Games and Economic Behavior 67 (2009) 542–551

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Evolutionary stability of discrimination under observability ☆

Florian Herold *, Christoph Kuzmics

Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL, United States

ARTICLE INFO

Article history:
Received 23 September 2008
Available online 5 April 2009

JEL classification:
C72
C73

Keywords:
Evolution of preferences
Observability
Discrimination
Stability

ABSTRACT

We study the evolution of preferences under perfect and almost perfect observability in symmetric 2-player games. We demonstrate that if nature can choose from a sufficiently general preference space, which includes preferences over outcomes that may depend on the opponent's preference-type, then, in most games, only discriminating preferences (treating different types of opponents differently in the same situation) can be evolutionary stable and some discriminating types are stable in a very strong sense in all games. We use these discriminating types to show that any symmetric outcome which gives players more than their minmax value in material payoffs (fitness) can be seen as equilibrium play of a player population with such strongly stable preferences.

1. Introduction

The literature on the evolution of preferences following the "indirect evolutionary approach" by Güth and Yaari (1992) and Güth (1995), finds, under the assumption of at least partial observability, evolutionary rationales for certain non-materialistic preferences such as altruistic, spiteful, or reciprocal preferences.[1] Dekel et al. (2007) highlight the role of restricting nature's choices to certain subsets of preferences in these models and show that only efficient outcomes can be supported by evolutionary stable preferences if nature can choose among all possible preferences over outcomes. It is thus key for an evolutionary analysis of preferences to allow for all possible preferences unless we have good reason, e.g. due to some biological constraints, to restrict preferences in a certain way. However, preferences over outcomes do not encompass all preferences that are relevant in economics. Recent experiments[2] suggest that individuals when playing a game do not

☆ There is a more comprehensive working paper version of this paper, which we will refer to repeatedly in the text as Herold and Kuzmics (2008). An older version of this paper has previously been circulated under the title of "Evolution of preferences under perfect observability: Almost anything is stable". We thank an anonymous advisory editor, two anonymous referees, Eddie Dekel, Jeff Ely, Okan Yilankaya, and participants at the workshop 'The Biological Basis of Economics', at 'Games 2008', and at 'ESEM 2008' for helpful comments and suggestions.

* Corresponding author.
E-mail addresses: f-herold@kellogg.northwestern.edu (F. Herold), c-kuzmics@kellogg.northwestern.edu (C. Kuzmics).

[1] For results of this nature see, e.g., Bester and Güth (1998), Koçkesen et al. (2000b), Koçkesen et al. (2000a), and Heifetz et al. (2007). If opponents' preferences are not observable (and players are selected randomly from a large population) evolutionary forces favor preferences, which coincide with the material payoff (evolutionary fitness) (see, e.g., Ok and Vega-Redondo, 2001) or at least lead to equilibrium play "as if" players were purely motivated by their fitness (see, e.g., Ely and Yilankaya, 2001).

[2] Charness and Levine (2007), in a lab-experiment, study "worker" responses to wages, where offered wages are composed of an intended wage choice by the "firm" and a random component. They show that workers react to the same wage offer differently if that wage is the sum of an observed low intended wage choice by the firm and a lucky random draw on the one hand or the sum of an observed high intended wage choice by the firm and an unlucky random draw on the other. Falk et al. (2003), study subjects' responses to offers in different ultimatum games where the games differ in the commonly known set of offers available to the proposers. They show that responders' responses to the same offer are typically highly sensitive to the proposers' set of available offers.

0899-8256/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.geb.2009.03.010

only care about (the distribution over) outcomes. They seem to care about their opponent's intentions (i.e. preferences) and discriminate between different types of opponents.[3] Also some important models of interdependent preferences consider preferences that depend on one's opponent's preferences. Levine (1998) introduces a class of such preferences, which are such that one individual's degree of altruism and spitefulness towards an opponent depends on the degree of altruism and spitefulness of this opponent. Levine (1998) models these preferences in a well-defined manner and finds that this specification is consistent with non-materialistic behavior in several experiments. How generally to model such truly interdependent preferences is discussed in Gul and Pesendorfer (2007). Sethi and Somanathan (2001) then subject a modified class of the preferences given in Levine (1998) to an evolutionary analysis. They provide conditions under which purely selfish preferences are not evolutionary stable, as well as conditions under which specific reciprocal types of preferences are stable. Thus preferences which depend on one's opponent's preference type are important from an experimental as well as from a theoretical perspective.

Once preferences that depend on the opponent's preferences are introduced, and for good reasons given the above quoted evidence, it is natural and important to consider all such preferences. In this paper we consider the evolution of preferences in normal form games under perfect and almost perfect observability, where nature can choose from any sufficiently general class of such preferences. Results change dramatically. We argue that not only selfish (or materialistic) preferences are not stable but also general non-discriminating preferences are typically not evolutionary stable except perhaps in a very weak form, while certain preferences which exhibit discrimination are always (in all games) evolutionary stable in a very strong sense. Discrimination is thus the key force in the evolution of preferences under observability, and an analysis of preference evolution needs to consider these discriminating preference types. This result is true for all sets of preference types[4] nature can choose from as long as certain specific (and simple) types are available to nature.

While we thus find a strong case for evolution to lead to discriminating behavior, the indirect evolutionary approach will not allow to predict the actual outcome of play under observability. In fact, under perfect and almost perfect observability any symmetric outcome with individuals obtaining more than their minmax fitness can be seen as equilibrium play of a population with strongly stable preferences. Thus the existing results in the literature on the evolution of preferences under observability change drastically if we extend the space of possible preferences by allowing also for discriminating preferences.[5] Also, there is no additional extension of the space of preferences that could override our results. To make our point as clearly as possible, we use the model of preference evolution almost exactly as in Dekel et al. (2007) with the single change that nature can choose preferences which directly depend on opponents' types. This and only this accounts for the difference in results.

The intuition behind the difference between the two results can be explained using Robson's (1990) metaphor of a secret hand-shake. Consider a population of incumbents who are all of the same preference type playing some inefficient outcome. Add a small fraction of mutants entering this population. These mutants come with a new "signal", the secret handshake, which helps mutants to identify each other but goes unnoticed by the incumbent. When two mutants meet they recognize each other by means of their secret hand-shake and can play a high-payoff strategy. When an entrant meets an incumbent the entrant tries her secret hand-shake, but does not receive one back. The incumbent does not realize she is facing an entrant and does not change her behavior. The entrant does recognize the incumbent and now behaves as an incumbent as well. Thus the entrant receives a higher payoff and successfully invades.

