tag{21}$$

The last inequality informs us that $\dot{\lambda} \leq 0$ if and only if $p_s\Big((1-p_{CCs})\big((1-\lambda)\beta + \lambda(1+\alpha)\big) - k_L^f\Big) \geq \lambda - k_H^f$, where the left-hand side is strictly positive in a neighborhood of λ^* by cyclicity condition (5a). As a consequence, $\dot{\lambda} > 0$ for all $(\mathbf{p}_H, \mathbf{p}_H, \lambda) \in \mathcal{A}_3$ for $p_s \approx 0$. Moreover, by cyclicity condition (5b) $\dot{\lambda} < 0$ for all $(\mathbf{p}_H, \mathbf{p}_H, \lambda) \in \mathcal{A}_3$ for $p_s \approx 1$ and $p_{CCs} \approx 0$, i.e., for sufficiently small perturbation ν_{CCs}. Taken together, $\mathcal{A}_3(1)$ is characterized by $\dot{p}_s < 0$, $\dot{\lambda} < 0$ for $p_s \approx 1$, and $\dot{\lambda} > 0$ for $p_s \approx 0$, $\mathcal{A}_3(2)$ by $\dot{p}_s = 0$. Finally, $\mathcal{A}_3(3)$ is characterized by $\dot{p}_s > 0$, $\dot{\lambda} < 0$ for $p_s \approx 1$, and $\dot{\lambda} > 0$ for $p_s \approx 0$.

Under the restriction to $p_{CCs} + p_{CDs} = 1$ payoffs for CDs and CCs are given by:

$$\Pi_H(CCs) = \lambda(p_{CCs} + p_{CDs})(1+m) + (1-\lambda)(-\beta) - k_H \tag{22}$$

$$\Pi_H(CDs) = \lambda(p_{CCs} + p_{CDs})(1+m) + (1-\lambda)p_s(-\beta) - k_H \tag{23}$$

Thus, $\Pi_H(CDs) > \Pi_H(CCs), \forall p_s \in [0,1), \lambda \in (0,1)$. For sufficiently small μ it follows that $\lambda \in (0,1)$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}_3$. Hence, equality in payoffs results only for $p_s = 1$. Under the restriction to $p_{CDs} + p_{CCs} = 1$ this translates into $\dot{p}_{CDs} > 0$ and $\dot{p}_{CCs} < 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}_3$ with $p_s < 1$.

Note that $\mu < \overline{\mu}$ and $\sigma > \sigma(\mu)$ guarantee that $\mathcal{A}_2(\mu, \sigma)$ is an asymptotically stable set of (6)–(7) by Lemma 3. Continuity of the dynamics implies then that for sufficiently small ν the attractor-property of $\mathcal{A}_2(\mu, \sigma)$ translates to $\mathcal{A}_3$ with respect to p_s and λ. Taken together with $\dot{p}_{CDs} > 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}_3$ with $p_s < 1$ this implies the convergence of $\xi_3(t, \mathbf{p}_H, \mathbf{p}_H, \lambda)$ to $\mathcal{A}_2(\mu, \sigma)$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}_3$. In other words, under the restriction to $p_{CDs} + p_{CCs} = 1$ small perturbations from $\mathcal{A}_2(\mu, \sigma)$ with respect to p_{CCs} vanish and the system will eventually be attracted by a stable fixed point (Lemma 1) or a stable (set of) limit cycle(s) (Lemma 3) in $\mathcal{A}_2(\mu, \sigma)$. □

Step Five

Lemma 5 *There exist $\varepsilon, \nu > 0$ and $\underline{p_{CDs}} \in (0, 1-\nu)$ such that $\dot{p}_{CDs} + \dot{p}_{CCs} \geq 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \big\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in C \big|\ p_{CDs} + p_{CCs} \geq 1 - \nu,\ p_{CDs} \geq \underline{p_{CDs}},\ p_s \in [0,1],\ \lambda \in (\lambda^* - \varepsilon, \lambda^* + \varepsilon)\big\}$, where $\dot{p}_{CDs} + \dot{p}_{CCs} = 0$ if and only if $p_{CDs} + p_{CCs} = 1$.*

Proof. We look at payoffs under the restriction to $p_{CDs} + p_{CCs} = 1$. We observe that for all $\lambda \in [0,1]$ and $p_s \in [0,1]$:

$$\Pi_H(CDs) \geq \Pi_H(CCs); \Pi_H(DDs) \geq \Pi_H(DCs) \tag{24}$$

$$\Pi_H(CDns) \geq \Pi_H(CCns); \Pi_H(DDns) \geq \Pi_H(DCns) \tag{25}$$

Note that all inequalities are strict for $p_s \in (0,1)$. Note further that

$$\Pi_H(CCs) > \Pi_H(DDs) \Leftrightarrow \lambda > \frac{\beta}{m - \alpha + \beta} \tag{26}$$

$$\Pi_H(CCs) > \Pi_H(CDns) \Leftrightarrow p_{CDs} > \frac{k_H + (1-\lambda)(1-p_s)\beta}{\lambda(1+m)} \tag{27}$$

$$\Leftarrow p_{CDs} > \frac{k_H + (1-\lambda)\beta}{\lambda(1+m)} \tag{28}$$

Moreover,

$$\Pi_H(CCs) > \Pi_H(DDns) \Leftrightarrow p_{CDs} > \frac{k_H + \beta - \lambda(m - \alpha + \beta)}{\lambda(1 + \alpha)} \tag{29}$$

Note that (26) is satisfied at λ^* because of the cyclicity condition (4). As a first step we want to determine the condition for the existence of a lower bound $\hat{p}_{CDs} < 1$ such that (28) and (29) are satisfied at λ^* for $p_{CDs} \geq \hat{p}_{CDs}$. Continuity in λ then implies that this also holds in a small neighborhood of λ^*. Evaluated at λ^* the RHSs of (28) and (29) are less than unity if and only if $k_H < \lambda^*(1+m+\beta)-\beta \equiv \overline{k}_H$. Thus, for $k_H < \overline{k}_H$ payoff inequalities in (27) and (29) are satisfied at λ^* if $p_{CDs} > \hat{p}_{CDs} \equiv \max\left\{\frac{k_H+(1-\lambda)\beta}{\lambda(1+m)}, \frac{k_H+\beta-\lambda(m-\alpha+\beta)}{\lambda(1+\alpha)}\right\} < 1$. Importantly, the cyclicity condition (4) implies that $\overline{k}_H > k_L$. Thus, although sufficient for a lower bound less than unity we don't have to assume that high types have a cost advantage, i.e., $k_H < k_L$.

Hence, for $k_H < \overline{k}_H$ there exists an $\varepsilon > 0$ and a $\hat{p}_{CDs} < 1$ such that the payoff-inequalities (26), (28), and (29) hold strictly for all $\lambda \in (\lambda^* - \varepsilon, \lambda^* + \varepsilon)$ and $p_{CDs} > \hat{p}_{CDs}$. Taken together this gives us for all $\lambda \in (\lambda^* - \varepsilon, \lambda^* + \varepsilon)$, $p_s \in [0, 1]$, and $p_{CDs} > \hat{p}_{CDs}$:

$$\Pi_H(CDs) \geq \Pi_H(CCs) > \Pi_H(DDs) \geq \Pi_H(DCs) \tag{30}$$

$$\Pi_H(CCs) > \Pi_H(CDns) \geq \Pi_H(CCns); \Pi_H(CCs) > \Pi_H(DDns) \geq \Pi_H(DCns) \tag{31}$$

That is, under the restriction $p_{CCs} + p_{CDs} = 1$ the strategies CDs and CCs earn the strictly highest payoffs as long as p_{CDs} is sufficiently high and λ is close to λ^*. Strictness implies that this is also true in a small neighborhood of $p_{CCs}+p_{CDs} = 1$. Thus, there exists a $\nu > 0$ such that strategies CDs and CCs earn the strictly highest payoffs as long as $p_{CCs}+p_{CDs} \geq 1-\nu$ and $p_{CDs} > \underline{p_{CDs}}$, for a sufficiently high $\underline{p_{CDs}} \in \left(\frac{k_H+\beta-(\lambda-\varepsilon)(m-\alpha+\beta)}{(\lambda-\varepsilon)(1+\alpha)}, 1-\nu\right)$. As a consequence, payoff-monotonicity implies that $\dot{p}_{CDs} + \dot{p}_{CCs} \geq 0$, where equality occurs if and only if $p_{CCs} + p_{CDs} = 1$. □

Step Six

If $\dot{p}_{CDs} + \dot{p}_{CCs} > 0$ would imply that $\dot{p}_{CDs} > 0$ we could simply choose a sufficiently small μ and $\sigma > \sigma(\mu)$ such that continuity of the dynamics and attractor-property of $\mathcal{A}_2(\mu, \sigma)$ would imply that any trajectory in a neighborhood would converge to $\mathcal{A}_2(\mu, \sigma)$. However, if $p_s = 1$ we have

$$\Pi_H(CDs) - \Pi_H(CCs) = \lambda\big((p_{DCns} + p_{DDns})\beta - (p_{CCns} + p_{CDns})(m - \alpha)\big), \tag{32}$$

which can be negative. Thus, given some small perturbation $\nu > 0$ the share p_{CCs} could grow while p_{CDs} decreases. As a consequence, the condition of $p_{CDs} > \underline{p_{CDs}}$ could be violated which in turn could compromise the convergence toward $\mathcal{A}_2(\mu, \sigma)$. In the proof of the following Lemma we show that the growth of p_{CCs} is bounded either because the dynamics are led by the attractor-property of $\mathcal{A}_2(\mu, \sigma)$ or the system is in the basin of attraction of the high-pooling equilibrium or by an argument which makes use of our assumption about the emergence of social innovations. To formalize this idea let CHPE(λ) denote the cooperative equilibrium set at λ, i.e., CHPE(λ) $= \big\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in C \big|\ p_{CDs} + p_{CCs} = 1,\ p_{CDs} \geq \frac{k_L}{\lambda(1+\alpha)},\ p_s = 1\big\}$.

Lemma 6 *Let $\mu < \overline{\mu}$ and $\sigma > \sigma(\mu)$. There exists a Lyapunov stable set $\mathcal{A}$ of the dynamic system (1)–(2) such that: (i) $\mathcal{A}_2(\mu, \sigma) \subset \mathcal{A}$, (ii) $\mathcal{A}^o$ is an open subset in $X_1 \times X_2 \times X_3$, and (iii) $\dot{p}_{CCs} + \dot{p}_{CDs} > 0,\ \forall (\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A} \setminus \mathcal{A}_2(\mu, \sigma)$.*

Proof. The strategy of the proof of Lyapunov stability is as follows. We will construct three disjointed sets $\mathcal{A}(i), i = 1, 2, 3$ and set $\mathcal{A} = \cup_i \mathcal{A}(i)$. Then we show that starting from *any* point in $\mathcal{A}(1)$ the trajectory will eventually enter $\mathcal{A}(2)$, starting from *any* point in $\mathcal{A}(2)$ the trajectory will eventually enter $\mathcal{A}(3)$ or reenter $\mathcal{A}(1)$, and starting from *any* point in $\mathcal{A}(3)$ the trajectory will eventually enter $\mathcal{A}(1)$ or $\mathcal{A}(2)$. This cycle establishes that $\mathcal{A}$ is a Lyapunov stable set.

