Center for European, Governance and Economic Development Research
cege

Discussion Papers

Number 219 – October 2014

THE EVOLUTION OF INEQUALITY AVERSION IN A SIMPLIFIED GAME OF LIFE

Stephan Müller

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

ISSN: 1439-2305

The evolution of inequality aversion in a simplified game of life.

Stephan Müller[1]

Abstract

This paper applies the indirect evolutionary approach to study the evolution of inequality aversion in a simplified game of life. The game comprises a dilemma, a problem of coordination, and a problem of distribution as a general framework for the evolution of preferences. In single-game environments, there emerges a global advantage for inequality-averse individuals in the dilemma and a global disadvantage for inequality-averse players who are favoured by the problem of distribution. The simplified game of life puts these strong predictions into perspective. In particular, selfish and inequality-averse individuals may coexist in the subpopulation, favoured in the problem of distribution.

Keywords: inequality aversion – evolution – preferences

JEL Classifications: C72, C73

[1] Stephan Müller: Göttingen University, Platz der Göttinger Sieben, 3, 37073 Göttingen, Germany (email stephan.mueller@wiwi.uni-goettingen.de). I am grateful to Georg v. Wangenheim, Werner Güth, Claudia Keser and Bertrand Munier for helpful discussion and comments.

1. Introduction

At the latest since the seminal work of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), an other-regarding preference in the form of inequality aversion has become a prominent explanation for many empirical and experimental findings which departure from the prediction of standard economic theory. The increasing importance calls for a rationalization for such preferences since it may otherwise be regarded as a rather ad-hoc adjustment of preferences in explaining empirical results. As Güth and Napel (2006) point out, such preferences should be compatible with the physical necessity to strive and compete for material rewards in an environment characterized by a scarcity of resources. In other words, it should be possible to rationalize such preferences from an evolutionary point of view.

Analysing the evolution of preferences offers a unifying framework for traditional microeconomic analysis concerned with forward-looking agents with fixed preferences. Yet it also incorporates evolutionary biology, focusing on the interplay of the social or biological environment and the success of certain behavioural strategies within that environment. In the past the evolution of preferences has been studied in highly artificial single-game environments (e.g. Huck and Oechssler 1999; Koçkesen et al. 2000a, 2000b; Sethi and Somanathan 2001 and Guttman 2003). Consequently, these studies were inconclusive in explaining the presence of certain preferences, because the behaviour induced by a certain preference might be advantageous in one environment, but disadvantageous in another. The agents' imperfect mental model of the world requires at least some link between the intrinsic motivations in different environments. Given this restriction, agents will have a limited possibility to develop game-specific or role-specific preferences. Hence, the decentralized results for the single environments need to be combined in a centralized picture in order to explain the success or failure of behavioural determinants such as inequality aversion, reciprocity and truthfulness in the complex social and biological environment that comprises seemingly endlessly many of those small worlds, the 'game of life' (Güth and Napel 2006). Therefore, this paper addresses as a first aim the rationalizability of a preference for equality in an environment that contains the major classes of games constituting the game of life.

More recently, some attempts were made to analyse the evolution of preferences in more complex environments. Güth and Napel (2006) analyse how the personal characteristic of inequality aversion evolves in a setting containing two well-studied and characteristic games: the Ultimatum game and the Dictator game. Poulsen and Poulsen (2006) study the evolution of other-regarding preferences in an environment that comprises a simultaneous and a sequential Prisoners' Dilemma. Their analysis illustrates that the study of evolution of preferences in a compound strategic environment yields more interesting and intuitive results than a game-specific analysis. However, the considered environments are not meant to estimate—and indeed aren't even rough approximations of—a game of life.

A prerequisite for the analysis of the evolution of preferences in the game of life is the structuring of the infinite set of potential games, which is the second aim of the paper. There is evidence that human behaviour is not game-specific, but behavioural responses are similar for entire, quite general classes of games (see Ashraf et al. 2006; Chaudhuri and Gangadharan 2007 and Slonim and Garbarino 2008; Blanco et al. 2011; Yamagishi et al. 2013). This raises hope that the overwhelming complexity of the real world might be reducible to these classes when the

evolution of preferences is considered. Many authors implicitly or explicitly share and express the viewpoint that there are two fundamentally different societal problems (see e.g. Sugden 1986; Milgrom et al. 1990), problems of coordination and social dilemmas. Apart from these two classes, Schotter (1981), Ullmann-Margalit (1977) and others share the view that there is (at least) a third type of social problem, one of redistributive nature. The notion of a game of life I suggest will comprise these three classes of games.

As a first step to address the first aim, I restrict this paper to the class of 2x2 games. 2x2 games are omnipresent as they serve as the workhorses in applied game theory and their simplicity is their power as they combine remarkable diversity[2] with minimal machinery. Despite this restriction, the analysis will reveal that the 2x2 case is representative in uncovering the major forces that in their interplay will determine the distribution of inequality aversion in the population. Furthermore, the purpose of the paper is to conduct an analysis for an environment that contains representatives of all classes present in my classification. In other words, the focus of the paper in terms of generality is on completeness within a certain world of games (2x2 games) rather than on the world of games as such (e.g. all finite games). I consider this as a first step in exploring the effects of considering a complete world, although restricted in size. I thus refine the first question in asking for the rationalizability of inequality aversion in what I will refer to as the 'simplified game of life'. With respect to the second goal, although definitions are given for the 2x2 case the classification of games readily translates to all finite normal-form games.

The remainder of the paper proceeds as follows. In Section 2 the precise definitions for the games that are comprised in the simplified game of life will be given. The evolution of a preference for equality in material outcomes for each of the single-game environments is studied in Section 3. Thereafter, the environment of the simplified game of life is considered in Section 4. Before I conclude in Section 6, I discuss the robustness of the results in Section 5.

2. Definition of terms

2.1. Dilemma and Problem of coordination

A non-cooperative strategic interaction between multiple agents is commonly considered to constitute a dilemma, if there exists a non-equilibrium outcome, that is Pareto-superior to a subset of all Nash equilibria. On the one end of the spectrum, one could define a dilemma if there is a Pareto improvement for at least one Nash equilibrium. In contrast, in the definition I suggest, a game is declared to be a dilemma only if there is a non-equilibrium Pareto-improvement relative to all equilibria, i.e. prior to the equilibrium selection. I incorporate the ex-ante viewpoint as it makes the classification of games and the analysis of the evolution of preferences less sensitive to assumptions regarding equilibrium selection. Furthermore, in the more general class of finite normal-form games, the majority of games would constitute a social dilemma following the alternative definition.

[2] The eight numbers that represent such a game yield a class of 144 problems of remarkable richness and complexity (Robinson and Goforth 2005).

Before I formally define a dilemma I introduce some notation. Let $\gamma\left(A^1, A^2\right)$ denote a generic 2x2 game with strategy spaces $S^1 = S^2 = \{0,1\} \equiv S$ and payoffs $A^1 = \left(a_{ij}^1\right)$ and $A^2 = \left(a_{ij}^2\right)$, $(i, j) \in S \times S$ for player 1 and 2 respectively. Let ΔS represent the mixed extension of S. Finally I write the expected payoff of player 1 for a pair of mixed strategy as $\pi^1\left(s^1, s^2\right) = \sum_{i=0}^{1} \sigma_i^1 a_{(i,j)}^1$, $s^n = \left(\sigma_0^n, \sigma_1^n\right) \in \Delta S$ and $\pi^2\left(s^1, s^2\right)$ accordingly. The set of (pure) Nash equilibria of $\gamma\left(A^1, A^2\right)$ is denoted by $NE(\gamma)\left(NE^{pure}(\gamma)\right)$. For symmetric games we have $A^1 = \left(A^2\right)^T \equiv A$ and I simply write $\gamma(A)$.

Definition A game $\gamma\left(A^1, A^2\right)$ is a *Dilemma* if

$$\exists\left(s^1, s^2\right) \in \Delta S^2 : \pi^n\left(s^1, s^2\right) > \pi^n\left(\left(s^1, s^2\right)^*\right),\ n \in \{1,2\}, \forall\left(s^1, s^2\right)^* \in NE(\gamma)$$

As problems of coordination are complementary to dilemmas and are characterized by the presence of multiple equilibria, I define them as follows.

Definition A game $\gamma\left(A^1, A^2\right)$ is a *problem of coordination* if $|NE(\gamma)| > 1$ and there exists no non-equilibrium outcome which Pareto-dominates all of these equilibria.

Note that all symmetric 2x2 games that constitute neither a dilemma nor a problem of coordination are exactly those with a unique equilibrium, which is not Pareto-dominated by some non-equilibrium outcome. In the world of symmetric games, such situations appear rather unproblematic since no dilemma and no problem of coordination is present. In other words, the set of symmetric games can be partitioned into three classes of games, dilemmas, problems of coordination, and unproblematic situations.

2.2. Problems of distribution

Any plausible definitions of distributional concern are related to a notion of asymmetry in payoffs. Again, one could take an *ex-ante* or an *ex-post* point-of-view. With an *ex-post* point-of-view, a game would constitute a problem of distribution if the selected equilibrium shows asymmetric payoffs. From an *ex-ante* perspective, a game would constitute a problem of distribution if all equilibria would show asymmetric payoffs, all in favour of the same player. In the former case, the game will only occasionally lead to asymmetries, whereas in the latter case, the game implies systematic asymmetries. It is more convincing, and in line with the corresponding decision with respect to the definition of social dilemmas, to take the *ex-ante* point-of-view. I will refer to those individuals (dis)favoured in the problem of distribution as (low) high types.

Definition A game $\gamma\left(A^1, A^2\right)$ is a *problem of distribution* if

$$\exists n \in \{1,2\} : \pi^n\left(\left(s^1, s^2\right)^*\right) > \pi^{-n}\left(\left(s^1, s^2\right)^*\right),\ \forall\left(s^1, s^2\right)^* \in NE(\gamma).$$

2.3. Inequality aversion

In the evolutionary analysis, I will make use of the standard evolutionary model, which deals with a large population. This population is structured by personal characteristics and by the way that individuals are matched. There are two sources of heterogeneity among individuals. The population is divided into two subpopulations that correspond to the two different roles assigned in the problem of distribution. There is also heterogeneity with respect to the evaluation of payoff distributions, i.e. agents show different levels of inequality aversion. Inequality aversion is modelled as follows. I will apply the definition suggested by (Fehr and Schmidt 1999) which in a 2x2 setting amounts to $u^n_{(i,j)} = a^n_{(i,j)} - \sigma^n \max\left\{a^{-n}_{(i,j)} - a^n_{(i,j)}, 0\right\} - \omega^n \max\left\{a^n_{(i,j)} - a^{-n}_{(i,j)}, 0\right\}$, $\sigma^n, \omega^n \in [0,1]$, i.e. σ^n and ω^n measure the degree of aversion of player n to inequality which disfavours or, respectively, favours him. I make the simplifying assumption that $\sigma^n = \omega^n \equiv \theta^n$ (see 5.4. for discussion). Hence, inequality aversion is parameterized by the one-dimensional space $[0,1]$. At time t agents' preferences regarding equality in material payoffs is distributed over $[0,1]$ according to the distribution function F^t_H and F^t_L for high types and low types, respectively. Initially, the density functions corresponding to F^t_H and F^t_L are assumed to have full support. I will drop the superscript t to represent equilibrium distributions, i.e. $F_{H,L} = \lim_{t\to\infty} F^t_{H,L}$.

2.4. The simplified game of life

As I will elaborate more deeply in the subsequent analysis, inequality aversion transforms the game $\gamma\left(A^1, A^2\right)$ into the game $\gamma\left(U^1, U^2\right)$. The latter and the former may well differ in the set of Nash equilibria. To ease reading and interpretation, I will make use of the following definitions.

Definition I say that an equilibrium (i, j) in the game $\gamma\left(A^1, A^2\right)$ is *contested* by player 1(2) if $u^1_{(\neg i,j)} > u^1_{(i,j)}$ $\left(u^2_{(i,\neg j)} > u^2_{(i,j)}\right)$, i.e. strategy *i(j)* loses its property of being a best response to strategy *j(i)* in the game $\gamma\left(U^1, U^2\right)$. An equilibrium in the game $\gamma\left(A^1, A^2\right)$ is *contestable*, if it may be contested by at least one player. I say that the strategy pair (i, j) is *stabilizable* if it is an equilibrium of $\gamma\left(U^1, U^2\right)$ for some levels of θ^1 and θ^2.

To simplify the analysis of the simplified game of life, I will restrict the included games in a way that ensures that in the game $\gamma\left(U^1, U^2\right)$ no situation with a unique mixed Nash equilibrium will occur. Since a unique mixed Nash equilibrium arises if a player who contests all pure Nash equilibria is matched with a purely selfish player, I employ the following definition.

Definition A game $\gamma\left(A^1, A^2\right)$ is called *strict* if there is no player who can contest all equilibria.

A player will not be able to contest all equilibria if at least one equilibrium is sufficiently strict for him, i.e. the material loss from unilateral deviations is sufficiently high. Note that in general finite normal-form games, this condition will be satisfied in the majority of the cases. Allowing the play of mixed equilibria has interesting consequences on the sharpness of the prediction regarding the

stable distributions of preferences. This will be outlined in Section 5.3. I am now able to define an environment that comprises all these classes.

Definition The *simplified game of life* is a game that comprises a symmetric dilemma, a strict symmetric problem of coordination and a strict problem of distribution.

The qualification for the dilemma and the problem of coordination to be symmetric is made in order to isolate the effects that the asymmetry of the problem of distribution implies.

2.5. Evolutionary framework

In what follows, I state the assumptions I make with respect to informational aspects, the matching process, evolutionary dynamics and the applied stability concept.

I assume that agents can mutually observe their attitude towards unequal payoff distributions. This assumption could be weakened to an awareness of the inequality aversion in a positive fraction of interactions, the availability of sufficiently accurate signals or sufficiently cheap screening technologies (see Güth 1995, Sethi and Somanathan 2001, Güth et al. 2003). The matching procedure takes place as follows. First, a random draw selects among the three types of games that constitute the simplified game of life. In case of a dilemma or a problem of coordination, individuals from the total population are randomly matched into pairs playing the selected game. Thereby each pair has the same probability in each short period of time. The interaction in the problem of distribution will be modelled as a 2-population model (see e.g. Weibull 1997), i.e. individuals interact across populations but not within. Again, each pairing has the same probability but the relative size of the subpopulations of high and low types matters for expected payoffs[3]. However, this will only amplify the advantage or disadvantage of high types over low types. For notational simplicity, I may thus assume that the two subpopulations are equal in size. Payoffs given by A^1 and A^2 represent the material payoffs of the stage game that will be decisive with respect to evolutionary success.

Whereas the belonging to one of the subpopulations due to role assignment in the problem of distribution is exogenous and common knowledge, the distribution of inequality-averse individuals in each of the two subpopulations is endogenous. Since inequality aversion reflects a particular evaluation of material payoffs, I will apply the indirect evolutionary approach[4] pioneered by Güth and Yaari (1992), i.e. preferences determine behaviour and behaviour in turn determines fitness. Fitness measured by material payoffs will determine the evolution of F^t. The evolutionary process is modelled by payoff monotone selection dynamics[5] (see e.g. Weibull 1997). With respect to stability, I apply the concept of asymptotic stability (see. e.g. Samuelson 1997 for definitions).

[3] If for instance, the subpopulation of low types is ten times as large as the subpopulation for high types, then any high type will play ten times as often as a low type.

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

[5] There are other forces than evolutionary selection shaping individual preferences. Bisin and Verdier (2001) for instance study intergenerational cultural transmission mechanisms.

Since I am interested in games that allow for multiple equilibria, an assumption with respect to equilibrium selection is needed. An appropriate equilibrium selection criterion should not *a priori* favour or disfavour a preference for equality with respect to evolutionary success. I therefore assume that if $\gamma\left(U^1, U^2\right)$ has multiple pure-strategy Nash equilibria, then players randomize over all pure-strategy Nash equilibria with equal probability. To clarify, it is not the players who randomize over strategies of different pure Nash equilibria independently, but pairs of players randomize jointly over the set of pure Nash equilibria. Indeed, as it will turn out (see 5.1 for discussion) neutrality of the equilibrium selection for all games requires a symmetric probability distribution over the set of equilibria, which for 2x2 games and if the set of pure Nash equilibria is considered amounts to uniformity.

Let $\Gamma, \left(\Gamma_{sym}\right)$ denote the set of (symmetric) 2x2 games, $\Gamma^{\circ}, \left(\Gamma^{\circ}_{sym}\right)$ the set of (symmetric) 2x2 games with neither weakly nor strictly dominated strategies. Games with weakly dominated strategies can be treated as the limiting case of games in Γ°. More precisely as $\Gamma \subset \mathbb{R}^8, \left(\Gamma_{sym} \subset \mathbb{R}^4\right)$ the subset of $\Gamma, \left(\Gamma_{sym}\right)$ containing no weakly dominated strategies is dense in $\Gamma, \left(\Gamma_{sym}\right)$. Since the critical level of inequality aversion are continuous in the parameters of a game $\gamma\left(A^1, A^2\right)$, the results for any game with weakly dominated strategies are a limit case of games in $\Gamma^{\circ}, \left(\Gamma^{\circ}_{sym}\right)$[6]. Given this technical note, I can concentrate on games with no weakly dominated strategies.

3. Inequality aversion in the separate environments

Symmetric dilemma For symmetric games there is always an equilibrium in pure strategies. Furthermore, games with multiple equilibria are free of the dilemma property. To see this, consider a symmetric game with two pure Nash-equilibria. A necessary condition for such a game to constitute a social dilemma would be that there is an outcome in pure strategies that gives each player more than the maximum of the two Nash equilibria in pure strategies. Nevertheless, the existence of such an outcome violates the Nash-equilibrium property in the first place because in 2x2 games this implies the existence of an alternative reply with higher payoffs than in equilibrium. Hence, a symmetric social dilemma must be in the set $\Gamma_{sym} \setminus \Gamma^{\circ}_{sym}$, the set of games with weakly or strictly dominated strategies. As a unilateral deviation from equilibrium can never lead to a strict Pareto-improvement, only the symmetric non-equilibrium outcome realized by bilateral deviation can yield strictly higher payoffs for both players. Hence, in 2x2 games, a symmetric dilemma corresponds to the classical Prisoners' Dilemma. Lemma 1 summarizes this insight. Let $AP_{(i,j)} \equiv \frac{a^1_{(i,j)} + a^2_{(i,j)}}{2}$ denote the average payoff if player one (two) plays $i\left(j\right)$. All proofs are given in the Appendix.

[6] More precisely, the mapping $\Phi : \Gamma \to \mathbb{R}$ which assigns to any game the critical value $\theta^{D,C,R}$ (see Section 3) is continuous.

Lemma 1 Let $\gamma(A) \in \Gamma_{sym}$. $\gamma(A)$ constitutes a dilemma if and only if $\gamma(A)$ is strictly dominance-solvable by the unique symmetric Nash equilibrium (i^*, i^*) and $AP_{(i^*,i^*)} < AP_{(\neg i^*,\neg i^*)}$.