The argument in Dekel et al. (2007) is similar, although the hand-shake is not really secret. While the incumbent does realize when she is facing an entrant (given the almost perfect observability of preferences), she does not change her behavior (much) in the relevant equilibrium, because she has preferences only over outcomes. In our setting, in contrast, preferences can be dependent on the opponent's type, and, because of the almost perfect observability of preferences (the hand-shake is actually not secret) the incumbent's behavior can change drastically when she faces an entrant. In particular, the incumbent can play spitefully against entrants and keep their fitness arbitrarily close to their minmax value. Thus a highly discriminating type, in an environment of almost perfect observability of preferences, can be stable even if that type does play an inefficient outcome.

To understand the gist of our argument consider the prisoners' dilemma, in which the efficient material outcome is mutual cooperation, while individually each player would maximize his fitness by playing defection. Consider a monomorphic population of incumbents who play an inefficient strictly mixed strategy against each other. In Dekel et al. (2007) these must be induced by some preferences over outcomes only. Suppose a small fraction of mutants enters this population. Suppose further that these mutants have coordination game preferences, such that they are just indifferent against the mixed strategy the incumbents play against each other. Thus these mutants when meeting an incumbent may play the same

[3] This is not the only "explanation" of the behavior in these experiments. For instance, it could be due to subjects punishing what they perceive as non-cooperative behavior. In a repeated version of the game this could be supported as equilibrium behavior and subjects might (mistakenly) act in the one shot game in the lab as if they were playing in a repeated game.

[4] In Herold and Kuzmics (2008) we demonstrate that these sets of preference types nature can choose from can be taken to be "valid" in the sense of Gul and Pesendorfer (2007). This means they can be taken to be isomorphic to a component of Gul and Pesendorfer's (2007) "canonical space of behavioral types" by their Theorem 2.

[5] Notice though, that our paper does reinforce the message of Dekel et al. (2007)'s methodological contributions: the results in the literature on the evolution of preferences depend crucially on the assumptions made on the degree of observability and on restricting attention to a subset of possible preferences.

mixed strategy as the incumbents do, but may cooperate against each other. Given that the incumbents' preferences are over outcomes only, it is one equilibrium behavior for incumbents, when meeting an entrant, to continue playing their mixed strategy. Thus, if incumbents play something (materially) inefficient, and have preferences over outcomes only, entrants can obtain higher material payoffs than incumbents and can successfully invade.

In this paper incumbents' preferences can depend on their opponents' preferences. One possible incumbent type is thus one who behaves in the same way as the incumbents above when meeting another incumbent, but plays defect against any non-incumbent, driven by the preference to minimize the material payoff of any mutant preference type. Upon entry, any mutant will thus receive a very low material payoff, arbitrarily close to his minmax payoff, while the expected payoff to incumbents hardly changes from the payoff he receives against his own type, which is strictly above the minmax value. Any mutant will thus quickly be driven to extinction. We thus obtain the very different result that every (symmetric) individually rational outcome[6] can be sustained by evolutionary stable (discriminating) preferences.[7]

Further literature. Discrimination, the driving force in this paper, also plays a role in Banerjee and Weibull (2000)'s analysis[8] of neutrally stable strategies in symmetric 2-player games in which players before playing the game send payoff-irrelevant messages, which can be interpreted as observable traits. Their neutrally stable outcomes, however, must lie in the convex hull of the base game Nash equilibrium payoffs, which is typically a much smaller set than the set of individually rational outcomes.

Finally, our argument has an interesting connection to the commitment-device folk theorem by Kalai et al. (2007).[9] In their model players can choose a commitment device that can condition on the commitment device chosen by their opponent. They find a commitment device folk theorem: every individually rational (correlated) strategy in a basic two player game G can be obtained as a (Nash equilibrium) of an extended commitment game. In essence this equilibrium of the commitment device game takes the following form: in equilibrium every player is supposed to choose a particular commitment device that plays the equilibrium strategy if everybody did choose his 'assigned' commitment device and minmaxes a player otherwise.

2. The model

The environment. We will use notation as closely as possible to that in Dekel et al. (2007), hereafter DEY, to facilitate a comparison. Let G be a symmetric 2-player game with finite action set $A = \{a_1, \ldots, a_n\}$ and (material) payoff function $\pi : A \times A \to \mathbb{R}$, which can be extended (by taking expectations) to the set of all mixed strategies Δ. Without loss of generality we will assume that payoffs π are between 0 and 1. Sometimes we use matrix notation. Let M denote the matrix of material payoffs, with entries all in $[0,1]$. I.e. for all $\sigma, \tau \in \Delta$ we have that $\pi(\sigma,\tau) = \sigma M \tau$. These material payoffs π represent fitness or evolutionary success and regulate the future occurrence of each preference type. Players can differ in their (subjective) preferences over outcomes. In particular, subjective preferences may differ from the material payoffs. Preferences determine players' strategies, strategies in turn determine outcomes, the material payoffs of each type, and thereby the evolutionary success.

A preference type in DEY is a function over outcomes in $A \times A$ into the real line. The set of all such preference types can be represented by the set $[0,1]^{n^2}$ (modulus affine transformations). Here we make our key departure from DEY. We extend their set of preference types and allow preferences to depend additionally on the opponent's preference type. The following approach of modeling preferences that condition on the opponent's preferences avoids any potential inconsistencies.[10]

Let Θ be a set of types. At this point this can be anything, later we will see that each type $\theta \in \Theta$ corresponds to a certain preference type. Now consider a function $u : \Theta \times \Theta \times A \times A \to [0,1]$. Again, at this point this can be any function. This function u induces a function $u_\theta : \Theta \times A \times A \to [0,1]$ for every $\theta \in \Theta$. We interpret u_θ as the preference-function of a type θ. Note that by assuming u_θ is constant in its first argument we could replicate all preference types of DEY. By allowing u_θ to vary also in its first argument, we permit that an individual's preferences over outcomes can depend on the type of the opponent.

[6] The original realm of this type of folk theorems are infinitely repeated games with very patient players. Some papers, e.g. Fudenberg and Maskin (1990) and Binmore and Samuelson (1992), employ concepts of evolutionary stability to refine the set of equilibria. Notice that the setting as well as the purpose of these papers is very different from ours.

