Co-Evolution of Deception and Preferences*

Preliminary and Incomplete

Yuval Heller and Erik Mohlin†
University of Oxford

April 14, 2014

Abstract

We study how preferences may co-evolve with the ability to detect other peoples' preferences, and the ability to deceive other people regarding one's preferences and intentions. An individual's type is a tuple consisting of a preference type and a cognitive type. Preferences are allowed to be defined not only over action profiles but also over the opponent's type. The cognitive type is an integer representing level of cognitive sophistication. The cognitive levels of the individuals in a match determine the probability that one of them observes the opponent's preferences and is able to deceive the opponent. For preferences defined solely over action profiles we find that, for low enough cognition costs, if a preference configuration is evolutionarily stable then all the induced outcomes are Nash equilibria, and in same-type matches, an efficient symmetric Nash equilibrium is played. Conversely any symmetric Nash equilibrium can be implemented as the outcome of an evolutionarily stable preference configuration. In contrast, for preferences defined over both actions and opponents' types, all Nash outcomes that give more than the minmax payoff can be implemented by evolutionarily stable preferences.

Keywords: Evolution of Preferences; Indirect Evolutionary Approach, Theory of Mind; Depth of Reasoning; Evolution of Cooperation.
JEL codes: C72, C73, D01, D03, D83.

*Valuable comments were provided by Larry Samuelson and Jörgen Weibull, as well as participants at presentations in Oxford, at G.I.R.L.13, the Toulouse Economics and Biology Workshop, and DGL13. Part of this project was carried out while Erik Mohlin was funded by the European Research Council, Grant no. 230251.

†Nuffield College and Department of Economics, University of Oxford. Address: Nuffield College, New Road, Oxford OX1 4PX, United Kingdom. E-mail: yuval.heller@nuffield.ox.ac.uk and erik.mohlin@nuffield.ox.ac.uk.

1 Introduction

For a long time economists took preferences as given. The study of their origin and formation was considered a question outside the scope of economics. Over the past twenty years this has changed dramatically. In particular there is now a large literature on the evolutionary foundations of preferences (for an overview see Robson and Samuelson (2011)). A prominent strand of this literature is the so-called "indirect evolutionary approach" to game theory, as pioneered by Güth and Yaari (1992).[1] It has been used to explain the existence of a number of "non-standard" preferences – preferences that do not coincide with material payoffs. For example Huck and Oechssler (1999) study the evolution of reciprocity in the form of a taste for rejecting unfair offers. Sethi and Somanthan (2001) consider reciprocity in the form of preferences that are conditional on the opponent's preference type. Bester and Güth (1998), Bolle (2000), and Possajennikov (2000) study combinations of altruism, spite and selfishness. Güth and Napel (2006) study preference evolution when players use the same preferences in both an ultimatum and a dictator game. Koçkcesen and Ok (2000) investigate survival of more general interdependent preferences in aggregative games. In all these models the authors find that various non-materialistic preferences are evolutionarily stable even under uniformly random matching.

A crucial feature of these models is that they explicitly or implicitly assume that preferences are at least partially observable. Consequently the results are vulnerable to the existence of mimics who signal that they have e.g. a preference for cooperation, but actually defect on cooperators, thereby earning the benefits of having the non-standard preference, while not having to pay the cost. The effect of varying the degree to which preferences can be observed has been investigated systematically by Ok and Vega-Redondo (2001), Dekel et al. (2007), and Herold and Kuzmics (2009). They confirm that the degree to which preferences are observed decisively influences the outcome of preference evolution. However, the degree to which preferences are observed is still exogenous in these models. In reality we would expect both the preferences and the ability to observe or conceal preferences to be the product of an evolutionary process. On this topic, Robson and Samuelson (2011) write:

> The standard argument is that we can observe preferences because people give signals – a tightening of the lips or flash of the eyes – that

[1] We do not find the term 'indirect' ideal since evolution selects on the basis of fitness and any feature of an organism, be it strategies, preferences, beliefs or something else, is selected for, only insofar as it (indirectly) contributes to fitness.

> provide clues as to their feelings. However, the emission of such signals and their correlation with the attendant emotions are themselves the product of evolution. [...] We cannot simply assume that mimicry is impossible, as we have ample evidence of mimicry from the animal world, as well as experience with humans who make their way by misleading others as to their feelings, intentions and preferences. [...] In our view, *the indirect evolutionary approach will remain incomplete until the evolution of preferences, the evolution of signals about preferences, and the evolution of reactions to these signals, are all analysed within the model.* [Emphasis added] (pp. 14-15)

This paper studies the missing link between evolution of preferences and evolution of how preferences are concealed and detected. In our model the ability to observe preferences as well as the ability to deceive and induce false beliefs about preferences, is endogenously determined by evolution, jointly with the evolution of preferences. Moreover we consider all preferences that are defined either on the set of action profiles or on the joint set of action profiles and opponents preference types.

As in standard evolutionary game theory we assume a population of individuals who are uniformly randomly matched to play a symmetric normal form game.[2] Each individual has a type, which is a tuple, consisting of a preference type and a cognitive type. A preference type is identified with a utility function. The cognitive type is simply a natural number, representing the level of cognitive sophistication of the individual. The marginal cost of increased cognition is strictly positive. The cognitive types of the individuals in a match determine the probability that one individual observes the opponent's preferences and is able to deceive the opponent.

