von Widekind, Sven

Working Paper

Evolution of non-expected utility preferences

Working Papers, No. 370

Provided in Cooperation with:
Center for Mathematical Economics (IMW), Bielefeld University

Suggested Citation: von Widekind, Sven (2005) : Evolution of non-expected utility preferences, Working Papers, No. 370, Bielefeld University, Institute of Mathematical Economics (IMW), Bielefeld, https://nbn-resolving.de/urn:nbn:de:hbz:361-7583

This Version is available at:
https://hdl.handle.net/10419/43796

Working Papers

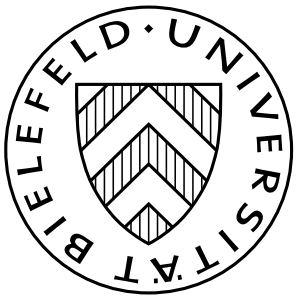
Institute of Mathematical Economics

370

July 2005

Evolution of Non-Expected Utility Preferences

Sven von Widekind



IMW · Bielefeld University
Postfach 100131
33501 Bielefeld · Germany

email: imw@wiwi.uni-bielefeld.de
http://www.wiwi.uni-bielefeld.de/~imw/Papers/showpaper.php?370
ISSN: 0931-6558

Evolution of Non-Expected Utility Preferences*

Sven von Widekind†

12th July 2005

Abstract

We investigate an extension of Dekel, Ely and Yilankaya's (2004) treatment of the evolution of preference to more general, possibly non-expected utility preferences. Along the lines of their analysis we consider a population of types that is repeatedly and randomly matched to play the mixed extension of any given symmetric two-player normal-form game with complete information. In our setup, a type is a generic best-response correspondence that is assumed to satisfy only standard assumptions. Preferences evolve according to the "success" of the player which is determined by the payoff she receives in the game. As in Dekel, Ely and Yilankaya (2004), the players observe the type of their opponent and a Nash equilibrium according to their best responses is played. We show that Dekel, Ely and Yilankaya's result that stability of an outcome implies efficiency is robust in this more general setup. However, in our model we obtain full equivalence between the two concepts for 2x2 games. We show that efficiency of any strategy also implies the stability of the outcome that it induces. This is in contrast to the former work in which only efficiency of a pure strategy leads to a stable outcome. The result implies the existence of a stable outcome in any 2x2 game. Considering the class of rank-dependent expected utility preferences as example we discuss the model's ability to embed specific types of non-expected utility theories. Moreover, we study implications for well-established games like the prisoner's dilemma.

JEL Classification Numbers: C72, D81.

Keywords: Evolution of Preferences, Non-Expected Utility Theory, Best-Response Correspondence, Stability, Efficient Strategy.

---

* Financial support from the German Academic Exchange Service (DAAD) and the German Science Foundation (DFG) is gratefully acknowledged. I would like to thank MEDS at Northwestern University for their hospitality during my visit when parts of this research were conducted. In particular, I thank Eddie Dekel and Peter Klibanoff for offering helpful comments and suggestions. However, all errors are mine.

† Institute of Mathematical Economics (IMW), Bielefeld University, Postfach 100 131, D-33501 Bielefeld, Germany, email: svenvon.widekind@uni-bielefeld.de.

1 Introduction

The evolution of preferences literature bases on the "indirect evolutionary approch" which originated in the articles of Güth and Yaari (1992) and Güth (1995). Players from a large population are randomly and repeatedly matched to play a given two-person normal-form game. In contrast to the prevailing literature on evolutionary game theory[1], where a 'type' is committed to play a certain strategy in the game, players are committed to certain behavior because they are endowed by nature with preferences over the set of outcomes of the game they are playing. The approach seems to be more sophisticated because preferences constitute the very basics of economic behavior and the choice of a particular strategy should come as a result of the decision process of a rational agent. As customary, the evoluionary "success" of a type is determined by its fitness, i.e. the payoff level that she receives in the game. In the course of the evolution types with low fitnesses will be driven out of the population and only the types with the highest expected fitness will survive. A fundamental discussion of the main features of these theories is beyond the scope of this work, though.

We investigate the extension of parts of the work of Dekel, Ely and Yilankaya (2004) to a richer set of possible types. As indicated a type in this setup consists of a preference relation over the set of mixed strategy profiles of the game. Preferences may very well differ from the true fitness values and the question arises whether agents that are endowed with preferences that deviate from payoff maximization can possibly have evolutionary advantages over payoff maximizers. We assume that preferences are observable, i.e. that in any matching between two types either player knows the type of her opponent, and we obtain an affirmative answer. Güth (1995) and Frank (1988) provide justifications for assuming observability. Moreover, if preferences were not observable, then preferences that maximize true fitness should perfom most successfully since every type faces the same distribution of opponents' actions.

Dekel, Ely and Yilankaya (2004) allow players to have any preference relation over the outcomes of the given game. However, they require these preferences to satify the axioms of von Neumann and Morgenstern's (1944) Expected Utility Theory in the sense that the agents are assumed to maximize their expected utilities in the mixed entension of the game. An individual might assign different von Neumann-Morgenstern utility values to two (pure strategy) outcomes which yield equal payoff values though. This latter seems indeed to be an appropriate and interesting assumption since it allows agents not only to take their own payoff (or "fitness") values into account, but also permits the consideration of the opponent's payoff, for instance. Hence, the model may among other things include preferences

[1]For a textbook introduction to the topic we refer, among others, to Weibull (1995).

for fairness and altruism[2]. However, in such an environment it does not seem obvious to us why these decision makers should necessarily maximize their expected utilities. In a related paper, Ely and Yilankaya (2001) themselves question the appropriateness of this assumption[3]. The assumption seems rather to having been made in order to simplify the analysis. In the past, numerous other authors have provided empirical support for the hypothesis that there actually is frequently observable choice behavior which systematically violates Expected Utility Theory and in particular its key component, the independence axiom[4]. In our setup, we consider the set of all preference relations on the set of distributions over outcomes, i.e. over the pairs of mixed strategies, in the mixed extension of the game, explicitly allowing for non-expected utility maximizing behavior. For technical reasons we impose continuity and quasi-concavity over the set of own actions. The type space can therefore possibly comprise preferences that conform to the recent alterative models to Expected Utility Theory. Rationality in our setup just requires a player in the type space to have a preference relation on the set of mixed strategy profiles of the game and to choose optimally in the sense that she always chooses the most preferred strategy available according to her preference relation. This is far less than claiming expected utility maximizing behavior. From a decision theorist's point of view it is still demanding as we at least require completeness and transitivity.

Along the lines of Dekel, Ely and Yilankaya's (2004) approach we analyze the implications of the evolution of preferences. Whenever two types are matched they play a Nash equilibrium of the game that results from their respective preferences. An outcome of the evolution is stable if all types in the corresponding population receive the same fitness and if no type (a 'mutant') can enter that population and receive a higher expected payoff from her matchings than the types already in place. Also, a strategy is efficient if no other strategy can yield a strictly higher fitness when played against itself. This concept is suggestive because the fitness that the types in any stable outcome can obtain must be equal to the fitness from such a symmetric strategy profile. In Dekel, Ely and Yilankaya (2004), this fitness must be equal to the 'efficient payoff'. This necessity of efficiency for stability of an outcome is robust to the current extension of the type space. However, it turns out that in our setup it is also sufficient for 2x2 games: any efficient strategy (whether pure or mixed) induces a stable outcome. This essentially differs from the Dekel et al.'s model, where generic 2x2 games

[2]See among many others, e.g. Levine (1998), Fehr and Schmidt (2000) and Charness and Rabin (2002).

[3]"We do not mean to advance to the position that rationality implies an expected utility representation of preferences" [Ely and Yilankaya (2001, p. 257)].

[4]Some well-known studies include Kahneman and Tversky (1979), Tversky and Kahneman (1992), MacCrimmon and Larsson (1979), Schoemaker (1982) and Camerer and Ho (1994).

with an efficient mixed strategy do not have stable outcomes.

In section 2, we present the generalized model and the resulting implications for the evolution of preferences. Section 3 discusses some examples like the prisoner's dilemma. In addition, the model's ability to embed specific non-expected utility theories is analyzed. The proofs are in the Appendix.