The symmetric Pareto-superior outcome in $\gamma(A)$ can be stabilised in $\gamma(U^1, U^2)$ if the degree of inequality aversion of both players exceeds a certain thresholds. The threshold is given by $\theta^D \equiv \frac{a_{(\neg i^*,i^*)} - a_{(\neg i^*,\neg i^*)}}{d_{(i^*,\neg i^*)}}$, $d_{(i,j)} \equiv \left| a_{(i,j)} - a_{(j,i)} \right|$ and has a straightforward economic interpretation. Since $a_{(\neg i^*,i^*)} - a_{(\neg i^*,\neg i^*)}$ measures the material gain of deviating from the non-equilibrium pair of strategies $(\neg i^*, \neg i^*)$ and $d_{(i^*,\neg i^*)}$ measures the implied loss in equality induced by such a deviation, θ^D measures the material price per unit of equality gained. Sufficient inequality aversion therefore translates into a sufficient willingness to pay for equality. Given this insight and the characterization of social dilemmas in Lemma 1, Proposition 1 characterizes the stable distributions of inequality aversion.

Proposition 1 Let $\gamma(A) \in \Gamma_{sym}$ be a social dilemma. If $(\neg s_1^*, \neg s_2^*)$ is stabilizable, then there exists a $\theta^D \in [0,1]$, such that the globally stable equilibrium is $F(\theta^D) = 0$[7]. Furthermore, the material advantage of sufficiently inequality-averse individuals is increasing in the share of individuals with $\theta \geq \theta^D$, i.e. $\operatorname{sgn}\left(\Pi^{\theta \geq \theta^D} - \Pi^{\theta < \theta^D}\right)_{(1-F(\theta^D))} = 1$, where $\left(\Pi^{\theta \geq \theta^D} - \Pi^{\theta < \theta^D}\right)_{(1-F(\theta^D))}$ denotes the derivative w.r.t. $1 - F(\theta^D)$, the share of inequality-averse individuals. Otherwise the share of inequality-averse individuals is determined by initial conditions and random shift.

Intuitively, the potential for an evolutionary advantage of inequality-averse individuals stems from the fact that a pair of sufficiently inequality-averse players will be able to transform the social dilemma into a coordination game. By definition of the dilemma the stabilized outcome yields Pareto-superior payoffs which benefits inequality-averse individuals as they randomize over all pure Nash equilibria.

Symmetric problem of coordination In games within the set of Γ°_{sym} which show multiple pure-strategy Nash equilibria either the two diagonal symmetric payoff-pairs or the two off-diagonal asymmetric payoff-pairs constitute the Nash equilibrium payoffs.

Lemma 2 Let $\gamma(A) \in \Gamma_{sym}$. $\gamma(A)$ constitutes a problem of coordination if and only if (1) $NE^{pure}(\gamma(A)) = \{(i,i)\}$ or (2) $NE^{pure}(\gamma(A)) = \{(i,j) | i \neq j\}$.

[7] As in the cases of a dilemma and a problem of coordination the roles of players are symmetric, in the corresponding subsections, I drop the subscripts reflecting types in the problem of redistribution.

I define a threshold θ^{C} that is the equivalent to θ^{D} in the symmetric dilemma. In the symmetric coordination game each of the off-diagonal equilibria of $\gamma(A)$ may be contestable for both players. Hence, θ^{C} will be the minimum of the two ratios measuring the material price per unit of equality gained for player one and two. These prices may differ as equilibria in $\gamma(A)$ may be asymmetric and players face different incentives to deviate. Formally, $\theta^{C} \equiv \min_{(i,j)\in NE^{pure}(\gamma(A))} \left\{ \frac{a_{(i,j)} - a_{(i,i)}}{d} \right\}$, where $d \equiv d_{(i,j)}\big|_{(i,j)\in NE^{pure}(\gamma(A))}$.

Let AP measure the average payoff of the equilibria in $\gamma(A)$, i.e. $AP \equiv AP_{(i,j)}\big|_{(i,j)\in NE^{pure}(\gamma(A))}$. For ease of readability, I will refer to individuals with $\theta \geq \theta^{C}$ $(\theta < \theta^{C})$ as inequality-averse individuals and selfish players, respectively.

Proposition 2 Let $\gamma(A) \in \Gamma_{sym}$ be a strict problem of coordination. Then:

If $NE^{pure}(\gamma(A)) = \{(i,i)\}$ or $NE^{pure}(\gamma(A)) = \{(i,j) | i \neq j\}$ and none of the material equilibria is contestable then the share of inequality-averse individuals in the population is determined by initial conditions and random shift.

If equilibria are contestable, then:

1. if the destabilized equilibrium is materially favourable for inequality-averse individuals then the globally stable equilibrium is characterized by $F(\theta^{C}) = 1$. Furthermore, $\operatorname{sgn}\left(\Pi^{\theta\geq\theta^{C}} - \Pi^{\theta<\theta^{C}}\right)_{(1-F(\theta^{C}))} \in \{-1,0,1\}$.
2. if the destabilized equilibrium is materially favourable for selfish individuals then the globally stable equilibrium is characterized by $F(\theta^{C}) = \theta^{C}\frac{d}{AP - AP_{(i,i)}}$. Furthermore, $\operatorname{sgn}\left(\Pi^{\theta\geq\theta^{C}} - \Pi^{\theta<\theta^{C}}\right)_{(1-F(\theta^{C}))} = -1$.

where $AP_{(i,i)}$ is the average payoff of the outcome that is stabilizable by two sufficiently inequality-averse individuals.

In case (1) of Lemma 2, the material equilibria are not contestable as any deviation from symmetric material payoffs not only reduces material payoff but also increases inequality. Consequently, no evolutionary pressure will emerge favouring or disfavouring inequality aversion. However, in case (2) with respect to utility, a deviation from materially asymmetric payoffs associated with a gain in equality might outweigh the material loss from deviation. Proposition 2 reveals that in strict problems of coordination, a strong preference for equality is weakly disadvantageous from an evolutionary point of view. If the destabilized equilibrium is materially favourable for inequality-averse individuals then not only do they suffer from deviating from material equilibrium, but they also lose relative individuals that are more selfish. This happens

because the equilibrium is destabilized where they gain more than selfish players do. Therefore, individuals with a strong preference for equality face an evolutionary disadvantage and will become extinct. If the reverse is true, then the disadvantage from unilaterally deviating from material equilibria is partially compensated by no longer playing a disadvantageous equilibrium and thereby increasing average payoffs. However, this effect diminishes as the share of sufficiently inequality-averse agents increases. This stabilizes a distribution of preference where selfish and inequality-averse individuals coexist.

Problem of distribution Lemma 3 below characterizes problems of redistribution and differentiates two cases that will become relevant in the course of the argument.

Lemma 3 Let $\gamma\left(\mathrm{A}^1,\mathrm{A}^2\right)\in\Gamma$. $\gamma\left(\mathrm{A}^1,\mathrm{A}^2\right)$ constitutes a strict problem of distribution if and only if all Nash equilibria favour the same individual and:

(1): $\gamma\left(\mathrm{A}^1,\mathrm{A}^2\right)$ has multiple equilibria that are not Pareto-ranked.

(2): $\gamma\left(\mathrm{A}^1,\mathrm{A}^2\right)$ has multiple equilibria that are Pareto-ranked.

Let $\widehat{\theta}^{\mathrm{R}}_{\mathrm{H,L}}\left(\breve{\theta}^{\mathrm{R}}_{\mathrm{H,L}}\right)$ denote the thresholds for high and low types respectively such that the more (less) equal material equilibrium is destabilized. A formal definition requires complicated notation and is not very insightful (see proof of Proposition 3). The economic meaning of the thresholds is the same as for the thresholds in the problem of coordination or the dilemma, i.e. they measure the price of deviation per unit equality gained. Let $\theta^{\mathrm{R}}_{\mathrm{H,L}}=\min\left\{\widehat{\theta}^{\mathrm{R}}_{\mathrm{H,L}},\breve{\theta}^{\mathrm{R}}_{\mathrm{H,L}}\right\}$. The type-contingent threshold $\theta^{\mathrm{R}}_{\mathrm{H,L}}$ plays the same role as θ^{D} and θ^{C} in the dilemma and the problem of coordination respectively, i.e. if the degree of inequality aversion for at least one player exceeds $\theta^{\mathrm{R}}_{\mathrm{H,L}}$ then at least one of the equilibria of $\gamma\left(\mathrm{A}^1,\mathrm{A}^2\right)$ loses its equilibrium property in $\gamma\left(\mathrm{U}^1,\mathrm{U}^2\right)$.

Proposition 3 Let $\gamma\left(\mathrm{A}^1,\mathrm{A}^2\right)$ constitute a strict problem of distribution.

1. If one of the material equilibria is contestable by low types, the unique globally stable equilibrium distribution is characterized by a homomorphic population with only inequality-averse individuals. $\mathrm{F_L}\left(\theta^{\mathrm{R}}_{\mathrm{L}}\right)=0,\ \mathrm{sgn}\left(\Pi^{\theta\geq\theta^{\mathrm{R}}_{\mathrm{L}}}_{\mathrm{L}}-\Pi^{\theta<\theta^{\mathrm{R}}_{\mathrm{L}}}_{\mathrm{L}}\right)_{\left(1-\mathrm{F_L}\left(\theta^{\mathrm{R}}_{\mathrm{L}}\right)\right)}\in\{-1,0,1\}$.
2. If one of the material equilibria is contestable by high types, with one exception, the globally stable equilibrium distribution is characterized by
$$\mathrm{F_H}\left(\theta^{\mathrm{R}}_{\mathrm{H}}\right)=1,\ \mathrm{sgn}\left(\Pi^{\theta\geq\theta^{\mathrm{R}}_{\mathrm{L}}}_{\mathrm{H}}-\Pi^{\theta<\theta^{\mathrm{R}}_{\mathrm{L}}}_{\mathrm{H}}\right)_{\left(1-\mathrm{F_H}\left(\theta^{\mathrm{R}}_{\mathrm{H}}\right)\right)}=1.$$
The exception arises in the case of two Pareto-ranked equilibria (case (2) of Lemma 3) with the Pareto-inferior equilibrium being contestable for both types. In that case, the globally stable equilibrium distribution is characterized by
$$\mathrm{F_H}\left(\theta^{\mathrm{R}}_{\mathrm{H}}\right)=0,\ \mathrm{sgn}\left(\Pi^{\theta\geq\theta^{\mathrm{R}}_{\mathrm{L}}}_{\mathrm{H}}-\Pi^{\theta<\theta^{\mathrm{R}}_{\mathrm{L}}}_{\mathrm{H}}\right)_{\left(1-\mathrm{F_H}\left(\theta^{\mathrm{R}}_{\mathrm{H}}\right)\right)}=-1.$$
Otherwise, the distribution is determined by initial conditions and random shift.

Intuitively, in case (1) of Lemma 3, one of the pure strategy equilibria shows strictly less inequality. Hence, the more (less) unequally distributed equilibrium is preferred by the high (low) type. It turns out that for the high type, the more equally distributed equilibrium is never contestable. I first consider the case where the more equally distributed equilibrium is not destabilized by the low type as in the first case of Proposition 3. On the one hand, if the more unequal equilibrium is destabilized by both players, then the more equally distributed equilibrium will become the unique equilibrium. In that case, such high types will play with certainty the less favourable equilibrium of $\gamma\left(A^1, A^2\right)$ and face an evolutionary disadvantage. Furthermore, the extent of the disadvantage for the high types increases with the share of sufficiently inequality-averse low types since more and more often they will end up playing the relative unfavourable equilibrium. The reverse argument applies for the low types. If on the other hand, the high type only destabilizes the more unequal equilibrium, the same argument applies for the high types but the disadvantage is now independent of the share of inequality-averse low types, as their best response behaviour is not altered by inequality aversion.

I second consider the case where the more equally distributed equilibrium is destabilized by the low type as in the second case of Proposition 3. If high types destabilize the more unequally distributed equilibrium, then this will result in an evolutionary disadvantage, as the relatively less favourable equilibrium will be selected. As no player can destabilize all equilibria, inequality-averse low types will face an evolutionary disadvantage as they destabilize the relative favourable one of the two pure Nash equilibria in $\gamma\left(A^1, A^2\right)$. In all other cases, the distribution of the preference parameter is undetermined. The major difference between case (1) and (2) of Lemma 3 responsible for the deviations in equilibrium distribution stems from the following fact. In case (1) of Lemma 3, the less unequally distributed equilibrium, which is relatively less favourable for the high type was not contestable. In case (2), however, the Pareto-superior equilibrium is not contestable. In this difference lies the potential for an evolutionary advantage of inequality-averse individuals among high types.

In summary, the analysis in separate environments makes relatively strong predictions (see also Figure 1)[8]. If inequality aversion has leverage on the set of equilibria played, then inequality aversion enjoys a global evolutionary advantage over more selfish preferences in a dilemma. In the class of problems of coordination, inequality aversion surprisingly faces a weak evolutionary disadvantage. This is the case in the sense that a stable inner equilibrium exists, at most, where inequality-averse and selfish players coexist. In all other cases, inequality-averse players will eventually disappear. In the problem of distribution, evolutionary selection dynamics will always favour the preference for equality among the disfavoured individuals. Among the individuals favoured by the problem of distribution in all cases except for one inequality aversion will eventually disappear.

[8] The three characteristics: the slope, the intercept and having a root in the open unit interval gives rise to eight different loci of the linear payoff differences. The analysis so far predicts that at most three of them are needed to describe the differences in payoffs between inequality-averse and selfish individuals (see Figure 1).

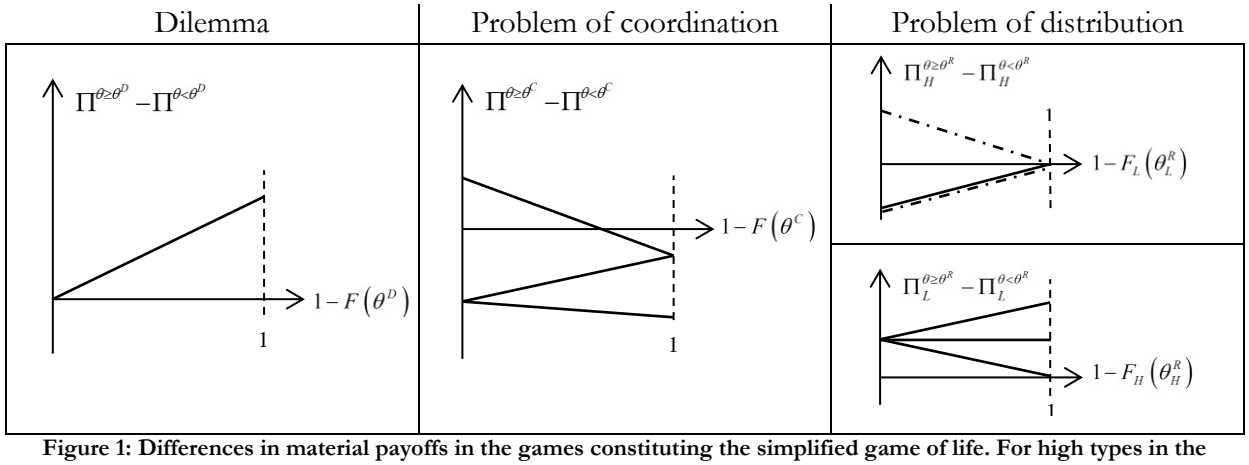


Figure 1: Differences in material payoffs in the games constituting the simplified game of life. For high types in the problem of distribution the continuous lines correspond to case A, the dotted lines to case (2) of Lemma 3.

4. Evolution of inequality aversion in the 2x2 simplified game of life

In this section, I analyse the interplay of the different types of interaction present in the simplified game of life. For ease of exposition I assume that the thresholds of the single environments coincide, i.e. $\theta^{\mathrm{D}}=\theta^{\mathrm{C}}=\theta^{\mathrm{R}}_{\mathrm{H,L}}\equiv\theta^{\mathrm{crit}}$. The profit for an individual in the simplified game of life is simply the weighted average of the profits earned in the single environments[9], i.e.:

$$\begin{aligned}\Pi^{\mathrm{S},\theta\geq\theta^{\mathrm{crit}}}_{\mathrm{H,L}}&=\mu\,\Pi^{\mathrm{D},\theta\geq\theta^{\mathrm{crit}}}\left(\mathrm{F}^{\mathrm{t}}\left(\theta^{\mathrm{crit}}\right)\right)+\nu\,\Pi^{\mathrm{C},\theta\geq\theta^{\mathrm{crit}}}\left(\mathrm{F}^{\mathrm{t}}\left(\theta^{\mathrm{crit}}\right)\right)+(1-\mu-\nu)\Pi^{\mathrm{R},\theta\geq\theta^{\mathrm{crit}}}_{\mathrm{H,L}}\left(\mathrm{F}^{\mathrm{t}}_{\mathrm{L,H}}\left(\theta^{\mathrm{crit}}\right)\right)\\ \Pi^{\mathrm{S},\theta<\theta^{\mathrm{crit}}}_{\mathrm{H,L}}&=\mu\,\Pi^{\mathrm{D},\theta<\theta^{\mathrm{crit}}}\left(\mathrm{F}^{\mathrm{t}}\left(\theta^{\mathrm{crit}}\right)\right)+\nu\,\Pi^{\mathrm{C},\theta<\theta^{\mathrm{crit}}}\left(\mathrm{F}^{\mathrm{t}}\left(\theta^{\mathrm{crit}}\right)\right)+(1-\mu-\nu)\Pi^{\mathrm{R},\theta<\theta^{\mathrm{crit}}}_{\mathrm{H,L}}\left(\mathrm{F}^{\mathrm{t}}_{\mathrm{L,H}}\left(\theta^{\mathrm{crit}}\right)\right)\end{aligned}\tag{1}$$

Hence, payoff differences are given by[10]:

$$\begin{aligned}&\Pi^{\mathrm{S},\theta\geq\theta^{\mathrm{crit}}}_{\mathrm{H,L}}-\Pi^{\mathrm{S},\theta<\theta^{\mathrm{crit}}}_{\mathrm{H,L}}=\\ &\mu\left(\underbrace{\Pi^{\mathrm{D},\theta\geq\theta^{\mathrm{crit}}}-\Pi^{\mathrm{D},\theta<\theta^{\mathrm{crit}}}}_{\geq 0}\right)+\nu\left(\underbrace{\Pi^{\mathrm{C},\theta\geq\theta^{\mathrm{crit}}}-\Pi^{\mathrm{C},\theta<\theta^{\mathrm{crit}}}}_{\leq 0}\right)+(1-\mu-\nu)\left(\underbrace{\Pi^{\mathrm{R},\theta\geq\theta^{\mathrm{crit}}}_{\mathrm{H,L}}-\Pi^{\mathrm{R},\theta<\theta^{\mathrm{crit}}}_{\mathrm{H,L}}}_{\mathrm{H}:\leq 0^{*},\ \mathrm{L}:\geq 0}\right)\end{aligned}\tag{2}$$

Let $\mathrm{d}\Pi$ denote the difference in payoffs between relatively inequality-averse and selfish players. Equation (2) can now be expressed in a more compact way as[11]:

$$\mathrm{d}\Pi^{\mathrm{S}}_{\mathrm{H,L}}=\mu\left(\underbrace{\mathrm{d}\Pi^{\mathrm{D}}\left(1-\mathrm{F}^{\mathrm{t}}_{\mathrm{H}}-\mathrm{F}^{\mathrm{t}}_{\mathrm{L}}\right)}_{\geq 0}\right)+\nu\left(\underbrace{\mathrm{d}\Pi^{\mathrm{C}}\left(1-\mathrm{F}^{\mathrm{t}}_{\mathrm{H}}-\mathrm{F}^{\mathrm{t}}_{\mathrm{L}}\right)}_{\leq 0}\right)+(1-\mu-\nu)\left(\underbrace{\mathrm{d}\Pi^{\mathrm{R}}_{\mathrm{H,L}}\left(1-\mathrm{F}^{\mathrm{t}}_{\mathrm{L,H}}\right)}_{\mathrm{H}:\leq 0^{*},\ \mathrm{L}:\geq 0}\right)\tag{3}$$

Making use of the linearity of the payoffs differences I write (3) as:

[9] D – dilemma; C – problem of coordination ; R – problem of distribution; S – simplified game of life.