When both individuals are of the same cognitive type they are assumed to play a Nash equilibrium of the complete information game induced by their preferences, just as in that standard indirect evolutionary approach. However when individuals in a match are of different cognitive level we allow the outcome to differ from a Nash equilibrium induced by the preferences of the matched individuals. We

[2]It is well-known that positive assortative matching, for instance due to spatially structured populations, is conducive to the evolution of altruism (Hines and Smith (1979)). Recently Alger and Weibull (2013) have shown that positive assortative matching allows for the evolution of non-materialistic preferences even when preferences are perfectly unobservable. It is also well-known that finite populations allow for the evolution of non-materialistic preferences, e.g. spite, even when preferences are perfectly unobservable (Schaffer (1988)). By assuming that individuals are uniformly randomly matched in an infinite population, we avoid confounding the effect of endogenously determined degree of observability, with the effect of non-uniform matching and finite population size.

assume a strong form of deception: The deceiver observes the opponent's type perfectly, and consequently the deceiver is able to tailor the deception to the current opponent's type. The deceiver is allowed to choose whatever she wants the deceived party to believe, both about the deceiver's type, and about the deceiver's intended action choice. In effect the deceiver is able to pick her favourite action profile, that is consistent with the deceived party behaving rationally given her preferences. In general the probability that one is able to deceive one's opponent should be increasing one's own cognitive type and decreasing in the opponent's cognitive type. In order to obtain tractable and general results we focus on the limiting case where the individual with the highest cognitive type in a match is always able to deceive the individual with the lowest cognitive type. The fact that we make these strong assumptions about deception allows us to interpret our results as providing an "upper bound" on the effect of endogenization of observation, and the introduction of deception.

Our way of modelling observation and deception implies that behaviour in a match is determined jointly by (i) the types of the matched individuals, (ii) the attempts at deception, summarised in a deception policy, and (ii) the actions that are taken as a function of the opponent's type and the attepted deception, summarised in a deception policy. A population state together with a set of action and deception policies is a called a configuration. We require configurations to be consistent in the sense that they induce Nash equilibrium when no deception takes place. In an *evolutionarily stable configuration* (ESC) all incumbents earn the same and if a small group of mutants enter they earn less than the incumbents in any post-entry state where the incumbents behave against each other in the same way as before the mutants entered.

We provide results both for preferences that are defined on the set of action profiles (type-neutral preferences) and for preferences defined or on the joint set of action profiles and opponents preference types (interdependent preferences). For type-neutral preferences our main result is that, for low enough cognition costs, if a configuration is evolutionarily stable then all the induced outcomes are Nash equilibria, and in same-type matches, an efficient symmetric Nash equilibrium is played. In contrast, for type-interdependent preferences, all Nash outcomes that give more than the minmax payoff can be implemented in an evolutionarily stable configuration.

Dekel et al. (2007) (see also Ok and Vega-Redondo (2001)) show that if preferences are unobservable, and if all preferences over outcomes are allowed, then only Nash equilibria of the game with material/fitness payoffs can supported by

evolutionarily stable preferences. If preferences are perfectly, or almost perfectly, observable, then only efficient outcomes can be supported by evolutionarily stable preferences. Herold and Kuzmics (2009) expand the framework of Dekel et al. (2007) to include interdependent preferences, i.e. preferences that depend on the opponent's preference type. Under perfect or almost perfect observability, if all preferences that depend on the opponent's type are considered, then any symmetric outcome above the minmax material payoff is stable in a strong sense, implying stability in the sense of Dekel et al. (2007). Furthermore, they find that non-discriminating preferences (including selfish materialistic preferences) are typically not evolutionary stable except in a very weak form. In contrast, certain preferences which exhibit discrimination are evolutionary stable in a very strong sense.

Within biology and evolutionary psychology there is a large literature on the evolution of theory of mind (Premack and Wodruff (1979)). According to the "Machiavellian intelligence" hypothesis (Humphrey (1976)), and "social brain" hypothesis (for an introduction see Dunbar (1998)), the extraordinary cognitive abilities of humans evolved as a result of the demands of social interactions, rather than the demands of the natural environment: In a single person decision problem there is a fixed benefit of being smart, but in a strategic situation it may be important to be smarter than the opponent. From an evolutionary perspective, the potential advantage of a better theory of mind has to be traded off against the cost of increased reasoning capacity. Increased cognitive sophistication, in the form of higher order beliefs, is associated with non-negligible costs (Holloway (1996), Kinderman et al. (1998)).[3]

The rest of the paper is organized as follows: Section 2 presents the model. Section 3 contains results on both type-interdependent and type-neutral preferences. Section 4 concludes. Additional results, and proofs not in the main text, can be found in the appendix.

2 Model

2.1 Game

Consider a symmetric two-player normal form game G with a finite pure action set A and mixed strategy set $\Delta(A)$. Payoffs are given by $\pi : A \times A \to \mathbb{R}$, where $\pi(a_i, a_j)$ is the payoff to a player using strategy a_i against strategy a_j. The payoff function is extended to mixed actions in the standard way. Let $\pi(\sigma_i, \sigma_j)$ denote

[3]There is also a smaller literature on the evolution of strategic sophistication in game theory; Stahl (1993), Banerjee and Weibull (1995), Stennek (2000), Mohlin (2012), and Heller (2013).

the payoff to a player, using strategy σ_i against strategy σ_j. With slight abuse of notation let a denote the degenerate mixed strategy that puts all weight on pure strategy a.