The disjointed sets $\mathcal{A}(i), i = 1, 2, 3$ correspond to their lower dimensional counterparts $\mathcal{A}_3(i), i = 1, 2, 3$. However, there are two caveats that need to be taken into account. First, as explained in detail above, considering the full-dimensional case introduces the possibility of a set of states $(\mathbf{p}_H, \mathbf{p}_L, \lambda)$ near $p_s = 1$ with positive measure such that the share of p_{CCs} increases while p_{CDs} decreases. Second, the loci of $\dot{p}_s = 0$ and $\dot{\lambda} = 0$ in Figure 6 tremble along the evolution of the aggregate share $p_{\neg CCs}$. We will take care of the first issue by adjusting the definitions of $\mathcal{A}_3(1)$ and $\mathcal{A}_3(3)$ appropriately. The second issue will be resolved by adjusting the definitions of the underlying sets which form the basis for the construction of $\mathcal{A}(i), i = 1, 2, 3$. Consider the following sets.

$$
\begin{aligned}
A_1 &= \left\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in C \middle| (p_s, \lambda) \in \mathcal{A}_2(\mu, \sigma),\ p_{CDs} < \frac{k_L}{\lambda(1+\alpha)} - \nu_{\neg CCs}\right\} \\
A_2 &= \left\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in C \middle| (p_s, \lambda) \in \mathcal{A}_2(\mu, \sigma),\ p_{CDs} > \frac{k_L}{\lambda(1+\alpha)} + \nu_{\neg CCs}\right\} \\
B &= \left\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in C \middle| (p_s, \lambda) \in \mathcal{A}_2(\mu, \sigma),\ p_{CDs} \geq 1 - \nu\right\} \\
D &= \left\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in C \middle| (p_s, \lambda) \in \mathcal{A}_2(\mu, \sigma),\ \frac{k_L}{\lambda(1+\alpha)} - \nu_{\neg CCs} \leq p_{CDs} \leq \frac{k_L}{\lambda(1+\alpha)} + \nu_{\neg CCs}\right\},
\end{aligned}
$$

where $0 < \nu, \nu_{CCs}, \nu_{\neg CCs} \ll 1$. That is, perturbations from $p_{CCs} = 0$ are denoted ν_{CCS}, the aggregate of all others by $\nu_{\neg CCS}$.

Essentially, the sets B, D are the full-dimensional counterparts of the homonymous sets of Lemma 4 whereas A_1, A_2 are full-dimensional perturbations of A.

To solve the aforementioned problems near $p_s = 1$ we introduce two additional sets:

$$E = \Big\{(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in C \Big| (p_s, \lambda) \in \mathcal{A}_2(\mu, \sigma),\ p_{CDs} + p_{CCs} \geq 1 - \nu_{\neg CCs},\ p_{CCs} \leq \nu_{CCs} + Lp_s \Big\}$$
$$F = \overline{U}_{\hat{\varepsilon}},$$

where $L > 0$ and $U_{\hat{\varepsilon}}$ denotes the $\hat{\varepsilon}$-neighborhood of $\bigcup_{\lambda} \text{CHPE}(\lambda)$ and $\hat{\varepsilon} < \hat{\varepsilon}(x) = \min_{\lambda} \varepsilon(\lambda)$, with $\lambda \in [\lambda^* + 2\mu - x, \lambda^* + 2\mu],\ x \in [0, 2\mu]$ and where $\varepsilon(\lambda)$ denotes the size of the basin of attraction of the CHPE(λ). Note that since the cooperative pooling equilibrium set is asymptotically stable the size of the basin of attraction, i.e., the minimum distance from all population states converging to the equilibrium set is strictly positive for all λ in the λ-support (see Table 1). Thus, $\hat{\varepsilon}(x)$ is well defined and strictly positive for $x < 2\mu$.

The partition $\mathcal{A}(i)$ is defined by setting $\mathcal{A}(1) = A_1 \cap B \cap E \cap F^c$, $\mathcal{A}(2) = D \cap B \cap F^c$, and $\mathcal{A}(3) = (A_2 \cap B) \cup F$. Since $D = A_1^c \cap A_2^c$, $\mathcal{A}(i) \cap \mathcal{A}(j) = \emptyset$ for $i \neq j$. Note that the locus of $\dot{p}_s = 0$ is contained in D for any $\nu_{\neg CCs}$ but might change its position with the composition of $\nu_{\neg CCs}$. In order to make use of Lemma 5 we set $2\mu < \varepsilon$ and choose ν such that $\dot{p}_{CDs} + \dot{p}_{CCs} > 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in B$. As a consequence, it follows that $\dot{p}_{\neg CCs} < 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A} \setminus F$. Since F is a subset of the aggregated basin of attraction of the CHPE we also have $\dot{p}_{\neg CCs} < 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in F$. We now turn to the details of the evolution of p_{CDs} and p_{CCs} which are relevant for states near $p_s \approx 1$. In full generality the payoff difference between CDs and CCs is given by:

$$\lambda\big((p_{DCns} + p_{DDns})\beta - (p_{CCns} + p_{CDns})(m - \alpha)\big) + (1 - \lambda)(1 - p_s)\beta. \tag{33}$$

Thus, for any small perturbation $\nu_{\neg CCs} > 0$ from $p_{CDs} + p_{CCs} = 1$ we have

$$\Pi_H(CDs) - \Pi_H(CCs) \geq -\lambda\nu_{\neg CCs}(m - \alpha) + (1 - \lambda)(1 - p_s)\beta. \tag{34}$$

Therefore CCs can earn strictly higher payoffs than CDs only if $p_s > 1 - \frac{\lambda\nu_{\neg CCs}(m-\alpha)}{(1-\lambda)\beta}$. Hence, for $\lambda \in (\lambda^* - 2\mu, \lambda^* + 2\mu)$ a necessary condition for CCs earning the highest payoffs is

$$p_s > 1 - \frac{(\lambda^* + 2\mu)(m - \alpha)}{(1 - \lambda^* - 2\mu)\beta}\nu_{\neg CCs} \equiv \underline{p}_s \xrightarrow{\nu_{\neg CCs} \to 0} 1. \tag{35}$$

Turning to the dynamics in $\mathcal{A}$ w.r.t. p_s and λ, note that in the full-dimensional case we have

$$\dot{p}_s < 0 \Leftrightarrow p_{CDs} < \frac{k_L}{\lambda(1+\alpha)} - p_{CDns} + p_{DCs} + p_{DCns} \tag{36}$$

$$\Rightarrow \begin{cases} p_{CDs} < \frac{k_L}{\lambda(1+\alpha)} - \nu_{\neg CCs} \Rightarrow \dot{p}_s < 0 \\ p_{CDs} > \frac{k_L}{\lambda(1+\alpha)} + \nu_{\neg CCs} \Rightarrow \dot{p}_s > 0 \end{cases} \tag{37}$$

$$\dot{\lambda} < 0 \Leftrightarrow \sum_{\mathfrak{s}_H \in \mathfrak{S}_H} p_{\mathfrak{s}_H} \Pi_H^f(\mathfrak{s}_H) < p_s \Pi_L^f(s) + (1-p_s)\Pi_L^f(ns) \tag{38}$$

For the same reason as given in the proof of Lemma 4 it holds that for sufficiently small ν, $\dot{\lambda} < 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}$ with $p_s \approx 1$, and $\dot{\lambda} > 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}$ with $p_s \approx 0$. In summary, it holds that $\dot{p}_s < 0$ in $\mathcal{A}(1)$, $\dot{p}_s > 0$ in $\mathcal{A}(3)$, and for both sets $\dot{\lambda} < 0$ if $p_s \approx 1$ and $\dot{\lambda} > 0$ if $p_s \approx 0$. Furthermore, $\dot{p}_{\neg CCs} < 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}$ and $\dot{p}_{CDs} > 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}$ if p_s satisfies (35).

We will now put these pieces together to prove the Lyapunov stability of $\mathcal{A}$. Note that for sufficiently small ν the continuity of the dynamics (1)–(2) and the attractor-property of $\mathcal{A}_2(\mu, \sigma)$ ensure that λ and p_s remain in $\mathcal{A}_2(\mu, \sigma)$ as long as $p_{CDs} \geq 1 - \nu$ holds. Let us start with any state $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}(1)$ with $p_s \approx 1$. If this state happens to be in the basin of attraction of the CHPE (though not being in F) the trajectory will eventually leave this basin since $\dot{\lambda} < 0$ in $\mathcal{A}(1)$ and the size of the basin of attraction converges to zero as λ approaches λ^*. This is because the CHPE exists only for $\lambda > \lambda^*$. Thus, let us assume that $(\mathbf{p}_H, \mathbf{p}_L, \lambda)$ is not in the aggregated basin of attraction of the CHPE. Hence, at this state not only $\dot{\lambda} < 0$ but also $\dot{p}_s < 0$.

We now turn to the details of the evolution of p_{CDs} and p_{CCs} which are relevant for states with $p_s \approx 1$ since in this case p_s might lie above the threshold given by the boundary condition in inequality (35). By definition $\lambda < \frac{k_L}{(1+\alpha)} \frac{1}{p_{CDs} + \nu_{\neg CCs}} \equiv \underline{\lambda}$ in $\mathcal{A}(1)$. As a consequence, for any $\lambda' < \underline{\lambda}$ there is a constant $L > 0$ such that $\left| \frac{\dot{p}_{CCs}}{\dot{p}_s} \right| = \left| \frac{\dot{p}_{CCs}}{\dot{p}_{ns}} \right| < L$ for all $p_s \in (\underline{p}_s, 1]$ and $\lambda \in [\lambda', \lambda - 2\mu]$. This is because the numerator is bounded above by the Lipschitz continuity of the growth-rate functions or by the social innovation assumption. With respect to the denominator $\Pi_L(s) - \Pi(ns) < 0$ and $p_s \approx 1$ imply by the social innovation assumption that $\dot{p}_{ns} > 0$ for all $\lambda < \underline{\lambda}$. Fixing some $\lambda' < \underline{\lambda}$ implies that $\dot{p}_{ns} > 0$ for all $\lambda \in [\lambda', \lambda - 2\mu]$ which in turn implies a lower bound above zero for $\dot{p}_{ns}$. Hence, if initially $p_{CCs} < \nu_{CCs}$ the share for strategy CCs in this region cannot grow above $\nu_{CCs} + Lp_s$. By setting ν_{CCs} and $\nu_{\neg CCs}$ sufficiently low we can ensure that p_{CCs} will not exceed ν. That is, we set $\nu_{CCs}, \nu_{\neg CCs} > 0$ such that $\nu > \nu_{CCs} + L(1 - \underline{p}_s)$.

Since $\mathcal{A}(1) \cap F = \emptyset$ we can indeed choose a λ' below $\underline{\lambda}$. Once p_s falls below $\underline{p}_s$ it follows that $\dot{p}_{CDs} > 0$. Thus, given the stability at the boundary with respect to p_s and λ and $\dot{p}_{\neg CCs} < 0$ any trajectory can leave $\mathcal{A}(1)$ only by entering $\mathcal{A}(2)$ which occurs once p_s has fallen sufficiently such that λ starts to increase, i.e., $\dot{\lambda} < 0$.

Now consider any state in $\mathcal{A}(2)$. If this state is in the aggregated basin of attraction of the CHPE then the trajectory will eventually reenter $\mathcal{A}(1)$ for the same reason as given above. If it is in the aggregated basin of attraction of the CSE then $\dot{\lambda} > 0$ and the trajectory will leave this basin to remain in $\mathcal{A}(2)$ or to enter $\mathcal{A}(3)$. Let us therefore assume that the trajectory is in neither of the two basins of attraction. The only potential attractor remaining is the fixed point p_s^*, λ^* in $\mathcal{A}_2(\mu, \sigma)$. Hence, either the trajectory converges to this fixed point or leaves $\mathcal{A}(2)$. In the latter case the trajectory either reenters $\mathcal{A}(1)$ or enters $\mathcal{A}(3)$.

Now consider any state in $\mathcal{A}(3)$. Again, if this state is in the aggregated basin of attraction of the CHPE then the trajectory will eventually reenter $\mathcal{A}(1)$. For any other state, $\dot{p}_s > 0$ by definition of $\mathcal{A}(3)$. The stability at the boundary with respect to p_s and λ guarantees that either the trajectory enters $\mathcal{A}(2)$ or it passes the threshold of p_s such that p_{CCS} could start to rise. However, we will choose the size of F captured by $\hat{\varepsilon}$ together with the magnitude of the perturbation, $\nu_{\neg CCS}$, such that any trajectory passing this threshold will be in the F.