[7] While we assume that (almost) perfect observability means that when an incumbent faces a mutant there is probability close to 1 that she in fact recognizes the exact type of her opponent, this is not always necessary for our result to go through. In the prisoners' dilemma, for instance, as long as the incumbent knows that the opponent is of any non-incumbent type, even though she may not know which, she can always simply play defect against such an opponent. This makes it impossible for the entrant to do well, which in turn makes it possible to stabilize essentially any outcome.

[8] The title of their original working paper, Banerjee and Weibull (1993), makes this explicit: 'Evolutionary selection with discriminating players'.

[9] In fact, our definition of a preference space, which guarantees the consistency of the interdependent preferences, was originally inspired by their definition of a device space.

[10] A 'naive' approach, in which the utility of a player can depend directly on the opponent's utility without further restrictions, could lead to circular statements and ill-defined preferences. Consider, for example, one player who is spiteful towards a second player and receives a utility of $+1$ if his counterpart receives a negative utility and -1 if his counterpart receives a nonnegative utility. The second player is altruistic and receives a positive utility of $+1$ if and only if his counterpart receives a positive utility. Then the spiteful player 1 is happy only if he is unhappy. Note that our definitions of a preference space makes sure that such problems cannot arise.

Let $\mathcal{U} = \{u_\theta : \Theta \times A \times A \to [0,1]\}$ denote the set of all preferences induced by u. If we assume, without loss of generality, that $u_\theta \neq u_{\theta'}$ for any $\theta, \theta' \in \Theta$, we then have a bijection between Θ and $\mathcal{U}$. Hence, Θ can again be thought of as the set of all preference types. Note that we could also assume, as in von Widekind (2004), that preferences are not necessarily of expected utility form, i.e. we could have $u : \Theta \times \Theta \times \Delta \times \Delta \to [0,1]$. This would generate even more preferences, but would not change our main results (see Herold and Kuzmics, 2008).

Definition 1 *(Preference space).* A space of preferences of G is a pair $(\Theta, u : \Theta \times \Theta \times A \times A \to [0,1])$. Θ is a nonempty set of possible preference types. We interpret $u_{\theta_1}(\theta_2, a_1, a_2) \equiv u(\theta_1, \theta_2, a_1, a_2) \in [0,1]$ as the subjective utility of a player of type θ_1 playing a_1 if he plays against an opponent of type θ_2 who plays a_2.

As in DEY, individuals observe the opponent's type (perfectly) with probability $p \in [0,1]$, while with remaining probability $1 - p$ an individual observes the uninformative signal ϕ.

The solution concept. The main point of this paper is to show that with a rich enough set of preference types any symmetric outcome above the minmax material payoff is stable. Hence, using a more demanding stability concept strengthens our results. We use an extremely demanding stability concept, that we call strong stability. In particular, strong stability of an outcome (as defined below) implies stability of that outcome according to the definition of DEY. Conveniently, many things will simplify.

Let $\mathcal{P}(\Theta)$ denote the set of all finite support probability distributions on Θ.[11] Let $\mu \in \mathcal{P}(\Theta)$. Let, as in DEY, $\Gamma_p(\mu)$ denote the Bayesian game in which nature first draws two types independently according to μ and then each individual independently observes the other's type with probability $p \in [0,1]$, while with probability $1-p$ a player observes the uninformative signal ϕ. Let $\Gamma(\mu)$ denote the complete information game corresponding to $p = 1$.

A strategy for preference type θ is a function $b_\theta : C(\mu) \cup \{\phi\} \to \Delta$, where $C(\mu)$ denotes the support of μ. Let $u_\theta(\theta', \sigma, \tau)$ denote the expected subjective utility a player with preference type θ receives when playing mixed strategy $\sigma \in \Delta$ against the observed type θ' who plays $\tau \in \Delta$. Let b denote the profile of all b_θ-functions. The profile b is an *equilibrium profile* if, for every $\theta, \theta' \in C(\mu)$:

$$b_\theta(\theta') \in \arg\max_{\sigma \in \Delta} \bigl(p u_\theta\bigl(\theta', \sigma, b_{\theta'}(\theta)\bigr) + (1-p) u_\theta\bigl(\theta', \sigma, b_{\theta'}(\phi)\bigr)\bigr),$$

and

$$b_\theta(\phi) \in \arg\max_{\sigma \in \Delta} \mathbb{E}_{\theta' \sim \mu}\bigl[p u_\theta\bigl(\theta', \sigma, b_{\theta'}(\theta)\bigr) + (1-p) u_\theta\bigl(\theta', \sigma, b_{\theta'}(\phi)\bigr)\bigr].$$

Let $B_p(\mu)$ denote the set of all such equilibrium profiles in $\Gamma_p(\mu)$. Let $\Pi_\theta(\mu|b)$ denote the expected material fitness of preference type $\theta \in C(\mu)$ given the distribution of types μ and the equilibrium profile $b \in B_p(\mu)$, i.e. as in DEY,

$$\begin{aligned}\Pi_\theta(\mu|b) = \sum_{\theta' \in C(\mu)} &\bigl[p^2 \pi\bigl(b_\theta(\theta'), b_{\theta'}(\theta)\bigr) + p(1-p)\pi\bigl(b_\theta(\theta'), b_{\theta'}(\phi)\bigr) \\ &+ p(1-p)\pi\bigl(b_\theta(\phi), b_{\theta'}(\theta)\bigr) + (1-p)^2 \pi\bigl(b_\theta(\phi), b_{\theta'}(\phi)\bigr)\bigr]\mu(\theta'),\end{aligned} \tag{1}$$

or in matrix notation

$$\begin{aligned}\Pi_\theta(\mu|b) = \sum_{\theta' \in C(\mu)} &\bigl[p^2 b_\theta(\theta') M b_{\theta'}(\theta) + p(1-p) b_\theta(\theta') M b_{\theta'}(\phi) \\ &+ p(1-p) b_\theta(\phi) M b_{\theta'}(\theta) + (1-p)^2 b_\theta(\phi) M b_{\theta'}(\phi)\bigr]\mu(\theta').\end{aligned}$$

For a configuration (μ, b) let $x(\mu, b)$ be the induced probability distribution over actions $A \times A$. Let $\mu \in \mathcal{P}(\Theta)$ be the incumbent preference distribution. Let $\mu' \in \mathcal{P}(\Theta)$ be a distribution over entering mutant preferences. Suppose that altogether this μ' distribution invades with a small fraction $\epsilon > 0$. The post-entry distribution of preferences is then given by $\tilde{\mu}_\epsilon = (1-\epsilon)\mu + \epsilon\mu'$. For a given configuration (μ, b), a parameter δ, and a post-entry population $\tilde{\mu}$ the *set of nearby equilibria* is given by

$$B_p^\delta(\tilde{\mu}|b) = \bigl\{\tilde{b} \in B_p(\tilde{\mu}): \bigl|x(\tilde{b}, \tilde{\mu}) - x(b, \mu)\bigr| < \delta\bigr\}.$$