2.2 Types

There is a large population of individuals who are randomly matched to play the game G. Each individual i in the population has a type

$$\theta = (\phi, \psi) \in \Theta = \Phi \times \Psi,$$

consisting of a preference type $\phi \in \Phi$ and a cognitive type $\psi \in \Psi$. Let $\mathcal{X}(\Theta)$ be the set of all finite support probability distributions on Θ. A *population state* is a point

$$x = \left(x_{\theta_1}, x_{\theta_2}, ...x_{\theta_{|C(x)|}} \right) \in \mathcal{X}(\Theta),$$

where $C(x)$ denotes the support (carrier) of $x \in \mathcal{X}(\Theta)$.

2.2.1 Cognitive Types

We let $\Psi = \mathbb{N}$. There is a fitness cost to increased cognition, represented by the positive and strictly increasing cost function $k : \mathbb{N} \to \mathbb{R}_+$. The fitness payoff of an individual equals the material payoff from the game, minus the cognitive cost. Let

$$\kappa^{\max} = \sup_{\psi \in \Psi} (k_{\psi+1} - k_\psi),$$
$$\kappa^{\min} = \inf_{\psi \in \Psi} (k_{\psi+1} - k_\psi).$$

In our view the most interesting case is that of a small but strictly positive marginal cognitive cost. By small we mean that the cost does not outweigh the improvement that arises from changing actions. We will make this more precise later.

2.2.2 Preference Types

Type-Interdependent Preferences We allow for *type-interdependent* (or interdependent) preferences. Each preference type ϕ is identified with a utility function defined over the set of types and action profiles

$$u^\phi : A \times A \times \Phi \to \mathbb{R}.$$

This formulation allows preferences that depend on the opponent's preference type but not the opponent's cognitive type.[4] The definition of u^{ϕ} extends to mixed actions in the obvious way. Thus we will let $u^{\phi_i}(\sigma_i, \sigma_j, \theta_j)$ denote the payoff that player i, of type θ_i, earns when she plays mixed action σ_i, against an opponent j who is of type θ_j and plays the mixed action σ_j.

Type-Neutral Preferences We will also study *type-neutral* preferences, i.e. preferences that can be represented with a utility function of the form

$$u^{\phi} : A \times A \to \mathbb{R}.$$

2.3 Observation and Deception

2.3.1 Probability of Observation and Deception

The cognitive types of the individuals in a match determine their ability to observe and deceive each other. Consider two individuals i and j with cognitive types ψ_i and ψ_j, who are are matched to play a game. If $\psi_i = \psi_j$ then neither individual is able to deceive the other, if $\psi_i > \psi_j$ then i is able to deceive j, and if $\psi_i < \psi_j$ then j is able to deceive i, i.e.

$$\Pr(i \text{ deceives } j) = \begin{cases} 1 & \text{if } \psi_i > \psi_j \\ 0 & \text{if } \psi_i \leq \psi_j \end{cases}. \tag{1}$$

When neither individual is able to deceive the other, we assume that they play a Nash equilibrium of the complete information game that is induced by their preference types.

2.3.2 Outcome of Deception

Type-Interdependent Preferences Consider *type-interdependent preferences*. If i makes j believe that i is of type $\hat{\phi}_i$ and will take action $\hat{\sigma}_i$ then rationality demands that j takes an action in $\arg\max_{\sigma_j} u^{\phi_j}\left(\sigma_j, \hat{\sigma}_i, \hat{\phi}_i\right)$. Moreover, rationality requires i to pick a pair $\left(\hat{\sigma}_i, \hat{\phi}_i\right)$ so as to maximize her utility. The only restriction on $\left(\hat{\sigma}_i, \hat{\phi}_i\right)$ is that the action $\hat{\sigma}_i$ should be rationalizable for an individual with

[4] For an explanation of why this way of defining interdependent preference types avoids inconsistencies, see Herold and Kuzmics (2009). See also Gul and Pesendorfer (2010).

preference type $\hat{\phi}_i$. The set of pairs $\left(\hat{\sigma}_i, \hat{\phi}_i\right)$ satisfying this restriction is

$$R = \left\{ \left(\hat{\sigma}_i, \hat{\phi}_i\right) \in \Delta(A) \times \Phi : \hat{\sigma}_i \text{ is rationalizable given } \hat{\phi} \right\}.$$

Fix a state $x \in \mathcal{X}(\Theta)$. For any type $\theta_i \in C(x)$ an *action policy* (at state x) is a mapping $\alpha^{\theta_i} : \Delta(A) \times C(x) \to \Delta(A)$. The action policy α^{θ_i} is *rational* if for all $\theta_j \in C(x)$ and all $(\sigma_j, \phi_j) \in R$ it holds that

$$\alpha^{\theta_i} \in \arg\max_{\sigma_i} u^{\phi_i}(\sigma_i, \sigma_j, \phi_j).$$

For any type $\theta_i \in C(x)$ a *deception policy* (at state x) is a mapping $\delta^{\theta_i} : C(x) \to \Delta(A) \times \Theta$ such that $\delta^{\theta_i} \in R$. The deception policy δ^{θ_i} is *rational*, given the rational action mappings $\{\alpha^{\theta}\}_{\theta \in C(x)}$, if for all $\theta_j \in C(x)$, it holds that

$$\delta^{\theta_i} \in \arg\max_{\left(\hat{\sigma}_i, \hat{\phi}_i\right)} u^{\phi_i}\left(\alpha^{\theta_i}\left(\alpha^{\theta_j}\left(\hat{\sigma}_i, \hat{\phi}_i\right), \phi_j\right), \alpha^{\theta_j}\left(\hat{\sigma}_i, \hat{\phi}_i\right), \phi_j\right).$$

This description assumes that when i deceives j then i knows how j will respond to different beliefs that i might give j. That is, when i deceives j then i knows α^{θ_j}.