We now turn to a critical point for the construction of $\mathcal{A}$, i.e., the selection of the appropriate level of x. There are two conditions in addition to $\hat{\varepsilon} < \hat{\varepsilon}(x)$ and $\nu > \nu_{CCs} + L(1 - \underline{p}_s)$ that need to be satisfied.

$$\hat{\varepsilon} > \frac{\lambda^* + 2\mu}{1 - \lambda^* - 2\mu} \frac{m - \alpha}{\beta} \nu_{\neg CCs} \tag{39}$$

$$\hat{\varepsilon} > 1 - \frac{k_L}{1 + \alpha} \frac{1}{\lambda^* + 2\mu - x} + \nu_{\neg CCs} - \nu_{CCs} \tag{40}$$

The first condition ensures that when the system is in $\mathcal{A}(3)$ and enters the region where CCs could grow, i.e., for p_s sufficiently close to one it will be in the aggregated basin of attraction of the CHPEs (see (35)). The second condition ensures that once the system is in F which implies that λ decreases and eventually reaches the boundary of F with respect to λ, i.e. $\lambda^* + 2\mu - x$, then the system will enter the set $\mathcal{A}(2)$. Since $\hat{\varepsilon}(x)$ is a decreasing function with $\hat{\varepsilon}(0) > 0$ and $\hat{\varepsilon}(\mu) = 0$ there exists an x which satisfies conditions (39), (40), and $\nu > \nu_{CCs} + L(1 - \underline{p}_s)$ if we set μ, ν_{CCs}, and $\nu_{\neg CCs}$ sufficiently low.

Thus, we have shown that starting from any state in $\mathcal{A}$ the trajectory stays in $\mathcal{A}$ which establishes its Lyapunov stability. We have also shown that $p_{\neg CCs} \xrightarrow{t \to \infty} 0$ since

$\dot{p}_{\neg CCs} < 0$ for all $(\mathbf{p}_H, \mathbf{p}_L, \lambda) \in \mathcal{A}$ and the only potential attractors in $\mathcal{A}$ are subsets of $\mathcal{A}_2(\mu, \sigma)$. Finally, by definition of $\mathcal{A}$, $\mathcal{A}_2(\mu, \sigma) \subset \mathcal{A}$ and $\mathcal{A}^o$ is open in $X_1 \times X_2 \times X_3$. □

Lemma 6 finishes the proof of the proposition. The central insight of it is that we constructed a set $\mathcal{A}$ which contains $\mathcal{A}_2(\mu, \sigma)$ such that starting from any point within this set the forward orbit is contained in $\mathcal{A}$. Moreover, the strategies CCs and CDs earn the strictly highest payoffs. Hence, as a consequence of payoff monotonicity the perturbation $\nu_{\neg CCs}$ will eventually vanish. Thus, we can concentrate on the dynamics of (1)–(2) under the restriction to $p_{CCs} + p_{CDs} = 1$ for which we have shown in Lemma 4 that for small perturbations from $p_{CDs} = 1$ any trajectory in $\mathcal{A}_3$ converges to $\mathcal{A}_2(\mu, \sigma)$. Taken together with the openness of $\mathcal{A}^o$ in $X_1 \times X_2 \times X_3$ this proves the local stability of the fixed point (Lemma 1) and the (set of) limit cycle(s) (Lemma 3) for the dynamic system (1)–(2). □

To separate the case of a (set of) limit cycle(s) from equilibria which are supported by a single BPE as in Proposition 2 and 3*(i)* we will refer to this equilibrium as the *transitional equilibrium*. Before we take a closer look at the conditions of Proposition 3, Corollary 1 provides comparative statics for λ^* and characterizes the transitional equilibrium of Proposition 3 in terms of type-contingent behavior and signaling. It also characterizes how the dynamic system converges to the equilibria.

Corollary 1 *(1) The equilibrium share of high types λ^* increases in the utility signaling cost for low types (k_L) and decreases in the incentive to defect (α). (2) In the transitional equilibrium, high-type individuals cooperate not only among each other but also with those low-type individuals who signal to be of the high type. (3) While approaching the stable equilibrium point or the transitional equilibrium any trajectory eventually circulates around the fixed point $(\mathbf{p}_H^*, \mathbf{p}_L^*, \lambda^*)$. As a consequence, the share of high types, the frequency of cooperation, and the proportion of low-type individuals who signal to be of the high type oscillate.*

Conditions (i) and (ii) in Proposition 3 separate the two cases regarding the stability of the semi-pooling equilibrium. Under condition (ii) it is unstable and therefore the cyclicity conditions (4)–(5) ensure the existence of a stable (set of) limit cycle(s). If, however, condition (i) holds, the semi-pooling equilibrium is stable. Without further specifying the dynamics, we can only show the local stability of this equilibrium. Therefore, a stable (set of) limit cycle(s) might still exist even for the case of a stable fixed point as in Proposition 3(i). Figures 7 and 8 depict an example of each of the cases (i) and (ii). The dynamics correspond to a modified version of the replicator dynamics.[20]

[20]The details of the modification can be found in Appendix A.

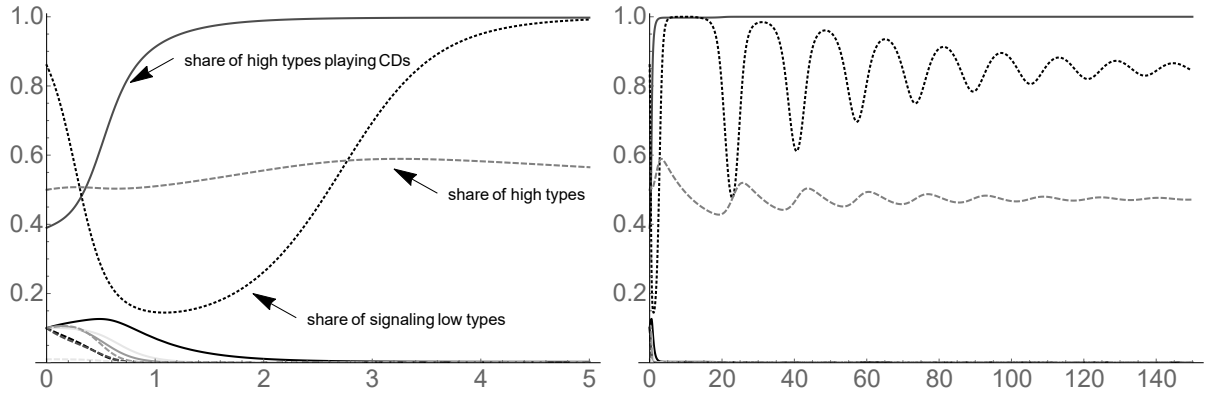

Figure 7: Proposition 3(i) – stable semi-pooling equilibrium: Evolution of population state and share of high types for modified replicator dynamics. Left panel, short run; right panel, long run. Parameters: $m = 2,\ \alpha = 0.9,\ \beta = 0.5,\ k_H = k_H^f = 0.25,\ k_L = k_L^f = 0.9,\ \sigma = 10,\ nn = 200,\ np = 0.1$.

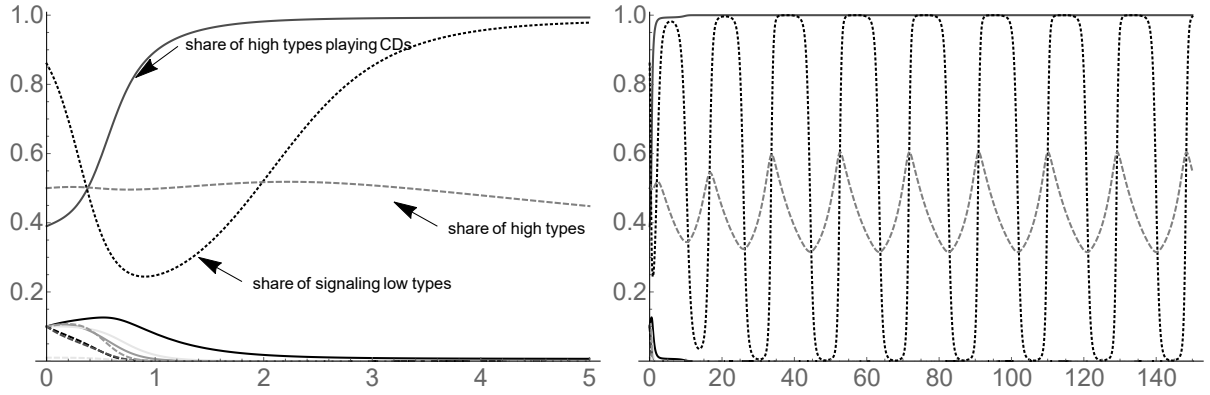

Figure 8: Proposition 3(ii) – stable limit cycle: Evolution of population state and share of high types for modified replicator dynamics. Left panel, short run; right panel, long run. Parameters: $m = 2,\ \alpha = 0.9,\ \beta = 0.5,\ k_H = k_H^f = 0.25,\ k_L = k_L^f = 0.8,\ \sigma = 10,\ nn = 200,\ np = 1$.

Turning to the existence of the transitional equilibrium, the cyclicity condition (4) reveals the importance of the strength of the cooperative preference, measured by m, also for the existence of the transitional equilibrium. More precisely, the higher the m, the more likely this condition is satisfied. Furthermore, if the utility cost of signaling for low types k_L are sufficiently high then conditions (4) and (5a) will be satisfied. If material signaling costs do not differ too much across types, conditions (5b) and (ii) in Proposition 3 will be satisfied. Conditions (5b) and (ii) in Proposition 3 also hold for a sufficiently high β. The Corollary 2 summarizes these insights.

Corollary 2 *If (1) for low types the utility costs of signaling are sufficiently high and do not differ too much across types, or (2) the preference for cooperation is sufficiently strong and the material signaling costs are sufficiently low for both types then the transitional equilibrium exists.*

Note that the conditions in Proposition 2 and Proposition 3 are mutually exclusive, i.e., there is at most one stable inner equilibrium. Contrary to the transitional equilibrium, the equilibrium in Proposition 2 and the semi-pooling equilibrium in Proposition 3 require $\alpha > \beta$, i.e., the temptation payoff needs to exceed the sucker's payoff in absolute terms. Both equilibria also require that high types have a sufficiently high cost advantage in signaling over low types in terms of fitness, i.e., $k_H^f < k_L^f$. Indeed, quite surprisingly the transitional equilibrium is consistent with high types bearing higher signaling cost in terms of fitness than low types. Moreover, if we do not impose any fitness cost for signaling, i.e., $k_H^f = k_L^f = 0$ then the transitional equilibrium still exists but no other cooperative equilibria.

6 Discussion

Since we place our analysis of the emergence of cooperation in social dilemmas in an environment which cannot rely on direct or indirect reciprocity, communication and the implied potential for conditional cooperators to recognize each other are necessary for cooperation to evolve. Thus, we will focus on discussing the nature of signaling costs and their relation across types.[21] We will also discuss the properties of the different cooperative equilibria in light of their potential to account for the three empirical regularities in cooperation in the realm of social dilemmas. Moreover, we discuss the relevance of the social innovation assumption for our results. We end this section with a brief comment on the endogeneity of the strength of the cooperative preference.