Definition 2 *(Strong stability).* A configuration (μ, b) is *strongly stable* if there exists a preference type $\theta \in \Theta$ such that:

1. $\mu = \mu_\theta$, where μ_θ is the Dirac distribution on θ, and b is the unique equilibrium of $\Gamma_p(\mu_\theta)$, i.e. $B_p(\mu_\theta) = \{b\}$.
2. for every $\delta > 0$ there is an $\bar{\epsilon} > 0$ such that for every $\epsilon \in (0, \bar{\epsilon})$ and for every $\mu' \in \mathcal{P}(\Theta)$ we have $B_p(\tilde{\mu}_\epsilon) = B_p^\delta(\tilde{\mu}_\epsilon|b) \neq \emptyset$ and $\Pi_\theta(\tilde{\mu}_\epsilon|b') > \Pi_{\theta'}(\tilde{\mu}_\epsilon|b')$ for every $\theta' \neq \theta$ with $\theta' \in C(\mu')$ and for every $b' \in B_p(\tilde{\mu}_\epsilon)$, where $\tilde{\mu}_\epsilon = (1-\epsilon)\mu_\theta + \epsilon\mu'$.

[11] We conjecture that it is not necessary to restrict attention to distributions over types with finite support.

An *outcome* x is *strongly stable* if there exists a strongly stable configuration with that outcome, i.e. there exists a strongly stable (μ, b) with $x = x(\mu, b)$.

In other words, we call an outcome and its supporting configuration strongly stable if and only if (1) it is induced by a configuration which consists of a single incumbent preference-type, (2) with a strategy which is the unique equilibrium given the game induced by this single type, (3) such that for any small fraction of entering mutant preference types there always exists an equilibrium, (4) while all resulting equilibria remain nearby, (5) and in all these equilibria the incumbent preference-type receives a strictly higher material payoff than any other type in the post-entry configuration. This definition may seem too demanding. Yet we can prove our main result for strong stability. As strong stability of an outcome implies stability of an outcome in the sense of DEY[12] this only strengthens our results.

3. On the (in-)stability of non-discriminating preferences

In this section we argue that a configuration supported only by non-discriminating preferences is not stable against the invasion of discriminating preference types except perhaps in a very weak sense. An evolutionary analysis of preferences, thus, needs to consider discriminating preferences lest it ignores an important evolutionary force.

Firstly, in any game there is certainly no non-discriminating preference type which supports a strongly stable configuration. Consider any non-discriminating incumbent type and a discriminating preference type whose utility over outcomes is identical to the incumbent type's whenever faced with either the incumbent type or his own type but whose utilities over outcomes differ when facing a third preference type that is not yet present in the support of the population. This discriminating type does equally well as the incumbent in contradiction to strong stability. While we will show that every outcome with individually rational material payoffs is strongly stable (supported by discriminating preferences) there are, thus, no non-discriminating preferences that support any outcome as strongly stable, if embedded in the richer space of possibly discriminating preferences.

Strong stability is a very demanding stability concept. This strengthens our existence results in the next section but makes the above non-existence of strongly stable non-discriminating preferences a weak result. In an effort to strengthen this result we provide an argument that for a broad class of games non-discriminating preferences are not immune to "evolutionary drift",[13] which will ultimately lead far away from these preferences (while there is no such drift back to these preferences). More precisely, let Θ be a sufficiently large preference type space, e.g. the canonical one of Gul and Pesendorfer (2007). Let $\mu \in \mathcal{P}(\Theta)$ be a given probability distribution over types with finite support with the interpretation that μ is a vector of proportions of preference types in Θ. Let the support of μ be denoted by $\Theta^{\mu} = C(\mu)$.

We call a configuration (μ, b) *not robust to drift* if it is either not stable in the sense of DEY or if there is a finite set of types $\widehat{\Theta} \subset \Theta$ with $\Theta^{\mu} \subset \widehat{\Theta}$ such that for all $\delta > 0$ there is a pair of functions $f : [0,1] \to \mathcal{P}(\widehat{\Theta})$, with $f(0) = \mu$ and f continuous, and $g : [0,1] \to \Delta^{|\widehat{\Theta}|}$, with $g(0) = b$ and $g(t) \in B_p^{\delta}(f(t)|b)$ for all $t \in [0,1]$, such that $(f(t), g(t))$ is balanced (see Dekel et al. (2007); i.e. all types in the support of $f(t)$ obtain the exact same material payoff given behavior $g(t)$), and finally $f(1)$ is not DEY-stable. Thus, a configuration is not robust to drift if there are mutant preference types which can gradually and persistently enter the population until eventually a configuration is reached which is not DEY-stable, while on this gradual path the mutants always earn the same fitness as the incumbents.

We here focus on games in which there is a single highest material payoff, which is strictly above the efficient outcome. To make this more precise consider any (symmetric 2-player) game G with action space A, mixed action space Δ, and material payoffs π. Define $\pi^{e}(G) = \max_{\sigma \in \Delta} \pi(\sigma, \sigma)$ as the efficient payoff, and $\pi^{*}(G) = \max_{\sigma, \tau \in \Delta} \pi(\sigma, \tau)$ as the highest possible payoff. Let $\mathcal{G}$ then be the class of games G for which $\pi^{e}(G) < \pi^{*}(G)$. Note that by definition we must have $\pi^{e}(G) \leqslant \pi^{*}(G)$. Thus we are only ruling out games in which $\pi^{e}(G) = \pi^{*}(G)$ such as coordination games. Thus, this broad class of games includes, in particular, any game in which the efficient outcome is not a Nash equilibrium (in material payoffs), such as the prisoners' dilemma. Also we construct just one (very simple) path evolutionary drift can take to illustrate our point. Often other paths would serve the same purpose.