2 Analysis and Results

The general setup of the model is analogous to Dekel, Ely and Yilankaya (2004). Occasionally, we use a different or adjusted notation. We consider the class of symmetric normal-form games with mixed strategies.

Definition 2.1 *A two-player normal-form game is a 3-tupel* $(N = \{1,2\}, (A_1, A_2), (\succsim_1, \succsim_2))$ *consisting of (i) a set of players* $N = \{1,2\}$*, (ii) for each player* $i = 1,2$ *a non-empty set of actions* A_i*, and (iii) for each player* $i = 1,2$ *a preference relation* $\succsim_i$ *over* $A_i \times A_{3-i}$*.*

Let $\bar{A} = \{a_1, ..., a_n\}$ be some finite set. Both players choose mixed strategies over $\bar{A}$. That is, their action sets consist of the set of all probability distributions over $\bar{A}$, i.e. $A_1 = A_2 = \Delta = \{(\sigma^1, ..., \sigma^n) \in \mathbb{R}^n \mid \sigma^i \geq 0$ for all $i = 1, ..., n$ and $\sum_{i=1}^{n} \sigma^i = 1\}$. Also, we assume the existence of a function $\pi : \bar{A} \times \bar{A} \to \mathbb{R}$ which is interpreted as the payoff (or "fitness") function, with $\pi(a, a')$ being the fitness that a player receives if she is playing the pure strategy a and her opponent is playing strategy a'. The function π is the same for both players and, thus, the game is henceforth called symmetric[5]. The evolutionary success of each player depends on her fitness values in the game. The fitness function π is extended as normal to the domain of pairs of mixed strategies, i.e. to $\Delta \times \Delta$, by the taking expected values. The following matrix is an example of the fitness values in a symmetric 2x2 normal-form game where $\bar{A} = \{A, B\}$:

	A	B
A	a,a	b,c
B	c,b	d,d

A large population of players is assumed to consist of different types, and we denote the set of these types by $\mathcal{T}$. In Dekel, Ely and Yilankaya (2004), the players have expected utility preferences with a von Neumann-Morgenstern utility function $u : \bar{A} \times \bar{A} \to [0,1]$ over pairs of pure strategies. That is, the authors allow each agent i to have preferences over outcomes

[5]We are aware that this terminology conflicts with the one in some of the literature where 'symmetry' of a game is sometimes defined in terms of the preferences, i.e. Osborne and Rubinstein (1994). Also, in abuse of notation, we refer to 'the' symmetric game although different preferences formally lead to different games, of course.

which differ from the true payoff values according to π. For type $T \in \mathcal{T}$ they have $(\sigma_T, \sigma_{-T}) \succsim_T (\sigma'_T, \sigma'_{-T})$ if and only if $\sum_i \sum_j \sigma_T^i \sigma_{-T}^j u_T(a_i, a_j) \geq \sum_i \sum_j \sigma'^i_T \sigma'^j_{-T} u_T(a_i, a_j)$ for all $\sigma_T, \sigma_{-T}, \sigma'_T, \sigma'_{-T} \in \Delta$.

In our setup, we enrich the type space and allow for non-expected utility preferences. Our type space $\mathcal{T}$ consists of all preference relations over $\Delta \times \Delta$ that are continuous and quasi-concave in the first component.

Definition 2.2 *A preference relation $\succsim_i$ on $A_i \times A_{-i}$ is quasi-concave on A_i if for every $a^* \in A_i \times A_{-i}$ the set $\{a_i \in A_i : (a_i, a^*_{-i}) \succsim_i a^*\}$ is convex.*

The reason for imposing continuity and quasi-concavity is technical. We would like to retain Dekel, Ely and Yilankaya's (2004) assumption that two players that are matched play a Nash equilibrium of the game. Continuity and quasi-concavity on the own strategy space of the preference relations are sort of the minimal requirements that guarantee the existence of a Nash equilibrium via Kakutani's fixed point theorem. Continuity basically secures the existence of a continuous utility function on $\Delta \times \Delta$ representing a player's preferences. As we discuss later in this paper, some specific modern alternatives to von Neumann and Morgenstern's (1944) Expected Utility Theory unfortunately do not satisfy the quasi-concavity assumption. However, with such preferences the game between two matched players may fail to have a Nash equilibrium. Hence, if one wishes to further generalize the model in order to include such preferences, then other concepts for the outcome of a matching between two players will be needed.

For convenience, we will often identify a "type", i.e. a preference relation over outcomes of the game, with the best-response correspondence it induces given the agent's rational behavior.

Definition 2.3 *Let type $T \in \mathcal{T}$ have a preference relation $\succsim$ over $\Delta \times \Delta$. The set-valued function $\beta_T : \Delta \to \Delta$ defined by $\beta_T(\sigma_{-T}) = \{\sigma_T \in \Delta \mid (\sigma_T, \sigma_{-T}) \succsim_T (\sigma'_T, \sigma_{-T})$ for all $\sigma'_T \in \Delta\}$ is called type T's best-response correspondence.*

Of course, a given best-response correspondence may be induced by different preference relations, but in terms of the evolution these types will coincide as will become apparent shortly. Therefore, we will henceforth concentrate on best responses. The structural assumptions on preferences we make restrict these best-response correspondences to satisfy the closed-graph criterion and the image sets to be non-empty and convex-valued.

Definition 2.4 *A Nash equilibrium of the two-player normal-form game $(N = \{1, 2\}, (\Delta, \Delta), (\succsim_1, \succsim_2))$ is a pair $(\sigma_1, \sigma_2) \in \Delta \times \Delta$ such that $\sigma_i \in \beta_i(\sigma_{3-i})$ for $i = 1, 2$.*

Existence of a Nash equilibrium for any matching of two types in $\mathcal{T}$ follows by Kakutani's fixed point theorem precisely because of our two structural assumptions of continuity and quasi-concavity of preferences.

Proposition 2.5 *The two-player normal-form game* $(N = \{1,2\}, (\Delta, \Delta), (\succsim_T, \succsim_{T'}))$ *played by any two types* $T, T' \in \mathcal{T}$ *has a Nash equilibrium.*

As in Dekel, Ely and Yilankaya (2004), we assume that whenever two player from the population are matched a Nash equilibrium of 'their' game is played. This is, if player T is characterized by a best-response correspondence β_T and she is matched with type T', then a strategy $(\sigma_T, \sigma_{T'})$ is played such that $\sigma_T \in \beta_T(\sigma_{T'})$ and $\sigma_{T'} \in \beta_{T'}(\sigma_T)$. We will not discuss the economic justification for this assumption. Rather, we refer to the exposition in Dekel, Ely and Yilankaya (2004), where an interpretation of the Nash equilibrium as the outcome of a learning process is provided. Given a probability distribution μ over the type space $\mathcal{T}$, we denote the support of μ by $C(\mu)$. We assume that $C(\mu)$ is finite.

Definition 2.6 *An equilibrium configuration is a function* $b : C(\mu) \times C(\mu) \to \Delta \times \Delta$ *such that* $b(T, T')$ *is a Nash equilibrium in the game between* T *and* T' *and* $b_1(T, T') = b_2(T', T)$ *for all* $T, T' \in C(\mu)$.

The set of all possible equilibirum configurations given μ is denoted by $B(\mu)$. The latter requirement in Definition 2.6 means that the players do not know their positions in the game and cannot condition their strategy on whether they are the row or the column player. In particular, if two players of the same type are matched, they need to play a symmetric Nash equilibrium. Under our assumptions, such a symmetric Nash equilibrium necessarily exists:

Proposition 2.7 *The game* $(N = \{1,2\}, (\Delta, \Delta), (\succsim, \succsim))$*, where* $\succsim$ *is continuous and quasi-concave in the first component, has a symmetric Nash equlibrium, i.e. there exists* $\sigma_{\succsim} \in \Delta$ *such that* $(\sigma_{\succsim}, \sigma_{\succsim})$ *is a Nash equilibrium.*

The proof of Proposition 2.7 is Exercise 20.4 in Osborne and Rubinstein (1994) and is based on Kakutani's fixed-point theorem.