Let us consider the relation of signaling costs first. The transitional equilibrium is least restrictive regarding the difference in the material signaling costs, it even allows for a higher fitness cost for high types. That is, the existence of the transitional equilibrium is consistent with both $k_L < k_H$, and $k_L^f < k_H^f$, which we find quite striking. We provide

[21] The literature also discusses alternative modes of communication: There are models (e.g., Güth, 1995 ; Sethi, 1996) which assume that cooperators can simply recognize each other. There is, however, mixed evidence as to what extent humans can unveil incomplete information about cooperative preferences (see Frank et al., 1993; Ockenfels and Selten, 2000; Brosig, 2002). Other models make use of an unsubverted signal like in Arthur Robson's 'secret handshake' model (Robson, 1990). These types of models are prone to what Ken Binmore calls the 'transparent disposition fallacy' (Binmore, 1994).

a numerical example in Appendix A for our modified replicator dynamics. Contrary to this surprising property, in the standard application of signaling theory it is necessary for a separating equilibrium to exist: that types with higher quality bear lower signaling costs. This is not true in our signaling-extended PD. The reason for this is our distinction between the material cost of signaling and the costs in utility terms. This distinction separates the conditions for the existence of the different stable population states which are based on the utility cost from the conditions on the evolutionary (dis)advantage of high types which also refer to material signaling costs.

Regarding the nature of signaling cost, it turned out that the transitional equilibrium, as the only cooperative equilibrium, might exist even if signaling is materially costless. However, a positive share of cooperators in equilibrium requires positive signaling cost in terms of utility for opportunists. A signal which presumably has no or a negligible material cost might consist of giving a smile or some other positive gesture, or a brief chat at the beginning of a pairwise encounter. According to Frank (1988), cooperators are endowed with an advanced emotional system. This system not only provides the motivation for the cooperative behavior, but also enables them to signal their cooperative attitude.[22] Thus, if it is at all possible for opportunists to send the signal, they would have a much higher non-material signaling cost. These costs may refer, for example, to the psychological cost resulting from the effort to fake or hide emotions[23] or cognitive dissonances caused by the discrepancy between preplay-communication and the action taken in the PD. Regarding the latter, there is evidence that lying cost are large and widespread (Abeler et al. 2014). Taken together, this would warrant the assumption of $k_L > k_H$, and in particular $k_L > 0$ which implies a positive share of high types in the equilibria of Proposition 3.

Additional to these psychological costs, a cooperative signal may also be associated with material cost, for instance, if the signal is too time-consuming and results in material opportunity costs. Another example of material signaling cost is provided by Gintis et al. (2001) where individuals can signal their type by the contribution to a multi-player public good game. In general, many acts of courtesy may indeed be understood as a signal for a cooperative attitude. Very often, such acts imply forgoing some (material) advantages for

[22] In a laboratory experiment Brosig (2002) finds that cooperative individuals are somewhat better at predicting their partner's decisions in one-shot prisoner's dilemma games than the individualistic ones. This, of course, is also consistent with a better ability to signal. Scharleman et al. (2001) and Eckel and Wilson (2003), for example, explored the reaction of individuals to seeing the faces of the people with whom they were supposedly interacting. Their results support the potential of smiles as a mechanism to allow subjects to read the intentions of others.

[23] It has been proposed by behavioral scientists that the display of spontaneous positive emotion can serve as a relatively honest signal to identify cooperators. See, for instance, Frank et al. (1993).

the benefit of others. It is *a priori* not clear which type bears higher opportunity costs, leaving the relation of k_L^f and k_H^f ambiguous.

Taken together, our model can capture any kind of costly behavior prior to the PD, which is socially accepted as the appropriate signaling device. The selection of any particular device appears to be a problem of coordination and is beyond the scope of this paper. However, such devices are apparently used.

With respect to the empirical regularities in cooperation, i.e., the heterogeneity of preferences, behavior, and pre-play communication, the semi-pooling equilibrium and the transitional equilibrium of Proposition 3 can account for full heterogeneity. Thus, our model provides a potential explanation for these qualitative properties. The transitional equilibrium is least demanding with respect to model parameters. First, all other cooperative equilibria require that the incentive to defect (α) exceeds the sucker's payoff in absolute terms (β). Second, all other cooperative equilibria only exist if low types face a disadvantage in material signaling cost. Thus, contrary to previous results in the literature, sustaining cooperation in the transitional equilibrium does not hinge on the assumption of some material cost advantage for cooperative types. These properties underline the appeal of the transitional equilibrium. Interestingly, since the transitional equilibrium is constituted by a (set of) limit cycle(s), it also offers a potential explanation for significant self-sustaining oscillations in signaling behavior, in the share of cooperators and therefore in the degree of cooperation in a population (see Figure 8).

Evidence from laboratory experiments on cooperation in social dilemmas provides two insights relevant to our analysis. First, conditional cooperators as identified in public goods experiments correspond to the individuals with the cooperative preference in our model. They account for around 50% of the participants.[24] This empirically identified share of conditional cooperators can be used to partially calibrate our model as it implies a condition for the relation of low types' signaling cost and the temptation to defect. Second, the variation in the incentive structure across various prisoner's dilemma experiments indicates that the share of conditional cooperators decreases in the magnitude of the incentive to defect (α) which is in line with our comparative statics results with respect to λ^* in Proposition 3. This is, for instance, demonstrated by the meta-analysis of Sally (1995) on prisoner's dilemmas where the probability of cooperation is shown to decrease in the temptation to defect. Experiments on social dilemmas with the opportunity for preplay-communication also indicate that (cooperative) signaling is more pronounced among subjects classified as cooperative types. This is consistent with the qualitative

[24] Fischbacher et al. (2001) report 50%; Herrmann and Thöni (2009), 47.7%–60%, and Fischbacher and Gächter (2010), 55%.

property of the semi-pooling and the transitional equilibrium where all high types signal but only a potentially oscillating fraction among low types do so.

Turning to our assumption that humans are capable of profitable social innovation, we want to emphasize that this assumption is sufficient but not necessary to prove our results in full generality for the large class of payoff-monotone and regular selection dynamics. We give an example in Appendix A which satisfies payoff monotonicity and regularity but violates our social innovation assumption. However, there are payoff-monotone and regular dynamics for which, for example, the cooperative high-pooling equilibrium is not stable or no transitional equilibrium exists.

Finally, in our model the size of the parameter m measuring the strength of the preference for cooperation is not driven by evolutionary forces, since no fitness payoff difference depends 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 m are considered, the one with the higher value is more likely to evolve toward a cooperative state. Thus, the population with the stronger preference for cooperation would have an evolutionary edge over the other. Furthermore, 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. Such an analysis could generate insights into the migrational effects on cooperation.

7 Conclusion

This paper aims at shedding light on three persistent patterns attributed to cooperative behavior in social dilemmas: heterogeneity in (1) preferences (coexistence of opportunists and conditional cooperators); (2) behavior (presence of cooperation and defection); and (3) communication. We study an evolutionary model where individuals are able to signal their preference for joint cooperation before engaging in a one-shot prisoner's dilemma. We locally derive the full set of Bayesian equilibria in the signaling-extended prisoner's dilemma and study their dynamic stability. This exhaustive search puts us in a position to study the transition across different Bayesian equilibria.

This dynamic perspective enables the identification of a richer set of cooperative evolutionary equilibria. As our main result we prove the existence of a stable equilibrium which is based on the dynamic interplay of a separating, a semi-pooling, and a pooling equilibria. This equilibrium which we refer to as the transitional equilibrium is constituted by a (set of) limit cycle(s). It is characterized by heterogeneity with respect to

all three dimensions: preferences, behavior, and signaling. More precisely, in the transitional equilibrium conditional cooperators collaborate not only among each other but also with those opportunists who signal they are a cooperator. These characteristics, but also the comparative statics for the transitional equilibrium, are consistent with experimental evidence for cooperation in social dilemmas.

The transitional equilibrium exists under mild conditions, and is least demanding in terms of differences in the signaling costs between conditional cooperators and opportunists. For a transitional equilibrium to exist it suffices, for instance, that the cooperative preference is sufficiently strong and the material signaling costs for both types of individuals are sufficiently low. Importantly, and quite surprisingly, the transitional equilibrium is consistent with conditional cooperators bearing higher signaling costs than opportunists both in terms of fitness and utility. Thus, our model provides an explanation for the emergence of cooperation which does not hinge on some sort of *ad hoc* advantage for cooperative types. Our analysis also revealed that this equilibrium is more likely to exist in societies with a strong cooperative norm among conditional cooperators and with members who are capable of profitable behavioral innovations.

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. Politics may provide support in the solution to the coordination problem of choosing a suitable signaling device. Furthermore, it may try to influence the cost structure of the signaling device in use such that the emergence of cooperation becomes more likely or is accelerated. 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 sucker's payoff may be sufficient to allow for cooperation to evolve. Finally, politics might have some leverage on strengthening the cooperative preference which will also increase the chance for cooperation.

As an additional insight, the oscillatory property of the transitional equilibrium provides an explanation for significant self-sustaining cycles in cooperative behavior, preferences, and signaling in equilibrium. Thus, it offers a micro-founded, alternative account for widely observed repeated historical changes in (social norms of) cooperation in societies which are not driven by social or environmental shocks.[25]. We leave a deeper

[25]Among such exogenous forces are: prominent leaders (e.g., Acemoglu and Jackson, 2014), conflict (e.g., Rohner et al., 2013), natural disasters (e.g., Quarantelli and Dynes, 1977), or technological innovations (e.g., Müller and von Wangenheim, 2017)

exploration of this relation for future research. The application of our theoretical framework to the study of migrational effects on social norm dynamics and its implication for the economic state of a society should also be considered as an insightful line of research.

Acknowledgments

We are grateful to Wolfram Elsner for helpful comments. Financial support is acknowledged to the Federal Ministry of Education and Research for the financial support from BMBF grant number 01UN1018C.

References

[1] Abeler, J., A. Becker, and A. Falk. (2014) "Representative evidence on lying costs." Journal of Public Economics 113(5), 96-104.

[2] Acemoglu, D., M. O. Jackson. (2014). "History, expectations, and leadership in the evolution of social norms." Review of Economic Studies, 82(2), 423-456.

[3] Alexander, R. D. (1987). The Biology of Moral Systems. New York: Aldine de Gruyter.

[4] Andreoni, J., J. H. Miller. (1993). "Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence." Economic Journal, 103(418), 570-585.

[5] Axelrod, R. (1984). The Evolution of Cooperation. New York, NY.: Basic Book. Inc.

[6] Bendixson, I. (1901). "Sur les courbes définies par des équations différentielles." Acta Mathematica, 24(1), 1-88.

[7] Bester, H., W. Güth. (1998). "Is altruism evolutionarily stable?" Journal of Economic Behavior & Organization, 34(2), 193-209.

[8] Binmore, K. (1994). Game Theory and the social contract I: Playing fair. Cambridge: MIT Press.

[9] Brosig, J. (2002). "Identifying cooperative behavior: some experimental results in a prisoner's dilemma game." Journal of Economic Behavior & Organization, 47(3), 275-290.

[10] 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.

[11] Dawes, R. M., J. McTavish, H. Shaklee. (1977). "Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation." Journal of Personality and Social Psychology, 35(1), 1.

[12] Dawes, R. M. (1980). "Social Dilemmas." Annual Review of Psychology, 31(1), 169-193.

[13] Dekel, E., J. C. Ely, O. Yilankaya. (2007). "Evolution of preferences." Review of Economic Studies, 74(3), 685-704.

[14] Eckel, C. C., R. K. Wilson. (2003). "The human face of game theory: Trust and reciprocity in sequential games." Trust and reciprocity: Interdisciplinary lessons from experimental research, 245-274.

[15] Fischbacher, U., S. Gächter, E. Fehr. (2001). "Are people conditionally cooperative? Evidence from a public goods experiment." Economics Letters, 71(3), 397-404.

[16] Fischbacher, U., S. Gächter. (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." American Economic Review, 100(1), 541-56.

[17] Fletcher, J. A., M. Zwick. (2004). "Strong altruism can evolve in randomly formed groups." Journal of Theoretical Biology, 228(3), 303-313.