[10] The asterisk in equation (2) and (3) refers to the exception in case (2) of Lemma 3 in which also among high types inequality-averse individuals enjoy an evolutionary advantage.

[11] Note that whereas the differences in the dilemma and the problem of coordination depend on the total share of inequality-averse individuals in the population, the according difference in payoffs for the problem of distribution depends only on the share in the subpopulation of the opposite type.

$$
\begin{aligned}
d\Pi_H^S &= \mu\beta^D\left(1-F_H^t-F_L^t\right)+\nu\left(\alpha^C+\beta^C\left(1-F_H^t-F_L^t\right)\right)+(1-\mu-\nu)\left(\alpha_H^R+\beta_H^R\left(1-F_L^t\right)\right)\\
d\Pi_L^S &= \mu\beta^D\left(1-F_H^t-F_L^t\right)+\nu\left(\alpha^C+\beta^C\left(1-F_H^t-F_L^t\right)\right)+(1-\mu-\nu)\left(\alpha_L^R+\beta_L^R\left(1-F_H^t\right)\right) \qquad (4)\\
&= d\Pi_H^S+(1-\mu-\nu)\underbrace{\left(\alpha_L^R+\beta_L^R\left(1-F_H^t\right)-\alpha_H^R-\beta_H^R\left(1-F_L^t\right)\right)}_{\geq 0^*}
\end{aligned}
$$

, where $\alpha^D=0$, $\alpha^C,\alpha_{H,L}^R$ and $\beta^D,\beta^C,\beta_{H,L}^R$ denote intercepts and slops of $d\Pi^D,d\Pi^C,d\Pi_{H,L}^R$ respectively.

In the case with two Pareto-ranked equilibria (case (2) of Lemma 3), if the Pareto-inferior equilibrium is destabilized by the low type, then inequality-averse players are favoured also among high types. In that case if the problem of coordination is played not too frequently or involves differences in payoffs that are comparably small, inequality-averse players in both sub-populations face an evolutionary advantage. In other words, the globally stable equilibrium distribution will be characterized by $F_{H,L}\left(\theta^{crit}\right)=0$, i.e. the population will consist only of inequality-averse individuals. Therefore, in the following, I focus on the non-exceptional cases with a problem of distribution being accompanied with a global disadvantage of inequality-averse players among high types. Note that in this case, $\beta_H^R=-\alpha_H^R$ (see Figure 1). Additionally, since low types and high types earn the same profits in the dilemma and the problem of coordination, a positive payoff difference for high types implies a positive difference for low types (see (4)). This has the immediate consequence that a locally stable equilibrium characterized by $F_H\left(\theta^{crit}\right)=0, F_L\left(\theta^{crit}\right)=1$, i.e. an equilibrium with only inequality-averse high types and only selfish low types does not exist in the simplified game of life.

The following theorem characterizes the equilibria that may emerge in the simplified game of life for the predominant case of a problem of distribution, which is disadvantageous for inequality-averse high types. For ease of readability, I abbreviate $F_H\left(\theta^{crit}\right)=F_H, F_L\left(\theta^{crit}\right)=F_L$.

<u>Theorem</u> Let $\theta^D=\theta^C=\theta_{H,L}^R\equiv\theta^{crit}$ and $d\Pi_H^R\leq 0$, then the set of equilibrium distributions of a preference for equality is characterized by:

$$
F_L=0, F_H=\begin{cases} 0 & , \quad \alpha^C>-\dfrac{\mu\beta^D+\nu\beta^C}{\nu} \\ 1+\dfrac{\nu\alpha^C}{\mu\beta^D+\nu\beta^C} & , \quad 0\leq\alpha^C\leq-\dfrac{\mu\beta^D+\nu\beta^C}{\nu} \end{cases}
$$

$$
F_H=1,\ F_L=\begin{cases} 0 & , \quad -\dfrac{1-\mu-\nu}{\nu}\alpha_L^R\leq\alpha^C\leq 0 \\ 1-\dfrac{\nu\alpha^C}{\mu\beta^D+\nu\beta^C}+\dfrac{1-\mu-\nu}{\mu\beta^D+\nu\beta^C}\alpha_L^R & , \quad -\dfrac{1-\mu-\nu}{\nu}\alpha_L^R+\dfrac{\mu\beta^D+\nu\beta^C}{\nu}<\alpha^C<-\dfrac{1-\mu-\nu}{\nu}\alpha_L^R \\ 1 & , \quad \alpha^C\leq-\dfrac{1-\mu-\nu}{\nu}\alpha_L^R+\dfrac{\mu\beta^D+\nu\beta^C}{\nu} \end{cases}
$$

Figure 2 illustrates the set of equilibria graphically. Only if the advantage of inequality-averse individuals increases or the disadvantage decreases in the share of inequality-averse individuals when the dilemma and the problem of coordination are considered alone can multiple equilibria occur ($\mu\beta^{D}+\nu\beta^{C}>0$). Inner equilibria with relative inequality-averse and selfish players in coexistence can only occur if the reverse is true. In such inner equilibria, only in one of the subpopulation that corresponds to the role assignment in the problem of redistribution inequality-averse and selfish players may coexist.

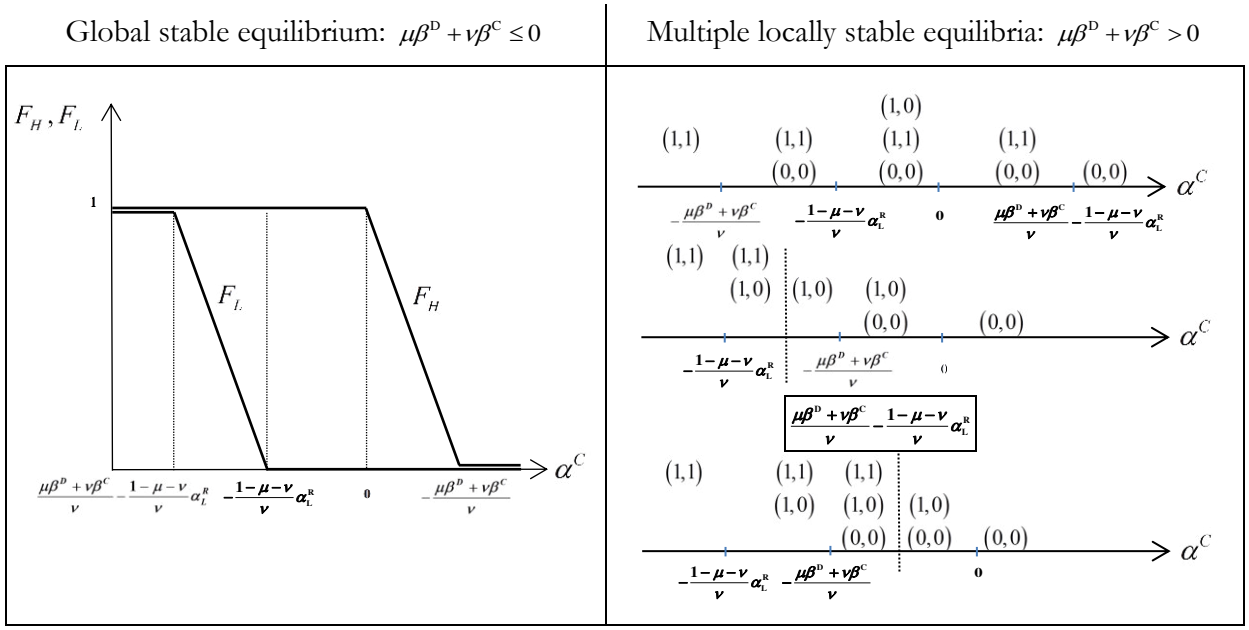


Figure 2: set of equilibria, (F_H, F_L) for the right column.

We saw in Proposition 3 that in almost all cases, a preference for equality above θ^{crit} cannot be sustained in equilibrium among high types in the problem of distribution. A particularly interesting question is therefore whether there is a stable equilibrium with a positive share of inequality-averse players among high types if the simplified game of life is considered, i.e. $F_H(\theta^{crit})\in(0,1)$. The theorem reveals that this is indeed the case. However, an inner equilibrium can only emerge in the case where the problem of coordination on its own would stabilize such a distribution of preferences ($0\le\alpha^{C}$, see Figure 2). On the other hand, for the subpopulation of low types the interplay of problem of coordination and the problem of distribution can induce a stable inner equilibrium (see Figure 2)[12].

In a population of only inequality-averse players, if selfish individuals would on average face an evolutionary disadvantage when only the dilemma and the problem of coordination are considered ($\mu\beta^{D}+\nu(\alpha^{C}+\beta^{C})>0$), then inequality aversion will be advantageous for high and low types and a stable equilibrium with $F_L=0, F_H=0$ exists. In all other cases, the inequality-averse high types are deemed to extinction also in the simplified game of life. If the problem of coordination is not too disadvantageous for inequality-averse individuals then the advantageousness for the dilemma and the problem of distribution carries over to the simplified

[12] A heteromorphic equilibrium population accords well with the experimental results of Andreoni and Miller (1993).

game of life and a stable equilibrium with only inequality-averse players exists. At an intermediate level of disadvantageousness, both inequality-averse and selfish players will coexist in the subpopulation of low types. Finally, if the disadvantage in the problem of coordination dominates then in both sub-populations only selfishness may be part of a stable distribution of preferences.

In summary, on the one hand the simplified game of life—as expected—gives rise to a greater variety in potential equilibrium distributions of preferences. In particular, the surprisingly strong predictions for the single environments are put into perspective. The global advantage of inequality-averse players in the dilemma and the global disadvantage for inequality-averse high types in almost all cases become subject to some qualification. On the other hand, the expected stabilization of inner equilibria for high types in which relatively inequality-averse individuals and relatively selfish individuals coexist occurs if and only if the single environment of a problem of coordination shows the same feature.

5. Discussion

In this section, I discuss the robustness of the results with regard to several issues. These issues consider the core assumptions of the paper: the equilibrium selection criteria, the equilibrium concept, the strictness property, and the model of inequality aversion.

5.1. Equilibrium selection

I now turn to the assumption concerning equilibrium selection that agents jointly randomize over the set of pure Nash equilibria with equal weight. I claimed in section 2.5 that when lacking a general theory of equilibrium selection, the requirement on the selection criteria to be *a priori* neutral with respect to the evolutionary success of inequality aversion amounts to a symmetric probability distribution over the set of equilibria. This requirement stems from the fact that I am solely interested in the evolutionary forces that follow from the impact of a particular preference on the set of Nash equilibria and not in forces that are based on selection bias. A symmetric probability distribution implies neutrality, because in that case any two matches of pairs of individuals with potentially different degrees of inequality aversion will earn the same expected material payoff as long as the set of pure Nash equilibria coincide. Symmetry is thus sufficient for neutrality. To see necessity, consider the following numerical example of a problem of coordination. Below, Table 1 presents the material payoffs of $\gamma\left(\mathrm{A}^1, \mathrm{A}^2\right)$ and their evaluation.

	0	1
0	3 3	4 2
1	2 4	0 0

Table 1: **Payoffs in** $\gamma\left(\mathrm{A}^1, \mathrm{A}^2\right)$

	0	1
0	3 3	$4-2\theta_2$ $2-2\theta_1$
1	$2-2\theta_2$ $4-2\theta_1$	0 0

Payoffs in $\gamma\left(\mathrm{U}^1, \mathrm{U}^2\right)$.

In a match of two individuals with inequality aversion $0<\theta_1<\theta_2<1/2$, i.e. when preferences of player two shows a higher degree of inequality aversion, the set of pure Nash equilibria of

$\gamma\left(\mathrm{A}^1,\mathrm{A}^2\right)$ and $\gamma\left(\mathrm{U}^1,\mathrm{U}^2\right)$ coincide. Any asymmetric probability distribution over the set $\{(0,1),(1,0)\}$ will (dis)favour the relative inequality-averse player if a (smaller) larger weight is put on $(0,1)$. Thus, an asymmetric distribution creates an evolutionary advantage or disadvantage for the relative inequality-averse player, but it is not neutral.

Note that the assumption of a uniform randomization over the set of pure Nash equilibria is equivalent to a play of the correlated equilibrium that assigns equal weights to each of the pure Nash equilibria. In other words if multiple pure Nash equilibria exist individuals play a particular correlated equilibrium. The implications of considering not one but the whole set of correlated equilibria is discussed in the next section.

5.2. Equilibrium concept

The reason why a preference for equality may be advantageous or disadvantageous from an evolutionary perspective lies in its leverage on the equilibrium set. How the set of equilibria is altered by transforming the underlying game in material payoffs by preferences evaluating these payoffs may depend on the applied notion of equilibrium. Most of applied game theory applies the Nash equilibrium as its reference point and deals with finer or coarser equilibrium concepts relative to the Nash concept. I will illustrate the effects for a concrete alternative, that of correlated equilibria[13], for the class of games that constitute a dilemma. The concept of correlated equilibria not only enlarges the set of equilibria but it also increases the set of achievable payoffs generated by the correlated strategies. An increasing set of achievable payoffs may in turn enlarge the class of dilemmas. Similar to the argument in Section 4, a symmetric social dilemma must be in the set $\Gamma_{\mathrm{sym}}\setminus\Gamma^{\circ}_{\mathrm{sym}}$. If a player has a strictly dominant strategy, then by symmetry, his opponent has the same strictly dominant strategy. Here, two cases can be distinguished. The first one corresponds to the classical Prisoners' Dilemma. The second, to which I will refer to as the non-PD-case, is given by payoffs where the equilibrium payoff is equal or even Pareto-superior to the non-equilibrium diagonal outcome. However, in the non-PD-case, a correlation of out-of-diagonal outcomes yields higher payoffs for both players.

Note that any strictly dominated strategy cannot be played with strictly positive probability in any correlated equilibrium of a finite game. Hence, the argument that only the symmetric non-equilibrium outcome of $\gamma(\mathrm{A})$ may be stabilized is still valid. In consequence, the definition of the critical threshold for the required inequality aversion carries over.

Proposition 4 Let $\gamma(\mathrm{A})\in\Gamma_{\mathrm{sym}}$ be a social dilemma. If $\left(\neg s_1^*,\neg s_2^*\right)$ is stabilizable, then there exists a $\theta^{\mathrm{D}}\in[0,1]$, such that the globally stable equilibrium in case of the Prisoners' Dilemma is characterized by $\theta=\theta^{\mathrm{D}}$ for all individuals in the population. In the non-PD-case the globally stable equilibrium is characterized by $\mathrm{F}\left(\theta^{\mathrm{D}}\right)=1$.

[13] There is plenty of theoretical (Aumann 1974, Brandenburger and Dekel 1987, Nyarko 1994, Lenzo and Sarver 2006 and Koch 2008), empirical (Duffy 2010) and experimental (van Huyck et al. 1992, Brandts and MacLeod 1995 and Seely et al. 2005) support for the relevance of the concept of correlated equilibrium.

Proposition 4 reveals that the qualitative results for the Prisoner's Dilemma type do not change, but gain in precision. In the case of the Prisoner's Dilemma, a precise value of inequality aversion is selected by evolutionary forces. This value corresponds to the lowest value that suffices to transform the dilemma into a coordination game. This gain in precision stems from the fact, that two individuals who are sufficiently inequality-averse to transform the Prisoner's Dilemma into a coordination-game no longer earn the same expected payoff when the concept of correlated equilibrium is applied (see Eq. (11) in the Appendix). However, in the non-PD-case, stabilization of the material non-equilibrium outcome implies an evolutionary disadvantage of inequality aversion, i.e. the reverse result. The intuition behind this is that it is relatively advantageous for inequality-averse individuals if a relatively low weight is put on the disadvantaged one of the two off-diagonal outcomes, which on average earns higher profits then the unique PD-outcome. In other words, it pays to be relatively opportunistic among the inequality-averse players because more weight is put on the off-diagonal outcome which is relatively advantageous. Consequently, while more successful inequality players are selected by evolution, less weight is put on the off-diagonals ultimately leading to a randomization among the two diagonals. This randomization is advantageous in the PD and disadvantageous in the non-PD-case.

Due to Proposition 4, with respect to the generalizability of the results of the Nash equilibrium concept, the preliminary results are ambiguous. A detailed analysis for all classes of games is left for future research. The effect on the precision of prediction regarding the equilibrium distribution of preferences will to some extent also be present when mixed strategies are allowed. This role of randomized play points for the assumption for the problems of coordination and distribution respectively to be strict, which is discussed in the next section.

5.3. Strictness

In this section, I first discuss symmetric problems of coordination. Since a game with two symmetric pure Nash equilibria is always strict, I focus on non-strict problems of coordination with off-diagonal equilibrium payoffs.

Proposition 5 Let $\gamma(A) \in \Gamma_{sym}$ be a problem of coordination such that both equilibria are contestable by one player.

(1) If the less strict equilibrium of $\gamma(A)$ and the favourable equilibrium coincide, then no additional stable equilibria arise. In particular, there is no stable distribution of preferences that assigns a positive share to players by whom both equilibria are contested.

(2) If the less strict equilibrium of $\gamma(A)$ and the favourable equilibrium diverge, then additional stable equilibria arise. In particular, there may be a stable distribution of preferences only with players by whom both equilibria are contested. Furthermore, there may be a stable distribution of preferences where players who contest none of the equilibria and players who contest both equilibria coexist. No stable equilibrium distributions exist with all three types of players: those who contest none of the equilibria, those who contest one equilibrium, and those who contest both equilibria.

In case (1) of Proposition 5 giving up strictness has no consequences with respect to the characterization of the stable distribution of preferences. However, in case (2), the results presented in Proposition 2 experience two qualifications. First, there is a minor qualification with

respect to the existence of an inner equilibrium where opportunistic and inequality-averse individuals coexist. In a non-strict problem of coordination there may also by a stable equilibrium with highly inequality-averse players who have so far been excluded from analysis and opportunistic players. Second, and this is a major qualification, the result implied by Proposition 2 that inequality-averse individuals may at most partially be present in equilibrium is put into perspective. In case (2) of Proposition 5, there may be a stable equilibrium with only (highly) inequality-averse individuals. However, it still holds for medium inequality-averse individuals, i.e. players who contest one equilibrium, that they may at most partially be present in equilibrium. Thus, the assumption for problems of coordination to be strict implies that the evolutionary success of inequality aversion is underestimated. This transfers to the simplified game of life and introduces another case for how inequality aversion could be stabilized among high types in the problem of distribution.