Given a distribution of types μ and an equilibrium configuration $b \in B(\mu)$, we can compute the expected fitness of every type $T \in C(\mu)$:

$$\Pi_T(\mu \mid b) = \sum_{T' \in C(\mu)} \mu(T')\ \pi(b(T, T')).$$

In the following definition of stability we use T to also denote a degenerate distribution that only consists of Ts.

Definition 2.8 *An outcome x is stable (with μ and $b \in B(\mu)$) if x is the outcome induced by the equilibrium configuration b under the distribution of types μ such that $\Pi_T(\mu \mid b) = \Pi_{T'}(\mu \mid b)$ for all $T, T' \in C(\mu)$ and we have:*

$$\forall T \in \mathcal{T}\ \exists \epsilon' > 0\ \forall \epsilon \in (0, \epsilon')\ \forall T_\mu \in C(\mu)\ \forall \bar{b} \in B((1-\epsilon)\mu + \epsilon T \mid b) : \\ \Pi_{T_\mu}((1-\epsilon)\ \mu + \epsilon\ T \mid \bar{b}) \geq \Pi_T((1-\epsilon)\ \mu + \epsilon\ T \mid \bar{b}),$$

where $B((1-\epsilon)\mu + \epsilon T \mid b) = \{\tilde{b} \in B((1-\epsilon)\mu + \epsilon T) : \tilde{b}(T_1, T_2) = b(T_1, T_2)$ whenever $T_1, T_2 \in C(\mu)\}$.

Note that this stability concept is static. Dynamic evolutionary processes that lead to a distribution of types associated with a stable outcome are not explicitly considered here.

As Dekel, Ely and Yilankaya (2004) we instead focus on characterizing stable outcomes using the concept of efficiency. A (mixed) strategy in a symmetric normal-form game is called efficient if no other strategy yields a higher fitness when played against itself. For a detailled discussion of the meaningfulness of this concept we again refer to the exposition in Dekel, Ely and Yilankaya (2004).

Definition 2.9 *Let G be any finite symmetric normal-form game. A strategy $\sigma^* \in \Delta$ is called efficient if $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma, \sigma)$ for all $\sigma \in \Delta$.*

Since Δ is compact and π is continuous, any finite symmetric normal-form game has an efficient strategy σ^*. We abbreviate $\pi(\sigma^*, \sigma^*)$ by π^*. By definition, π^* is uniquely determined.

Proposition 2.10 *Let G be any finite symmetric normal-form game. Suppose that x is stable with a distribution μ and an equilibrium configuration $b \in B(\mu)$. Then, we have $\Pi_T(\mu \mid b) = \pi(b(T, T')) = \pi^*$ for all $T, T' \in C(\mu)$.*

Proof. The proofs of Propositions 1 and 3 in Dekel, Ely and Yilankaya (2004) can be directly brought forward to our setup. □

Proposition 2.10 shows that in a stable outcome all types in the distribution must receive the same fitness from any matching with any other type. Also, this fitness value must be equal to the efficient payoff. The result illustrates the robustness of our setup with respect to the efficiency implications of stability as shown in Dekel, Ely and Yilankaya (2004). The intuition is that with our richer type space there are more types that could possibly invade the population. Hence, stability is now even a stronger requirement as the conditions in Definition 2.8 have to hold against more potential entrants. Therefore, the expected fitness that every type needs to obtain in a stable outcome cannot be lower than before. It will be more interesting to see what we can say about sufficient conditions for stability. As we will see

shortly, the results change substantially here, in particular regarding stability implications of efficient mixed strategies in 2x2 games. The results for pure strategies can be carried over.

Proposition 2.11 *Let G be any symmetric 2x2 normal form-game. If a pure strategy $a_i \in \Delta$ is efficient and $\pi(a_i, a_i) > \pi(a_j, a_i)$ for all pure strategies $a_j \neq a_i$ in Δ, then (a_i, a_i) is stable.*

Proof. The proof of Proposition 2 in Dekel, Ely and Yilankaya (2004) can be directly brought forward to our setup. □

In the case of 2x2 games, we conveniently identify a strategy with an element σ of $[0, 1]$. It is meant as a shortcut for $(\sigma, 1-\sigma)$, where A is played with probability σ and B is played with probability $1-\sigma$. In the following, we assume w.l.o.g. that $a \geq d$, where the fitness values in the 2x2 game are as in the example given at the beginning of this section.

Proposition 2.12 *Let G be any symmetric 2x2 normal form-game. Suppose that A is efficient. Then, the outcome (a, a) is stable.*

Proof. The proof of Proposition 4 a) in Dekel, Ely and Yilankaya (2004) can be directly brought forward to our setup. □

If, in addition, some strategy $\sigma^* \neq A$ is efficient, then we must necessarily have $\pi(\sigma^*, \sigma^*) = \pi(A, A)$. Next we show that in this case the outcome induced by (σ^*, σ^*) is also stable.

Proposition 2.13 *Let G be any symmetric 2x2 normal-form game such that A is efficient. Then, if $\sigma^* \in \Delta$ is efficient, the outcome induced by (σ^*, σ^*) is stable.*

Proof. See Appendix. □

Stability of outcomes induced by an efficient mixed strategy conditional on the co-existence of the efficient pure strategy A is not discussed in Proposition 4 in Dekel, Ely and Yilankaya (2004). However, an example shows that in their setup such an outcome need not necesserily be stable. Consider the following 2x2 symmetric normal-form game:

	A	B
A	1,1	0,2
B	2,0	1,1

We have $\pi(\sigma,\sigma) = \sigma^2 + (1-\sigma)^2 + 2\sigma(1-\sigma) = 1$ for all $\sigma \in [0,1]$. Therefore, all strategies are efficient. Let $\sigma' \in (0,1)$. For any expected utility maximizer, we must have $\beta(\sigma') = [0,1]$ in order to obtain the outcome induced by (σ',σ'). Consider an entrant with the following best-response correspondence:

$$\beta_e(\sigma) = \begin{cases} [0,1] & \text{if } \sigma = 1 \\ 1 & \text{otherwise} \end{cases}.$$

This best-response correspondence comes from expected utility preferences. Hence, the entrant is in Dekel et al.'s type space. Assume that in the post-entry equilibrium configuration $\bar{b}$ the equilibrium that is played when an entrant and any incumbent are matched is $(\sigma_e,\sigma_i) = (\sigma',1)$. The entrant's expected fitness from these matches is $\sigma' + 2(1-\sigma') > 1 = \pi(\sigma',\sigma')$. Hence, if entering in any proportion, the entrant can successfully invade the population. The outcome induced by (σ',σ') is not stable. Note that theses arguments for non-stability can be brought forward to all cases of co-existence of efficient mixed and efficient pure strategies unless $a = b = c = d$.

In contrast to Dekel, Ely and Yilankaya (2004), the stabibilty implication can generally be carried over to all efficient mixed strategies in our setup. The co-efficiency of the pure strategy A is not required.

Proposition 2.14 *Let G be any symmetric 2x2 normal form-game such that A is not efficient. Suppose that $\sigma^* \in \Delta$ is efficient. Then, the outcome induced by (σ^*,σ^*) is stable.*

Proof. See Appendix. □

Combining Propositions 2.12, 2.13 and 2.14 we have the following result.

Theorem 2.15 *Let G be any symmetric 2x2 normal-form game. If $\sigma^* \in \Delta$ is efficient, then the outcome induced by (σ^*,σ^*) is stable.*

Every finite symmetric normal-form game has an efficient strategy. The following existence result for stable outcomes is now an immediate consequence of Theorem 2.15.

Corollary 2.16 *Any symmetric 2x2 normal-form game has a stable outcome.*

In contrast to Dekel, Ely and Yilankaya's (2004) setup with expected utility maximizing agents, in which generic 2x2 games with an efficient mixed strategy do not have a stable outcome, existence is always warranted in our non-expected utility setup.