[18] Frank, M. G., P. Ekman, W. V. Friesen. (1993). "Behavioral markers and recognizability of the smile of enjoyment." Journal of Personality & Social Psychology, 64(1), 83-93.

[19] Frank, R. H., (1988). Passions Within Reason. The Strategic Role of the Emotions. New York: WW Norton & Co.

[20] Frank, R. H., T. Gilovich, D. T. Regan. (1993). "The evolution of one-shot cooperation: an experiment." Ethology and Sociobiology 14, 247-256.

[21] Frey, B. S., S. Meier. (2004). "Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment." American Economic Review, 94(5), 1717-1722.

[22] Fudenberg, D., E. Maskin. (1986). "The folk theorem in repeated games with discounting or with incomplete information." Econometrica, 54(3), 533-554.

[23] Gintis, H., E. A. Smith, S. Bowles. (2001). "Costly signaling and cooperation." Journal of Theoretical Biology, 213(1), 103-119.

[24] Grafen, A. (1990). "Biological signals as handicaps." Journal of Theoretical Biology, 144(4), 517-546.

[25] Güth, W. (1995). "An evolutionary approach to explaining cooperative behavior by reciprocal incentives." International Journal of Game Theory, 24(4), 323-344.

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

[27] 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.

[28] 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.

[29] Guttman, J.M. (2003). "Repeated Interaction and the Evolution of Preferences For Reciprocity." Economic Journal, 113(489), 631-656.

[30] Guttman, J. M. (2013). "On the evolution of conditional cooperation." European Journal of Political Economy, 30, 15-34.

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

[32] Hamilton, W. D. (1964a). The genetical evolution of social behavior. I." Journal of Theoretical Biology, 7(1), 1-16.

[33] Hamilton, W. D. (1964b). "The genetical evolution of social behaviour. II." Journal of Theoretical Biology, 7(1), 17-52.

[34] 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.

[35] Harsanyi, J. C. (1968a). "Games with Incomplete Information Played by 'Bayesian' Players, I-III. Part III. The Basic Probability Distribution of the Game." Management Science, 14(7), 486-502.

[36] Harsanyi, J. C. (1968b). "Games with Incomplete Information Played by 'Bayesian' Players, I-III. Part II. Bayesian Equilibrium Points." Management Science, 14(5), 320-334.

[37] Herrmann, B., C. Thöni. (2009). "Measuring conditional cooperation: a replication study in Russia." Experimental Economics, 12(1), 87-92.

[38] Hopkins, E. (2014). "Competitive Altruism, Mentalizing, and Signaling." American Economic Journal: Microeconomics, 6(4), 272-92.

[39] Huberman, B. A., N. S. Glance. (1993). "Evolutionary games and computer simulations." Proceedings of the National Academy of Sciences, 90(16), 7716-7718.

[40] Huck, S., J. Oechssler. (1999). "The indirect evolutionary approach to explaining fair allocations." Games and Economic Behavior, 28(1), 13-24.

[41] Janssen, M. A. (2008). "Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents." Journal of Economic Behavior & Organization, 65(3), 458-471.

[42] Johnstone, R. A. (1995). "Sexual selection, honest advertisement and the handicap principle." Biological Reviews, 70(1), 65.

[43] Kandori, M. (1992). "Social norms and community enforcement." Review of Economic Studies, 59(1), 63-80.

[44] Keser, C., F. Van Winden. (2000). "Conditional cooperation and voluntary contributions to public goods." Scandinavian Journal of Economics, 102(1), 23-39.

[45] Killingback, T., M. Doebeli, N. Knowlton. (1999). "Variable investment, the continuous prisoner's dilemma, and the origin of cooperation." Proceedings of the Royal Society of London B: Biological Sciences, 266(1430), 1723-1728.

[46] 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.

[47] Lotem, A., M. A. Fishman, L. Stone, L. (2003). "From reciprocity to unconditional altruism through signalling benefits." Proceedings of the Royal Society of London B: Biological Sciences, 270(1511), 199-205.

[48] Macfarlan, S. J., R. Quinlan, M. Remiker. (2013). "Cooperative behaviour and prosocial reputation dynamics in a Dominican village." Proceedings of the Royal Society of London B: Biological Sciences, 280(1761), 20130557.

[49] Maynard Smith, J., G. R. Price. (1973). "The Logic of Animal Conflict." Nature, 246, 15-18

[50] Maynard Smith, J. (1991). "Honest signalling: the Philip Sidney game." Animal Behaviour, 42(6), 1034-1035.

[51] McElreath, R., R. Boyd. (2007). Modeling the Evolution of Social Behavior. Princeton: Princeton University Press.

[52] Müller, S., G. von Wangenheim. (2017). "The impact of market innovations on the dissemination of social norms: the sustainability case." Journal of Evolutionary Economics, 27(4), 663-690.

[53] Nowak, M. A., R. M. May. (1992). "Evolutionary games and spatial chaos." Nature, 359(6398), 826-829.

[54] Nowak, M. A., S. Bonhoeffer, R. M. May. (1994). "Spatial games and the maintenance of cooperation." Proceedings of the National Academy of Sciences, 91(11), 4877-4881.

[55] Nowak, M. A., K. Sigmund. (1998). "Evolution of indirect reciprocity by image scoring." Nature, 393(6685), 573-577.

[56] Ockenfels, A., J. Weimann. (1999). "Types and patterns: an experimental East-West-German comparison of cooperation and solidarity." Journal of Public Economics, 71(2), 275-287.

[57] Ockenfels, A., R. Selten. (2000). "An experiment on the hypothesis of involuntary truth-signalling in bargaining." Games and Economic Behavior, 33, 90-116.

[58] Ostrom, E., J. Walker. (1991). "Communication in a commons: cooperation without external enforcement." Laboratory Research in Political Economy, 287-322.

[59] Panchanathan, K., R. Boyd. (2004). "Indirect reciprocity can stabilize cooperation without the second-order free rider problem." Nature, 432(7016), 499-502.

[60] Possajennikov, A. (2000). "On the evolutionary stability of altruistic and spiteful preferences." Journal of Economic Behavior & Organization, 42(1), 125-129.

[61] Quarantelli, E. L., R. R. Dynes (1977). "Response to social crisis and disaster." Annual Review of Sociology, 3(1), 23-49.

[62] Queller, D. C. (1985). "Kinship, reciprocity and synergism in the evolution of social behaviour." Nature, 318, 366-367.

[63] Rapaport, A., A. M. Chammah. (1965). Prisoner's Dilemma. Ann Arbor: Univ. of Michigan Press.

[64] Robson, A. J. (1990). "Efficiency in evolutionary games: Darwin, Nash and the secret handshake." Journal of Theoretical Biology, 144, 376-396.

[65] Rohner, D., M. Thoenig, F. Zilibotti. (2013). "War signals: A theory of trade, trust, and conflict." Review of Economic Studies, 80(3), 1114-1147.

[66] Roth, A. (1988). "Laboratory experimentation in economics: A methodological overview." Economic Journal, 98, 974-1031.

[67] Sally, D. (1995). "Conversation and cooperation in social dilemmas a meta-analysis of experiments from 1958 to 1992." Rationality and Society, 7(1), 58-92.

[68] Samuelson, L. (1997). Evolutionary games and equilibrium selection. Cambridge: MIT Press.

[69] Samuelson, L., J. Zhang. (1992). "Evolutionary stability in asymmetric games." Journal of Economic Theory, 57(2), 363-391.

[70] Scharlemann, J. P., C. C. Eckel, A. Kacelnik, R. K. Wilson. (2001). "The value of a smile: Game theory with a human face." Journal of Economic Psychology, 22(5), 617-640.

[71] Sethi, R. (1996). "Evolutionary stability and social norms." Journal of Economic Behavior & Organization, 29(1), 113-140.

[72] Spence, M. (1973). "Job market signaling." Quarterly Journal of Economics, 87(3), 355-374.

[73] Trivers, R. L. (1971). "The evolution of reciprocal altruism." Quarterly Review of Biology, 46(1), 35-57.

[74] Wärneryd, K. (2002). "Rent, risk, and replication: Preference adaptation in winner-take-all markets." Games and Economic Behavior, 41(2), 344-364.

[75] Wedekind, C., M. Milinski. (2000). "Cooperation through image scoring in humans." Science, 288(5467), 850-852.

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

[77] Zahavi, A. (1977). "The cost of honesty: further remarks on the handicap principle." Journal of Theoretical Biology, 67(3), 603-605.

A Dynamics and Examples

A.1 Modified replicator dynamics

We modify the standard replicator dynamics by setting:

$$\dot{p}_{\mathfrak{s}_\theta} = \sigma \cdot \left(\Pi_\theta(\mathfrak{s}_\theta) - \overline{\Pi}_\theta\right) \cdot p_{\mathfrak{s}_\theta} + in_{\mathfrak{s}_\theta}(\mathbf{p}_H, \mathbf{p}_L, \lambda) - p_{\mathfrak{s}_\theta} \sum_{\mathfrak{s}'_\theta \in \mathfrak{S}_\theta} in_{\mathfrak{s}'_\theta}(\mathbf{p}_H, \mathbf{p}_L, \lambda) \quad \text{(A.1)}$$

$$\dot{\lambda} = \left(\Pi_H^f - \Pi_L^f\right) \cdot \lambda \cdot (1-\lambda), \quad \text{(A.2)}$$

where $\overline{\Pi}_\theta = \sum_{\mathfrak{s}_\theta \in \mathfrak{S}_\theta} p_{\mathfrak{s}_\theta} \Pi_\theta(\mathfrak{s}_\theta)$ and $in_{s_\theta} : X_1 \times X_2 \times X_3 \to \mathbb{R}$ with $in_{\mathfrak{s}_\theta}(\mathbf{p}_H, \mathbf{p}_L, \lambda) = np \cdot \max\{0, (1 - nn \cdot p_{s_\theta})^3\} \cdot \max\{0, \Pi_{s_\theta} - \max_{s'_\theta \neq s_\theta} \Pi_{s'_\theta}\}$.

A.2 A stable limit cycle with $k_L < k_H$ and $k_L^f < k_H^f$

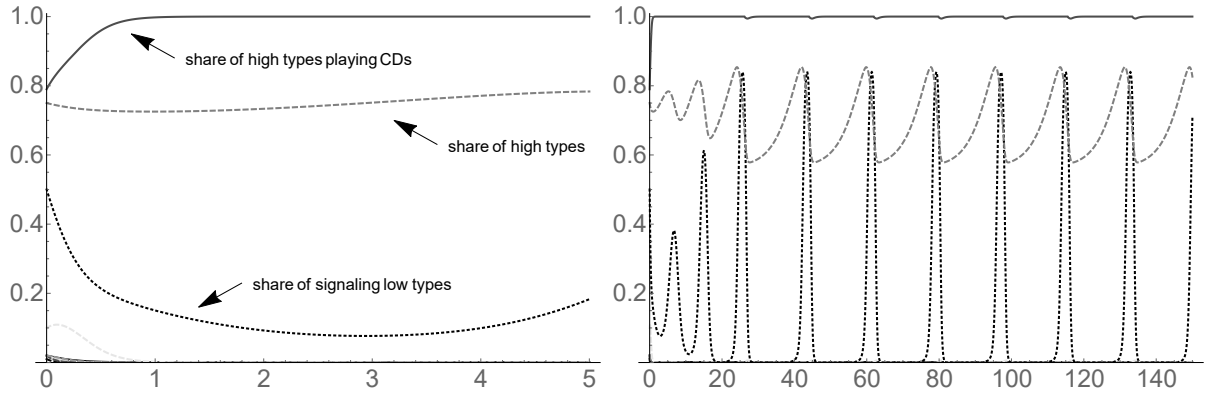

Figure 9: Proposition 3(ii) – stable limit cycle: Evolution of population state and share of high types for modified replicator dynamics. Left panel, short run; right panel, long run. Parameters: $m = 1,\ \alpha = 1/4,\ \beta = 1,\ k_H = 35/32,\ k_L = 15/16,\ k_L^f = 1/8,\ k_H^f = 9/16,\ \sigma = 20,\ nn = 200,\ np = 0.1$.