I now turn to problems of distribution. In particular, I am interested in whether the strong prediction of an evolutionary disadvantage for inequality-averse high types carries over to non-strict problems of distribution. Proposition 3 revealed that with one exception, the distribution of inequality aversion among high types is characterized by $F_H\left(\theta_H^R\right)=1$, i.e. only relatively opportunistic players are present in the equilibrium. This exception occurs if the two pure Nash equilibria are Pareto-ranked. If the equilibria are not ranked, then the distribution always exhibits the property of an evolutionary disadvantage of inequality aversion among high types.

Proposition 6 Let $\gamma\left(A^1, A^2\right)$ constitute a non-strict problem of distribution, such that the pure Nash equilibria are not Pareto-ranked. Then the globally stable equilibrium distribution is characterized by $F_H\left(\theta_H^R\right)=1$.

Proposition 6 shows that the disadvantage of inequality-averse high types transfers to non-strict problems of redistribution if equilibria are not Pareto-ranked. However, next to the two cases distinguished in Lemma 3, there is a third class of games that may constitute a problem of distribution if strictness is relinquished, namely that of a game with the unique Nash equilibrium being in mixed strategies. This case and the one with Pareto-ranked equilibria are left for future research.

5.4. Modelling inequality aversion

Finally, I discuss the assumption that individuals care about favourable and unfavourable inequality in the same way. In what follows, I elaborate on the consequence of a more complex model of inequality aversion proposed by Fehr and Schmidt (1999)[14], i.e. $u_{(i,j)}^n = a_{(i,j)}^n - \sigma^n \max\left\{a_{(i,j)}^{-n} - a_{(i,j)}^n, 0\right\} - \omega^n \max\left\{a_{(i,j)}^n - a_{(i,j)}^{-n}, 0\right\}$, $\sigma^n, \omega^n \in [0,1]$. Thus, an individual's preference for equality is no longer characterized by the single parameter θ, but by a pair (σ, ω).

In a dilemma, the Pareto-superior outcome can be stabilized by sufficiently inequality-averse players as they devaluate the material gain from defecting on a cooperative opponent due to the

[14] Note that the concept of inequality aversion according to Bolton and Ockenfels (2000) implies symmetry, but it is left for further research as to whether this notion will change qualitative results of the evolutionary analysis.

induced inequality generated by such a defection. Hence, in case of a symmetric dilemma, it is not inequality aversion *per se* but aversion against favourable outcomes that is required to support cooperation. With respect to problems of coordination, two cases were distinguished in Proposition 2. In the first case, the destabilized equilibrium is materially favourable for inequality-averse players. Hence, an aversion against favourable inequality is decisive. In the second case, the reverse holds, i.e. the destabilized equilibrium is materially favourable for selfish individuals. Therefore, an aversion against unfavourable inequality becomes relevant. Below, Proposition 7 summarizes these insights.

Proposition 7 Let $\gamma(\mathrm{A}) \in \Gamma_{\mathrm{sym}}$ be a social dilemma, then $\theta^{\mathrm{D}} = \omega^{\mathrm{D}}$. Let $\gamma(\mathrm{A}) \in \Gamma_{\mathrm{sym}}$ be a strict problem of coordination. If the equilibria are contestable then:

1. if the destabilized equilibrium is materially favourable for inequality-averse individuals then $\theta^{\mathrm{D}} = \omega^{\mathrm{D}}$.
2. if the destabilized equilibrium is materially favourable for selfish individuals then $\theta^{\mathrm{C}} = \sigma^{\mathrm{C}}$.

For problems of distribution, there is no such clear assignment for the thresholds of Proposition 3. To see this, consider the example given in Table 2 which belongs to the first case in Proposition 3. The game presented in Table 2 has two pure non-Pareto-ranked Nash equilibria on the diagonal. I consider the case where none of the equilibria is contestable by high types (column player) and the $(0,0)$ is contestable by low types (row player).

	0	1
0	$\mathrm{A}-\theta^2\lvert\mathrm{A}-\mathrm{a}\rvert$ $\mathrm{a}-\theta^1\lvert\mathrm{A}-\mathrm{a}\rvert$	$\mathrm{B}-\theta^2\lvert\mathrm{B}-\mathrm{b}\rvert$ $\mathrm{b}-\theta^1\lvert\mathrm{B}-\mathrm{b}\rvert$
1	$\mathrm{C}-\theta^2\lvert\mathrm{C}-\mathrm{c}\rvert$ $\mathrm{c}-\theta^1\lvert\mathrm{C}-\mathrm{c}\rvert$	$\mathrm{D}-\theta^2\lvert\mathrm{D}-\mathrm{d}\rvert$ $\mathrm{d}-\theta^1\lvert\mathrm{D}-\mathrm{d}\rvert$

Table 2: $\mathrm{A}>\mathrm{B},\ \mathrm{D}>\mathrm{C},\ \mathrm{a}>\mathrm{c},\ \mathrm{d}>\mathrm{b},\ \mathrm{a}<\mathrm{d}<\mathrm{D}<\mathrm{A}$

The $(0,0)$-equilibrium is contested by a low type if and only if:

$$\mathrm{a}-\theta^1\lvert\mathrm{A}-\mathrm{a}\rvert<\mathrm{c}-\theta^1\lvert\mathrm{C}-\mathrm{c}\rvert. \tag{5}$$

The example implies that $\mathrm{A}-\mathrm{a}>0$, but there is no relation for $\mathrm{C}-\mathrm{c}$. If the outcome of playing $(1,0)$ also favours high types, i.e. $\mathrm{C}-\mathrm{c}>0$ then (5) becomes

$$\mathrm{a}-\sigma^1(\mathrm{A}-\mathrm{a})<\mathrm{c}-\sigma^1(\mathrm{C}-\mathrm{c}). \tag{6}$$

This suggests that if high types are favoured, no matter which strategies are played, then the threshold $\theta_{\mathrm{L}}^{\mathrm{R}}$ in Proposition 3 refers to inequality aversion concerning unfavourable outcomes.

If however, the reverse is true, i.e. $\mathrm{C}-\mathrm{c}<0$, then (5) becomes

$$a-\sigma^1\left(A-a\right)<c-\omega^1\left(c-C\right). \tag{7}$$

In this case, both parameters become relevant and no clear assignment to the thresholds in Proposition 3 is possible. Rewriting (7) as

$$\sigma^1>\frac{a-c}{A-a}+\frac{c-C}{A-a}\omega^1 \tag{8}$$

reveals that the threshold θ_L^R needs to be substituted by a linear condition, which separates the two dimensional parameter space characterizing the preference for equality by (σ,ω)-pairs. Thus, individuals with (σ,ω) located above (below) that line can (not) contest the equilibrium. A similar argument applies to high types. If $B-b>0$ then $(0,0)$ is destabilized by a high type if and only if $A-\omega^2\left(A-a\right)<B-\omega^2\left(B-b\right)$. Thus, for the example given in Table 2, θ_H^R in Proposition 3 refers to inequality aversion concerning favourable outcomes. As for low types, if the reverse holds, i.e. $B-b<0$, then both parameters become relevant and θ_H^R needs to be substituted by a linear condition in the fashion of (8).

In short, regarding the assumption of a uniform distribution over the set of all pure Nash equilibria, it turns out that the neutrality of the distribution with respect to the evolutionary success of inequality aversion implies symmetry, and symmetry implies uniformity when 2x2 games are considered. With respect to generalizability of the results for the Nash equilibrium concept (Proposition 1-Proposition 3) the preliminary results (Proposition 4) are ambiguous and further research is needed to fully understand the sensitivity of the results regarding the coarseness of the applied equilibrium concept relative to the Nash equilibrium. In terms of the assumption regarding the problem of coordination to be strict, the degree of disadvantageousness of inequality aversion (Proposition 2) is put into perspective by the possible existence of a stable equilibrium with only inequality-avers players. However, this phenomenon may only occur for case (2) in Lemma 2 and requires that the equilibria are less strict for those players who are disfavoured in the equilibria (see case (2) in Proposition 5). However, if the reverse is true, no additional equilibria arise if the assumption of strictness is relaxed. Proposition 6 proves that the strong prediction of an evolutionary disadvantage for inequality-averse high types also holds for non-strict problems of distribution if equilibria of $\gamma\left(A^1,A^2\right)$ are not Pareto-ranked. Finally, if a model of preferences that distinguishes between aversion against favourable and unfavourable inequality is applied, then the results of Proposition 1 (dilemma) and Proposition 2 (problem of coordination) carry over. However, the parameter measuring inequality aversion in the simplified model (θ) is replaced by either the parameter for aversion against favourable (ω) or by the one for unfavourable (σ) inequality. For problems of distribution the discussion in 5.4 suggests that the thresholds of Proposition 3 are either replaced by a threshold referring to aversion against favourable (high types) or unfavourable (low types) inequality or by a linear constraint relating the two parameters of the alternative model of inequality aversion.

6. Conclusion

Following the argument for a requirement to analyse the evolution of preference in an environment that comprises at best all relevant classes of games individuals engage in, I have suggested a particular notion of a simplified game of life. Within that framework, I have analysed the evolution of a particular type of other-regarding preference, namely that of inequality aversion.

The analysis in the separate environments revealed that if inequality aversion has leverage on the set of equilibria being played, then inequality aversion enjoys a global evolutionary advantage over more selfish preferences in a dilemma. In the class of problems of coordination inequality, aversion surprisingly faces a weak evolutionary disadvantage in the sense that a stable inner equilibrium exists at most where relative inequality-averse and relative selfish players coexist. In all other cases, relatively inequality-averse players will eventually disappear. In the problem of distribution, a preference for equality will always be favoured by evolutionary selection dynamics among those individuals disfavoured by the problem. For those individuals favoured in the problem of distribution, in all cases up to one, inequality aversion will eventually disappear. I consider these predictions in light of the considered generality as rather strong. Furthermore, due to the exemplary variations of assumptions discussed in Section 5, these predictions appear quite robust.

The simplified game of life that comprises all three types of interaction, as expected, gives rise to a greater variety in potential equilibrium distributions of preferences. In particular, the surprisingly strong predictions for the single environments are put into perspective. The global advantage of inequality-averse players in the dilemma and the global disadvantage for inequality-averse high types in almost all cases experiences significant qualification. In particular, whenever the interplay of the dilemma and the problem of distribution allows for a locally stable equilibrium with only inequality-averse players, then this transfers to the simplified game of life, i.e. inequality aversion may also be present among high types. On the other hand, the expected stabilization of inner equilibria in which relatively inequality-averse individuals and relatively selfish individuals coexist occurs if and only if the problem of coordination shows the same feature, i.e. the coexistence of both types.

The contribution of the paper is threefold. First, the different results in the single-game environments and in the simplified game of life again underpin the necessity to carefully select the relevant game environment in any study of the evolution of preferences. Otherwise, any negative or positive results with respect to the rationalization of a particular preference may only point to a potential evolutionary force, which however may not be decisive if all relevant environments are considered. Second, the paper methodologically contributes to the field of evolutionary economics by offering a precise suggestion of an evolutionary framework for the study of the evolution of preferences. Third, this paper provides an evolutionary rationale for the presence of inequality aversion within the compound environment of the simplified game of life.

References

Andreoni, J. A., J. H. Miller. 1993. Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence. *Economic Journal* **103**(418) 570–585.

Ashraf, N., I. Bohnet, N. Piankov. 2006. Decomposing trust and trustworthiness. *Experimental Economics* **9**(3) 193–208.

Aumann, R. J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**(1) 67–96.

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

Blanco, M., D. Engelmann, H. T. Normann. 2011. A within-subject analysis of other-regarding preferences. *Games and Economic Behavior* **72**(2) 321–338.

Bolton, G. E., A. Ockenfels. 2000. ERC: A theory of equity, reciprocity, and competition. *American Economic Review* 166–193.

Brandenburger, A., E. Dekel. 1987. Rationalizability and correlated equilibria. *Econometrica: Journal of the Econometric Society* **55** 1391–1402.

Brandts, J., MacLeod, W. B. 1995. Equilibrium selection in experimental games with recommended play. *Games and Economic Behavior* **11**(1) 36–63.

Calvo-Armengol, A. 2006. The Set of Correlated Equilibria of 2x2 Games. Mimeo.

Chaudhuri, A., L. Gangadharan. 2007. An experimental analysis of trust and trustworthiness. *Southern Economic Journal* 959–985.

Duffy, J., Feltovich, N. 2010 Correlated Equilibria, Good and Bad: An Experimental Study. *International Economic Review* **51**(3) 701–721.

Fehr, E., K. M. Schmidt. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* **114**(3) 817–868.

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

Güth, W., S. Napel. 2006. Inequality Aversion in a Variety of Games-An Indirect Evolutionary Analysis*. *The Economic Journal* **116**(514) 1037–1056.

Güth, W., C. Schmidt, M. Sutter. 2003. Fairness in the mail and opportunism in the internet: A newspaper experiment on ultimatum bargaining. *German Economic Review* **4**(2) 243–265.

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.

Guttman, J. M. 2003. Repeated interaction and the evolution of preferences for reciprocity. *Economic Journal* **113**(489) 631-656.

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

Koch, L. P. 2008. *Evolution and correlated equilibrium,* Bonn econ discussion papers.

Koçkesen, L., E. A. Ok, R. Sethi. 2000a. Evolution of interdependent preferences in aggregative games. *Games and Economic Behavior* **31**(2) 303–310.

Koçkesen, L., E. A. Ok, R. Sethi. 2000b. The strategic advantage of negatively interdependent preferences. *Journal of Economic Theory* **92**(2) 274–299.

Lenzo, J., T. Sarver. 2006. Correlated equilibrium in evolutionary models with subpopulations. *Games and Economic Behavior* **56**(2) 271–284.

Milgrom, P. R., D. C. North, B. R. Weingast. 1990. The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics & Politics* **2**(1) 1–23.

Nyarko, Y. 1994. Bayesian learning leads to correlated equilibria in normal form games. *Economic Theory* **4**(6) 821–841.

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

Poulsen, A., O. Poulsen. 2006. Endogenous preferences and social-dilemma institutions. *Journal of Institutional and Theoretical Economics JITE* **162**(4) 627–660.

Robinson, D., D. Goforth. 2005. *The Topology of the 2× 2 Games*. New York: Routledge.

Samuelson, L. 1997. Evolutionary games and equilibrium selection. MIT Press (Cambridge, Mass.).

Schotter, A. 1981. The economic theory of social institutions. Cambridge University Press (Cambridge Eng. and New York).

Seely, B., van Huyck, J., Battalio, R. 2005. Credible assignments can improve efficiency in laboratory public goods games. *Journal of Public Economics* **89**(8) 1437–1455.

Sethi, R., E. Somanathan. 2001. Preference evolution and reciprocity. *Journal of Economic Theory* **97**(2) 273–297.

Slonim, R., E. Garbarino. 2008. Increases in trust and altruism from partner selection: Experimental evidence. *Experimental Economics* **11**(2) 134–153.

Sugden, R. 1986. The economics of rights, co-operation and welfare. Basil Blackwell Oxford.

Ullmann-Margalit, E. 1977. *The emergence of norms*. Clarendon Press Oxford.

Van Huyck, J. B., Gillette, A. B., Battalio, R. C. 1992. Credible assignments in coordination games. *Games and Economic Behavior* **4**(4) 606–626.

Weibull, J. W. 1997. *Evolutionary game theory*. MIT press.

Yamagishi, T., N. Mifune, Y. Li, M. Shinada, H. Hashimoto, Y. Horita, A. Miura, K. Inukai, S. Tanida, T. Kiyonari. 2013. Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes* **120**(2) 260–271.

Appendix

The proofs of Lemma 1 - Lemma 3 and the proof of Proposition 7 are omitted as the argument is given in detail in the paper.

Proof of Proposition 1:

A symmetric dilemma can be represented by the following matrix $A=\begin{pmatrix} a & c \\ b & d \end{pmatrix}$, showing the payoffs for the row player. Without loss of generality I assume $b > a,\ d > c$, i.e. '1' is the dominant strategy. According to Lemma 1, $a > d$ must hold. This implies the following ordering of parameters $b > a > d > c$. Define $\theta^D = \frac{b-a}{b-c} \in (0,1)$. In terms of utility, two individuals with inequality aversion θ^1 and θ^2 respectively give rise to the following bimatrix:

	0	1
0	a a	$b-\theta_2(b-c)$ $c-\theta_1(b-c)$
1	$c-\theta_1(b-c)$ $b-\theta_2(b-c)$	d d

Table 3: Payoffs in the dilemma $\gamma(U^1,U^2)$.

In the following I will distinguish two cases. The first case corresponds to a match of two players with a degree of inequality aversion above the threshold θ^D. In the second case, for at least one player this condition is violated.

(i) $\theta^1,\theta^2 \geq \theta^D$

$(0,0),(1,1)$ are the two pure Nash equilibria over which individuals randomize with equal weight and both gain a material payoff of: $\left(\frac{a+d}{2}\right)$

(ii) $\theta^1 \vee \theta^2 < \theta^D$

'1' remains for at least one agent the dominant strategy. Hence, both individuals will earn: (d)

Note that all individuals with $\theta \geq \theta^D$ earn the same expected payoff $\Pi^{\theta \geq \theta^D} = F(\theta^D)d + (1-F(\theta^D))\frac{a+d}{2}$, whereas individuals with $\theta < \theta^D$ earn $\Pi^{\theta<\theta^D} = F(\theta^D)d + (1-F(\theta^D))d = d$. Hence, as long as there are some individuals with a degree of inequality aversion above θ^D those players face an evolutionary advantage because $\Pi^{\theta \geq \theta^D} - \Pi^{\theta<\theta^D} = (1-F(\theta^D))\frac{a-d}{2} > 0$. Hence, the globally stable distribution of inequality aversion is characterized by $F^{\infty}(\theta^D) = 0$. The advantage increases with the share of sufficiently inequality-averse players, i.e. $\frac{\partial(\Pi^{\theta \geq \theta^D} - \Pi^{\theta<\theta^D})}{\partial(1-F(\theta^D))} = \frac{a-d}{2} > 0$. QED

Proof of Proposition 2:

A symmetric problem of coordination can be represented by $A=\begin{pmatrix} a & c \\ b & d \end{pmatrix}$, showing the payoffs for the row player. For a game with the Nash equilibria on the diagonal $a>b,\ d>c$ holds. Hence, any degree of inequality aversion leaves the set of pure Nash equilibria unchanged. Thus, any match of two players will generate the same payoff, the average of the two pure Nash equilibria. Therefore the distribution of preferences will be determined by initial conditions and random shifts. Hence, I shall assume for the Nash equilibria to lie on the off-diagonal, i.e. w.l.o.g. $b>a,\ d<c$. In terms of utility, two individuals with inequality aversion θ^1 and θ^2 respectively give rise to a bimatrix as depicted in Table 3.