Type-Neutral Preferences The above treatment of deception can be adapted quite naturally to the case of *type-neutral preferences*. If i makes j believe that i will take action $\hat{\sigma}_i$ then rationality demands that j takes an action in $\arg\max_{\sigma_j} u^{\phi_j}(\sigma_j, \hat{\sigma}_i)$. Moreover, rationality requires i to pick $\hat{\sigma}_i$ so as to maximize her utility.

Fix a state $x \in \mathcal{X}(\Theta)$. For any type $\theta_i \in C(x)$ an *action policy* is a mapping $\alpha^{\theta_i} : \Delta(A) \to \Delta(A)$. The action policy α^{θ_i} is *rational* if for all σ_j it holds that

$$\alpha^{\theta_i} \in \arg\max_{\sigma_i} u^{\phi_i}(\sigma_i, \sigma_j).$$

For any type $\theta_i \in C(x)$ a *deception policy* (at state x) is a mapping $\delta^{\theta_i} : C(x) \to \Delta(A)$. The deception policy δ^{θ_i} is *rational*, given the rational action mappings $\{\alpha^{\theta}\}_{\theta \in C(x)}$, if for all $\theta_j \in C(x)$, it holds that

$$\delta^{\theta_i} \in \arg\max_{\hat{\sigma}_i} u^{\phi_i}\left(\alpha^{\theta_i}\left(\alpha^{\theta_j}(\hat{\sigma}_i)\right), \alpha^{\theta_j}(\hat{\sigma}_i)\right).$$

This description assumes that when i deceives j then i knows how j will respond to different beliefs that i might give j. That is, when i deceives j then i knows α^{θ_j}.

2.3.3 Outcome of Mutual Observation

As mentioned above, if i and j are matched and both observe each other's preferences then they are assumed to play a Nash equilibrium of the complete information game induced by their preference types (like in the standard indirect evolutionary approach). We need to impose restriction on the action profiles in order to ensure that they are consistent with Nash equilibrium in matches where both individuals observe the opponent's type. There may be more than one such equilibrium. Let $NE(\phi_i, \phi_j)$ be the set of Nash equilibria when the two players have preferences ϕ_i and ϕ_j.

Again, fix a state $x \in \mathcal{X}(\Theta)$. An *equilibrium selection* is a mapping $\nu : C(x) \times C(x) \to \Delta(A) \times \Delta(A)$ such that $\nu(\phi_i, \phi_j) = \nu(\phi_j, \phi_i) \in NE(\phi_i, \phi_j)$. Let $\nu^{(\phi_i,\phi_j)}$ and $\nu^{(\phi_j,\phi_i)}$ be the action played by i and j, respectively, in $\nu(\phi_i, \phi_j)$.

For *type-interdependent preferences*, we say that the equilibrium selection ν is *consistent* with the action policies $\{\alpha^\theta\}_{\theta \in C(x)}$ if for all $\theta_i \in C(x)$, and $\theta_j \in C(x)$,

$$\nu^{(\phi_i,\phi_j)} = \alpha^{\theta_i}\left(\nu^{(\phi_j,\phi_i)}, \phi_j\right).$$

For *type-neutral preferences* The equilibrium selection ν is *consistent* with the action policies $\{\alpha^\theta\}_{\theta \in C(x)}$ if for all $\theta_i \in C(x)$, and $\theta_j \in C(x)$,

$$\nu^{(\phi_i,\phi_j)} = \alpha^{\theta_i}\left(\nu^{(\phi_j,\phi_i)}\right).$$

2.3.4 Configurations and Payoffs

We combine our rationality and consistency requirements:

Definition 1 *A policy and selection profile* $\omega(x) = \left(\{\alpha^\theta\}_{\theta \in C(x)}, \{\delta^\theta\}_{\theta \in C(x)}, \nu\right)$ *is* ***rational and consistent*** *if (i) for each type* $\theta \in C(x)$ *the action policy* α^{θ_i} *is rational, (ii) for any type* $\theta \in C(x)$ *the deception policy* δ^{θ_i} *is rational given* $\{\alpha^\theta\}_{\theta \in C(x)}$*, and (iii) the equilibrium selection* ν *is consistent with* $\{\alpha^\theta\}_{\theta \in C(x)}$*.*

A policy and selection profile $\omega(x) = \left(\{\alpha^\theta\}_{\theta \in C(x)}, \{\delta^\theta\}_{\theta \in C(x)}, \nu\right)$ induces an *outcome* mapping $\eta^{\omega(x)} : C(x) \times C(x) \to \Delta(A) \times \Delta(A)$, such that in a match between types θ_i and θ_j the outcome is $\eta^{\omega(x)}(\theta_i, \theta_j)$. The payoff to individual i when facing individual j is

$$w(\theta_i, \theta_j) = \pi\left(\eta^{\omega(x)}(\theta_i, \theta_j), \eta^{\omega(x)}(\theta_j, \theta_i)\right).$$

To compute expected payoff we need to combine the information in the policy and selection profile with information about the state.