3 Examples

In this section we illustrate the significance of the results considering some examples. First, we look at a classical prisoner's dilemma game, where the fitness values are given as follows:

	A	B
A	3,3	0,10
B	10,0	1,1

Strategy A (Cooperation) is strictly dominated by B (Defection). By straightforward calculation, one finds that the unique efficient strategy is $\sigma^* = \frac{2}{3}$. The associated payoff is $\pi(\sigma^*, \sigma^*) = 3\frac{2}{3}$. Note that this fitness is strictly higher than 3, the fitness obtained through cooperation. This is naturally true since A is not efficient. We can clearly see that in the prisoner's dilemma neither an outcome induced by a population of cooperating players nor one induced by a population of defecting players can be stable. This is also true in Dekel, Ely and Yilankaya's (2004) model. However, in their setup this game does not have a stable outcome (Proposition 4b in Dekel, Ely and Yilankaya (2004)). Corollary 2.16 shows the existence of such stable outcome in the more general model. Stability is, for instance, induced by a monomorphic population of types that always play B if their opponent is not playing the efficient strategy σ^* and that are indifferent between all strategies in $[0, \sigma^*]$ if her opponent is playing σ^*. The associated equilibrium configuration is obviously (σ^*, σ^*). We refer to case *iii)* of the proof of Proposition 2.14 in the Appendix for the details.

At this point, we would like to remark that the type space that we have chosen in our setup is very general and permits a large range of preferences to occur in the population. Non-expected utility maximizing behavior is explicitly embedded. However, we assume that preferences are continuous and satisfy the quasi-concavity condition. The latter is crucial since is still excludes some types of preferences that have been developed as alternatives to von Neumann and Morgenstern's (1944) Expected Utility Theory. To illustrate this, consider decision makers with Rank-dependent Expected Utility (RDEU) preferences[6]. For simplicity, assume that the utility function over outcomes coincides with the fitness function. That is, the agents' preferences satisfy the conditions of Yaari's (1987) Dual Theory of Choice under Risk. The preference relation $\succsim_Y$ of such a player in a game is then represented by a strictly increasing, continuous function $f : [0,1] \to [0,1]$

[6]RDEU models have been first introduced by Quiggin (1982). Axiomatizations have been provided by Wakker (1994) and Yaari (1987), the latter with a linear utility function over the set of outcomes that coincides with the fitness function.

with $f(0) = 0$ and $f(1) = 1$ such that $(\sigma_Y, \sigma_{-Y}) \succsim_Y (\sigma'_Y, \sigma'_{-Y})$ if and only if $\sum_{i=1}^{n^2} \pi(\bar{a}_i)[f(\sum_{j=i}^{n^2} p_j) - f(\sum_{j=i+1}^{n^2} p_j)] \geq \sum_{i=1}^{n^2} \pi(\bar{a}_i)[f(\sum_{j=i}^{n^2} p'_j) - f(\sum_{j=i+1}^{n^2} p'_j)]$, where $\bar{a}_i = (a_k, a_l)$ for some $k, l \in \{1, ..., n\}$ (and $\bar{a}_i \neq \bar{a}_{i'}$ for $i \neq i'$) such that $\pi(\bar{a}_1) \leq ... \leq \pi(\bar{a}_{n^2})$ and $p_i = \sigma_Y^k \sigma_{-Y}^l$ for the specified k and l. The $\bar{a}_i$s are just an re-ordering of the outcomes in terms of the resulting fitness values. Such an agent is simply maximizing her rank-dependent expected fitness.

Consider a symmetric normal-form game where no two pure strategy pairs receive the same fitness value. If the preferences of all such players are represented by strictly convex fs, then the resulting game which is played by any two of these players can only have pure strategy Nash equilibria (Ritzberger (1996), Proposition 1). For example, consider the 2x2 game with the following fitness values:

	A	B
A	1,1	20,20
B	10,10	0,0

For any two players with preferences as just described, the game has two Nash equilibria, (A, B) and (B, A). However, it has no symmetric Nash equilibrium. That means that in order to play a Nash equilibrium two players of the same type would need to condition their choice on their position in the game when matched with each other, which contradicts our assumption that they cannot. The reason for the non-existence of mixed strategy equilibria is that the players are not willing to randomize between the pure strategies. The image sets of the best-response correspondences are not convex-valued. Obviously, the requirements of Kakutani's fixed point theorem are not fulfilled. Note that the case of convex probability transformation functions corresponds to a situation where the decision makers are risk-averse, in any sense of the world[7]. The existence of the appropriate Nash equilibria can be guaranteed only for types with (weakly) concave probability transformation functions. We omit the details. But with concave transformation functions none of the agents can actually be risk-averse. Indeed, a setup with (possibly) risk-averse agents would probably be the more interesting case. If one aims at embedding types with such preferences, concepts for the outcome of a matching of two types other than Nash equilibrium will be needed. Alternatively, the assumption that the players cannot condition their choices on their positions in the game must be relaxed.

[7]For the argument see Yaari (1987) and Röell (1987).

4 Appendix

Proof of Proposition 2.13. For lucidity define $\pi^* \equiv \pi(\sigma^*, \sigma^*)$. We consider several possible cases with monomorphic populations, respectively, and equilibrium configurations b such two members of the respective population play the equilibrium (σ^*, σ^*) whenever they are matched.

In the following steps, we assume that the fitness values of the game are given as follows:

	A	B
A	a,a	b,c
B	c,b	d,d

Define $f : [0,1] \to \mathbb{R}$ by $f(\sigma) = \sigma^2 a + \sigma(1-\sigma)(b+c) + (1-\sigma)^2 d - a$. Since A and σ^* are efficient, we have $f \leq 0$ and $f(1) = 0 = f(\sigma^*)$. Rewriting f leads to $f(\sigma) = (a + d - b - c)\ \sigma^2 + (b + c - 2d)\ \sigma + d - a$.

If $\sigma^* = 0$, then we must have $a = d$ and $f \leq 0$ implies $2a = a + d \geq b + c$. If $\sigma^* \in (0,1)$, then 1 and σ^* can only both maximize f if $f = 0$, in which case all strategies $\sigma \in [0,1]$ are efficient. Consequently, we have $a = d$ and $b + c = 2a$. Hence, we need only consider the cases where $a = d$ and $2a \geq b + c$.

i) $a = d = b$: First, assume that $\sigma^* \in (0,1)$. We must have $b + c = 2d$ which implies that $a = b = c = d$. The outcome induced by (σ^*, σ^*) is trivially stable, for instance with a monomorphic population of types $T(\sigma^*)$ that always play $\sigma = \sigma^*$.

Second, assume that $\sigma^* = 0$, i.e. B is efficient. Then, we have $2a = a + d \geq b + c$, i.e. $a \geq c$. Consider a monomorphic population of types $T(0,1)$ with the following best-response correspondence:

$$\beta_{0,1}(\sigma) = \begin{cases} [0,\ 1] & \text{if } \sigma = 0 \\ 1 & \text{otherwise} \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{0,1}((1-\epsilon)\ T(0,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*, \sigma^*) + \epsilon\ \pi(\sigma_i, \sigma_e)$$

and

$$\Pi_e((1-\epsilon)\ T(0,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e, \sigma_i) + \epsilon\ \pi(\sigma_3, \sigma_3).$$

If $\sigma_e \neq 0$, then $\sigma_i = 1$. Hence, $\pi(\sigma_e, \sigma_i) = \sigma_e\ a + (1 - \sigma_e)\ c \leq a = \pi^*$ and $\pi(\sigma_i, \sigma_e) = \sigma_e\ a + (1 - \sigma_e)\ b = a = \pi^*$. Therefore $\Pi_{0,1}(.) = \pi^* \geq \Pi_e(.)$ for all

ϵ (remember that $\pi(\sigma_3, \sigma_3) \leq \pi^*$ by the definition of efficiency). If $\sigma_e = 0$, then $\pi(\sigma_e, \sigma_i) = \sigma_i c + (1 - \sigma_i)d \leq d = \pi^*$ and $\pi(\sigma_i, \sigma_e) = \sigma_i b + (1 - \sigma_i)d = a = \pi^*$. Again, the expected fitness of the entrant never exceeds that of the incumbent.