Define $\theta^C_{(0,1),2}=\theta^C_{(1,0),1}\equiv\frac{b-a}{|b-c|}>0,\ \theta^C_{(0,1),1}=\theta^C_{(1,0),2}\equiv\frac{c-d}{|b-c|}>0$. These thresholds represent the ratio of the material incentive to stick to the considered (material) equilibrium and the gain in non-material terms from deviation stemming from an increasing equality. A threshold above one represents a situation where the maximum gain in equality is smaller than the material loss from deviating from (material) equilibrium behaviour. In other words, no level of inequality aversion can destabilize this equilibrium. If for a player $\theta\geq\theta^C_{(0,1),2}\left(\theta^C_{(1,0),2}\right)$, then for this player the equilibrium $(0,1)((1,0))$ is contestable. In the following subsections (1)-(3), I consider the different possible matches according to the relation of the thresholds and the involved players' inequality aversion.

(1) $\theta^1>\theta^C_{(1,0),1},\theta^2>\theta^C_{(0,1),2}$

, i.e. both equilibria are contestable (by different players) and are indeed destabilized. The strategy-tuple $(0,0)$ is stabilized. Now two cases can be distinguished. First '0' has become the dominant strategy for at least one player (subcases a and c) or $(1,1)$ is also stabilized (subcase b).

a) $\theta^1<\theta^C_{(0,1),1},\theta^2<\theta^C_{(1,0),2}\vee\theta^C_{(0,1),1},\theta^C_{(1,0),2}>1$

, i.e. $(1,1)$ is not stabilized either because inequality aversion is too weak or the equilibria are not contestable by the considered players. In that case '0' becomes the dominant strategy and the unique Nash equilibrium is given by $(0,0)$. (a,a)

b) $\theta^1>\theta^C_{(0,1),1},\theta^2>\theta^C_{(1,0),2}\wedge\theta^C_{(0,1),1},\theta^C_{(1,0),2}<1$

, i.e. $(1,1)$ also becomes an equilibrium. There are now the two pure Nash equilibria $(1,1)$ and $(0,0)$.

$$\left(\frac{a+d}{2},\frac{a+d}{2}\right)$$

c) $\theta^1<\theta^C_{(0,1),1}\vee\theta^C_{(0,1),1}>1,\theta^2>\theta^C_{(1,0),2}\wedge\theta^C_{(1,0),2}<1$

, i.e. '0' is the dominant strategy for player one and '0' is the best response for player two.

(a,a)

(2) $\theta^1 < \theta^C_{(1,0),1}, \theta^2 < \theta^C_{(0,1),2} \vee \theta^C_{(0,1),2}, \theta^C_{(1,0),1} > 1$

, i.e. $(0,0)$ is not stabilized either because inequality aversion is too weak or the equilibria are not contestable by the considered players.

a) $\theta^1 < \theta^C_{(0,1),1}, \theta^2 < \theta^C_{(1,0),2} \vee \theta^C_{(0,1),1}, \theta^C_{(1,0),2} > 1$

, i.e. $(1,1)$ is not stabilized either because inequality aversion is too weak or the equilibria are not contestable by the considered players. The sets of Nash equilibria of $\gamma(A)$ and $\gamma(U^1, U^2)$ coincide. $\left(\frac{b+c}{2}, \frac{b+c}{2}\right)$

b) $\theta^1 > \theta^C_{(0,1),1}, \theta^2 > \theta^C_{(1,0),2} \wedge \theta^C_{(0,1),1}, \theta^C_{(1,0),2} < 1$

, i.e. both material equilibria are contestable and are indeed destabilized. In that case '1' becomes the dominant strategy. (d,d)

(3) w.l.o.g. $\theta^1 < \theta^C_{(1,0),1}$, $\theta^2 > \theta^C_{(0,1),2} \vee \theta^C_{(0,1),2}, \theta^C_{(1,0),1} < 1$ (player 1 is selfish, player 2 is inequality-averse), i.e. one players' inequality aversion makes one equilibrium contestable.

a) $\theta^2 < \theta_{(1,0),2} \vee \theta_{(1,0),2} > 1$

, i.e. this player inequality aversion is either too weak or the remaining equilibrium is not contestable by this player. In that case '0' is the dominant strategy of this player. Two cases can be distinguished for the remaining player.

(i) $\theta^1 < \theta_{(0,1),1} \vee \theta_{(0,1),1} > 1$

, i.e. this opponents' inequality aversion is either too weak to or the remaining equilibrium is not contestable from this perspective. (b,c)

(ii) $\theta^1 > \theta_{(0,1),1} < 1$

, i.e. the remaining equilibrium is also contestable and indeed destabilized. (a,a)

b) $\theta^2 > \theta_{(1,0),2} < 1$

, i.e. this player makes both equilibria contestable and indeed both equilibria are destabilized.

(i) $\theta^1 > \theta_{(0,1),1} < 1$

, i.e. '1' becomes the dominant strategy of this player (d,d)

(ii) $\theta^1 < \theta_{(0,1),1} \vee \theta_{(0,1),1} > 1$

There is a unique mixed equilibrium which is played (Π^{mix}, Π^{mix})

Note that strictness excludes the cases 1b), 1c), 3b). Table 4 depicts equilibrium payoffs in the various matches for the case of $\theta^C_{(0,1),2} = \theta^C_{(1,0),1} \equiv \theta^C < 1$, i.e. the case where both equilibria are contestable by different players.

	$\theta_2 < \theta^C$	$\theta^C < \theta_2$
$\theta_1 < \theta^C$	$\left(\frac{b+c}{2}, \frac{b+c}{2}\right)$	(b,c)
$\theta^C < \theta_1$	(c,b)	(a,a)

Table 4: Equilibrium payoffs according to the degree of inequality aversion of the matched players.

Note that individuals with $\theta \geq \theta^C$ earn the same expected payoff $\Pi^{\theta \geq \theta^C} = F(\theta^C)c + (1 - F(\theta^C))a$, whereas individuals with $\theta < \theta^C$ earn $\Pi^{\theta < \theta^C} = F(\theta^C)\frac{b+c}{2} + (1 - F(\theta^C))b$. Hence, $\Pi^{\theta \geq \theta^C} - \Pi^{\theta < \theta^C} = F(\theta^C)\frac{c-b}{2} + (1 - F(\theta^C))(a-b)$. Note that $(\Pi^{\theta \geq \theta^C} - \Pi^{\theta < \theta^C})(F(\theta^C) = 0) = a - b < 0$. If $b > c$, then $(\Pi^{\theta \geq \theta^C} - \Pi^{\theta < \theta^C})(F(\theta^C) = 1) = \frac{c-b}{2} < 0$ and the globally stable equilibrium is characterized by $F(\theta^C) = 1$. Furthermore $\frac{\partial(\Pi^{\theta \geq \theta^C} - \Pi^{\theta < \theta^C})}{\partial(1 - F(\theta^C))} = a - \frac{b+c}{2}$.

If the reverse holds, i.e. $b < c$, then $(\Pi^{\theta \geq \theta^C} - \Pi^{\theta < \theta^C})(F(\theta^C) = 1) = \frac{c-b}{2} > 0$ and as a consequence there exist a globally stable inner equilibria characterized by $F(\theta^C) = \frac{b-a}{\frac{b+c}{2} - a} = \theta^C \frac{c-b}{\frac{b+c}{2} - a}$. Furthermore, $\frac{\partial(\Pi^{\theta \geq \theta^C} - \Pi^{\theta < \theta^C})}{\partial(1 - F(\theta^C))} = a - \frac{b+c}{2} < 0$. The case $\theta^C_{(0,1),1} = \theta^C_{(1,0),2} < 1$ is analysed in the analogue way $(a \leftrightarrow d, b \leftrightarrow c)$. With the definitions for $AP_{(i,i)}$ and d in the text the claim follows. QED

Proof of Proposition 3:

Let me first consider case 1 of Lemma 3 with payoffs given by $A^1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $A^2 = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$. W.l.o.g. I will consider a game with Nash equilibria on the diagonal (relabeling the strategies for one player transforms such a game in a game with equilibria on the off-diagonal and vice versa), i.e. $A > B$, $D > C$ and $d > b$, $a > c$. W.l.o.g. let player two be the type who is favoured by the problem of distribution, i.e. $A > a$ and $D > d$. The assumption that the two pure Nash equilibria are not Pareto-ranked leaves us with two possibilities, either $a < d < D < A$ or $d < a < A < D$. W.l.o.g. I will assume the first relations to hold. This implies that the equilibrium $(1,1)$ is characterized by a strictly lower degree of inequality. In terms of utility, two individuals with inequality aversion θ_1 and θ_2 respectively give rise to the following bimatrix:

	0	1
0	$A - \theta_2\lvert A - a\rvert$ $a - \theta_1\lvert A - a\rvert$	$B - \theta_2\lvert B - b\rvert$ $b - \theta_1\lvert B - b\rvert$
1	$C - \theta_2\lvert C - c\rvert$ $c - \theta_1\lvert C - c\rvert$	$D - \theta_2\lvert D - d\rvert$ $d - \theta_1\lvert D - d\rvert$

Table 5: Payoffs in the problem of distribution $\gamma(U^1, U^2)$.

Note that:

- (i) $D - \theta_2\lvert D - d\rvert \geq d > a > c$
- (ii) $D - \theta_2\lvert D - d\rvert > C - \theta_2\lvert C - c\rvert$, because for $\theta_2 = 0$ $D > C$ and for $\theta_2 = 1$ $d > c \geq C - \lvert C - c\rvert$
- (iii) $\lvert A - a\rvert > \lvert D - d\rvert$

Before I analyse the different types of matches, I will define the following thresholds: $\theta^{R}_{(0,0),2} \equiv \frac{A-B}{|A-a|-|B-b|}, \theta^{R}_{(1,1),2} \equiv \frac{D-C}{|D-d|-|C-c|}, \theta^{R}_{(0,0),1} \equiv \frac{a-c}{|A-a|-|C-c|}, \theta^{R}_{(1,1),1} \equiv \frac{d-b}{|D-d|-|B-b|}$

Note that due to (ii) $\theta^{R}_{(1,1),2} > 1$, i.e. the equilibrium $(1,1)$ is not contestable for player two. Let for all other thresholds $\theta^{R}_{(0,0),2}, \theta^{R}_{(0,0),1}, \theta^{R}_{(1,1),1} \in (0,1)$, i.e. both equilibria are contestable, $(1,1)$ only by player one, $(0,0)$ by both players.

1. $\theta > \theta^{R}_{(0,0),2}$ ('1' is the dominant strategy for player 2)
   (1) $\theta < \theta^{R}_{(1,1),1}$ (d, D)
   (2) $\theta > \theta^{R}_{(1,1),1}$ (b, B)
2. $\theta < \theta^{R}_{(0,0),2}$

a) $\theta < \theta^{R}_{(0,0),1}$
   (1) $\theta < \theta^{R}_{(1,1),1}$: $(0,0), (1,1)$ remain both equilibria $\left(\frac{a+d}{2}, \frac{A+D}{2}\right)$
   (2) $\theta > \theta^{R}_{(1,1),1}$: '0' is the dominant strategy for player 1 (a, A)

b) $\theta > \theta^{R}_{(0,0),1}$
   (1) $\theta < \theta^{R}_{(1,1),1}$: '1' is the dominant strategy for player 1 (d, D)
   (2) $\theta > \theta^{R}_{(1,1),1}$: there is a unique mixed equilibrium $\left(\Pi_1^{mixed}, \Pi_2^{mixed}\right)$

Note that all other values of threshold can be analysed via 1. and 2., because for $\theta^{R}_{(i,i),j} \geq 1$ simply the subcase $\theta > \theta^{R}_{(i,i),j}$ is left out of the analysis. The same holds for negative values, i.e. $\theta^{R}_{(i,i),j} < 0$ simply the case $\theta > \theta^{R}_{(i,i),j}$ is left out of the analysis. The last statement may need some clarification. A negative threshold implies that a deviation from an equilibrium (not only decreases the material payoff, but also) increases inequality. In that case, for no level of inequality aversion a deviation from equilibrium becomes profitable in utility terms. This is equivalent to a situation where an equilibrium is contestable, but inequality aversion is too weak to indeed destabilize the equilibrium, i.e. $\theta > \theta^{R}_{(i,i),j}$ is left out. Note that strictness of the problem of distribution excludes case 2b (2). Table 6 depicts equilibrium payoffs in the various matches.

	$0 \leq \theta \leq \theta^{R}_{(0,0),2}, \theta^{R}_{(0,0),2} < 0, \theta^{R}_{(0,0),2} \geq 1$	$\theta > \theta^{R}_{(0,0),2}$
$0 \leq \theta \leq \theta^{R}_{(0,0),1}$ $\theta^{R}_{(0,0),1} < 0$ $\theta^{R}_{(0,0),1} \geq 1$	(1) $\left(\frac{a+d}{2}, \frac{A+D}{2}\right)$ (2) (a, A)	(1) (d, D) (2) (b, B)
$\theta > \theta^{R}_{(0,0),1}$	(1) (d, D)	(1) (d, D)

Table 6: Equilibrium payoffs according to the degree of inequality aversion of the matched players.

$\theta^{R}_{H,L} = \min\left\{\hat{\theta}^{R}_{H,L}, \breve{\theta}^{R}_{H,L}\right\}$, $\theta^{R}_{(1,1),2} > 1$ and $(1,1)$ being the materially more equal distributed equilibrium imply that $\theta^{R}_{H} = \breve{\theta}^{R}_{H} = \theta^{R}_{(0,0),2}$. Furthermore, $\theta^{R}_{L} = \min\left\{\theta^{R}_{(0,0),1}, \theta^{R}_{(1,1),1}\right\}$.

1. $\theta^{R}_{(0,0),2} \geq 1 \vee \theta^{R}_{(0,0),2} < 0$ (materially more unequal distributed equilibrium is not contestable for high type)

Obviously all high types will earn the same payoff. Hence, the distribution of inequality aversion among high types is determined by initial conditions and random shift. With respect to low types, let me first consider the case when the materially more unequal distributed equilibrium is contestable, i.e. $\theta^{R}_{L} = \theta^{R}_{(0,0),1}$. In that case, payoffs for low types are given by $\Pi_{L}^{\theta<\theta^{R}_{L}} = \frac{a+d}{2}$ and $\Pi_{L}^{\theta\geq\theta^{R}_{L}} = d$. Hence, payoff difference is given by $\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}} = \frac{d-a}{2} > 0$ and the globally stable equilibrium is characterized by $F_{L}\left(\theta^{R}_{L}\right) = 0$. Furthermore, $\frac{\partial\left(\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}}\right)}{\partial\left(1-F_{H}\left(\theta^{R}_{H}\right)\right)} = 0$. I will now turn to the case, where the materially less unequal equilibrium is contestable for the low type, i.e. $\theta^{R}_{L} = \theta^{R}_{(1,1),1}$. Payoff difference is given by $\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}} = d - a > 0$. Hence, the globally stable equilibrium is characterized by $F_{L}\left(\theta^{R}_{L}\right) = 0$. Furthermore, $\frac{\partial\left(\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}}\right)}{\partial\left(1-F_{H}\left(\theta^{R}_{H}\right)\right)} = 0$.

Finally, if none of the material equilibria is contestable for the low type, the distribution of inequality aversion among low types is determined by initial conditions and random shift.

2. $\theta^{R}_{(0,0),2} \in (0,1)$

Let me first consider the case, when the materially more unequal distributed equilibrium is contestable for the low type, i.e. $\theta^{R}_{L} = \theta^{R}_{(0,0),1}$. In that case, payoffs for low types are given by $\Pi_{L}^{\theta<\theta^{R}_{L}} = F_{H}\left(\theta^{R}_{H}\right)\frac{a+d}{2} + \left(1-F_{H}\left(\theta^{R}_{H}\right)\right)d$ and $\Pi_{L}^{\theta\geq\theta^{R}_{L}} = F_{H}\left(\theta^{R}_{H}\right)d + \left(1-F_{H}\left(\theta^{R}_{H}\right)\right)d = d$, for high types $\Pi_{H}^{\theta<\theta^{R}_{H}} = F_{L}\left(\theta^{R}_{L}\right)\frac{A+D}{2} + \left(1-F_{L}\left(\theta^{R}_{L}\right)\right)D$ and $\Pi_{H}^{\theta\geq\theta^{R}_{H}} = F_{L}\left(\theta^{R}_{L}\right)D + \left(1-F_{H}\left(\theta^{R}_{H}\right)\right)D = D$. Hence, differences are given by $\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}} = F_{H}\left(\theta^{R}_{H}\right)\frac{d-a}{2} > 0$ and $\Pi_{H}^{\theta\geq\theta^{R}_{H}} - \Pi_{H}^{\theta<\theta^{R}_{H}} = -F_{L}\left(\theta^{R}_{L}\right)\frac{A-D}{2} < 0$. Hence, the globally stable equilibrium is characterized by $F_{H}\left(\theta^{R}_{H}\right) = 1, F_{L}\left(\theta^{R}_{L}\right) = 0$. Furthermore, $\frac{\partial\left(\Pi_{H}^{\theta\geq\theta^{R}_{H}} - \Pi_{H}^{\theta<\theta^{R}_{H}}\right)}{\partial\left(1-F_{L}\left(\theta^{R}_{L}\right)\right)} = \frac{A-D}{2} > 0$ and $\frac{\partial\left(\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}}\right)}{\partial\left(1-F_{H}\left(\theta^{R}_{H}\right)\right)} = -\frac{d-a}{2} < 0$.

I will now turn to the case, where the materially less unequal equilibrium is contestable for the low type, i.e. $\theta^{R}_{L} = \theta^{R}_{(1,1),1}$. In that case, payoffs are given by $\Pi_{L}^{\theta<\theta^{R}_{L}} = F_{H}\left(\theta^{R}_{H}\right)a + \left(1-F_{H}\left(\theta^{R}_{H}\right)\right)b$ and $\Pi_{L}^{\theta\geq\theta^{R}_{L}} = F_{H}\left(\theta^{R}_{H}\right)d + \left(1-F_{H}\left(\theta^{R}_{H}\right)\right)d = d$, for high types $\Pi_{H}^{\theta<\theta^{R}_{H}} = F_{L}\left(\theta^{R}_{L}\right)A + \left(1-F_{L}\left(\theta^{R}_{L}\right)\right)D$ and $\Pi_{H}^{\theta\geq\theta^{R}_{H}} = F_{L}\left(\theta^{R}_{L}\right)B + \left(1-F_{H}\left(\theta^{R}_{H}\right)\right)D$. Hence, differences are given by $\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}} = d - b - F_{H}\left(\theta^{R}_{H}\right)(a-b)$ and $\Pi_{H}^{\theta\geq\theta^{R}_{H}} - \Pi_{H}^{\theta<\theta^{R}_{H}} = F_{L}\left(\theta^{R}_{L}\right)(B-A) \leq 0$. Note that $\left(\Pi_{L}^{\theta\geq\theta^{R}_{L}} - \Pi_{L}^{\theta<\theta^{R}_{L}}\right)\left(F_{H}\left(\theta^{R}_{H}\right) = 0\right) = d - b > 0$ and

$\left(\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}\right)\left(F_H\left(\theta_H^R\right)=1\right)=d-a>0$. Hence, the globally stable equilibrium is given by $F_H\left(\theta_H^R\right)=1, F_L\left(\theta_L^R\right)=0$. Furthermore, $\frac{\partial\left(\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}\right)}{\partial\left(1-F_L\left(\theta_L^R\right)\right)}=A-B>0$ and $\frac{\partial\left(\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}\right)}{\partial\left(1-F_H\left(\theta_H^R\right)\right)}=a-b$.