Definition 2 *A **configuration** $(x, \omega(x))$, is the combination of a state $x \in \mathcal{X}(\Theta)$ and a rational and consistent policy and selection profile $\omega(x) = \left(\{\alpha^\theta\}_{\theta \in C(x)}, \{\delta^\theta\}_{\theta \in C(x)}, \nu\right)$.*

Thus in configuration $(x, \omega(x))$ the expected (fitness) payoff to an individual of type θ is

$$\Pi_\theta(x) = \sum_{\theta' \in \Theta} x_{\theta'} w(\theta, \theta') - k_\theta = \sum_{\theta' \in \Theta} x_{\theta'} \pi\left(\eta^{\omega(x)}(\theta, \theta'), \eta^{\omega(x)}(\theta', \theta)\right) - k_\theta.$$

2.4 Evolutionary Stability

Recall the definition of an evolutionarily stable strategy, due to Maynard Smith and Price (1973) (see also Taylor and Jonker (1978)).

Definition 3 *A mixed strategy $\sigma \in \Delta(A)$ is an **evolutionarily stable strategy (ESS)** if for every $\sigma' \in \Delta(A)$, $\sigma' \neq \sigma$, there is some $\bar{\varepsilon} \in (0,1)$ such that if $\varepsilon \in (0, \bar{\varepsilon})$, then $\tilde{\pi}(\sigma', (1-\varepsilon)\sigma + \varepsilon\sigma') < \tilde{\pi}(\sigma, (1-\varepsilon)\sigma + \varepsilon\sigma')$. If the strict inequality is replaced by weak inequality then $\sigma \in \Delta(A)$ is a **neutrally stable state (NSS)**.*

We extend the notion of an ESS to an evolutionarily stable configuration (ESC). Note that a configuration completely determines payoffs $w(\theta_i, \theta_j)$ for all $\theta_i, \theta_j \in C(x)$. We use this fact to define a type game as follows:

Definition 4 *For any configuration $(x, \omega(x))$ the corresponding **type game** is the symmetric two-player game, where each player's strategy space is $C(x)$, and the payoff to type-strategy θ, against type-strategy θ', is $w(\theta, \theta') - k_\theta$.*

The definition of a type game allows us to apply notions and results from standard evolutionary game theory, where evolution acts upon strategies, to the present setting where evolution acts upon types.[5]

We want to capture robustness with respect to small groups of mutants. Suppose that a fraction ε of the population is replaced by mutants and suppose that the distribution of types within the group of mutants is $x' \in \mathcal{X}(\Theta)$. Consequently the post-entry population state is $\tilde{x} = (1-\varepsilon)x + \varepsilon x'$. In line with the rest of the literature on the indirect evolutionary approach we assume that adjustment of policies is infinitely much faster than the adjustment of the distribution of preferences (i.e. the movement between population states).[6] Thus we require that behaviour at

[5] A similar notion was defined in Mohlin (2012).

[6] Mohlin (2010) examines a model where the speed of learning relative to evolution is less extreme.

the new state $\tilde{x}$ is described by a rational and consistent policy and selection profile $\tilde{\omega}(\tilde{x}) = \left(\{\tilde{\alpha}^{\theta}\}_{\theta\in C(\tilde{x})}, \{\tilde{\delta}^{\theta}\}_{\theta\in C(\tilde{x})}, \nu\right)$ defined for the set of types in the support of $\tilde{x}$. There is no reason for an incumbent type $\theta \in C(x)$ to change its behaviour towards other incumbent types. A post-entry profile that exhibits this kind of conservativeness relative to the initial profile is called focal:

Definition 5 *Given an initial state x, and a post-entry state $\tilde{x} = (1-\varepsilon)x + \varepsilon x'$, a post-entry profile $\tilde{\omega}(\tilde{x}) = \left(\{\tilde{\alpha}^{\theta}\}_{\theta\in C(\tilde{x})}, \{\tilde{\delta}^{\theta}\}_{\theta\in C(\tilde{x})}, \nu\right)$ is **focal** relative to the initial profile $\omega(x) = \left(\{\alpha^{\theta}\}_{\theta\in C(x)}, \{\delta^{\theta}\}_{\theta\in C(x)}, \nu\right)$, if for all $\theta_i, \theta_j \in C(x)$ and all $a_i \in A$, it holds that $\tilde{\alpha}^{\theta_i}(a_i, \theta_j) = \alpha^{\theta_i}(a_i, \theta_j)$, $\tilde{\delta}^{\theta_i}(\theta_j) = \delta^{\theta_i}(\theta_j)$, and $\nu(\phi_i, \phi_j) = \nu(\phi_i, \phi_j)$.*

Our stability notion requires that incumbents outperform all mutants for all focal profiles that are consistent at $\tilde{x}$.

Definition 6 *A configuration $(x, \omega(x))$ constitutes an **evolutionarily stable configuration (ESC),** if for every $x' \in \mathcal{X}(\Theta)$, $x' \neq x$, and every rational and consistent post-entry profile $\tilde{\omega}(\tilde{x})$ which is focal relative to the initial profile, the state x is an ESS of the type game induced by the configuration $(\tilde{x}, \tilde{\omega}(\tilde{x}))$. If ESS is substituted for NSS then x is a **neutrally, stable configuration (NSC).***

3 Results

3.1 Some Preference Types

We wish to consider the largest possible set of preference types. For the case of type-interdependent preferences we may let Φ contain one type for each utility function $u : A \times A \to \mathbb{R}$. This means that $\Phi = \mathbb{R}^{n^2}$. Since von Neuman-Morgenstern utility functions are unique up to affine transformations we may assume $\Phi = [0,1]^{n^2}$.