If the game is such that there are no DEY-stable non-discriminating preferences, then our point is already made. Consider, thus, a game $G \in \mathcal{G}$ and a DEY-stable configuration (μ, b) with finite support of μ given by Θ^{μ} and all types $\theta \in \Theta^{\mu}$ have preferences over outcomes only. Then, we know from Dekel et al. (2007) that under observability efficiency of the induced outcome is a necessary condition for DEY-stability within the space of non-discriminating preferences over outcomes. Thus, efficiency of the induced outcome must also be necessary for the stability of non-discriminating preferences if we enlarge the space of potential entrants to include discriminating preference types. In such a configuration incumbents earn a material payoff of $\pi^{e}(G)$. Now we construct a continuous path along which entrants earn the same payoff as incumbents and which is such that it leads to a configuration in which a new entrant earns strictly more than all incumbents. Say there are k types in Θ^{μ}. Then let $\widehat{\Theta}$ consist of $2k + 1$ preference types, all types in Θ^{μ} plus another $k + 1$ types. In fact, let us write $\widehat{\Theta} = \Theta^{\mu} \cup \Theta^{\mu,*} \cup \{\theta^{*}\}$ as follows. For every $\theta \in \Theta^{\mu}$ let there be a $\theta' \in \Theta^{\mu,*}$ with the property that θ' shares the same preferences as θ when facing any type in $\Theta^{\mu} \cup \Theta^{\mu,*}$, yet θ' "loves" preference type θ^{*}, where "love" means that her

[12] See Herold and Kuzmics (2008) for details.

[13] See, e.g., Binmore and Samuelson (1999) for a discussion of evolutionary drift and its consequences.

preferences over outcomes are her opponent's material payoffs. Let type θ', thus derived from type θ, be called the discriminating type corresponding to type θ. Let $\Theta^{\mu,*}$ be the minimal such set, i.e. it has exactly k elements. Finally, let preference type θ^* have some arbitrary preferences when meeting her own type, material preferences when meeting any type in $\Theta^{\mu,*}$, and preferences to minimize the opponent's material payoff against all other types. Let $f(0) = \mu$ and $f(1) = \mu'$ with $\mu'(\theta') = \mu(\theta)$, where $\theta' \in \Theta^{\mu,*}$ is the discriminating type that corresponds to type $\theta \in \Theta^{\mu}$. We can then easily construct a continuous function $f(t)$ from $f(0)$ and $f(1)$, which can be interpreted as a process of gradual replacement of all types $\theta \in \Theta^{\mu}$ by their corresponding discriminating types $\theta' \in \Theta^{\mu,*}$. Then, by construction, a function $g(t) \in B_p^{\delta}(f(t)|b)$ can be found for any $\delta > 0$. In fact aggregate behavior along the path $g(t)$ can be taken to be identical to behavior $g(0) = b$. Finally, the configuration $(f(1), g(1))$ is not DEY-stable as type θ^*, now "loved" by everyone, can successfully invade, because her material payoff is arbitrarily close to $\pi^*(G)$, whereas incumbents' payoffs are arbitrarily close to the strictly smaller payoff $\pi^{\mathrm{e}}(G)$. Thus the original configuration (μ, b), consisting of purely non-discriminating preference types, is not robust to drift.

4. Stable outcomes and discriminating preferences

We now turn to the question as to which outcomes can and cannot be sustained with strongly stable discriminating preferences. Let $\bar{\pi}$ denote the expected material payoff each of the two players could guarantee for him- or herself, i.e. $\bar{\pi} = \max_{\sigma\in\Delta} \min_{\tau\in\Delta} \sigma M \tau$. First, an outcome with material payoff below the material minmax value of $\bar{\pi}$ cannot be sustained in even a DEY-stable configuration, and, hence, can also not be sustained in a strongly stable configuration. The intuition behind this result is that a mutant whose preferences coincide with the material payoffs will always do at least as well as the minmax payoff and, hence, is able to invade successfully. The formal statement and proof of this result is in Herold and Kuzmics (2008).

In order to prove our main results the following definitions and lemma are useful. For any $y \in \operatorname{int}(\Delta)$, with all $y_i > 0$, consider the following subjective payoff matrix

$$A^y = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & \frac{1}{c(y)y_n} \\ \frac{1}{c(y)y_1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{c(y)y_2} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \vdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & 0 & \frac{1}{c(y)y_{n-1}} & 0 \end{pmatrix},$$

where $c(y) = \sum_{i=1}^{n} \frac{1}{y_i}$. This payoff matrix is such that whatever the opponent plays the other player always strictly prefers to play one strategy higher, except if the opponent plays the last strategy, then the other player strictly prefers to play the first strategy. If both players share this payoff matrix then the resulting game has a unique symmetric Nash equilibrium which is given by exactly y. If $n = 2$ this game is the Hawk–Dove game. For $n \geqslant 3$ one might call this a generalized Hawk–Dove game.

For $y \notin \operatorname{int}(\Delta)$, i.e. $y_i = 0$ for some $i \in A$, we have to modify the definition of A^y somewhat. Intuitively we just define A^y as before but only for those rows and columns i which are such that $y_i > 0$. The rest of the matrix is then filled with zeros and occasional 1's. To be more precise there are two cases which need separate treatment. First, let $y \in \Delta$ be such that $y_i = 1$ for some $i \in A$. Then A^y is such that $A^y_{ij} = 1$ for all j for i such that $y_i = 1$ and all other A^y_{ij}'s are equal to 0 (action i is a dominant strategy). Second, and without loss of generality,[14] let $y \in \Delta$ be such that $y_i = 0$ for all $i \leqslant l$ and $y_i > 0$ for all $i > l$ for some $l \leqslant n-2$. Let $\bar{c}(y) = \sum_{i:\ y_i>0} \frac{1}{y_i}$, i.e. $\bar{c}(y) = \sum_{i=l+1}^{n} \frac{1}{y_i}$. Then for rows and columns $l+1$ and above define A^y just as above, but replacing $c(y)$ with $\bar{c}(y)$. All other rows, the first l, shall be zeros only. To then ensure uniqueness of the symmetric equilibrium y in this case we need to have at least one positive element in

[14] Alternatively, one could define A^y for such y, with $y_i = 0$ for some i and $y_i < 1$ for all i, as follows: Let $i' = \min\{j{:}\ y_j > 0\}$ and let $i'' = \max\{j{:}\ y_j > 0\}$. Then $A^y_{i'i''} = \frac{1}{\bar{c}(y)y_{i''}}$. Furthermore $A^y_{ij} = \frac{1}{\bar{c}(y)y_i}$ if $y_i > 0$ and $j = \max\{j'{:}\ j' < i \wedge y_{j'} > 0\}$. Finally $A^y_{ji} = 1$ for $j = \min\{j' > i{:}\ y_{j'} > 0\}$ when i is such that $y_i = 0$ and all remaining A^y_{ij}'s are equal to 0.