A.3 A stable limit cycle without social innovations

$$\dot{p}_{\mathfrak{s}_\theta} = \sigma \cdot \left(\Pi_\theta(\mathfrak{s}_\theta) - \overline{\Pi}_\theta\right) \cdot p_{\mathfrak{s}_\theta} \cdot e^{(p^*_{s_\theta} - p_{s_\theta})(\lambda - \lambda^*)} \quad \text{(A.3)}$$

$$\dot{\lambda} = \left(\Pi_H^f - \Pi_L^f\right) \cdot \lambda \cdot (1-\lambda), \quad \text{(A.4)}$$

where $p^*_{s_\theta} = p^*_s$ for $s_L = s$ and zero otherwise.

For Online Publication

B Separating and Pooling Equilibria – Existence and Stability

In this section we derive the separating and the pooling equilibria with respect to the signaling behavior of the signaling-extended PD for a given share of high types $\lambda \in (0,1)$. That is, we study for a given λ the existence and asymptotic stability of (sets of) population states $(\mathbf{p}_H, \mathbf{p}_L)$ for the dynamics given by equation (1). We will also study the existence and stability of semi-pooling equilibria at λ^*.

To prove stability or instability of an equilibrium we will rely on phase diagrams. We will prove instability by arguing that the system cannot be Lyapunov-stable. In case of an equilibrium point in the interior of the support of the equilibrium the involved strategies earn strictly higher payoffs then non-equilibrium strategies. Small perturbations will not alter this property. Payoff-monotone dynamics will decrease the share of the non-equilibrium strategies. Hence, in that case, to analyzing the stability properties it suffices to consider the involved equilibrium strategies and whether the dynamics will reestablish the equilibrium values given a small perturbation. At the boundaries of the support of an equilibrium point a non-equilibrium strategy will earn the same profits as the equilibrium strategies. In that case these strategies need to be included in the analysis. However, with respect to all other strategies the previous argument still applies.

Note that by the assumption of independence of the evolution of strategy shares across types, the expected payoff for each type-contingent strategy is additively separable in the payoffs for the two types. We will make use of this property when discussing the stability of equilibria. That is, to prove the (in)stability of a certain equilibrium (set) we will consider contingent-wise changes of behavior. Furthermore, we will write expected payoff as linear combination of type-contingent payoffs. That is, if a type-contingent strategy $s \in \mathscr{S}$ is identified with the pair $(\mathfrak{s}_H, \mathfrak{s}_L)$ of type-contingent strategies then $\Pi_s = \lambda \Pi_H(\mathfrak{s}_H) + (1-\lambda)\Pi_L(\mathfrak{s}_L)$. Finally, in the phase diagrams thick solid lines or points correspond to equilibrium sets or points, respectively. Iso-profit lines are depicted by thick dotted lines.

Before we restrict to the different cases of equilibria separately we present the payoffs $\Pi_H(\mathfrak{s}_H)$ and $\Pi_L(\mathfrak{s}_L)$ for all type-contingent strategies in full generality.

Type-contingent payoffs for a generic population state:

$$\Pi_H(CCs) = \lambda\Big((1+m)\sum p_{C\cdot\cdot,\cdot} + (-\beta)\sum p_{D\cdot\cdot,\cdot}\Big) + (1-\lambda)(-\beta) - k_H$$
$$\Pi_H(CDs) = \lambda\Big((1+m)\sum p_{C\cdot s,\cdot} + (1+\alpha)\sum p_{C\cdot ns,\cdot} - \beta\sum p_{D\cdot s,\cdot}\Big) - (1-\lambda)\beta\sum p_{\cdot\cdot\cdot,s} - k_H$$
$$\Pi_H(DCs) = \lambda\Big((1+\alpha)\sum p_{C\cdot s,\cdot} + (1+m)\sum p_{C\cdot ns,\cdot} - \beta\sum p_{D\cdot ns,\cdot}\Big) - (1-\lambda)\beta\sum p_{\cdot\cdot\cdot,ns} - k_H$$
$$\Pi_H(DDs) = \lambda(1+\alpha)\sum p_{C\cdot\cdot,\cdot} - k_H$$
$$\Pi_H(CCns) = \lambda\Big((1+m)\sum p_{\cdot C\cdot,\cdot} + (-\beta)\sum p_{\cdot D\cdot,\cdot}\Big) + (1-\lambda)(-\beta) - k_H$$
$$\Pi_H(CDns) = \lambda\Big((1+m)\sum p_{\cdot Cs,\cdot} + (1+\alpha)\sum p_{\cdot Cns,\cdot} - \beta\sum p_{\cdot Ds,\cdot}\Big) - (1-\lambda)\beta\sum p_{\cdot\cdot\cdot,s} - k_H$$
$$\Pi_H(DCns) = \lambda\Big((1+\alpha)\sum p_{\cdot Cs,\cdot} + (1+m)\sum p_{\cdot Cns,\cdot} - \beta\sum p_{\cdot Dns,\cdot}\Big) - (1-\lambda)\beta\sum p_{\cdot\cdot\cdot,ns} - k_H$$
$$\Pi_H(DDns) = \lambda(1+\alpha)\sum p_{\cdot C\cdot,\cdot} - k_H$$
$$\Pi_L(s) = \lambda(1+\alpha)\sum p_{C\cdot\cdot,\cdot}$$
$$\Pi_L(ns) = \lambda(1+\alpha)p_{\cdot C\cdot,\cdot}$$

B.1 Separating Equilibria

A separating equilibrium is defined by $\sum p_{\cdot\cdot s,ns} = 1$ or $\sum p_{\cdot\cdot ns,s} = 1$.

B.1.1 High types signal, low types do not signal: $\sum p_{\cdot\cdot s,ns} = 1$

Existence

Note that for $\lambda \in (0,1)$ and $p_{CCs,ns}, p_{CDs,ns}, p_{DCs,ns}, p_{DDs,ns} > 0$, it follows that $\Pi_{CCs,ns} < \Pi_{CDs,ns}$, $\Pi_{DCs,ns} < \Pi_{DDs,ns}$, $\Pi_{CCns,ns} < \Pi_{CDns,ns}$, and $\Pi_{DCs,ns} < \Pi_{DDs,ns}$. After deletion of these strictly dominated strategies, payoffs of the remaining strategies are given by:

$$\Pi_{CDs,ns} = \lambda\big[\lambda[p_{CDs,ns}(1+m) - p_{DDs,ns}\beta] - k_H\big] \;;\; \Pi_{DDs,ns} = \lambda\big[\lambda(p_{CDs,ns})(1+\alpha) - k_H\big]$$

$$\Pi_{CDns,ns} = \lambda\big[-(p_{CDs,ns} + p_{DDs,ns})\beta\big] \;;\; \Pi_{DDns,ns} = 0$$

For a separating equilibrium where high types send the signal and low types do not, only two undominated strategies are left, CDs, ns and DDs, ns , i.e. $p_{CDs,ns} + p_{DDs,ns} = 1$. Thus, $CDns, ns$ would earn strictly less than $DDns, ns$.

1. Let us first analyze the case $p_{CDs,ns} = 1$. In that case the following three conditions are necessary and sufficient for this to constitute an equilibrium:

(i) $\Pi_{CDs,ns} > \Pi_{DDs,ns}$, which is always satisfied, because of $m > \alpha$.

(ii) $\Pi_{CDs,ns} \geq \Pi_{DDns,ns} \Leftrightarrow \lambda \geq \frac{k_H}{1+m}$.

(iii) $\Pi_{CDs,ns} \geq \Pi_{CDs,s} \Leftrightarrow \lambda \leq \frac{k_L}{1+\alpha}$.

Thus, the three conditions are equivalent to $\frac{k_H}{1+m} \leq \lambda \leq \frac{k_L}{1+\alpha}$. Note that for $k_H < \bar{k}_H$ the λ-support for this equilibrium is not empty.

2. Let us now analyze the case $p_{DDs,ns} = 1$. In that case $DDns, ns$ would earn strictly higher payoffs, hence such an equilibrium cannot exist.

3. Finally, we consider a mixed equilibrium, i.e., $p_{CDs,ns} + p_{DDs,ns} = 1$. In that case the following three conditions are necessary and sufficient for this to be an equilibrium:

(i) $\Pi_{CDs,ns} = \Pi_{DDs,ns} \Leftrightarrow p_{CDs,ns} = \frac{\beta}{\beta+m-\alpha}$.

(ii) $\Pi_{DDs,ns} \geq \Pi_{DDns,ns} \Leftrightarrow p_{CDs,ns} \geq \frac{\lambda\beta+k_H}{\lambda(1+m+\beta)}$.

(iii) $\Pi_{CDs,ns} \geq \Pi_{CDs,s} \Leftrightarrow p_{CDs,ns} \leq \frac{k_L}{\lambda(1+\alpha)}$.

At $p_{CDs,ns} = \frac{\beta}{\beta+m-\alpha}$ the last two conditions are equivalent to $\frac{\beta+m-\alpha}{\beta(1+\alpha)}k_H \leq \lambda \leq \frac{\beta+m-\alpha}{\beta(1+\alpha)}k_L$.

Stability

Case 1: This equilibrium is certainly stable in the interior range $\frac{k_H}{1+m} < \lambda < \frac{k_L}{1+\alpha}$ since all payoff inequalities hold strictly. At the upper bound $\lambda = \frac{k_L}{1+\alpha}$, the strategies CDs, ns and CDs, s earn the same profits, i.e., low types are indifferent between signaling and not sending the signal. Consider a small perturbation such that CDs, s is played with a small positive probability. To reestablish $p_{CDs,ns} = 1$, the share of high types playing CDs must decrease, because $\Pi_L(s) - \Pi_L(ns) = \lambda p_{CDs}(1+\alpha) - k_L$. However, for small perturbations CDs is still dominant for high types. Hence, CDs persists as part of the equilibrium strategy and there is no force reestablishing the non-signaling contingency for low types. Thus, the separating equilibrium is not stable at the upper bound. A similar argument establishes that it is also not stable at the lower bound. At the lower bound $\frac{k_H}{1+m}$, the strategies CDs, ns and $DDns, ns$ earn the same profits, i.e., high types are indifferent between cooperating and incurring the cost of the signal on the one hand, and defecting and no signaling on the other. Consider a random drift, such that $p_{DDns,ns} > 0$. This drift will lower profits for CDs, ns and leaves profits for $DDns, ns$ unchanged. Hence, the equilibrium will not be restored. In other words, this equilibrium is not stable at $\lambda = \frac{k_H}{1+m}$.

Case 3: In this equilibrium with $p_{CDs,ns} = \frac{\beta}{\beta+m-\alpha}$, $p_{DDs,ns} = \frac{m-\alpha}{\beta+m-\alpha}$ we have the following differences in type-specific payoffs:

$$\Pi_H(CDs) - \Pi_H(DDs) \geq 0 \Leftrightarrow p_{CDs} \geq \frac{\beta - (1-\lambda)\beta p_{ns}}{\lambda(m-\alpha-\beta)}$$
$$\Pi_L(ns) - \Pi_L(s) \geq 0 \Leftrightarrow p_{CDs} \leq \frac{k_L}{\lambda(1+\alpha)},$$

we obtain the following phase diagram. Note that for the support of that equilibrium $\frac{k_L}{\lambda(1+\alpha)} > \frac{k_H+\lambda\beta-(1-\lambda)\beta p_{ns}}{\lambda(m-\alpha-\beta)}$ holds. Note further that the upper bound of the support $\lambda \leq \frac{m-\alpha+\beta}{\beta}\frac{k_L}{1+\alpha}$ implies $\frac{\beta}{m-\alpha+\beta} \leq \frac{k_L}{\lambda(1+\alpha)}$. Additionally, $\frac{k_H+\lambda\beta}{\lambda(1+m+\beta)} \leq \frac{\beta}{m-\alpha+\beta} \Leftrightarrow \frac{k_H}{\lambda(1+\alpha)} \leq \frac{\beta}{m-\alpha+\beta}$. As the diagram clearly indicates, this equilibrium is unstable for all λ in the support.