In the case of type-interdependent preferences we need to be more careful in specifying the content of Φ. It turns out that we are able to prove all our results below provided that Φ contains a set of types that we are able to list explicitly. Individuals with preferences that coincide with fitness /material payoffs will be called *materialistic*.

Definition 7 *The set of materialist (M) types Θ^M is the set of types such that if $\theta_i = (\phi_i, \psi_i) \in \Theta^M$ then $u^{\phi_i}(a_i, a_j, \phi_j) = \pi(a_i, a_j)$ for all a_i, a_j, and ϕ_j.*

Next we define the *minmaxing discriminator* type. The minmax action is

$$a^m \in \min_{a_i} \max_{a_j} \pi(a_j, a_i).$$

The minmaxed player earns

$$m = \max_{a_j} \pi\left(a_j, a^m\right),$$

and the player that attempts to minmax her opponent earns at least $n = \min_{a_j} \pi\left(a^m, a_j\right)$.

We also define a type which has preferences that induce play of a profile (a^1, a^2) when meeting someone of the same type, but when meeting an opponent of another type it strictly prefers the action that minmaxes the opponent (in terms of fitness). We call this the *minmaxing discriminator* preference type.

Definition 8 *For a given profile $a = (a^1, a^2)$, the set of $MMDa$-types, Θ^{MMDa}, is the set of types such that for $\theta_i = (\phi_i, \psi_i) \in \Theta^{MMDa}$:*

(a) If $\phi_i \neq \phi_j$ then $u^{\phi_i}\left(a^m, a_j, \phi_j\right) > u^{\phi_i}\left(a_i, a_j, \phi_j\right)$ for all a_i and a_j.

(b) If $\phi_i = \phi_j$ then $u^{\phi_i}\left(a^1, a^2, \phi_j\right) > u^{\phi_i}\left(a^2, a^1, \phi_j\right) > u^{\phi_i}\left(a_i, a_j, \phi_j\right)$ for all $(a_i, a_j) \notin \left\{\left(a^1, a^2\right), \left(a^2, a^1\right)\right\}$.

The type which combines ϕ-preferences with cognitive type ψ, is denoted $\phi\psi$. When there is no risk of confusion we will simply write $\phi\psi$.

Throughout the paper we will assume that each of the above mentioned types is contained in the set of type-interdependent preference types Φ. Moreover we assume that any combination of such a preference type with a cognitive type $\psi \in \Psi$ is contained in the set of types Θ. Formally we assume:

$$\left\{M\psi\right\}_{\psi\in\Psi} \cup \left\{MMDa\psi\right\}_{\psi\in\Psi, a\in A\times A} \subseteq \Theta.$$

3.2 Results for Single Pure Outcome Configurations

In this subsection we restrict attention to configurations that induce play of one single pure outcome in all matches. That is, there is some pure action a such that $\eta^{\omega(x)}\left(\theta, \theta'\right) = a$ for all $\theta, \theta' \in C\left(x\right)$. The fact that there is a single outcome implies that the outcome has to be symmetric, since (according to the way our model is set up) when two indvividuals of the same type play they will play a symmetric outcome

The first result concerns the cognitive types that are present in an NSC.

Proposition 1 *Consider either **type-interdependent** or **type-neutral** preferences: Suppose a is the only outcome in $\left(x, \omega\left(x\right)\right)$. If $\left(x, \omega\left(x\right)\right)$ is NSC then all individuals have the same cognitive type. Moreover, if preferences are **type-neutral** then all individuals are of the lowest cognitive type.*

Proof. Since both players earn the same payoff in a they must also incur the same cognitive cost for the state to be part of an NSC configuration. In the case of type-neutral preferences this level must be the lowest level. Otherwise a mutant of a lower level, who prefers to play a against all actions would be able to invade. ■

Next result relates stable outcomes to Nash equilibrium and efficiency.

Proposition 2 *Consider* ***type-interdependent*** *preferences: If a is a symmetric Nash equilibrium, then there is an NSC in which each match results in a, and if $\pi(a) > m$ then there is an ESC in which each match results in a.*

Consider ***type-neutral*** *preferences:*

With one kind of mutant at a time*: If a is a symmetric Nash equilibrium (in fitness payoffs), and there is no symmetric profile with a higher payoff, then there is an NSC in which each match results in a.*

*[*With more than one kind of mutant at a time*: If a is a symmetric Nash equilibrium (in fitness payoffs), and there is no profile with a higher average payoff per player, then there is an NSC in which each match results in a.]*

Proof. Let $a = (a^*, a^*)$ be a Nash equilibrium such that $\pi(a) = \pi(a^*, a^*) > m$.

Type-interdependent preferences: Consider a population consisting entirely of the $MMDa1$-type. Suppose a mutant enters. A mutant cannot earn more than $\pi(a)$ when deceiving an incumbent. When a mutant is deceived by an incumbent the mutant earns at most m. When a mutant and an incumbent observe each others' preferences the mutant earns at most m. Thus a mutant of cognitive type 1 will earn at most expected payoff m against incumbents. A mutant of higher cognitive type will earn at most $\pi(a)$ against incumbents, and will have to pay a higher cognitive cost than the incumbents.