ii) $a = d > b$*:* First, assume that $\sigma^* \in (0, 1)$. We have $c = 2a - b > a$. Consider a monomorphic population of types $T(0, \sigma^*)$ with the following best-response correspondence:

$$\beta_{0,\sigma^*}(\sigma) = \begin{cases} [0,\ \sigma^*] & \text{if } \sigma = \sigma^* \\ 0 & \text{otherwise} \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{0,\sigma^*}((1 - \epsilon)\ T(0, \sigma^*) + \epsilon\ T(e) \mid \bar{b}) = (1 - \epsilon)\ \pi(\sigma^*, \sigma^*) + \epsilon\ \pi(\sigma_i, \sigma_e)$$

and

$$\Pi_e((1 - \epsilon)\ T(0, \sigma^*) + \epsilon\ T(e) \mid \bar{b}) = (1 - \epsilon)\ \pi(\sigma_e, \sigma_i) + \epsilon\ \pi(\sigma_3, \sigma_3).$$

If $\sigma_e \neq \sigma^*$, then $\sigma_i = 0$. Hence, $\pi(\sigma_e, \sigma_i) = \sigma_e\ b + (1 - \sigma_e)\ d \leq d = \pi^*$ and $\pi(\sigma_i, \sigma_e) = \sigma_e\ c + (1 - \sigma_e)\ d \geq d = \pi^*$. Therefore $\Pi_{0,\sigma^*}(.) \geq \pi^* \geq \Pi_e(.)$ for all ϵ (remember that $\pi(\sigma_3, \sigma_3) \leq \pi^*$ by the definition of efficiency). If $\sigma_e = \sigma^*$, then

$$\begin{aligned} \pi(\sigma_e, \sigma_i) &= \sigma_i\ [\sigma^* a + (1 - \sigma^*)c] + (1 - \sigma_i)\ [\sigma^* b + (1 - \sigma^*)d] \\ &\leq \pi^* \end{aligned}$$

and

$$\begin{aligned} \pi(\sigma_i, \sigma_e) &= \sigma_i\ [\sigma^* a + (1 - \sigma^*)b] + (1 - \sigma_i)\ [\sigma^* c + (1 - \sigma^*)d] \\ &\geq \pi^*, \end{aligned}$$

for any $\sigma_i \in [0, \sigma^*]$ (with equalities for $\sigma_i = \sigma^*$). Again, the expected fitness of the entrant never exceeds that of the incumbent.

Second, assume that $\sigma^* = 0$, i.e. B is efficient. We have that (d, d) is stable with a monomorphic population of $T(0)$s, i.e. with types that always play B. For, an entrant can obtain a payoff of at least π^* from a matching with a $T(0)$ only if she plays $\sigma_e = 0$. But in this case the incumbent also receives π^* from this matching and the entrant cannot succesfully invade the population.

iii) $b > a = d$*:* First, assume that $\sigma^* \in (0,1)$. We have $c = 2a - b < a$. Consider a monomorphic population of types $T(\sigma^*, 1)$ with the following best-response correspondence:

$$\beta_{\sigma^*,1}(\sigma) = \begin{cases} [\sigma^*,\ 1] & \text{if } \sigma = \sigma^* \\ 1 & \text{otherwise} \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{\sigma^*,1}((1-\epsilon)\ T(\sigma^*,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*,\sigma^*) + \epsilon\ \pi(\sigma_i,\sigma_e)$$

and

$$\Pi_e((1-\epsilon)\ T(\sigma^*,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e,\sigma_i) + \epsilon\ \pi(\sigma_3,\sigma_3).$$

If $\sigma_e \neq \sigma^*$, then $\sigma_i = 1$. Hence, $\pi(\sigma_e,\sigma_i) = \sigma_e\ a + (1-\sigma_e)\ c \leq a = \pi^*$ and $\pi(\sigma_i,\sigma_e) = \sigma_e\ a + (1-\sigma_e)\ b \geq a = \pi^*$. Therefore $\Pi_{\sigma^*,1}(.) \geq \pi^* \geq \Pi_e(.)$ for all ϵ (remember that $\pi(\sigma_3,\sigma_3) \leq \pi^*$ by the definition of efficiency). If $\sigma_e = \sigma^*$, then

$$\begin{aligned} \pi(\sigma_e,\sigma_i) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma_i)\ [\sigma^* b + (1-\sigma^*)d] \\ &\leq \pi^* \end{aligned}$$

and

$$\begin{aligned} \pi(\sigma_i,\sigma_e) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)b] + (1-\sigma_i)\ [\sigma^* c + (1-\sigma^*)d] \\ &\geq \pi^*, \end{aligned}$$

for any $\sigma_i \in [\sigma^*, 1]$ (with equalities for $\sigma_i = \sigma^*$). The expected fitness of the entrant never exceeds that of the incumbent.

Second, assume that $\sigma^* = 0$, i.e. B is efficient. With similar arguments as in the case $\sigma^* \in (0,1)$ one verifies that (d,d) is stable with a monomorphic population of $T(0,1)$s, i.e. with the following best-response correspondence:

$$\beta(\sigma) = \begin{cases} [0,\ 1] & \text{if } \sigma = 0 \\ 1 & \text{otherwise} \end{cases}.$$

This gives us the desired result. □

Proof of Proposition 2.14. Since A is not efficient, B cannot be efficient either because we have assumed that $a \geq d$. Hence, $\sigma^* \in (0,1)$. By definition, $\sigma^* = \arg\max_{\sigma\in[0,1]} \sigma^2\ a + \sigma(1-\sigma)\ (b+c) + (1-\sigma)^2\ d$, which yields

$$\sigma^* = \frac{b + c - 2d}{2(b + c - a - d)}$$

via the first-order condition. The efficient payoff is

$$\begin{aligned}\pi^* \equiv \pi(\sigma^*, \sigma^*) &= (\sigma^*)^2\ a + \sigma^*(1-\sigma^*)\ (b+c) + (1-\sigma^*)^2\ d \\ &= d + \frac{(b+c-2d)^2}{4(b+c-a-d)}.\end{aligned}$$

In the following, we consider populations, in each of which (σ^*, σ^*) is an equilibrium that is played in b whenever two incumbents are matched. We show that in a each such case no type can successfully invade the respective population. We consider several possible cases:

i) $a \geq c,\ d \geq b$*:* Since $a = \max\{a, b, c, d\}$, we have that A is efficient. Thus, this case cannot occur.

ii) $a \geq c,\ b > d$*:* If $c \geq b$, then $a = \max\{a, b, c, d\}$, and therefore A is efficient. This case cannot occur. Hence, we must have $b > c$. Consider a monomorphic population of types $T(\sigma^*, 1)$ with the following best-response correspondence:

$$\beta_{\sigma^*,1}(\sigma) = \begin{cases} [\sigma^*,\ 1] & \text{if } \sigma = \sigma^* \\ 1 & \text{otherwise} \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{\sigma^*,1}((1-\epsilon)\ T(\sigma^*, 1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*, \sigma^*) + \epsilon\ \pi(\sigma_i, \sigma_e)$$

and

$$\Pi_e((1-\epsilon)\ T(\sigma^*, 1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e, \sigma_i) + \epsilon\ \pi(\sigma_3, \sigma_3).$$

If $\sigma_e \neq \sigma^*$, then $\sigma_i = 1$. Hence, $\pi(\sigma_e, \sigma_i) = \sigma_e\ a + (1-\sigma_e)\ c \leq a < \pi^*$. We can find a sufficiently small $\epsilon' > 0$ such that for all $\epsilon \in (0, \epsilon')$ we have $\Pi_{\sigma^*,1}(.) > \Pi_e(.)$. If $\sigma_e = \sigma^*$, then $\sigma_i \in [\sigma^*, 1]$ and

$$\begin{aligned}\pi(\sigma_e, \sigma_i) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma_i)\ [\sigma^* b + (1-\sigma^*)d] \\ &\leq \pi^*,\end{aligned}$$

and

$$\begin{aligned}\pi(\sigma_i, \sigma_e) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)b] + (1-\sigma_i)\ [\sigma^* c + (1-\sigma^*)d] \\ &\geq \pi^*,\end{aligned}$$

for any $\sigma_i \in [\sigma^*, 1]$ (with equalities for $\sigma_i = \sigma^*$). The expected fitness of the entrant never exceeds that of the incumbent. With this population the

outcome induced by (σ^*, σ^*) is therefore stable.