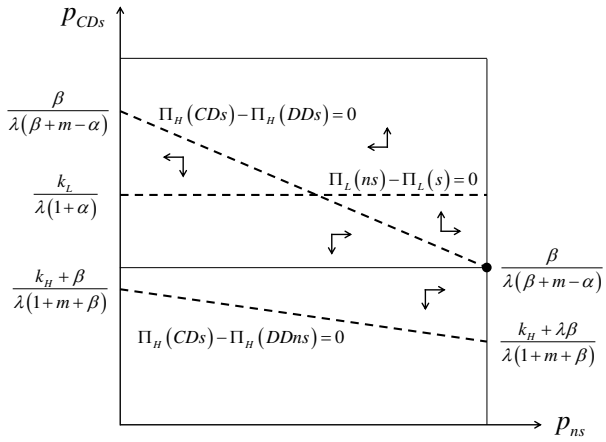

Figure 10: Dynamics for a separating equilibrium where high types signal and low types don't.

B.1.2 High types do not signal, low types signal: $\sum p_{..ns,s} = 1$

Existence

Let us again first study the signaling contingency for high types. Note that for $\lambda \in (0,1)$ and $p_{CCns,s}, p_{CDns,s}, p_{DCns,s}, p_{DDns,s} > 0$, it follows that $\Pi_{CCs,s} < \Pi_{DCs,s}$, $\Pi_{CDs,s} < \Pi_{DDs,s}$, $\Pi_{CCns,s} < \Pi_{DCns,s}$, and $\Pi_{CDns,s} < \Pi_{DDns,s}$. After deletion of these strictly dominated strategies, payoffs for low types to signal is $-k_L$, whereas signaling yields an expected payoff of $\lambda p_{DCns,s}(1+\alpha)$. Thus, strategies that imply no signaling for low types generate strictly higher payoffs. Hence, such an equilibrium cannot exist.

B.2 Pooling Equilibria

B.2.1 High types and low types do not signal: $\sum p_{\cdot\cdot ns,ns} = 1$

Existence

Let us again first study the signaling contingency for high types. Note that in a pooling equilibrium where nobody sends the signal, $CCns, ns$ and $DCns, ns$ ($CDns, ns$ and $DDns, ns$) will always earn the same profits irrespective of the chosen signal and the particular composition. We will denote profits by $\Pi_{CCns,ns/DCns,ns}$, and $\Pi_{CDns,ns/DDns,ns}$. Since those pairs are indistinguishable we only have to consider the following cases:

1. $p_{CCns,ns} + p_{DCns,ns} = 1$.

   (i) In that case $\Pi_{CCns,ns/DCns,ns} > \Pi_{CCs,ns/DCs,ns}$, $\Pi_{CDns,ns/DDns,ns} > \Pi_{CDs,ns/DDs,ns}$, and $\Pi_{CCns,ns/DCns,ns} > \Pi_{CCns,s/DCns,s}$, because $p_{CDns,ns} = 0$.

   (ii) $\Pi_{CCns,ns/DCns,ns} \geq \Pi_{CDns,ns/DDns,ns} \Leftrightarrow \lambda \geq \frac{\beta}{\beta+m-\alpha}$. Because of $\Pi_{CCns,ns/DCns,ns} > \Pi_{CDns,ns/DDns,ns} > \Pi_{CDs,ns/DDs,ns}$, the condition $\lambda \geq \frac{\beta}{\beta+m-\alpha}$ is necessary and sufficient.

2. $p_{CDns,ns} + p_{DDns,ns} = 1$.

   (i) In that case $\Pi_{CCns,ns/DCns,ns} < \Pi_{CDns,ns/DDns,ns}$, because $p_{CCns,ns} + p_{CDns,ns} = 0$.

   (ii) $\Pi_{CDns,ns/DDns,ns} \geq \Pi_{CDs,ns/DDs,ns} \Leftrightarrow \lambda p_{CDns,ns} \leq \frac{k_H}{1+\alpha}$.

   (iii) $\Pi_{CDns,ns/DDns,ns} \geq \Pi_{CCs,ns/DCs,ns} \Leftrightarrow \lambda p_{CDns,ns} \leq \frac{\beta+k_H}{1+m+\beta}$.

   (iv) $\Pi_{CDns,ns/DDns,ns} \geq \Pi_{CDns,s/DDns,s} \Leftrightarrow \lambda p_{CDns,ns} \leq \frac{k_L}{1+\alpha}$.

   Note that, $\frac{k_L}{1+\alpha} > \frac{k_H}{1+\alpha}$. Thus, (ii) and (iii) are necessary and sufficient.

3. $p_{CDns,ns} + p_{DDns,ns} + p_{CCns,ns} + p_{DCns,ns} = 1$.

   (i) In that case all no-signaling strategies earn the same payoff: $\lambda\big[\lambda[(p_{CCns,ns} + p_{CDns,ns})(1+m+\beta)-\beta]+(1-\lambda)\big[\lambda(p_{CCns,ns}+p_{DCns,ns})(1+\alpha)\big] = \lambda\big[\lambda(p_{CCns,ns}+ p_{CDns,ns})(1+\alpha)\big] + (1-\lambda)\big[\lambda(p_{CCns,ns} + p_{DCns,ns})(1+\alpha)\big] \Leftrightarrow \lambda(p_{CCns,ns} + p_{DCns,ns}) = \frac{\beta}{\beta+m-\alpha}$.

   (ii) $\Pi_{CCns,ns/DCns,ns} \geq \Pi_{CCs,ns/DCs,ns} \Leftrightarrow \lambda(p_{CDns,ns} - p_{DCns,ns}) \leq \frac{k_H}{1+m+\beta}$.

   (iii) $\Pi_{CDns,ns/DDns,ns} \geq \Pi_{CDs,ns/DDs,ns} \Leftrightarrow \lambda(p_{CDns,ns} - p_{DCns,ns}) \leq \frac{k_H}{1+\alpha}$.

(iv) $\Pi_{CDns,ns/DDns,ns/CCns,ns/DCns,ns} \geq \Pi_{CDns,s/DDns,s/CCns,s/DCns,s} \Leftrightarrow \lambda(p_{CDns,ns} - p_{DCns,ns}) \leq \frac{k_L}{1+\alpha}$.

Note that, because of $k_L \geq k_H$ and $m > \alpha$, (ii) implies (iii) and (iv). Thus, such an equilibrium exists if and only if $\lambda(p_{CCns,ns} + p_{DCns,ns}) = \frac{\beta}{\beta+m-\alpha}$ and $\lambda(p_{CDns,ns} - p_{DCns,ns}) \leq \frac{k_H}{1+m+\beta}$.

Stability

Case 1: The equilibrium set is stable for $\lambda > \frac{\beta}{m-\alpha+\beta}$ since all inequalities hold strictly, i.e., for any small perturbation the equilibrium strategies earn strictly more than any other strategy. Note that the pre-perturbation shares are not necessarily reestablished, but that the sum of their shares equals unity. At the boundary $\lambda = \frac{\beta}{m-\alpha+\beta}$ there are too few high types and the agents become indifferent between cooperation and defection, i.e., $\Pi_{CCns,ns/DCns,ns} = \Pi_{CDns,ns/DDns,ns}$. Note that it is still a strictly best response not to signal contingent on being a low type. Given the following differences in type-specific payoffs:

$$\Pi_H(CCns) - \Pi_H(CDns) \geq 0 \Leftrightarrow p_{CCns} \geq \frac{\beta}{\lambda(m-\alpha)+\beta} - p_{DCns},$$

we obtain the following phase diagram.

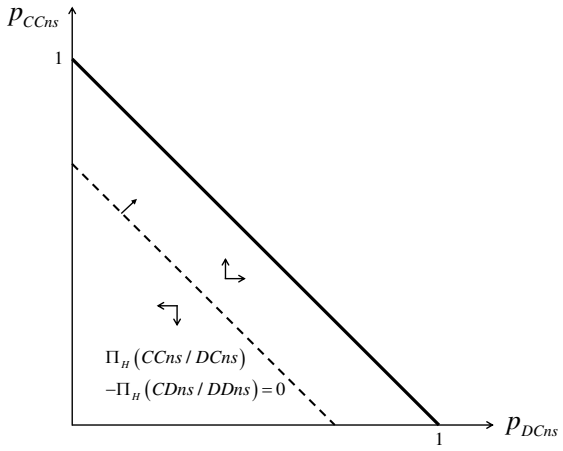

Figure 11: Dynamics I for a pooling equilibrium where no type signals.

Note that at $\lambda = \frac{\beta}{m-\alpha+\beta}$ a perturbation from $CCns, ns$ towards $DDns, ns$ decreases the payoffs for the equilibrium strategies strictly more than for $DDns, ns$ and decreases profits for all other strategies weakly more, i.e., those strategies still earn

strictly less than $DDns, ns$, and the share of $DDns, ns$ increases. Hence, there is no force reestablishing the equilibrium set. Note that the iso-profit line is shifted toward the boundary as λ approaches the lower limit of the support $\frac{\beta}{m-\alpha+\beta}$. As the diagram clearly indicates, this equilibrium is stable for all $\lambda > \frac{\beta}{m-\alpha+\beta}$ in the support.

Case 2: The equilibrium set is stable for $p_{CDns,ns} < \frac{1}{\lambda}\min\{\frac{k_H+\beta}{1+m+\beta}, \frac{k_H}{1+\alpha}\}$ since all inequalities strictly hold, i.e., for any small perturbation the equilibrium strategies earn strictly more than any other strategy.

Given the following differences in type-specific payoffs:

$$\Pi_H(CDns) - \Pi_H(DDns) = -\beta(1-\lambda)p_s \leq 0$$
$$\Pi_L(ns) - \Pi_L(s) = k_L - \lambda(1+\alpha)p_{CDns} \geq 0 \Leftrightarrow p_{CDns} \leq \frac{k_L}{\lambda(1+\alpha)},$$

we obtain the following phase diagram. As the diagram clearly indicates, this equilibrium set is stable for all λ in the support.

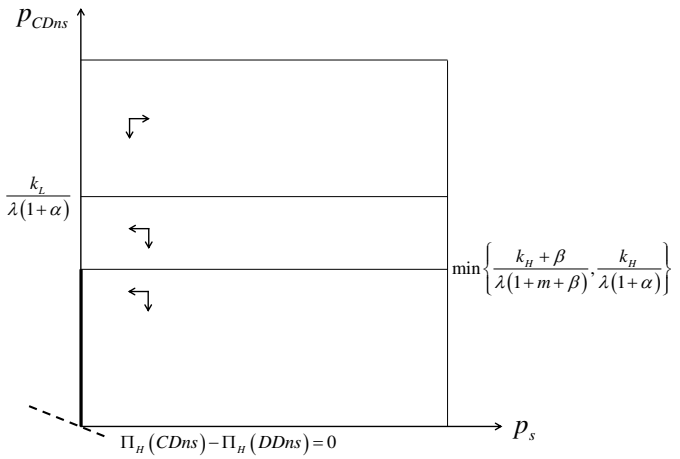

Figure 12: Dynamics II for a pooling equilibrium where no type signals.

Case 3: Observe that the payoffs for the equilibrium strategies can be written as linear functions in $p_{CCns,ns} + p_{DCns,ns}$.