Finally, if none of the material equilibria is contestable for the low type, the distribution of inequality aversion among low types is determined by initial conditions and random shift. Payoff difference for high types is given by $\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}=F_L\left(\theta_L^R\right)(B-A)\leq 0$ with $\frac{\partial\left(\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}\right)}{\partial\left(1-F_L\left(\theta_L^R\right)\right)}=A-B>0$. Hence, the globally stable equilibrium is given by $F_H\left(\theta_H^R\right)=1$.

Let us now turn to case (2) of Lemma 3 with two Pareto-ranked equilibria. Given the assumption parallel to case A this leaves us with two possibilities, either $d<a\wedge D<A$ or $a<d\wedge A<D$. For ease of comparability to case (1) of Lemma 3, I will w.l.o.g. assume $a<d\wedge A<D$ to hold. Hence, the only relation that has changed in comparison to case (1) is the one between parameters A and D. Note that inequalities (i) and (ii) still hold. Again, due to (ii) $\theta_{(1,1),2}^R>1$, i.e. the equilibrium $(1,1)$ is not contestable for player two. That is, in case (2) the Pareto-superior equilibrium is not contestable for high types. The equilibrium analysis is equivalent to case A and equilibrium payoffs correspond to those in Table 6, their relation to each other may have changed though.

1. $\theta_{(0,0),2}^R\geq 1\vee\theta_{(0,0),2}^R<0$ (Pareto-inferior equilibrium is not contestable for high type)

Parameters A and D are not involved, hence the results are identical to those in case A.

2. $\theta_{(0,0),2}^R\in(0,1)$

Let me first consider the case when the Pareto-inferior equilibrium is contestable for the low type, i.e. $\theta_L^R=\theta_{(0,0),1}^R$. Payoffs are equivalent to case (1). Differences in payoffs among low types are given by $\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}=F_H\left(\theta_H^R\right)\frac{d-a}{2}>0$ and by $\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}=-F_L\left(\theta_L^R\right)\frac{A-D}{2}>0$ among high types. Hence, the globally stable equilibrium is given by $F_H\left(\theta_H^R\right)=F_L\left(\theta_L^R\right)=0$. Furthermore, $\frac{\partial\left(\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}\right)}{\partial\left(1-F_L\left(\theta_L^R\right)\right)}=\frac{A-D}{2}<0$ and $\frac{\partial\left(\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}\right)}{\partial\left(1-F_H\left(\theta_H^R\right)\right)}=-\frac{d-a}{2}<0$.

I will now turn to the case, where the Pareto-superior equilibrium is contestable for the low type, i.e. $\theta_L^R=\theta_{(1,1),1}^R$. In that case, payoffs are given by $\Pi_L^{\theta<\theta_L^R}=F_H\left(\theta_H^R\right)a+\left(1-F_H\left(\theta_H^R\right)\right)b$ and $\Pi_L^{\theta\geq\theta_L^R}=F_H\left(\theta_H^R\right)d+\left(1-F_H\left(\theta_H^R\right)\right)d=d$, for high types $\Pi_H^{\theta<\theta_H^R}=F_L\left(\theta_L^R\right)A+\left(1-F_L\left(\theta_L^R\right)\right)D$ and $\Pi_H^{\theta\geq\theta_H^R}=F_L\left(\theta_L^R\right)B+\left(1-F_H\left(\theta_H^R\right)\right)D$. Hence, differences are given by $\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}=d-b-F_H\left(\theta_H^R\right)(a-b)$ and $\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}=F_L\left(\theta_L^R\right)(B-A)\leq 0$. Note that $\left(\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}\right)\left(F_H\left(\theta_H^R\right)=0\right)=d-b>0$ and

$\left(\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}\right)\left(F_H\left(\theta_H^R\right)=1\right)=d-a>0$. Hence, the globally stable equilibrium is given by $F_H\left(\theta_H^R\right)=1, F_L\left(\theta_L^R\right)=0$. Furthermore, $\frac{\partial\left(\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}\right)}{\partial\left(1-F_L\left(\theta_L^R\right)\right)}=A-B>0$ and $\frac{\partial\left(\Pi_L^{\theta\geq\theta_L^R}-\Pi_L^{\theta<\theta_L^R}\right)}{\partial\left(1-F_H\left(\theta_H^R\right)\right)}=a-b$.

Finally, if none of the material equilibria is contestable for the low type, the distribution of inequality aversion among low types is determined by initial conditions and random shift. Payoff difference for high types is given by $\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}=F_L\left(\theta_L^R\right)(B-A)\leq 0$ with $\frac{\partial\left(\Pi_H^{\theta\geq\theta_H^R}-\Pi_H^{\theta<\theta_H^R}\right)}{\partial\left(1-F_L\left(\theta_L^R\right)\right)}=A-B>0$. Hence, the globally stable equilibrium is given by $F_H\left(\theta_H^R\right)=1$. QED

Proof of Theorem:

Let $\theta^D=\theta^C=\theta_{H,L}^R\equiv\theta^{crit}$ and $d\Pi_H^R\leq 0$. $d\Pi_H^R\leq 0$ implies that $\beta_H^R=-\alpha_H^R$.

Payoff differences are given by

$$\begin{aligned} d\Pi_H^S\geq 0 &\Leftrightarrow \left(1-F_L^t\right)\left(\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_H^R\right)\geq \mu\beta^D+\nu\beta^C-\nu\alpha^C-(1-\mu-\nu)\alpha_H^R-\left(1-F_H^t\right)\left(\mu\beta^D+\nu\beta^C\right)\\ d\Pi_L^S\geq 0 &\Leftrightarrow \left(1-F_L^t\right)\left(\mu\beta^D+\nu\beta^C\right)\geq \mu\beta^D+\nu\beta^C-\nu\alpha^C-(1-\mu-\nu)\alpha_L^R-\left(1-F_H^t\right)\left(\mu\beta^D+\nu\beta^C+(1-\mu-\nu)\beta_L^R\right) \end{aligned} \quad (*)$$

I will distinguish 3 cases:

(i) $0<\mu\beta^D+\nu\beta^C<\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_H^R$, (ii) $\mu\beta^D+\nu\beta^C<\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_H^R<0$ and (iii) $\mu\beta^D+\nu\beta^C<0<\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_H^R$.

(i) $0<\mu\beta^D+\nu\beta^C<\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_H^R$

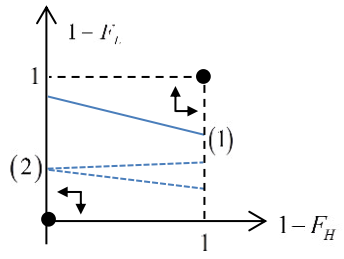
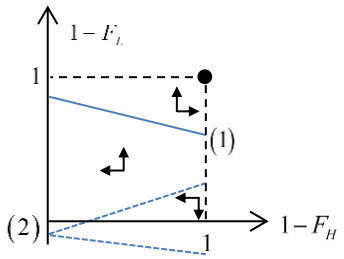
(*) can be written as:

$$(1):\left(1-F_L^t\right)\geq 1-\frac{\nu\alpha^C}{\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_H^R}-\frac{\mu\beta^D+\nu\beta^C}{\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_H^R}\left(1-F_H^t\right)$$

$$(2):\left(1-F_L^t\right)\geq 1-\frac{\nu\alpha^C+(1-\mu-\nu)\alpha_L^R}{\mu\beta^D+\nu\beta^C}-\frac{\mu\beta^D+\nu\beta^C+(1-\mu-\nu)\beta_L^R}{\mu\beta^D+\nu\beta^C}\left(1-F_H^t\right)$$

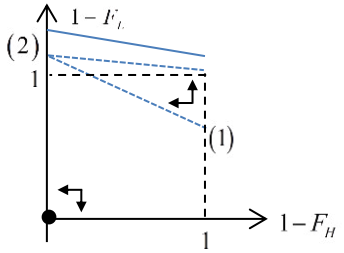
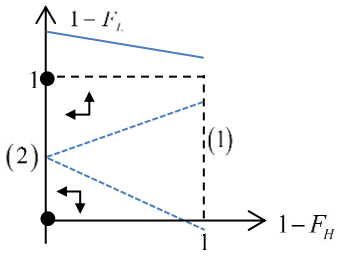
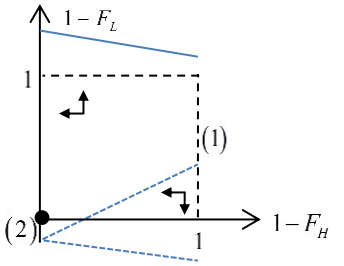
a) $\alpha^C>0$:

It follows that the intercept of (1) in below one and above the intercept of (2). Given the negative slope of (1) essentially 2 cases can be distinguished. The following table depicts the phase diagrams which clearly indicate the stable equilibria. The last row states the precise condition for the case considered.

$F_H = F_L = 0,\ F_H = F_L = 1$	$F_H = F_L = 0$
	
$1-\frac{\nu\alpha^C+(1-\mu-\nu)\alpha_L^R}{\mu\beta^D+\nu\beta^C}>0\Leftrightarrow$ $\alpha^C<\frac{\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_L^R}{\nu}$	$1-\frac{\nu\alpha^C+(1-\mu-\nu)\alpha_L^R}{\mu\beta^D+\nu\beta^C}\le 0\Leftrightarrow$ $\alpha^C\ge\frac{\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_L^R}{\nu}$

b) $\alpha^C \le 0$:

It follows that the intercept of (1) is above one and above the intercept of (2). Given the negative slope of (1), essentially 4 cases can be distinguished. The following table depicts the phase diagrams which clearly indicate the stable equilibria. The last row states the precise condition for the case considered.

$F_H = F_L = 1$	$F_H = 1,\ F_L = 0;\ F_H = F_L = 1$	$F_H = 1,\ F_L = 0$
		
$1-\frac{\nu\alpha^C+(1-\mu-\nu)\alpha_L^R}{\mu\beta^D+\nu\beta^C}>1\Leftrightarrow$ $\alpha^C<-\frac{1-\mu-\nu}{\nu}\alpha_L^R$	$1-\frac{\nu\alpha^C+(1-\mu-\nu)\alpha_L^R}{\mu\beta^D+\nu\beta^C}\in(0,1)$	$1-\frac{\nu\alpha^C+(1-\mu-\nu)\alpha_L^R}{\mu\beta^D+\nu\beta^C}<0\Leftrightarrow$ $\alpha^C>\frac{\mu\beta^D+\nu\beta^C-(1-\mu-\nu)\alpha_L^R}{\nu}$

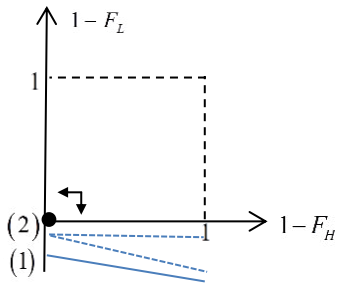
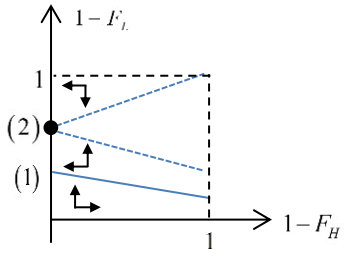
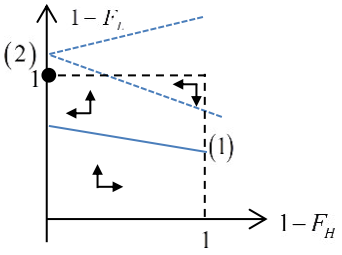
The same three cases emerge, if (1) has a value of below one at $1-F_H^t=1$. However, in that case an additional locally stable equilibrium arises, that of $F_H = F_L = 0$. The condition for this is $\mu\beta^D+\nu\left(\alpha^C+\beta^C\right)>0$.

(ii) $\mu\beta^{D}+\nu\beta^{C}<\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{H}^{R}<0$

Note that the slope of (1) is again negative.

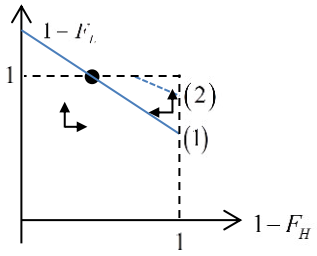
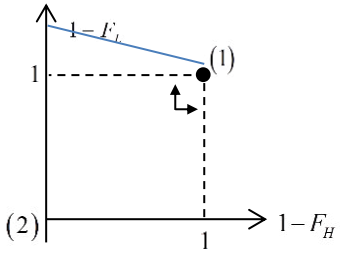
a) $\alpha^{C}<0$:

It follows that the intercept of (1) in below one and below the intercept of (2). Given the negative slope of (1), essentially 3 cases can be distinguished. The following table depicts the phase diagrams which clearly indicate the stable equilibria. The last row states the precise condition for the case considered.

$F_H = F_L = 1$	$F_H = 1,\ F_L = 1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}$	$F_H = 1,\ F_L = 0$
		
$1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}\le 0 \Leftrightarrow$ $\alpha^{C}\le\frac{\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{L}^{R}}{\nu}$	$1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}\in(0,1)$	$1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}\ge 1 \Leftrightarrow$ $\alpha^{C}\ge-\frac{(1-\mu-\nu)}{\nu}\alpha_{L}^{R}$

b) $\alpha^{C}\ge 0$:

In this case, the intercept of (1) is above one. The following two cases can be distinguished.

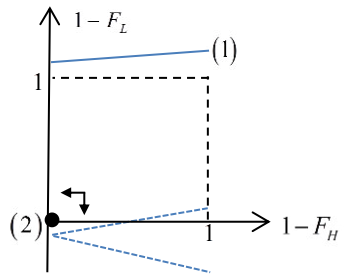
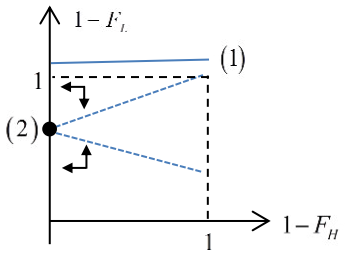
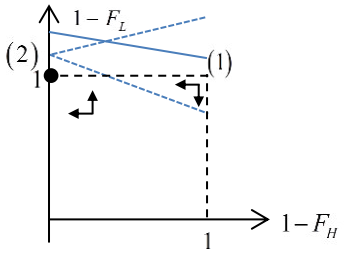
$F_L = 0,\ F_H = 1+\frac{\nu\alpha^{C}}{\mu\beta^{D}+\nu\beta^{C}}$	$F_H = F_L = 0$
	
$1-\frac{\nu\alpha^{C}}{\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{H}^{R}}-\frac{\mu\beta^{D}+\nu\beta^{C}}{\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{H}^{R}}<1$ $\Leftrightarrow \mu\beta^{D}+\nu\left(\alpha^{C}+\beta^{C}\right)<0$	$\mu\beta^{D}+\nu\left(\alpha^{C}+\beta^{C}\right)\ge 0$

(iii) $\mu\beta^{D}+\nu\beta^{C}<0<\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{H}^{R}$

Note that the intercept of (1) is above one.

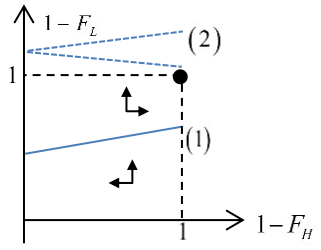
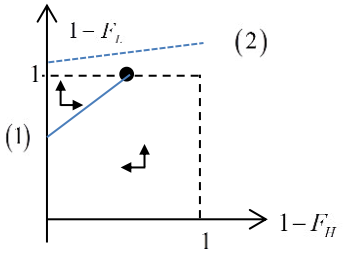
a) $\alpha^{C}<0$:

It follows that the slope of (1) is positive. Essentially 3 cases can be distinguished.

$F_H = F_L = 1$	$F_H = 1,\ F_L = 1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}$	$F_H = 1,\ F_L = 0$
		
$1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}\le 0 \Leftrightarrow$ $\alpha^{C}\le\frac{\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{L}^{R}}{\nu}$	$1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}\in(0,1)$	$1-\frac{\nu\alpha^{C}+(1-\mu-\nu)\alpha_{L}^{R}}{\mu\beta^{D}+\nu\beta^{C}}\ge 1 \Leftrightarrow$ $\alpha^{C}\ge-\frac{(1-\mu-\nu)}{\nu}\alpha_{L}^{R}$

b) $\alpha^{C}\ge 0$:

In this case, the intercept of (1) is below one whereas the intercept of (2) is above one. The following two cases can be distinguished.

$F_H = F_L = 0$	$F_L = 0,\ F_H = 1+\frac{\nu\alpha^{C}}{\mu\beta^{D}+\nu\beta^{C}}$
	
$1-\frac{\nu\alpha^{C}}{\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{H}^{R}}-\frac{\mu\beta^{D}+\nu\beta^{C}}{\mu\beta^{D}+\nu\beta^{C}-(1-\mu-\nu)\alpha_{H}^{R}}<1$ $\Leftrightarrow \mu\beta^{D}+\nu\left(\alpha^{C}+\beta^{C}\right)>0$	$\mu\beta^{D}+\nu\left(\alpha^{C}+\beta^{C}\right)\le 0$

QED

Proof of Proposition 4:

As a preparation for the proof I will first restate some results by Calvó-Armengol (2006). Thereafter, I will present a lemma which presents a necessary and sufficient condition for a player to earn a higher payoff then his opponent.

For $\gamma\left(A^1, A^2\right) \in \Gamma^\circ$ define $\alpha_1 = \left|a^1_{00} - a^1_{10}\right| \Big/ \left|a^1_{11} - a^1_{01}\right|$ and $\alpha_2 = \left|a^2_{00} - a^2_{01}\right| \Big/ \left|a^2_{11} - a^2_{10}\right|$. In the absence of neither weakly nor strictly dominated strategies α_1 and α_2 are well defined and strictly positive. The defined values give rise to three different types of games:

	0	1
0	α_1, α_2	$0,0$
1	$0,0$	$1,1$

$\gamma_{\mathrm{I}}\left(\alpha_1, \alpha_2\right)$: coordination

	0	1
0	$-\alpha_1, -\alpha_2$	$0,0$
1	$0,0$	$-1,-1$

$\gamma_{\mathrm{II}}\left(\alpha_1, \alpha_2\right)$: anti-coordination

	0	1
0	$-\alpha_1, \alpha_2$	$0,0$
1	$0,0$	$-1,1$

$\gamma_{\mathrm{III}}\left(\alpha_1, \alpha_2\right)$: competitive

Table 7: Classification of 2x2 games by Calvó-Armengol (2006)

Lemma 4 (Calvó-Armengol 2006, Lemma 1) Let $\gamma\left(A^1, A^2\right) \in \Gamma^\circ$. Then, for the set of correlated equilibria (CE) of $\gamma\left(A^1, A^2\right) \in \Gamma^\circ$ holds: $\mathrm{CE}\left(\gamma\left(A^1, A^2\right)\right) = \mathrm{CE}\left(\gamma_1\left(\alpha, \beta\right)\right)$, for some $1 \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$.