every one of the first l columns. Let, in fact $A^y_{ni} = 1$ for all $1 \leqslant i \leqslant l$. The matrix A^y, in this case, can then be written as follows:

$$A^y = \left(\begin{array}{ccc|cccccc}
0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\hline
0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{\bar{c}(y)y_n} \\
\vdots & \cdots & \vdots & \frac{1}{\bar{c}(y)y_{l+1}} & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \cdots & \vdots & 0 & \frac{1}{\bar{c}(y)y_{l+2}} & 0 & \cdots & 0 & 0 \\
\vdots & \cdots & \vdots & 0 & 0 & \ddots & \ddots & \vdots & 0 \\
0 & \cdots & 0 & \vdots & \vdots & \ddots & \ddots & 0 & \vdots \\
1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \frac{1}{\bar{c}(y)y_{n-1}} & 0
\end{array}\right).$$

This defines A^y for all $y \in \Delta$. The following result is about the game induced by these generalized Hawk–Dove preferences and shows that the game has a unique symmetric Nash equilibrium.

Lemma 1. *The symmetric 2-player game with payoff-matrix A^y, with $y \in \Delta$, has a unique symmetric Nash equilibrium, which is y.*

Proof. We will do the proof for the case $y \in \mathrm{int}(\Delta)$ only. The proof extends to all $y \in \Delta$ straightforwardly. Let $y \in \mathrm{int}(\Delta)$ and let A^y be defined as above. Then y is obviously the only symmetric Nash equilibrium with full support as it is the only vector which equalizes the payoff for all strategies of the opponent. Suppose there is a symmetric NE $z \in \Delta$ with non-full support. I.e. let $z_j = 0$ for some $j \in \{1, \dots, n\}$. Suppose first that $j = n$. Then strategy a_1, being good only against a_n, is strictly dominated by any mixture with full support in $A \setminus \{a_n\}$, and, hence, we must have $z_1 = 0$. Suppose now that $j < n$, the only other case. But then strategy a_{j+1} is strictly dominated by any mixture with full support in $A \setminus \{a_j\}$, and, hence, we must have $z_{j+1} = 0$. Iterating this argument provides us with $z_i = 0$ for all $i \in \{1, \dots, n\}$, which provides a contradiction. □

Let M^{t} denote the transpose of matrix M, the matrix of material payoffs.

Definition 3. A *fully discriminating preference type*, indexed by $y \in \Delta$, denoted by θ^y, is such that $u_{\theta^y}(\theta^y, \sigma, \tau) = \sigma A^y \tau$ and $u_{\theta^y}(\theta, \sigma, \tau) = \sigma(I^n - M^{\mathrm{t}})\tau$ for all $\theta \in \Theta$, $\theta \neq \theta^y$, where I^n denotes the $n \times n$-matrix of all 1's.

The fully discriminating preference type with index y, therefore, has the following preferences over outcomes. When facing her own type θ^y her preferences are of the generalized Hawk–Dove variety with subjective payoff matrix A^y as described above, and when facing any other type her preferences are spiteful with subjective payoff matrix $I^n - M^{\mathrm{t}}$, implying that she will seek to minimize her opponent's material payoff in this case. Now we can state our main result.

Proposition 1. *Let Θ be an arbitrary preference space, except that it contains the fully discriminating preference type θ^y, as defined in Definition* 3*, for some $y \in \Delta$ with $\pi(y, y) > \bar{\pi}$. Then, there exists a $\bar{p} \in (0, 1)$ such that for all degrees of observability p with $\bar{p} \leqslant p \leqslant 1$ the configuration (μ, b) is strongly stable, where μ is the Dirac measure putting probability* 1 *on θ^y and b is such that $b_{\theta^y}(\theta^y) = b_{\theta^y}(\phi) = y$.*

Proof. First, given the preferences of the fully discriminating type θ^y and by Lemma 1 we have that $B_p(\mu_\theta) = \{b_{\theta^y}\}$. Second, consider, without loss of generality, any $\mu' \in \mathcal{P}(\Theta)$ such that $\theta^y \notin C(\mu')$. Let $\tilde{\mu}_\epsilon = (1 - \epsilon)\mu + \epsilon\mu'$. Note that we must have $B_p(\tilde{\mu}_\epsilon) \neq \emptyset$. Now we need to characterize any $b' \in B_p(\tilde{\mu}_\epsilon)$.

We need to determine $b'_\theta : \Theta \cup \{\phi\} \to \Delta$ for all $\theta \in C(\mu') \cup \{\theta^y\}$. First of all, we know there exists such a symmetric $b' \in B_p(\tilde{\mu}_\epsilon)$. We really only care about how type θ^y behaves in any such b'. There are 3 components to this. We need to determine how type θ^y behaves when meeting and observing its own type, when meeting and observing another type, and when observing ϕ. The subjective payoff to type θ^y when recognizing its own type and using strategy $z \in \Delta$, while everyone else plays according to b', is given by $p u_{\theta^y}(\theta^y, z, b'_{\theta^y}(\theta^y)) + (1 - p)u_{\theta^y}(\theta^y, z, b'_{\theta^y}(\phi))$, which, by the fact that u_θ is of the expected utility form, equals

$$u_{\theta^y}\bigl(\theta^y, z, p b'_{\theta^y}\bigl(\theta^y\bigr) + (1 - p)b'_{\theta^y}(\phi)\bigr).$$

Consider first the case $y \in \mathrm{int}(\Delta)$. If p is sufficiently close to 1, by the same argument as in the proof of Lemma 1, we cannot have that $(b'_{\theta^y}(\theta^y))_i = 0$ for any $i \in \{1, \dots, n\}$. Hence, the only possibility is that $(b'_{\theta^y}(\theta^y))_i > 0$ for all $i \in \{1, \dots, n\}$. This implies that type θ^y must be indifferent between all strategies in A. This implies that $p(b'_{\theta^y}(\theta^y))_i + (1 - p)(b'_{\theta^y}(\phi))_i = y_i$, or equivalently $(b'_{\theta^y}(\theta^y))_i = \frac{y_i - (1-p)(b'_{\theta^y}(\phi))_i}{p}$, which, for p close to 1, is close to y_i. This alone is sufficient to show that for any $\delta > 0$ there is an $\bar{\epsilon} > 0$ such that for all $\epsilon \in (0, \bar{\epsilon})$ we have that any such $b' \in B_p(\tilde{\mu}_\epsilon)$ also satisfies $b' \in B_p^\delta(\tilde{\mu}_\epsilon)$ for p sufficiently close to 1.