Type-neutral preferences: Consider a state where all incumbents are of cognitive level 1, and of the same preference type, which prefers to play a^* regardless of what the opponent plays. A mutant could not earn more than an incumbent against an incumbent. Suppose *a single kind of mutant* enters. When a mutant meets a mutant of the same type they play a symmetric profile, and by assumption, there is no symmetric profile that gives a higher payoff than what the incumbents get. [If we allow for *more than one kind of mutant at the same time* then we need to make sure that the the group does not earn more than the incumbents. Unless we assume that $\pi(a)$ is larger than the average payoff in any profile one can construct counter examples in which a group of mutants is able to invade.] ■

Is it possible to say that if (i) a is not a Nash equilibrium, or (ii) a is not efficient, then then there is no NSC (and hence no ESC) in which a is the outcome of every match? Yes this holds for a sufficiently low marginal cost. Let

Proposition 3 *Consider either* ***type-interdependent*** *or* ***type-neutral*** *preferences: Suppose*

$$\kappa^{\max} = \sup_{\psi \in \Psi} \left(k\left(\psi + 1\right) - k\left(\psi\right) \right) < \min_{a_i, a_i', a_j} \left| \pi\left(a_i, a_j\right) - \pi\left(a_i', a_j\right) \right| = \delta.$$

Suppose a is the only outcome in $(x, \omega(x))$. If $(x, \omega(x))$ is NSC then a is a Nash equilibrium.

Proof. We want to prove that $\pi(a^*, a^*) = \pi(a)$ is a Nash equilibrium. To prove this by contradiction assume that a^* is not a best response to itself, i.e. $\max_{a_i} \pi(a_i, a^*) > \pi(a^*, a^*) = \pi(a)$. If all incumbents are of cognitive level ψ^* then the average payoff in the population is $\bar{\Pi} = \pi(a) - k_{\psi^*}$. Consider a materialist mutant θ' of level $\psi' = \psi^* + 1$. For a (vanishingly) small fraction of θ' we have

$$\Pi_{\theta'} = \max_{a_i} \pi\left(a_i, a^*\right) - k_{\psi^*+1} \geq \pi\left(a\right) + \delta - k_{\psi^*+1}.$$

Hence $\Pi_{\theta'} - \bar{\Pi} \geq \delta - (k_{\psi^*+1} - k_{\psi^*})$. By the assumption assumption that $0 < \kappa^{\min} \leq \kappa^{\max} < \delta$. we have $\delta > k_{\psi^*+1} - k_{\psi^*}$. Thus $\Pi_{\theta'} > \bar{\Pi}$. ■

For the case of type-interdependent preferences we are also interested in the following result, which provides a partial converse to proposition 2.

Proposition 4 *Consider* ***type-interdependent*** *preferences. Suppose a is the only outcome in $(x, \omega(x))$. If $(x, \omega(x))$ is ESC, and $\kappa^{\max} < \delta$ as defined in proposition 3, then $\pi(a) > m$.*

Proof. Suppose $(x, \omega(x))$ is an ESC. It follows from proposition 3 that a is NE. If it were the case that $\pi(a) \leq m$ then there are mutants that earn weakly more than the incumbents. ■

Furthermore, for the case of type-neutral preferences, we add the following result, which completes the converse to proposition 2.

Proposition 5 *Consider* ***type-neutral*** *preferences:*

With one kind of mutant at a time*: Suppose a is the only outcome in $(x, \omega(x))$. If $(x, \omega(x))$ is NSC then there is no symmetric profile a' with a higher payoff, i.e. it holds that $\pi(a) \geq \pi(a')$ for all a'.*

*[*With more than one kind of mutant at a time*: Suppose a is the only outcome in* $(x, \omega(x))$*. If* $(x, \omega(x))$ *is NSC then there is no profile with a higher average payoff per player]*

Proof. Suppose x is an NSC. Suppose only a *single type of mutant* at a time is allowed. If there is a symmetric profile $a' = \left(a'_i, a'_j\right)$ such that $\pi(a) < \pi(a')$ then a mutant can invade which prefers to play a_i against a_j and a'_i against a'_j (which is the same thing as prefering a_j against a_i and a'_j against a'_i since the profiles are symmetric). If *more than one kind of mutant* is allowed to enter at the same time then the condition needs to be strenghthened to rule out asymmetric profiles $a' = \left(a'_i, a'_j\right)$ such that $\pi(a) < \pi(a')$. To see this consider two mutants both of whom prefer a_i against a_j. One of them prefers a'_i against a'_j, and the other prefers a'_j against a'_i (which is not the same thing since the profile is asymmetric). ■

Example 9 *Consider the* ***Stag Hunt*** *game*

$$\begin{array}{ccc} & S & H \\ S & 3,3 & 0,1 \\ H & 1,0 & 2,2 \end{array}, \tag{2}$$

where there are two Nash equilibria (S,S) *and* (H,H) *with payoffs* $\pi(S,S) = 3$ *and* $\pi(H,H) = 2$*. The minmax payoff is* 2*. Consider type-neutral preferences. Proposition 2 implies that there is an NSC in which* (S,S) *is the outcome of every match. Propositions 3 and 5 imply that if marginal cognitive cost is sufficiently low then* (S,S) *is the only outcome that can be part of an NSC.*

The above propositions together imply the following characterization.