iii) $c > a \geq d \geq b$*:* Consider a monomorphic population of types $T(0, \sigma^*)$ with the following best-response correspondence:

$$\beta_{0,\sigma^*}(\sigma) = \begin{cases} [0,\ \sigma^*] & \text{if } \sigma = \sigma^* \\ 0 & \text{otherwise} \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{0,\sigma^*}((1-\epsilon)\ T(0,\sigma^*) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*, \sigma^*) + \epsilon\ \pi(\sigma_i, \sigma_e),$$

and

$$\Pi_e((1-\epsilon)\ T(0,\sigma^*) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e, \sigma_i) + \epsilon\ \pi(\sigma_3, \sigma_3).$$

If $\sigma_e \neq \sigma^*$, then $\sigma_i = 0$. Hence, $\pi(\sigma_e, \sigma_i) = \sigma_e\ b + (1-\sigma_e)\ d \leq a < \pi^*$. We can find a sufficiently small $\epsilon' > 0$ such that for all $\epsilon \in (0, \epsilon')$ we have $\Pi_{0,\sigma^*}(.) > \Pi_e(.)$. If $\sigma_e = \sigma^*$, then $\sigma_i \in [0, \sigma^*]$ and

$$\begin{aligned} \pi(\sigma_e, \sigma_i) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma_i)\ [\sigma^* b + (1-\sigma^*)d] \\ &\leq \pi^* \end{aligned}$$

and

$$\begin{aligned} \pi(\sigma_i, \sigma_e) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)b] + (1-\sigma_i)\ [\sigma^* c + (1-\sigma^*)d] \\ &\geq \pi^* \end{aligned}$$

for any $\sigma_i \in [0, \sigma^*]$ (with equalities for $\sigma_i = \sigma^*$). The expected fitness of the entrant never exceeds that of the incumbent. With this population the outcome induced by (σ^*, σ^*) is therefore stable.

iv) $c > a,\ b > d,\ b = c$*:* This is the only case (i.e., a non-generic Hawk-Dove game) in which in Dekel, Ely and Yilankaya's (2004) setup efficiency of a mixed strategy implies stability. Their "stable" population can be used here as well. Consider a monomorphic population of types $T(0, \sigma^*, 1)$ with the following best-response correspondence:

$$\beta_{0,\sigma^*,1}(\sigma) = \begin{cases} 1 & \text{if } \sigma > \sigma^* \\ [0,\ 1] & \text{if } \sigma = \sigma^* \\ 0 & \text{if } \sigma < \sigma^* \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium

between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{0,\sigma^*,1}((1-\epsilon)\ T(0,\sigma^*,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*,\sigma^*) + \epsilon\ \pi(\sigma_i,\sigma_e),$$

and

$$\Pi_e((1-\epsilon)\ T(0,\sigma^*,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e,\sigma_i) + \epsilon\ \pi(\sigma_3,\sigma_3).$$

If $\sigma_e < \sigma^*$, then $\sigma_i = 0$. Hence,

$$\begin{aligned}
\pi(\sigma_e,\sigma_i) &= \sigma_e\ b + (1-\sigma_e)\ d \\
&< \sigma^*\ b + (1-\sigma^*)\ d \\
&= d + \sigma^*(b-d) \\
&= d + \frac{2b-2d}{2(2b-a-d)}(b-d) \\
&= d + \frac{(2b-2d)^2}{4(2b-a-d)} \\
&= \pi^*.
\end{aligned}$$

If $\sigma_e > \sigma^*$, then $\sigma_i = 1$. Hence,

$$\begin{aligned}
\pi(\sigma_e,\sigma_i) &= \sigma_e\ a + (1-\sigma_e)\ c \\
&< \sigma^*\ a + (1-\sigma^*)\ b \\
&= \pi^*
\end{aligned}$$

(the latter equality holds because we have $\sigma^*\ b + (1-\sigma^*)\ d = \pi^* = \sigma^*\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma^*)\ [\sigma^* b + (1-\sigma^*)d]$). In either case, we can find a sufficiently small $\epsilon' > 0$ such that for all $\epsilon \in (0,\epsilon')$ we have $\Pi_{0,\sigma^*,1}(.) > \Pi_e(.)$. If $\sigma_e = \sigma^*$, then $\sigma_i \in [0,1]$ and we have

$$\begin{aligned}
\pi(\sigma_e,\sigma_i) &= \pi(\sigma^*,\sigma_i) \\
&= \sigma_i\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma_i)\ [\sigma^* b + (1-\sigma^*)d] \\
&= \pi^*
\end{aligned}$$

and

$$\begin{aligned}
\pi(\sigma_i,\sigma_e) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)b] + (1-\sigma_i)\ [\sigma^* c + (1-\sigma^*)d] \\
&= \pi^*
\end{aligned}$$

since $b = c$. The expected fitness of the entrant never exceeds that of the incumbent. With this population the outcome induced by (σ^*,σ^*) is therefore stable.

v) $b > c > a > d,\ c \leq \pi^*$*:* Consider a monomorphic population of types $T(\sigma^*, 1)$ with the following best-response correspondence:

$$\beta_{\sigma^*,1}(\sigma) = \begin{cases} [\sigma^*,\ 1] & \text{if } \sigma = \sigma^* \\ 1 & \text{otherwise} \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{\sigma^*,1}((1-\epsilon)\ T(\sigma^*, 1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*, \sigma^*) + \epsilon\ \pi(\sigma_i, \sigma_e),$$

and

$$\Pi_e((1-\epsilon)\ T(\sigma^*, 1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e, \sigma_i) + \epsilon\ \pi(\sigma_3, \sigma_3).$$

If $\sigma_e \neq \sigma^*$, then $\sigma_i = 1$. Hence, $\pi(\sigma_e, \sigma_i) = \sigma_e\ a + (1-\sigma_e)\ c \leq c \leq \pi^*$ with equality only if $\sigma_e = 0$. However, then $\pi(\sigma_i, \sigma_e) = b > \pi^*$ and the expected fitness of the incumbent would therefore be higher than that of the entrant, whose expected payoff is π^* at most. If $\sigma_e \neq \sigma^*$ and $\sigma_e \neq 0$, then $\pi(\sigma_e, \sigma_i) < c \leq \pi^*$ and we can find a sufficiently small $\epsilon' > 0$ such that for all $\epsilon \in (0, \epsilon')$ we have $\Pi_{\sigma^*,1}(.) > \Pi_e(.)$. If $\sigma_e = \sigma^*$, then $\sigma_i \in [\sigma^*, 1]$ and

$$\begin{aligned} \pi(\sigma_e, \sigma_i) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma_i)\ [\sigma^* b + (1-\sigma^*)d] \\ &\leq \pi^* \end{aligned}$$

and

$$\begin{aligned} \pi(\sigma_i, \sigma_e) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)b] + (1-\sigma_i)\ [\sigma^* c + (1-\sigma^*)d] \\ &\geq \pi^* \end{aligned}$$

for any $\sigma_i \in [\sigma^*, 1]$ (with equalities for $\sigma_i = \sigma^*$). The expected fitness of the entrant never exceeds that of the incumbent. With this population the outcome induced by (σ^*, σ^*) is therefore stable.

vi) $c > b > a > d,\ b \leq \pi^*$*:* Consider a monomorphic population of types $T(0, \sigma^*)$ with the following best-response correspondence:

$$\beta_{0,\sigma^*}(\sigma) = \begin{cases} [0,\ \sigma^*] & \text{if } \sigma = \sigma^* \\ 0 & \text{otherwise} \end{cases}.$$

Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{0,\sigma^*}((1-\epsilon)\ T(0, \sigma^*) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*, \sigma^*) + \epsilon\ \pi(\sigma_i, \sigma_e),$$

and

$$\Pi_e((1-\epsilon)\ T(0,\sigma^*)+\epsilon\ T(e)\mid \bar{b})=(1-\epsilon)\ \pi(\sigma_e,\sigma_i)+\epsilon\ \pi(\sigma_3,\sigma_3).$$