This is also sufficient to show that, for p close to 1 and for small $\epsilon > 0$, the material payoff of type θ^y is strictly above that of any other type $\theta \in \mu'$ for any $\mu' \in \mathcal{P}(\Theta)$. To see this let $y' = b'_{\theta^y}(\theta^y)$. The material payoff to type θ^y is bounded from below by $(1-\epsilon)[p^2\pi(y', y') + O(1-p)]$, where $O(1-p)$ is a term that converges to 0 as $1-p$ tends to 0. This lower material payoff bound, hence, tends to $\pi(y, y)$ if ϵ tends to 0 and p to 1. The material payoff to any other type θ is bounded from above by $(1-\epsilon)[p^2\bar{\pi} + O(1-p)] + \epsilon$, where again $O(1-p)$ is a term that converges to 0 as $1-p$ tends to 0. This upper material payoff bound, hence, tends to $\bar{\pi}$ if ϵ tends to 0 and p to 1. Hence, for p sufficiently close to 1 and ϵ sufficiently close to 0 we have that $\Pi_\theta(\tilde{\mu}_\epsilon | b') > \Pi_{\theta'}(\tilde{\mu}_\epsilon | b')$ for every $\theta \in C(\mu)$ and every $\theta' \in C(\mu')$ and for every $b' \in B_p(\tilde{\mu}_\epsilon)$.

In the general case $y \in \Delta$ the same arguments apply. Notice that all actions i with $y_i = 0$ are strictly dominated for (and thus not played by) type θ^y when facing his own type. □

Proposition 1 shows that $\pi(y, y) > \bar{\pi}$ is a sufficient condition for strong stability of an outcome y. Thus, under observability the indirect evolutionary approach has almost no predictive power with respect to outcomes beyond that the outcome y has to satisfy the necessary condition $\pi(y, y) \geqslant \bar{\pi}$. Any outcome induced by a single type choosing strategy $y \in \Delta$ is strongly stable for sufficiently large p as long as the resulting material payoff $\pi(y, y)$ is individually rational.

Proposition 1 also provides specific, relatively simple preferences for which this outcome is strongly stable. These preferences are such that the incumbent type has generalized Hawk–Dove preferences when her opponent is of the same type, but when her opponent is of another type she has preferences which are diametrically opposed to the material payoffs of the opponent. These are our fully discriminating preference types. Typically there are other strongly stable preferences. It is for instance not necessary for the incumbent to minimize her mutant opponent's fitness as long as she plays something that is materially worse for the entrant than what the incumbents obtain. Any strongly stable preferences, however, must be of a discriminating type.

Finally, note that the preference space Θ in Proposition 1 is arbitrary as long as it contains the fully discriminating preference type θ^y. This implies that even if we added all possible other preference types (i.e. even if we consider Gul and Pesendorfer's (2007) canonical space of behavioral preferences) this discriminating preference type θ^y is still strongly stable. In fact, nothing changes in our argument even if we also allow preferences of a non-expected utility form as in von Widekind (2004). Thus, there are no additional preference types whose presence could overturn our result.

5. Extensions

All claims in this section are substantiated in Herold and Kuzmics (2008), the working paper version of this paper. Given that our definition of strong stability requires that an outcome be supported by a single type we can only obtain outcomes $x \in \mathcal{P}(A \times A)$ which are of the symmetric product form, i.e. $x = y \cdot y$, where $y \in \Delta$. This derives from the simple fact that if there is only one type all players of this type must choose the same mixed strategy when playing against each other. These results, however, can easily be extended to asymmetric outcomes if players can also condition on their player position.

For the case of no and almost no observability Dekel et al. (2007) show that being a Nash equilibrium of G, i.e. in material payoffs, is a necessary condition for an outcome to be DEY-stable. Their argument still goes through in our setting: Under no and almost no observability being a Nash equilibrium is a necessary condition for DEY-stability (and therefore also necessary for strong stability). DEY show also that being a strict Nash equilibrium is a sufficient condition for DEY-stability under no observability. This also remains valid in our setting. Under almost no observability the conditions for DEY-stability are somewhat different in DEY and in our setting.

For the case of any arbitrary degree of observability p between 0 and 1, one might ask whether the transition from only Nash equilibria being stable for small p to anything above the minmax-value being stable for large p is a continuous one or whether there is a jump at some level of p. It turns out that either of these can be the case depending on the game at hand.

Finally, consider a random distribution over games instead of a single one, where, however, players always know which game they are playing. Our argument hardly changes. Consider a meta-outcome (a combination of outcomes in all games) and a type that against his own type has preferences such that each of these outcomes constitutes a subjective equilibrium in the respective game, yet minimizes the expected material payoff of any other type in each game. This can be constructed very much as we construct our discriminating types in this paper. Again, this fully discriminating type stabilizes any meta-outcome with expected payoffs above the overall minmax payoff. Notice that in this scenario the material payoff of the incumbent can even be below the minmax value for some games, as long as the average payoff is greater than the expected payoff of a player who receives the minmax payoff in all games.

6. Discussion

A wide range of evidence from casual introspection to lab-experiments strongly suggests that individuals do not only care about outcomes. They also care about the motives or intentions (i.e. preferences) that lead to these outcomes. This point, as experimentally made in e.g. Charness and Levine (2007), is already present in Adam Smith's (1976) Theory of Moral Sentiments:

> We do not, therefore, thoroughly and heartily sympathize with the gratitude of one man towards another, merely because this other has been the cause of his good fortune, unless he has been the cause of it from motives which we entirely go along with. [Smith, The Theory of Moral Sentiments, II.I.18.]

For instance, we may well treat someone who seems to be a nice, altruistic person very differently from someone considered opportunistic or even spiteful. We, thus, discriminate between players with different preferences.

Such discriminating preferences have been used in important work on interdependent preferences such as Levine (1998) and Gul and Pesendorfer (2007) and analyzed as to their evolutionary merit in Sethi and Somanathan (2001). Especially under observability we see no reason why nature would be restricted to only choose preferences over outcomes. Once we allow nature to choose preferences that depend on opponent's preferences there is also no obvious natural restriction as to the particular form these preferences can take. We, thus, also in the spirit of Dekel et al. (2007), consider all possible preferences. Furthermore, the fully discriminating preference type which we use to establish our folk-theorem like result is rather simple. Thus, even if we want to consider only a subclass of plausible discriminating preferences, these preferences should be included and our results remain valid.