The restated result of Calvó-Armengol (2006) proves that the set Γ° of 2x2 games can be partitioned into three equivalence classes for the set of correlated equilibrium strategies. It is easily verified that $\mathrm{CE}\left(\gamma_{\mathrm{III}}\left(\alpha_1, \alpha_2\right)\right) = \mathrm{NE}\left(\gamma_{\mathrm{III}}\left(\alpha_1, \alpha_2\right)\right)$, i.e. the sets of correlated equilibria and Nash equilibria coincide and the set consist of a single point in Δ_3, the 3-dimensional simplex of $\mathbb{R}^4$.

Lemma 5 (Calvó-Armengol 2006, Lemma 2) $\mu \in \mathrm{CE}\left(\gamma_{\mathrm{I}}\left(\alpha_1, \alpha_2\right)\right)$ if and only if $\tau(\mu) \in \mathrm{CE}\left(\gamma_{\mathrm{II}}\left(\alpha_1, \frac{1}{\alpha_2}\right)\right)$, where $\tau(\mathrm{x}) = \left(\mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_1, \mathrm{x}_2\right)$ for $\left(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4\right) \in \mathbb{R}^4$.

Lemma 5 reveals that the class of coordination games and the class of anti-coordination games are isomorphic to one another. It thus suffices to characterize the set of correlated equilibria for one class. I will restate the result for the class of coordination games. A game $\gamma_{\mathrm{I}}\left(\alpha_1, \alpha_2\right)$ of that class has three Nash equilibria and two correlated equilibria, the probability measures of which are given in Table 8.

μ	μ_{00}	μ_{11}	μ_{10}	μ_{01}
$\mu_{\mathrm{C}}^{*}(\alpha,\beta)$	1	0	0	0
$\mu_{\mathrm{D}}^{*}(\alpha,\beta)$	0	1	0	0
$\mu_{\mathrm{E}}^{*}(\alpha,\beta)$	$\frac{1}{(1+\alpha_1)(1+\alpha_2)}$	$\frac{\alpha_1\alpha_2}{(1+\alpha_1)(1+\alpha_2)}$	$\frac{\alpha_2}{(1+\alpha_1)(1+\alpha_2)}$	$\frac{\alpha_1}{(1+\alpha_1)(1+\alpha_2)}$
$\mu_{\mathrm{F}}^{*}(\alpha,\beta)$	$\frac{1}{1+\alpha_2+\alpha_1\alpha_2}$	$\frac{\alpha_1\alpha_2}{1+\alpha_2+\alpha_1\alpha_2}$	$\frac{\alpha_2}{1+\alpha_2+\alpha_1\alpha_2}$	0
$\mu_{\mathrm{G}}^{*}(\alpha,\beta)$	$\frac{1}{1+\alpha_1+\alpha_1\alpha_2}$	$\frac{\alpha_1\alpha_2}{1+\alpha_1+\alpha_1\alpha_2}$	0	$\frac{\alpha_1}{1+\alpha_1+\alpha_1\alpha_2}$

Table 8: Probability measures for correlated equilibria and Nash equilibria for a game $\gamma_{\mathrm{I}}(\alpha_1,\alpha_2)$.

Proposition 8 relates the 5 vertices given in Table 8 and the set of correlated equilibria.

Proposition 8 (Calvó-Armengol 2006, Proposition 1) $\mathrm{CE}\left(\gamma_{\mathrm{I}}(\alpha_1,\alpha_2)\right)$ is a polytope of Δ_3 with five vertices given in Table 8.

Due to the linearity of the inner product, the calculation of the expected payoff amounts to the determination of the centre of mass of P, the polytope of Proposition 8. Equation (9) states this property formally, where $\pi=(\mathrm{a}_{00},\mathrm{a}_{11},\mathrm{a}_{10},\mathrm{a}_{01})$ denotes the payoffs associated with the payoff matrix A of the game.

$$\mathrm{E}\pi=\int_{\mu\in\mathrm{P}}\frac{1}{\mathrm{Vol}(\mathrm{P})}\langle\pi,\mu\rangle\,\mathrm{d}\mu=\left\langle\pi,\frac{1}{\mathrm{Vol}(\mathrm{P})}\int_{\mu\in\mathrm{P}}\mu\,\mathrm{d}\mu\right\rangle=\left\langle\pi,\mu^{\mathrm{CM}}\right\rangle \tag{9}$$

The centre of mass of the polytope given by the vertices presented in Table 8 can be calculated as the average with relative volume as weights of the centres of mass of the two pyramids DCGE and DCFE, i.e. $\mathrm{CM}_{\mathrm{P}_{\mathrm{CDEFG}}}=\frac{\mathrm{V}_{\mathrm{P}_{\mathrm{DCGE}}}}{\mathrm{V}_{\mathrm{P}_{\mathrm{DCGE}}}+\mathrm{V}_{\mathrm{P}_{\mathrm{DCFE}}}}\mathrm{CM}_{\mathrm{P}_{\mathrm{DCGE}}}+\frac{\mathrm{V}_{\mathrm{P}_{\mathrm{DCFE}}}}{\mathrm{V}_{\mathrm{P}_{\mathrm{DCGE}}}+\mathrm{V}_{\mathrm{P}_{\mathrm{DCFE}}}}\mathrm{CM}_{\mathrm{P}_{\mathrm{DCFE}}}$.

The centre of mass for these two pyramids is located on the line segment connecting the centre of the (any) triangular base and the top of the pyramids. Some elementary algebra yields $\mathrm{V}_{\mathrm{P}_{\mathrm{DCGE}}}=\frac{\alpha_1}{6}\frac{1}{1+\alpha_1+\alpha_1\alpha_2}\frac{\alpha_2}{(1+\alpha_1)(1+\alpha_2)}$, $\mathrm{V}_{\mathrm{P}_{\mathrm{DCFE}}}=\frac{\alpha_1}{6}\frac{1}{1+\alpha_2+\alpha_1\alpha_2}\frac{\alpha_2}{(1+\alpha_1)(1+\alpha_2)}$.
Furthermore, the centres of mass of the two pyramids are given by:

$$\mathrm{CM}_{\mathrm{P_{DCGE}}} = \begin{pmatrix} \mu_{00}^{\mathrm{CM_{P_{DCGE}}}} \\ \mu_{10}^{\mathrm{CM_{P_{DCGE}}}} \\ \mu_{01}^{\mathrm{CM_{P_{DCGE}}}} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1+\dfrac{1}{1+\alpha_1+\alpha_1\alpha_2}+\dfrac{1}{(1+\alpha_1)(1+\alpha_2)} \\ \alpha_2\dfrac{1}{(1+\alpha_1)(1+\alpha_2)} \\ \alpha_1\left(\dfrac{1}{1+\alpha_1+\alpha_1\alpha_2}+\dfrac{1}{(1+\alpha_1)(1+\alpha_2)}\right) \end{pmatrix} \quad \text{and}$$

$$\mathrm{CM}_{\mathrm{P_{DCFE}}} = \begin{pmatrix} \mu_{00}^{\mathrm{CM_{P_{DCFE}}}} \\ \mu_{10}^{\mathrm{CM_{P_{DCFE}}}} \\ \mu_{01}^{\mathrm{CM_{P_{DCFE}}}} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1+\dfrac{1}{1+\alpha_2+\alpha_1\alpha_2}+\dfrac{1}{(1+\alpha_1)(1+\alpha_2)} \\ \alpha_2\left(\dfrac{1}{1+\alpha_2+\alpha_1\alpha_2}+\dfrac{1}{(1+\alpha_1)(1+\alpha_2)}\right) \\ \alpha_1\dfrac{1}{(1+\alpha_1)(1+\alpha_2)} \end{pmatrix}$$

Plugging in values and rearranging terms yields:

$$\begin{pmatrix} \mu_{00}^{\mathrm{CM}} \\ \mu_{10}^{\mathrm{CM}} \\ \mu_{01}^{\mathrm{CM}} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1+\dfrac{1}{1+\alpha_1+\alpha_1\alpha_2}+\dfrac{1}{1+\alpha_2+\alpha_1\alpha_2}+\dfrac{1}{(1+\alpha_1)(1+\alpha_2)}-\dfrac{2}{2+\alpha_1+\alpha_2+2\alpha_1\alpha_2} \\ \alpha_2\left(\dfrac{1}{1+\alpha_2+\alpha_1\alpha_2}+\dfrac{1}{(1+\alpha_1)(1+\alpha_2)}-\dfrac{1}{2+\alpha_1+\alpha_2+2\alpha_1\alpha_2}\right) \\ \alpha_1\left(\dfrac{1}{1+\alpha_1+\alpha_1\alpha_2}+\dfrac{1}{(1+\alpha_1)(1+\alpha_2)}-\dfrac{1}{2+\alpha_1+\alpha_2+2\alpha_1\alpha_2}\right) \end{pmatrix}. \qquad (10)$$

Again, let the symmetric dilemma be represented by the following matrix $\mathrm{A}=\begin{pmatrix} \mathrm{a} & \mathrm{c} \\ \mathrm{b} & \mathrm{d} \end{pmatrix}$. Expected payoffs are then given by: $\mathrm{E}\pi_1 = \mu_{00}\mathrm{a}+\mu_{01}\mathrm{c}+\mu_{10}\mathrm{b}+\mu_{11}\mathrm{d}$, $\mathrm{E}\pi_2 = \mu_{00}\mathrm{a}+\mu_{01}\mathrm{b}+\mu_{10}\mathrm{a}+\mu_{11}\mathrm{d}$ and hence the difference by: $\mathrm{E}\pi_1 - \mathrm{E}\pi_2 = (\mu_{10}-\mu_{01})(\mathrm{b}-\mathrm{c})$. Plugging in the values for the centre of mass given by (10) yields:

$$\mathrm{E}\pi_1 - \mathrm{E}\pi_2 = (\alpha_2-\alpha_1)(\mathrm{b}-\mathrm{c})$$
$$\underbrace{\left(\frac{1+\alpha_1\alpha_2}{(1+\alpha_1+\alpha_1\alpha_2)(1+\alpha_2+\alpha_1\alpha_2)}+\frac{1}{(1+\alpha_1)(1+\alpha_2)}-\frac{1}{(1+\alpha_1+\alpha_1\alpha_2)+(1+\alpha_2+\alpha_1\alpha_2)}\right)}_{>0}$$

Thus $\mathrm{E}\pi_1 - \mathrm{E}\pi_2 > 0 \Leftrightarrow (\mu_{10}^{\mathrm{CM}} - \mu_{01}^{\mathrm{CM}})(\mathrm{a}_{10}-\mathrm{a}_{01}) > 0 \Leftrightarrow (\alpha_2-\alpha_1)(\mathrm{a}_{10}-\mathrm{a}_{01}) > 0$. (11)

Given this preparation I can now proof the claim of Proposition 4:

The line of argument in the proof of Proposition 1 is still valid, i.e. if and only if two individuals are matched who are sufficiently inequality-averse the set of equilibria changes. In case of the concept of correlated equilibria, the vertices of the set are given in Table 8. According to (11) for

two such individuals the one with the lower degree of inequality-aversion earns higher profits. Furthermore, the difference in profits is monotonic decreasing in the difference in the degrees of inequality-aversion. Hence, the highest profit is earned by individuals with $\theta = \theta^{D}$ and the lowest profit such an individual can earn is realized when matched with another individual with $\theta = \theta^{D}$. Again, let the symmetric dilemma be represented by the following matrix $A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$. W.l.o.g. $a > d$, in that case $\theta_1 = \theta_2 = \theta^{D}$ implies $\alpha_1 = \alpha_2 = 0$ and $\mu_{00}^{CM} = \frac{3}{4}, \mu_{11}^{CM} = \frac{1}{4}, \mu_{01}^{CM} = \mu_{10}^{CM} = 0$ yielding expected payoff $E\pi = \frac{3}{4}a + \frac{1}{4}d > d$ strictly greater than the payoff received by opportunistic individuals. Hence, the only stable equilibrium that can emerge is the singular distribution with all agents sharing the same degree of inequality-aversion.

I turn now to the non-PD-case. In that case, the results with respect to profits for individuals with $\theta \geq \theta^{D}$ also hold. No two different values θ_1, θ_2 with $\theta_1, \theta_2 \geq \theta^{D}$ can be part of an equilibrium, because both individual earn the same profit when matched with an opportunistic opponent and the one with the lower degree of inequality-aversion earns higher profits than the one with the higher value in any match with some other agent with $\theta \geq \theta^{D}$. Hence, only types with $\theta = \theta^{D}$ could be part of an equilibrium. However, the same calculation of expected payoffs as in the PD-case applies, but in the non-PD-case this amount to a disadvantage because w.l.o.g. $b > a,\ d > c,\ \frac{b+c}{2} > d,\ a \leq d$ and thereby $E\pi = \frac{3}{4}a + \frac{1}{4}d \leq d$. Hence, the globally stable equilibrium distribution is characterized by $F\left(\theta^{D}\right) = 1$. QED

Proof of Proposition 5:

Given the definition of thresholds and the derivation of different equilibria in the proof of Proposition 2, I focus herein on the case where one player alone can destabilize both pure Nash equilibria. By symmetry, potentially both players can thus destabilize all equilibria individually. Again, since inequality aversion has no leverage on coordination games, I study anti-coordination games. In other words, I am concerned with games represented by a matrix $A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ such that $b < a,\ c > d$. Both equilibria being contestable is equivalent to $\theta_{(0,1),2}^{C}, \theta_{(1,0),1}^{C} < 1,\ \theta_{(0,1),1}^{C}, \theta_{(1,0),2}^{C} < 1$ and $\theta_{(0,1),2}^{C}, \theta_{(1,0),1}^{C} < 1,\ \theta_{(0,1),1}^{C}, \theta_{(1,0),2}^{C} < 1 \Leftrightarrow \begin{cases} c < \left\{a, \frac{b+d}{2}\right\},\ b > c \Rightarrow b > a > c > d \\ b < \left\{d, \frac{a+c}{2}\right\},\ b < c \Rightarrow c > d > b > a \end{cases}$.

I first study the case $\theta_{(0,1),2}^{C} = \theta_{(1,0),1}^{C} < \theta_{(0,1),1}^{C} = \theta_{(1,0),2}^{C} \left(\Leftrightarrow \alpha < \beta \Leftrightarrow \frac{b-a}{c-d} < 1\right)$.

Table 9 below presents the payoffs depending on the two level of inequality aversion being matched. I will refer to an individual in lowest interval, medium and high interval as A, B and C-types respectively.

	A	B	C
	$\theta_2 < \theta^C_{(0,1),2} < \theta^C_{(1,0),2}$	$\theta^C_{(0,1),2} < \theta_2 < \theta^C_{(1,0),2}$ '0' is dominant str.	$\theta^C_{(0,1),2} < \theta^C_{(1,0),2} < \theta_2$
$\theta_1 < \theta^C_{(1,0),1} < \theta^C_{(0,1),1}$	$\left(\frac{b+c}{2}, \frac{b+c}{2}\right)$	(b,c)	$\left(\Pi_1^{mix}(\alpha_1,\alpha_2), \Pi_2^{mix}(\alpha_1,\alpha_2)\right)$
$\theta^C_{(1,0),1} < \theta_1 < \theta^C_{(0,1),1}$	(c,b)	(a,a)	(a,a)
$\theta^C_{(1,0),1} < \theta^C_{(0,1),1} < \theta_1$	$\left(\Pi_1^{mix}(\alpha_1,\alpha_2), \Pi_2^{mix}(\alpha_1,\alpha_2)\right)$	(a,a)	$\left(\frac{a+d}{2}, \frac{a+d}{2}\right)$

Table 9: Payoffs in the various matches.

For the mixed equilibrium: $\alpha_1 = \frac{|a-b+\theta_1|b-c||}{|c-\theta_1|b-c|+d|}$, $\alpha_2 = \frac{|a-b+\theta_2|b-c||}{|c-\theta_2|b-c|+d|}$, and

$\alpha < \beta \Leftrightarrow \frac{b-a}{c-d} < 1 \Leftrightarrow \theta^C_{(0,1),2} = \theta^C_{(1,0),1} < \theta^C_{(0,1),1} = \theta^C_{(1,0),2}$.

I will first consider the case $b > c$. Note that in that case B-types destabilize the equilibrium that favors them, but not the equilibrium that disfavors them. This suggests an evolutionary disadvantage for B-types. If there exists only a mixed Nash equilibrium, i.e. in a match between A-types and C-types, profits are given by:

$$\Pi_1^{mix}(\alpha_1,\alpha_2) = \frac{1}{(1+\alpha_1)(1+\alpha_2)}(a+\alpha_1 b+\alpha_2 c+\alpha_1\alpha_2 d), \quad \Pi_2^{mix}(\alpha_1,\alpha_2) = \frac{1}{(1+\alpha_1)(1+\alpha_2)}(a+\alpha_1 c+\alpha_2 b+\alpha_1\alpha_2 d)$$

$$\Pi_1^{mix}(\alpha_1,\alpha_2) - \Pi_2^{mix}(\alpha_1,\alpha_2) = \frac{(\alpha_2-\alpha_1)}{(1+\alpha_1)(1+\alpha_2)}(b-c) > 0 \Leftrightarrow \alpha_1 < \alpha_2$$

Consider a match between type A as player one and type C as player two, i.e. player one is opportunistic and player two is highly inequality-averse. In that case,

$$\alpha_1 < \alpha_2 \Leftrightarrow \frac{|a-b+\theta_1|b-c||}{|c-\theta_1|b-c|+d|} < \frac{|a-b+\theta_2|b-c||}{|c-\theta_2|b-c|+d|} \Leftrightarrow \frac{a-b+\theta_1|b-c|}{d-c+\theta_1|b-c|} < \frac{a-b+\theta_2|b-c|}{d-c+\theta_2|b-c|}.$$
$$\Leftrightarrow (a-b)\theta_2|b-c| + (d-c)\theta_1|b-c| > (a-b)\theta_1|b-c| + (d-c)\theta_2|b-c| \underset{b-a<c-d}{\Leftrightarrow} \theta_1 < \theta_2$$

When I considered a strict and symmetric problem of coordination type C player were simply left out of analysis. Thus, I will focus on equilibria with type C players. Note that there can be no B, C equilibrium, because C players would be worse off. For the same reason there cannot be an equilibrium with only C players, since B players could successfully invade.