Given the following differences in type-specific payoffs:

$$\Pi_H(CCns) - \Pi_H(CDns) \geq 0 \Leftrightarrow p_{DCns} \leq \frac{\beta}{\lambda(m-\alpha+\beta)} - p_{CCns},$$

we obtain the following phase diagram. All other payoff differences of equilibrium strategies vanish. The figure incorporates the two conditions for existence, i.e., $\lambda(p_{CCns,ns}+p_{DCns,ns}) = \frac{\beta}{\beta+m-\alpha}$ and $\lambda(p_{CDns,ns}-p_{DCns,ns}) \leq \frac{k_H}{1+m+\beta}$. As the diagram clearly indicates, this equilibrium set is unstable.

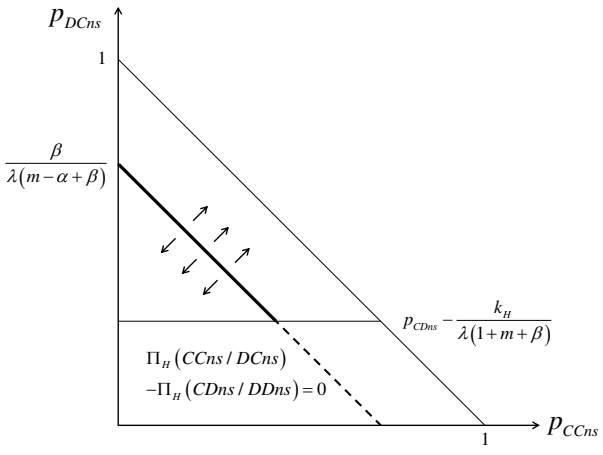

Figure 13: Dynamics III for a pooling equilibrium where no type signals.

B.2.2 High types and low types signal: $\sum p_{\cdot\cdot s,s} = 1$

Existence

Let us again first study the signaling contingency for high types. Note that in a pooling equilibrium where everybody sends the signal, CCs, s and CDs, s (DCs, s and DDs, s) will always earn the same profits irrespective of the chosen signal and the particular composition. We will denote profits by $\Pi_{CCs,s/CDs,s}$, and $\Pi_{DCs,s/DDs,s}$. Since those pairs are indistinguishable we only have to consider the following cases:

1. $p_{CCs,s} + p_{CDs,s} = 1$.

   (i) $\Pi_{CCs,s/CDs,s} \geq \Pi_{CCns,s/CDns,s} \Leftrightarrow \lambda p_{CD_{s,s}} \geq \frac{k_H}{1+m+\beta}$.

   (ii) $\Pi_{CCs,s/CDs,s} \geq \Pi_{DCs,s/DDs,s} \Leftrightarrow \lambda \geq \frac{\beta}{\beta+m-\alpha}$.

   (iii) $\Pi_{CCs,s/CDs,s} \geq \Pi_{DCns,s/DDns,s} \Leftrightarrow \lambda \geq \frac{\beta+k_H}{\beta+m-\alpha+p_{CDs,s}(1+\alpha)}$.

   (iv) $\Pi_{CCs,s/CDs,s} \geq \Pi_{CCs,ns/CDs,ns} \Leftrightarrow \lambda \geq \frac{k_L}{p_{CDs,s}(1+\alpha)}$.

   Note that (iv) implies (i), for (iv) to be satisfied a strictly positive share needs to play CDs, s. Furthermore, (ii) and (iv) imply (iii). Hence, since $p_{CDs,s} \in [0, 1]$ such an equilibrium exists for $\lambda \geq \max\{\frac{k_L}{1+\alpha}, \frac{\beta}{\beta+m-\alpha}\}$.

2. $p_{DCs,s}+p_{DDs,s}=1$. This cannot constitute an equilibrium, because not sending the signal contingent on being a low type yields strictly higher payoffs.

3. $p_{CDs,s}+p_{DDs,s}+p_{CCs,s}+p_{DCs,s}=1$.

   In that case all signaling strategies earn the same payoff: $\lambda\big[\lambda(p_{CCs,s}+p_{CDs,s})(1+m+\beta)-\beta-k_H\big]+(1-\lambda)\big[\lambda(p_{CCs,s}+p_{CDs,s})(1+\alpha)-k_L\big]=\lambda\big[\lambda(p_{CCs,s}+p_{CDs,s})(1+m+\beta)-\beta-k_H\big]+(1-\lambda)\big[\lambda(p_{CCs,s}+p_{CDs,s})(1+\alpha)-k_L\big]\Leftrightarrow\lambda(p_{CCs,s}+p_{CDs,s})=\frac{\beta}{\beta+m-\alpha}$. The following conditions are necessary and sufficient for existence.

   (i) $\Pi_{CCs,s/CDs,s}\geq\Pi_{CCns,s/CDns,s}\Leftrightarrow\lambda(p_{CDs,s}-p_{DCs,s})\geq\frac{k_H}{1+m+\beta}$.

   (ii) $\Pi_{DCs,s/DDs,s}\geq\Pi_{DCns,s/DDns,s}\Leftrightarrow\lambda(p_{CDs,s}-p_{DCs,s})\geq\frac{k_H}{1+\alpha}$.

   (iii) $\Pi_{CDs,s/DDs,s/CCs,s/DCs,s}\geq\Pi_{CDs,ns/DDs,ns/CCs,ns/DCs,ns}\Leftrightarrow\lambda(p_{CDs,s}-p_{DCs,s})\geq\frac{k_L}{1+\alpha}$.

   Note that (ii) implies (i), and (iii) implies (ii). Hence, such an equilibrium exists if and only if $\lambda(p_{CCs,s}+p_{CDs,s})=\frac{\beta}{\beta+m-\alpha}$,and $\lambda(p_{CDs,s}-p_{DCs,s})\geq\frac{k_L}{1+\alpha}$.

Stability

Case 1: Note that at $p_{CDs}=\frac{k_L}{\lambda(1+\alpha)}$ low types are indifferent between signaling and no signaling. As soon as low types start not to signal which is ensured by our social innovation assumption, CDs earns strictly higher payoffs than CCs such that the incentive for low types to signal will be restored. However, at $\lambda=\frac{k_L}{(1+\alpha)}$, p_{CDs} equals 1 and therefore cannot increase. Thus, this equilibrium is unstable at the upper bound $\frac{k_L}{(1+\alpha)}$. If $\lambda=\frac{\beta}{m-\alpha+\beta}$, then high types given a received signal are indifferent between cooperative and defective play. For a small increase in the share $p_{DCs,s}+p_{DDs,s}$, the profits for the equilibrium strategies will decline more than the profits for $DCs,s/DDs,s$. Since the equilibrium strategies and $DCs,s/DDs,s$ will still earn higher profits than any other, there is no force bringing back the system to $p_{CCs,s}+p_{CDs,s}=1$. Hence, the equilibrium is unstable at $\lambda=\frac{\beta}{m-\alpha+\beta}$.

Given the following differences in type-specific payoffs:

$$\Pi_H(CCs)-\Pi_H(CDs)=-\beta(1-\lambda)p_{ns}\leq 0$$
$$\Pi_L(ns)-\Pi_L(s)\geq 0\Leftrightarrow p_{CDs}\leq\frac{k_L}{\lambda(1+\alpha)},$$

we obtain the following phase diagram.

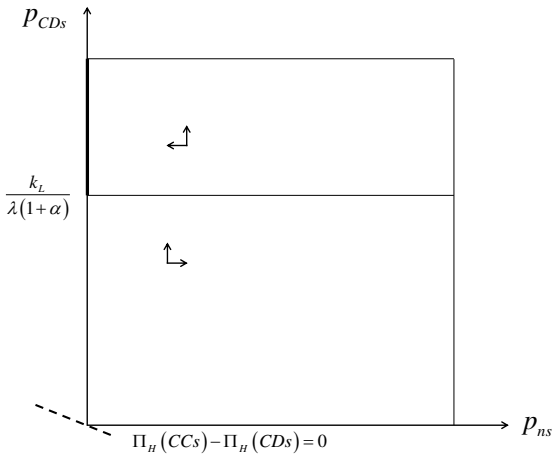

Figure 14: Dynamics for a pooling equilibrium where both types signals.

As the diagram clearly indicates, this equilibrium set is stable for $\lambda > \max\{\frac{k_L}{p_{CDs,s}(1+\alpha)}, \frac{\beta}{\beta+m-\alpha}\}$.

Case 2: Let $p_{CCs} + p_{CDs} = x$ and $p_{DCs} + p_{DDs} = y$. Note that $y = 1 - x$, because of $p_{CCs} + p_{CDs} + p_{DCs} + p_{DDs} = 1$. Thus, we can write payoffs for high types as: $\Pi_H(CCs) = \Pi_H(CDs) = \lambda x(1 + m + \beta) - \beta - k_H$, and $\Pi_H(DCs) = \Pi_H(DDs) = \lambda x(1 + \alpha) - k_H$. Given any perturbation that violates the equilibrium condition $\lambda(p_{CCs} + p_{CDs}) = \frac{\beta}{m-\alpha+\beta}$ the equilibrium set will not be restored because $\Pi_H(CCs/CDs) - \Pi_H(DCs/DDs) = \lambda x(m - \alpha + \beta) - \beta \geq 0 \Leftrightarrow \lambda x \geq \frac{\beta}{m-\alpha+\beta}$. Thus, an increase in x is self-enforcing.

B.3 Semi-Pooling Equilibria at λ^*

We can focus on the strategies involved in the cooperative separating and in the cooperative high-pooling equilibrium, i.e., the strategies CDs, ns, CCs, ns, and CDs, s. This is because we show in the proof of Proposition 3 that any perturbation from $p_{CCs} + p_{CDs} = 1$ will eventually vanish. Since for these strategies high types always signal a semi-pooling equilibrium can only arise in case of pooling among low types.

Under the restriction to $p_{CDs,ns} + p_{CCs,ns} + p_{CDs,s} = 1$ the difference in profits for low types at λ^* is given by $\Pi_L(ns) - \Pi_L(s) = \lambda^*(1 + \alpha - k_L)(1 - p_{CDs,ns} - p_{CDs,s})$. Thus, indifference of low types implies $p_{CDs,ns} + p_{CDs,s} = 1$. Evaluating the profits of high types under this restriction reveals that all non-signaling strategies except DDs with zero profits earn strictly negative payoffs. The remaining three signaling strategies at λ^* earn

the following profits: $\Pi(CCs) = \lambda^*(1+m) - \beta(1-\lambda^*) - k_H \leq \lambda^*(1+m) - \beta(1-\lambda^*)p_s - k_H = \Pi(CDs)$, and $\Pi(DCs) = k_L - k_H - \beta(1-\lambda^*)p_{ns} \leq \Pi(DDs)$, where inequalities are strict if low types pool. Thus the relevant comparison is between CDs and DDs. Note that $\Pi(CDs) > \Pi(DDs) \Leftrightarrow \frac{k_L}{1+\alpha} > \frac{\beta p_s}{\beta p_s + m - \alpha}$. This inequality holds because of the cyclic condition (4), i.e., $\frac{k_L}{1+\alpha} > \frac{\beta}{\beta+m-\alpha}$. Thus, the remaining payoff constraint is $\Pi(CDs) \geq \Pi(DDns) = 0$ which reduces to an upper bound for the signaling cost for high types, i.e., $k_H \leq \frac{k_L}{1+\alpha}(1+m+\beta p_s) - \beta p_s$. Note that this inequality follow from $k_H < \overline{k}_H$.

Hence, for $k_H < \overline{k}_H$ there exists a semi-pooling equilibrium at λ^* characterized by $p_{CDs,s} + p_{CDs,ns} = 1$. This equilibrium is clearly stable since CDs earns the strictly highest payoff and low types are indifferent between sending the signal or not if $p_{CDs} = 1$. Any perturbation from $p_{CDs} = 1$ will vanish. This might in the meantime induce a shift in the share p_s but this has no impact on the dominance of CDs. Hence, the system will eventually be led back to the equilibrium set $p_{CDs,s} + p_{CDs,ns} = 1$.