This leads to a related question we want to address before we discuss the implications of our findings. Could it be that such a richer preference space makes the observability assumption more demanding relative to the existing literature? If we want her to reveal her relevant preferences in a setting where preferences are over outcomes only, it would in principle be enough to ask for her best response correspondence. If preferences depend on the opponent's preferences this might be more difficult.[15]

In any case it is very questionable in both settings whether players would be willing to reveal their types by answering such questions truthfully, in particular if the opponent's future play depends on this information. Observability is therefore often interpreted in a way that two individuals with different preferences are indeed visibly different. For the sake of the argument let us say they have different "genetic codes" and this fact cannot be hidden. I.e. one can indeed see or smell or generally sense that any two individuals with different preferences are different. But then any two different types of individuals in this paper are just as different as any two different types of individuals with preferences over outcomes only. Finally, note that perfect observability is not needed for our results and examples, given in Herold and Kuzmics (2008), can be found in which the degree of observability can be far from perfect and yet our results go through.

We, thus, motivated by the above arguments, study the evolution of preferences under observability when nature is essentially unrestricted in the set of preferences, allowed to depend on opponent preferences, she can choose from.

The results in this paper can be separated into three main findings. First, once we allow nature to choose among this large set of preferences, we find that those preferences, which depend on outcomes only, i.e. which are non-discriminating, are typically not evolutionary stable, unless perhaps in a very weak sense. Second, there are, however, strongly stable preferences which exhibit discriminating behavior. Third and finally, almost any (symmetric) outcome can be sustained by such discriminating and strongly stable preferences.

Methodologically our findings reinforce the point made by Dekel et al. (2007) that the choice of preference space at nature's disposal is crucial in the evolution of preferences.

Our results, however, are very different from any result in the literature. The first and second result imply that evolutionary forces such as commitment effects and the drive towards efficiency are typically overpowered by the force of discrimination, a force so far neglected in the evolutionary literature. Furthermore, the fully discriminating preference type which we use to prove our folk-theorem type result is a fairly simple and natural type, who does whatever it does against its own type and minimizes the material payoff of everyone else. Negative discrimination against someone perceived as being not of your own kind seems to be an unfortunate, yet common, human trait. Given the simplicity of these types we would also like to argue that any model of preference evolution should, thus, either consider these types or have a good reason why nature is not able to endow an individual with these preferences.

The third result, furthermore, implies that very different ways how to play a game can be part of a stable equilibrium. This suggests that the evolutionary force of discrimination enables a potentially huge heterogeneity of "cultures" in different, separated, societies. Thus, in the light of our findings it would not be surprising if the same game is played very differently in these different societies.

Note also that we obtain these results for any preference space nature has at her disposal as long as the above identified simple type is among it. Finally, there is no additional generalization of the preference space that could overturn these results.[16]

[15] Note, however, that we can take our preference space to be valid in the sense of Gul and Pesendorfer (2007). This implies that you could identify preference types sequentially from responses to similar questions.

[16] Even introducing non-expected utility preferences as in von Widekind (2004) would not change the result.

References

Banerjee, A., Weibull, J.W., 1993. Evolutionary selection with discriminating players. Working paper, Department of Economics, Harvard University, WP 1637.

Banerjee, A., Weibull, J.W., 2000. Neutrally stable outcomes in cheap-talk coordination games. Games Econom. Behav. 32, 1–24.

Bester, H., Güth, W., 1998. Is altruism evolutionary stable? J. Econom. Behav. Organization 34, 193–209.

Binmore, K.G., Samuelson, L., 1992. Evolutionary stability in repeated games played by finite automata. J. Econom. Theory 57, 278–305.

Binmore, K.G., Samuelson, L., 1999. Evolutionary drift and equilibrium selection. Rev. Econom. Stud. 66, 363–393.

Charness, G., Levine, D.I., 2007. Intention and stochastic outcomes: An experimental study. Econom. J. 117, 1051–1072.

Dekel, E., Ely, J.C., Yilankaya, O., 2007. Evolution of preferences. Rev. Econom. Stud. 74, 685–704.

Ely, J.C., Yilankaya, O., 2001. Nash equilibrium and evolution of preferences. J. Econom. Theory 97, 255–272.

Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Econom. Inquiry 41, 20–26.

Fudenberg, D., Maskin, E., 1990. Evolution and cooperation in noisy repeated games. Amer. Econom. Rev. Papers Proc. 80, 274–279.

Gul, F., Pesendorfer, W., 2007. The canonical space for behavioral types. Mimeo.

Güth, W., 1995. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. Internat. J. Game Theory 24, 323–344.

Güth, W., Yaari, M., 1992. Explaining reciprocal behavior in a simple strategic game. In: Explaining Process and Change—Approaches to Evolutionary Economics. Univ. Michigan Press, pp. 23–24.

Heifetz, A., Shannon, C., Spiegel, Y., 2007. The dynamic evolution of preferences. Econom. Theory 32 (2), 251–286.

Herold, F., Kuzmics, C., 2008. Evolution of preferences under perfect observability: Almost anything is stable. Mimeo, SSRN.

Kalai, A., Kalai, E., Lehrer, E., Samet, D., 2007. A commitment folk theorem. Mimeo.

Koçkesen, L., Ok, E.A., Sethi, R., 2000a. Evolution of interdependent preferences in aggregative games. Games Econom. Behav. 31, 303–310.

Koçkesen, L., Ok, E.A., Sethi, R., 2000b. The strategic advantage of negatively interdependence preferences. J. Econom. Theory 92, 274–299.

Levine, D.K., 1998. Modelling altruism and spitefulness in experiments. Rev. Econom. Dynamics 1, 593–622.

Ok, E., Vega-Redondo, F., 2001. On the evolution of individualistic preferences: An incomplete information scenario. J. Econom. Theory 97, 231–254.

Robson, A.J., 1990. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. J. Theoret. Biol. 144, 379–396.

Sethi, R., Somanathan, E., 2001. Preference evolution and reciprocity. J. Econom. Theory 97, 273–297.

Smith, A., 1976. The Theory of Moral Sentiments. Oxford University Press. Originally published in 1759.

von Widekind, S., 2004. Evolution of non-expected utility preferences, Bielefeld University. IMW working paper #370.