In an equilibrium with both types A and C present, only players with minimal $\alpha_1 = \alpha_1\left(\theta_1 = \frac{b-a}{b-c}\right) = 0$ among A types and those with minimal $\alpha_2 = \alpha_2(\theta_2 = 1) = \frac{a-c}{d+b-2c}$ can be part of the equilibrium, because $\frac{\partial \Pi_1^{mix}(\alpha_1,\alpha_2)}{\partial \alpha_1} = -\frac{a-c+(b-d)\alpha_2}{(1+\alpha_1)^2(1+\alpha_2)} < 0$ and

$\frac{\partial \Pi_2^{mix}(\alpha_1,\alpha_2)}{\partial \alpha_2} = -\frac{a-c+(b-d)\alpha_1}{(1+\alpha_1)(1+\alpha_2)^2} < 0$. Due to $\frac{\partial \alpha_1}{\partial \theta_1} = \frac{b-c}{(c-d-\theta_1(b-c))^2}(b-a-(c-d))\underset{b-a<c-d}{<} 0$ and

$\frac{\partial \alpha_2}{\partial \theta_2} = \frac{b-c}{\left(c-d-\theta_2(b-c)\right)^2}\left(b-a-(c-d)\right) \underset{b-a<c-d}{<} 0$, minimal α translates into maximal θ. Thus, Table 9 simplifies to:

	A	**B**	**C**
A	$\left(\frac{b+c}{2}, \frac{b+c}{2}\right)$	(b,c)	$\left(\frac{-bc+a(2b-2c+d)}{a+b-3c+d}, \frac{-c^2+a(b-c+d)}{a+b-3c+d}\right)$
B	(c,b)	(a,a)	(a,a)
C	$\left(\frac{-c^2+a(b-c+d)}{a+b-3c+d}, \frac{-bc+a(2b-2c+d)}{a+b-3c+d}\right)$	(a,a)	$\left(\frac{a+d}{2}, \frac{a+d}{2}\right)$

It turns out that type A types earn strictly higher payoffs than type C players, because

$\Pi_2^{mix}\left(\alpha_1 = 0, \alpha_2 = \frac{a-c}{d+b-2c}\right) = \frac{-c^2+a(b-c+d)}{a+b-3c+d} < \frac{b+c}{2}$ and

$\Pi_1^{mix}\left(\alpha_1 = 0, \alpha_2 = \frac{a-c}{d+b-2c}\right) = \frac{-bc+a(2b-2c+d)}{a+b-3c+d} > \frac{a+d}{2}$. Hence, such an equilibrium cannot exist.

Intuitively, if $b > c$, then weighting the outcome (0,1) and (1,1) less reduced payoffs for player two. For the lowest weight payoffs for player two are a weighted average of a and c, and therefore higher than c.

Finally, I analyze whether there exists a A,B,C equilibrium. It turns out that for the most profitable type A player an even stronger inequality holds: $\Pi_1^{mix}\left(\alpha_1 = 0, \alpha_2 = \frac{a-c}{d+b-2c}\right) > a$. Hence A-types would earn strictly higher profits than B-types in an A,B,C equilibrium.

Thus no additional equilibria arise.

a) $b < c$.

To summarize conditions: $c > d > b > a$, $b-a < c-d$, $b < \frac{a+c}{2}$.

These conditions imply: $\frac{\partial \Pi_2^{mix}(\alpha_1,\alpha_2)}{\partial \alpha_2} = \frac{c-a+(d-b)\alpha_1}{(1+\alpha_1)(1+\alpha_2)^2} > 0$ and $\frac{\partial \Pi_1^{mix}(\alpha_1,\alpha_2)}{\partial \alpha_1} = \frac{c-a+(d-b)\alpha_2}{(1+\alpha_1)^2(1+\alpha_2)} > 0$.

I focus again on equilibria with C types being present. I first consider the case with only C-types present in equilibrium. Such a distribution cannot be invaded by B types. The fittest A type is the one with maximal α_1, which transfers to a minimal θ_1. Note that the profit of the fittest A type is independent of the degree of inequality aversion of the C type, $\Pi_1^{mix}\left(\alpha_1 = \frac{b-a}{c-d}, \alpha_2\right) = \frac{bc-ad}{b-a+c-d}$.

Hence, a locally stable equilibrium with only inequality averse players of type C exists if $\frac{bc-ad}{b-a+c-d} < \frac{a+d}{2}$.

I now study whether there is an equilibrium with A and C types present. Again, this demands $\alpha_1 = \frac{b-a}{c-d}, \alpha_2 = \infty$ implying the following profits: $\Pi_1^{mix}\left(\alpha_1 = \frac{b-a}{c-d}, \alpha_2\right) = \frac{bc-ad}{b-a+c-d}$ and $\Pi_2^{mix}\left(\alpha_1 = \frac{b-a}{c-d}, \alpha_2 \to \infty\right) = \frac{-c^2+(a-b)d+cd}{a-b-c+d}$. Thus, Table 9 simplifies to:

	A	**B**	**C**
A	$\left(\frac{b+c}{2}, \frac{b+c}{2}\right)$	(b,c)	$\left(\frac{bc-ad}{b-a+c-d}, \frac{-c^2+(a-b)d+cd}{a-b-c+d}\right)$
B	(c,b)	(a,a)	(a,a)
C	$\left(\frac{-c^2+(a-b)d+cd}{a-b-c+d}, \frac{bc-ad}{b-a+c-d}\right)$	(a,a)	$\left(\frac{a+d}{2}, \frac{a+d}{2}\right)$

Let Π^A and Π^C denote the payoffs of A-types and C-types respectively. Let $F(A)$ denote the share of A-types in equilibrium, then $\Pi^A = F(A)\frac{b+c}{2} + (1-F(A))\frac{bc-ad}{b-a+c-d}$ and $\Pi^C = F(A)\frac{-c^2+(a-b)d+cd}{a-b-c+d} + (1-F(A))\frac{a+d}{2}$.

An A,C equilibrium exists if and only if $\frac{bc-ad}{b-a+c-d} > \frac{a+d}{2}$, because $\frac{b+c}{2} < \frac{-c^2+(a-b)d+cd}{a-b-c+d}$ holds. In that case the equilibrium share of A-types is given by $F(A) = \frac{a^2+2bc-a(b+c)-(b+c)d+d^2}{a^2-b^2+(c-d)^2-2ad+2bd}$. The equilibrium is locally stable if the profits of B are smaller than equilibrium payoffs, given the equilibrium share of A and C-types. Note that a parameterization with $a=0,\ b=\frac{1}{5},\ d=\frac{1}{3},\ c=1$ indeed satisfies all condition, i.e. $c>d>b>a,\ b-a<c-d,\ b<\frac{a+c}{2},\ \frac{bc-ad}{b-a+c-d} > \frac{a+d}{2},\ F(A)\in(0,1)$, and $\Pi^B < \Pi^A$, thus such a stable equilibrium indeed exists.

I will finally analyze the existence of an A,B,C equilibrium. Payoffs of the different types are given by: $\Pi^A = F(A)\frac{b+c}{2} + F(B)b + (1-F(A)-F(B))\frac{bc-ad}{b-a+c-d}$,

$\Pi^B = F(A)c + F(B)a + (1-F(A)-F(B))a$, and

$\Pi^C = F(A)\frac{-c^2+(a-b)d+cd}{a-b-c+d} + F(B)a + (1-F(A)-F(A))\frac{a+d}{2}$. The two equations $\Pi^A = \Pi^B$ and $\Pi^B = \Pi^C$ imply the following equilibrium values for $F(A)$ and $F(B)$:

$$F(A)=\frac{2(a-b)(a-d)(a-b-c+d)^2}{\left(\begin{array}{l}2a^4+4b^3c-(b-c)(b+c)(3b+c)d+2(b^2-2bc-c^2)d^2+(b+c)d^3+a^3(-5(b+c)+2d)\\+a^2(4(b^2+3bc+c^2)-5(b+c)d+2d^2)-a(b^3+b^2(11c-6d)+(c-d)(c^2+3cd-2d^2)+b(3c^2-4cd+3d^2))\end{array}\right)}$$

$$F(B)=1-F(A)\frac{(a^2-a(b+3c-2d)+b(2c-d)+(c-d)d)}{(a-d)(a-b-c+d)}.$$

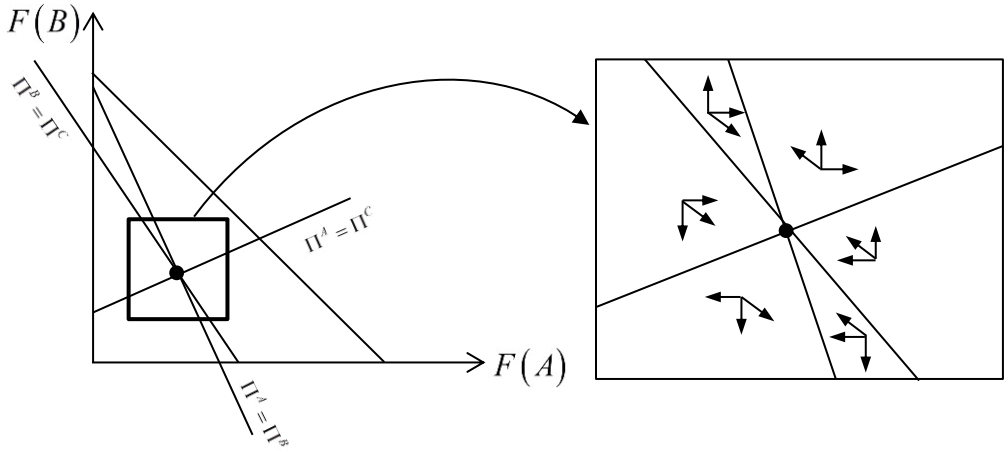
Given the summarizing conditions of this case $c>d>b>a,\ b-a<c-d,\ b<\frac{a+c}{2}$, it turns out that for the slopes of the three equation the following ordering holds:

$\frac{\partial F(B)}{\partial F(A)}^{\Pi^A=\Pi^B}<\frac{\partial F(B)}{\partial F(A)}^{\Pi^B=\Pi^C}<-1<0<\frac{\partial F(B)}{\partial F(A)}^{\Pi^A=\Pi^C}$, where

$$\frac{\partial F(B)}{\partial F(A)}^{\Pi^A=\Pi^B}=\frac{-2a^2+b^2+a(b+3c)+c(-c+d)-b(2c+d)}{2(a-b)(b-d)},$$

$$\frac{\partial F(B)}{\partial F(A)}^{\Pi^B=\Pi^C}=\frac{-a^2+a(b+3c-2d)+b(-2c+d)+d(-c+d)}{(a-d)(a-b-c+d)}, \text{ and}$$

$$\frac{\partial F(B)}{\partial F(A)}^{\Pi^A=\Pi^C}=\frac{a^2-b^2+(c-d)^2-2ad+2bd}{a^2+2b^2-a(3b+c-2d)-bd+(c-d)d}.$$ This gives rise to the following phase diagram.



As the diagram clearly indicates this equilibrium is unstable.

Finally, if $\theta^C_{(0,1),2}=\theta^C_{(1,0),1}>\theta^C_{(0,1),1}=\theta^C_{(1,0),2}\left(\Leftrightarrow\frac{b-a}{c-d}>1\right)$, the role of *b* and *c* and the role of *a* and *d* are simply reversed. QED

Proof of Proposition 6:

The set of equilibrium payoffs can be found in the proof of Proposition 3. If the two Nash equilibria are not Pareto-ranked then I may w.l.o.g. assume that $a < d < D < A$ (see Table 5). Two cases with respect to the thresholds for low types may be distinguished. I first consider $\theta^{R}_{(0,0),1} < \theta^{R}_{(1,1),1}$

1. $0 < \theta^{R}_{(0,0),1} < \theta^{R}_{(1,1),1} < 1$

Table 10 depicts equilibrium payoffs in the various matches.

	$\theta < \theta^{R}_{(0,0),2}$	$\theta > \theta^{R}_{(0,0),2}$
$\theta < \theta^{R}_{(0,0),1}$	$\left(\frac{a+d}{2}, \frac{A+D}{2}\right)$	(d, D)
$\theta^{R}_{(0,0),1} < \theta < \theta^{R}_{(1,1),1}$	(d, D)	(d, D)
$\theta^{R}_{(1,1),1} < \theta$	$\left(\Pi_1^{mix}(\alpha_1, \alpha_2), \Pi_2^{mix}(\alpha_1, \alpha_2)\right)$	(b, B)

Table 10: Equilibrium payoffs: $0 < \theta^{R}_{(0,0),1} < \theta^{R}_{(1,1),1} < 1$.

I will refer to individuals with $\theta < \theta^{R}_{(0,0),1}$, $\theta^{R}_{(0,0),1} < \theta < \theta^{R}_{(1,1),1}$, and $\theta^{R}_{(1,1),1} < \theta$ as A types, B types and C types respectively.

There can be no equilibrium with B types only as the more opportunistic A type would earn strictly higher profits as long as some high types are opportunistic. In an equilibrium with A and B types opportunistic high types would earn strictly higher payoffs. I will now consider the case of C types, who give rise to the play of a mixed equilibrium when matched with an opportunistic high type. I will show that $\Pi_2^{mix}(\alpha_1, \alpha_2) > B$, thus in such an equilibrium only opportunistic high types can be present.

Note that $\Pi_2^{mix}(\alpha_1, \alpha_2) > B \Leftrightarrow A + \alpha_1 + \alpha_2 C + \alpha_1 \alpha_2 D > (1+\alpha_1)(1+\alpha_2) B \Leftrightarrow A - B + \alpha_1 \alpha_2 (D - B) > \alpha_2 (B - C)$

Consider first $b \geq B$, then $D > B$ and hence $A - B + \alpha_1 \alpha_2 (D - B) > \alpha_2 (B - C) \Leftarrow A - B > \alpha_2 (B - C)$.

Note that $\frac{\partial \alpha_2}{\partial \theta_2} = \frac{A(c-d) + B(-c + 2C + d - 2D) - (a+b)(C-D)}{(D + C(-1+\theta 2) - (c - d + D)\theta_2)^2}$ if $C > c$. This derivative is negative if and only if the numerator is negative which can be written as $-(A-B)(d-c) + (D-C)(a+b-2B)$. This term is negative because $\theta^{R}_{(0,0),2} < 1 \Leftrightarrow a + b < 2B$. If $C < c$, then $\frac{\partial \alpha_2}{\partial \theta_2} = \frac{-A(c - 2C + d) + B(c + d - 2D) - (a+b)(C-D)}{(D(-1+\theta_2) - (c+d)\theta_2 + C(1+\theta_2))^2}$. This derivative is negative if and only if the numerator is negative which can be written as $-(A-B)(d+c) + 2AD - 2BD + (D-C)(a+b)$. Note that:

$$-(A-B)(d+c) + 2AD - 2BD + \underbrace{(D-C)}_{>0}(a+b) < 0 \underset{\theta^{R}_{(0,0),2} < 1 \Leftrightarrow a+b<2B}{\Leftarrow}$$

$$-(A-B)(d+c) + 2AD - 2BD + 2B(D-C) < 0 \Leftrightarrow \underbrace{(A-B)}_{>0}(2C - (d+c)) < 0.$$

This term is negative, because $C < c < d$ implies $\left(2C-\left(d+c\right)\right)<0$. Thus, $\frac{\partial\alpha_2}{\partial\theta_2}<0, b \geq B$.

Consider second $B > b$, then $D > B$ because $\theta^R_{(1,1),1}<1 \Leftrightarrow 2\underbrace{\left(d-b\right)}_{>0}<D-B$. Hence, still $\Pi_2^{mix}\left(\alpha_1,\alpha_2\right)>B \Leftarrow A-B>\alpha_2\left(B-C\right)$ holds. I show that also in this case $\frac{\partial\alpha_2}{\partial\theta_2}<0$.

$\theta^R_{(0,0),2}<1 \Leftrightarrow a+b<2B$. If $C < c$, then $\frac{\partial\alpha_2}{\partial\theta_2}=\frac{-A\left(c-2C+d\right)+B\left(c-2C+d\right)-\left(a-b\right)\left(C-D\right)}{\left(D\left(-1+\theta_2\right)-\left(c+d\right)\theta_2+C\left(1+\theta_2\right)\right)^2}$. This derivative is negative if and only if the numerator is negative which can be written as $-\underbrace{\left(A-B\right)}_{>0}\underbrace{\left(c+d-2C\right)}_{>0,\ C<c<d}+\left(a-b\right)\underbrace{\left(D-C\right)}_{>0}$. Note that: $\theta^R_{(0,0),2}<1 \Leftrightarrow a-b<0$, hence $\frac{\partial\alpha_2}{\partial\theta_2}<0$.

If $C > c$, then $\frac{\partial\alpha_2}{\partial\theta_2}=\frac{\left(A-B\right)\left(c-d\right)-\left(a-b\right)\left(C-D\right)}{\left(D+C\left(-1+\theta_2\right)-\left(c-d+D\right)\theta_2\right)^2}$. This derivative is negative if and only if the numerator is negative which can be written as $-\underbrace{\left(A-B\right)\left(d-c\right)}_{>0}+\left(a-b\right)\underbrace{\left(D-C\right)}_{>0}$. Note that: $\theta^R_{(0,0),2}<1 \Leftrightarrow a-b<0$, hence $\frac{\partial\alpha_2}{\partial\theta_2}<0$.

In summary, if $0<\theta^R_{(0,0),1},\theta^R_{(1,1),1}<1$ and one equilibrium is contestable for the high type, i.e. $0<\theta^R_{(0,0),2}<1$, then $\frac{\partial\alpha_2}{\partial\theta_2}<0$. Note that I did not make use of $\theta^R_{(0,0),1}<\theta^R_{(1,1),1}$. Hence, the result also applies for the second case $0<\theta^R_{(1,1),1}<\theta^R_{(0,0),1}<1$ which will be considered next. Hence,

$$\Pi_2^{mix}\left(\alpha_1,\alpha_2\right)>B \Leftarrow A-B>\alpha_2\left(B-C\right) \Leftarrow A-B>\alpha_2^{max}\left(B-C\right) \overset{\frac{\partial\alpha_2}{\partial\theta_2}<0}{=} \alpha_2\left(\theta_2=0\right)\left(B-C\right)=\frac{A-B}{D-C}\left(B-C\right)$$
$$\Leftrightarrow D>B$$

Since the last inequality holds, the claim $\Pi_2^{mix}\left(\alpha_1,\alpha_2\right)>B$ is established.

2. $0<\theta^R_{(1,1),1}<\theta^R_{(0,0),1}<1$

In that case Table 10 becomes:

	$\theta<\theta^R_{(0,0),2}$	$\theta>\theta^R_{(0,0),2}$
$\theta<\theta^R_{(0,0),1}$	$\left(\frac{a+d}{2},\frac{A+D}{2}\right)$	$\left(d,D\right)$
$\theta^R_{(0,0),1}<\theta<\theta^R_{(1,1),1}$	$\left(a,A\right)$	$\left(b,B\right)$
$\theta^R_{(1,1),1}<\theta$	$\left(\Pi_1^{mix}\left(\alpha_1,\alpha_2\right),\Pi_2^{mix}\left(\alpha_1,\alpha_2\right)\right)$	$\left(b,B\right)$

Table 11: Equilibrium payoffs: $0<\theta^R_{(1,1),1}<\theta^R_{(0,0),1}<1$.

Since $\Pi_2^{mix}\left(\alpha_1,\alpha_2\right)>B$ also holds and since $A>B$ dominance of relative opportunistic players among high types is even strict. Thus, also in this case no inequality-averse individuals can be part of a stable equilibrium. QED