If $\sigma_e \neq \sigma^*$, then $\sigma_i = 0$. Hence, $\pi(\sigma_e,\sigma_i) = \sigma_e\ b + (1-\sigma_e)\ d \leq b \leq \pi^*$ with equality only if $\sigma_e = 1$. However, then $\pi(\sigma_i,\sigma_e) = c > \pi^*$ and the expected fitness of the incumbent would therefore still be higher than that of the entrant, whose expected payoff is π^* at most. If $\sigma_e \neq \sigma^*$ and $\sigma_e \neq 1$, then $\pi(\sigma_e,\sigma_i) < b \leq \pi^*$ and we can find a sufficiently small $\epsilon' > 0$ such that for all $\epsilon \in (0,\epsilon')$ we have $\Pi_{0,\sigma^*}(.) > \Pi_e(.)$. If $\sigma_e = \sigma^*$, then $\sigma_i \in [0,\sigma^*]$ and

$$\begin{aligned}\pi(\sigma_e,\sigma_i) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma_i)\ [\sigma^* b + (1-\sigma^*)d] \\ &\leq \pi^*\end{aligned}$$

and

$$\begin{aligned}\pi(\sigma_i,\sigma_e) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)b] + (1-\sigma_i)\ [\sigma^* c + (1-\sigma^*)d] \\ &\geq \pi^*\end{aligned}$$

for any $\sigma_i \in [0,\sigma^*]$ (with equalities for $\sigma_i = \sigma^*$). The expected fitness of the entrant never exceeds that of the incumbent. With this population the outcome induced by (σ^*,σ^*) is therefore stable.

The remaining cases are $b > c > \pi^* > a > d$ and $c > b > \pi^* > a > d$. In the following, define $\bar{\sigma}$ such that $\pi^* = \bar{\sigma}a + (1-\bar{\sigma})c$, i.e. $\bar{\sigma} = \frac{c-\pi^*}{c-a}$. First, we investigate the algebraic sign of $\bar{\sigma} - \sigma^*$. We have

$$\begin{aligned}&4(c-a)(b+c-a-d)(\bar{\sigma}-\sigma^*) \\ &\quad= 4(c-a)(b+c-a-d)\left(\frac{c-\pi^*}{c-a} - \frac{b+c-2d}{2(b+c-a-d)}\right) \\ &\quad= 4(c-d)(b+c-a-d) - (b+c-2d)^2 - 2(b+c-2d)(c-a) \\ &\quad= (c-b)(b+c-2a).\end{aligned}$$

As $c > a$, $b + c > a + d$ and $b + c > 2a$ it follows that $\bar{\sigma} > \sigma^*$ if $c > b$ and $\bar{\sigma} < \sigma^*$ if $b > c$.

Further define $\bar{\bar{\sigma}}$ such that $\pi^* = \bar{\bar{\sigma}}b + (1-\bar{\bar{\sigma}})d$, i.e. $\bar{\bar{\sigma}} = \frac{\pi^*-d}{b-d}$. Again, we investigate the algebraic sign of $\bar{\bar{\sigma}} - \bar{\sigma}$ and have

$$\begin{aligned}(c-a)(b-d)(\bar{\bar{\sigma}}-\bar{\sigma}) &= (\pi^*-d)(c-a) - (c-\pi^*)(b-d) \\ &= \frac{1}{4}(b-c)^2 > 0.\end{aligned}$$

As $c > a$ and $b > d$, we conclude that $\bar{\bar{\sigma}} > \bar{\sigma}$.

vii) $b > c > \pi^* > a > d$*:* Consider a monomorphic population of types $T(0, \bar{\sigma}, \sigma^*, 1)$ with the following best-response correspondence:

$$\beta_{0,\bar{\sigma},\sigma^*,1}(\sigma) = \begin{cases} 1 & \text{if } \sigma > \bar{\sigma} \text{ and } \sigma \neq \sigma^* \\ [\sigma^*,\, 1] & \text{if } \sigma = \sigma^* \\ [0,1] & \text{if } \sigma = \bar{\sigma} \\ 0 & \text{if } \sigma < \bar{\sigma} \end{cases} .$$

As $b > c$ we know from above that $\sigma^* > \bar{\sigma}$. Hence, there is a type with this best-response correspondence in our type space $\mathcal{T}$. For, $\beta(\sigma)$ is non-empty and convex for all $\sigma \in [0,1]$, we have $\sigma^* \in \beta(\sigma^*)$, and the closed graph criterion is also satisfied (β is upper hemi-continuous). Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{0,\bar{\sigma},\sigma^*,1}((1-\epsilon)\ T(0,\bar{\sigma},\sigma^*,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*,\sigma^*) + \epsilon\ \pi(\sigma_i,\sigma_e),$$

and

$$\Pi_e((1-\epsilon)\ T(0,\bar{\sigma},\sigma^*,1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e,\sigma_i) + \epsilon\ \pi(\sigma_3,\sigma_3).$$

If $\sigma_e < \bar{\sigma}$, then $\sigma_i = 0$. Hence,

$$\begin{aligned} \pi(\sigma_e,\sigma_i) &= \sigma_e\ b + (1-\sigma_e)\ d \\ &< \bar{\sigma}\ b + (1-\bar{\sigma})\ d \\ &< \bar{\bar{\sigma}}\ b + (1-\bar{\bar{\sigma}})\ d \\ &= \pi^*. \end{aligned}$$

If $\sigma_e > \bar{\sigma}$ and $\sigma_e \neq \sigma^*$, then $\sigma_i = 1$. Hence,

$$\begin{aligned} \pi(\sigma_e,\sigma_i) &= \sigma_e\ a + (1-\sigma_e)\ c \\ &< \bar{\sigma}\ a + (1-\bar{\sigma})\ c \\ &= \pi^*. \end{aligned}$$

In either case, we can find a sufficiently small $\epsilon' > 0$ such that for all $\epsilon \in (0, \epsilon')$ we have $\Pi_{0,\bar{\sigma},\sigma^*,1}(.) > \Pi_e(.)$. If $\sigma_e = \sigma^*$, then $\sigma_i \in [\sigma^*, 1]$ and we have

$$\begin{aligned} \pi(\sigma_e,\sigma_i) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)c] + (1-\sigma_i)\ [\sigma^* b + (1-\sigma^*)d] \\ &\leq \pi^* \end{aligned}$$

and

$$\begin{aligned} \pi(\sigma_i,\sigma_e) &= \sigma_i\ [\sigma^* a + (1-\sigma^*)b] + (1-\sigma_i)\ [\sigma^* c + (1-\sigma^*)d] \\ &\geq \pi^* \end{aligned}$$

for any $\sigma_i \in [\sigma^*, 1]$ (with equalities for $\sigma_i = \sigma^*$). The expected fitness of the entrant cannot be higher than that of the incumbent. If $\sigma_e = \bar{\sigma}$, then $\sigma_i \in [0, 1]$ and we have

$$\begin{aligned}
\pi(\sigma_e, \sigma_i) &= \pi(\bar{\sigma}, \sigma_i) \\
&= \sigma_i\ [\bar{\sigma}a + (1-\bar{\sigma})c] + (1-\sigma_i)\ [\bar{\sigma}b + (1-\bar{\sigma})d] \\
&= \sigma_i\ \pi^* + (1-\sigma_i)\ [\bar{\sigma}b + (1-\bar{\sigma})d] \\
&\leq \sigma_i\ \pi^* + (1-\sigma_i)\ [\bar{\bar{\sigma}}b + (1-\bar{\bar{\sigma}})d] \\
&= \pi^*
\end{aligned}$$

with equality if and only if $\sigma_i = 1$. However, then we have

$$\begin{aligned}
\pi(\sigma_i, \sigma_e) &= \bar{\sigma}a + (1-\bar{\sigma})b \\
&> \bar{\sigma}a + (1-\bar{\sigma})c \\
&= \pi^*.
\end{aligned}$$

The expected fitness of the entrant never exceeds that of the incumbent. With this population the outcome induced by (σ^*, σ^*) is therefore stable.

viii) $c > b > \pi^* > a > d$*:* Consider a monomorphic population of types $T(0, \sigma^*, \bar{\sigma}, 1)$ with the following best-response correspondence:

$$\beta_{0,\sigma^*,\bar{\sigma},1}(\sigma) = \begin{cases} 1 & \text{if } \sigma > \bar{\sigma} + \eta \\ [0,1] & \text{if } \sigma = \bar{\sigma} + \eta \\ [0,\ \sigma^*] & \text{if } \sigma = \sigma^* \\ 0 & \text{if } \sigma < \bar{\sigma} + \eta \text{ and } \sigma \neq \sigma^* \end{cases},$$

where $\eta \equiv \frac{\bar{\bar{\sigma}} - \bar{\sigma}}{2} > 0$. As $c > b$ we know from above that $\sigma^* < \bar{\sigma} < \bar{\sigma} + \eta$. Hence, there is a type with this best-response correspondence in our type space. For, $\beta(\sigma)$ is non-empty and convex for all $\sigma \in [0, 1]$, we have $\sigma^* \in \beta(\sigma^*)$, and the closed graph criterion is also satisfied (β is upper hemi-continuous). Consider any entrant in the population such that in $\bar{b}$ the equilibrium which is played between an entrant and an incumbent is (σ_e, σ_i) and the equilibrium between two entrants is (σ_3, σ_3). The expected fitnesses of the incumbent and the entrant are, respectively,

$$\Pi_{0,\sigma^*,\bar{\sigma},1}((1-\epsilon)\ T(0, \sigma^*, \bar{\sigma}, 1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma^*, \sigma^*) + \epsilon\ \pi(\sigma_i, \sigma_e),$$

and

$$\Pi_e((1-\epsilon)\ T(0, \sigma^*, \bar{\sigma}, 1) + \epsilon\ T(e) \mid \bar{b}) = (1-\epsilon)\ \pi(\sigma_e, \sigma_i) + \epsilon\ \pi(\sigma_3, \sigma_3).$$

If $\sigma_e < \bar{\sigma} + \eta$ and $\sigma_e \neq \sigma^*$, then $\sigma_i = 0$. Hence,

$$\begin{aligned}
\pi(\sigma_e, \sigma_i) &= \sigma_e\ b + (1-\sigma_e)\ d \\
&< (\bar{\sigma} + \eta)\ b + (1 - \bar{\sigma} - \eta)\ d \\
&< \bar{\bar{\sigma}}\ b + (1 - \bar{\bar{\sigma}})\ d \\
&= \pi^*.
\end{aligned}$$

If $\sigma_e > \bar{\sigma} + \eta$, then $\sigma_i = 1$. Hence,

$$\begin{aligned} \pi(\sigma_e, \sigma_i) &= \sigma_e\ a + (1 - \sigma_e)\ c \\ &< \bar{\sigma}\ a + (1 - \bar{\sigma})\ c \\ &= \pi^*. \end{aligned}$$

If $\sigma_e = \bar{\sigma} + \eta$, then $\sigma_i \in [0, 1]$ and we have

$$\begin{aligned} \pi(\sigma_e, \sigma_i) &= \pi(\bar{\sigma} + \eta, \sigma_i) \\ &= \sigma_i\ [(\bar{\sigma} + \eta)a + (1 - \bar{\sigma} - \eta)c] + (1 - \sigma_i)\ [(\bar{\sigma} + \eta)b + (1 - \bar{\sigma} - \eta)d] \\ &< \sigma_i\ [\bar{\sigma}a + (1 - \bar{\sigma})c] + (1 - \sigma_i)\ [\bar{\bar{\sigma}}b + (1 - \bar{\bar{\sigma}})d] \\ &= \pi^*. \end{aligned}$$

In either case, we can find a sufficiently small $\epsilon' > 0$ such that for all $\epsilon \in (0, \epsilon')$ we have $\Pi_{0,\sigma^*,\bar{\sigma},1}(.) > \Pi_e(.)$. If $\sigma_e = \sigma^*$, then $\sigma_i \in [0, \sigma^*]$ and we have

$$\begin{aligned} \pi(\sigma_e, \sigma_i) &= \sigma_i\ [\sigma^* a + (1 - \sigma^*)c] + (1 - \sigma_i)\ [\sigma^* b + (1 - \sigma^*)d] \\ &\leq \pi^* \end{aligned}$$

and

$$\begin{aligned} \pi(\sigma_i, \sigma_e) &= \sigma_i\ [\sigma^* a + (1 - \sigma^*)b] + (1 - \sigma_i)\ [\sigma^* c + (1 - \sigma^*)d] \\ &\geq \pi^* \end{aligned}$$

for any $\sigma_i \in [0, \sigma^*]$ (with equalities for $\sigma_i = \sigma^*$). The expected fitness of the entrant never exceeds that of the incumbent. With this population the outcome induced by (σ^*, σ^*) is therefore stable. $\square$

References

[1] Camerer, C. F. and T.-H. Ho (1994): Violations of the Betweenness Axiom and Nonlinearity in Probability, *Journal of Risk and Uncertainty*, 8, pp. 167-196.

[2] Charness, G. and M. Rabin (2002): Understanding Social Preferences with Simple Tests, *Quarterly Journal of Economics*, 117(3), pp. 817-869.

[3] Dekel, E.; J. C. Ely and O. Yilankaya (2004): Evolution of Preferences, *mimeo*, Northwestern University.

[4] Ely, J. C. and O. Yilankaya (2001): Nash Equilibrium and the Evolution of Preferences, *Journal of Economic Theory*, 97, pp. 255-272.

[5] Fehr, E. and K. Schmidt (2000): Theories of Fairness and Reciprocity - Evidence and Economic Applications, (paper prepared for the invited session of the *8th World Congress of the Econometric Society*).

[6] Frank, R. H. (1988): *Passions within Reason - The Strategic Role of the Emotions*, New York: Norton.

[7] Güth, W. (1995): An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives, *International Journal of Game Theory*, 24, pp. 323-344.

[8] Güth, W. and M. E. Yaari (1992): An evolutionary approach to explain reciprocal behavior in a simple strategic game, in *Explaining Process and Change: Approaches in Evolutionary Economics*, ed. by U. Witt, Ann Arbor: The University of Michigan Press, pp. 23-34.

[9] Kahneman, D. and A. Tversky (1979): Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, 47:2, pp. 263-291.

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

[11] MacCrimmon, K. R. and S. Larsson (1979): Utility Theory: Axioms vs. 'Paradoxes', in *Expected Utility Hypotheses and the Allais Paradox*, ed. by M. Allais and O. Hagen, Dordrecht, Holland: D. Reidel Publishing Company, pp. 333-409.

[12] von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

[13] Osborne, M. J. and A. Rubinstein (1994): *A Course in Game Theory*, Cambridge, MA and London: MIT Press.

[14] Ritzberger, K. (1996): On Games under Expected Utility with Rank Dependent Probabilities, *Theory and Decision*, 40, 1-27.

[15] Röell, A. (1987): Risk Aversion in Quiggin and Yaari's Rank-Order Model of Choice under Uncertainty, *The Economic Journal*, 97 (Suppl), pp. 143-159.

[16] Quiggin, J. (1982): A Theory of Anticipated Utility, *Journal of Economic Behaviour and Organization*, 3, pp. 323-343.

[17] Schoemaker, P. (1982): The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations, *Journal of Economic Literature*, 20:2, pp. 529-563.

[18] Tversky, A. and D. Kahneman (1992): Advances in Prospect Theory: Cumulative Representation of Uncertainty, *Journal of Risk and Uncertainty*, 5:4, pp. 297-323.

[19] Wakker, P. P. (1994): Separating Marginal Utility and Probabilistic Risk Aversion, *Theory and Decision*, 36, pp. 1-44.

[20] Weibull, J. W. (1995): *Evolutionary Game Theory*, Cambridge, MA and London: MIT Press.

[21] Yaari, M. E. (1987): The Dual Theory of Choice under Risk, *Econometrica*, 55, pp. 95-115.