Corollary 1 *Consider* ***type-neutral*** *preferences: Suppose*

$$\kappa^{\max} = \sup_{\psi \in \Psi} \left(k(\psi + 1) - k(\psi)\right) < \min_{a_i, a'_i, a_j} \left|\pi(a_i, a_j) - \pi(a'_i, a_j)\right| = \delta.$$

With one kind of mutant at a time: Suppose a (symmetric) is the only outcomes in $(x, \omega(x))$*. The configuration* $(x, \omega(x))$ *is NSC if and only if a is a Nash equilibrium and there is no symmetric profile* a' *with a higher payoff, i.e. it holds that* $\pi(a) \geq \pi(a')$ *for all* a'*.*

[With more than one kind of mutant at a time: The configuration $(x, \omega(x))$ *is NSC if and only if a is a Nash equilibrium and there is no profile with a higher average payoff per player]*

Consider ***type-interdependent*** *preferences: Suppose a (symmetric) is the only outcomes in $(x, \omega(x))$. The configuration $(x, \omega(x))$ is ESC if and only if a is a Nash equilibrium and $\pi(a) \geq m$.*

3.3 Results for Multiple Outcomes Configurations

When trying to establish what is implied by a state being ESC (or NSC) we need to distinguish more carefully different levels of cognitive cost. For any configuration $(x, \omega(x))$ let $\delta(x)$ denote the smallest gain that some type can obtain by unilaterly switching action;

$$\delta(x, \omega(x)) = \min_{\theta, \theta'} \left| \max_{a} \pi\left(a, \eta^{\omega(x)}(\theta', \theta)\right) - \pi\left(\eta^{\omega(x)}(\theta, \theta'), \eta^{\omega(x)}(\theta', \theta)\right) \right|.$$

Note that $\delta((x, \omega(x))) > 0$ if and only if $(x, \omega(x))$ induces some outcome that is not a Nash equilibrium. Furhermore let

$$\bar{\kappa}(x, \omega(x)) = \delta(x, \omega(x)) \times \frac{\min_{\theta \in C(x)} x_\theta}{\max_{\theta \in C(x)} \psi}.$$

Recall $\kappa^{\max} = \sup_{\psi \in \Psi}(k_{\psi+1} - k_\psi)$. If $\kappa^{\max}$ is low enough relative to $\bar{\kappa}(x, \omega(x))$ then $(x, \omega(x))$ cannot be an ESC unless it induces only Nash equilibrium outcomes:

Proposition 6 *Consider either* ***type-interdependent*** *or* ***type-neutral*** *preferences: For any $(x, \omega(x))$, if $\kappa^{\max} < \bar{\kappa}(x, \omega(x))$, and if $(x, \omega(x))$ is an NSC (and hence if $(x, \omega(x))$ is ESC), then the outcome of each match is a Nash equilibrium in fitness payoffs, and in every match between two individuals of the same type, a symmetric Nash equilibrium (in fitness payoffs) is played.*

Proof. Suppose we have an NSC $(x, \omega(x))$ in which there are at least two types $\theta' = (\phi', \psi')$ and $\theta'' = (\phi'', \psi'')$, such that θ' does not play a best response (in terms of fitness) against θ'' (either when deceived or when deceiving or both). In this case there is a mutant θ^* with cognitive type $\psi^* = \max_{\theta \in C(x)} \psi + 1$ such that

$$w(\theta^*, \theta'') - w(\theta', \theta'') \geq \delta(x, \omega(x)).$$

Suppose a vanishingly small fraction ε of the mutant θ^* enter the population.[7] At the new state $\tilde{x}$ we have (for $\varepsilon \to 0$),

$$\Pi_{\theta^*}(\tilde{x}) - \Pi_{\theta'}(\tilde{x}) \geq \delta(x, \omega(x)) x_{\theta''} - \kappa^{\max}(\psi + 1 - \psi').$$

[7] Alternatively a group of mutant may enter such that at the post-entry state the fraction $x_{\theta''}$ is unaltered.

Thus if $\delta(x,\omega(x))\,x_{\theta''} > \kappa^{\max}(\psi+1-\psi')$ then the mutant θ^* outperforms the incumbent θ'.

In matches between two individuals of the same type a symmetric profile is played. Hence in all matches between two individuals of the same type a symmetric Nash equilibrium (in fitness payoff) is played. ■

Next we show that for any given configuration of type-interdependent preferences, if we push the marginal cost low enough then we ensure that if the configuration is ESC then in all outcomes in the configuration, all players earn a payoff strictly above the minmax payoff.

Proposition 7 *Consider* ***type-interdependent*** *preferences. For any x, if $\kappa^{\max} < \bar{\kappa}(x,\omega(x))$, and if $(x,\omega(x))$ is an ESC, then in each match both individuals earn strictly more than m.*

Proof. Suppose $\kappa^{\max} < \bar{\kappa}$ and that $(x,\omega(x))$ is an ESC with at least two types $\theta' = (\phi',\psi')$ and $\theta'' = (\phi'',\psi'')$. It follows from proposition 6 that all matches result in Nash equilibria, and in every match between types θ' and θ'', the payoff to type θ' is the same (and the payoff to θ'' is the same). Suppose that (σ^1,σ^2) and (σ^2,σ^1) are the outcomes when θ' and θ'' play. To obtain a contradiction assume $\pi(\sigma^1,\sigma^2) = \pi(\sigma^2,\sigma^1) \leq m$ so that. In this case there is a mutant θ^* with cognitive type $\psi^* = \max_{\theta\in C(x)}\psi + 1$ such that

$$w(\theta^*,\theta'') - w(\theta',\theta'') \geq \delta(x,\omega(x)).$$

Suppose a vanishingly small fraction ε of the mutant θ^* enter the population.[8] At the new state $\tilde{x}$ we have (for $\varepsilon \to 0$),

$$\Pi_{\theta^*}(\tilde{x}) - \Pi_{\theta'}(\tilde{x}) \geq \delta(x,\omega(x))\,x_{\theta''} - \kappa^{\max}(\psi+1-\psi').$$

Thus if $\delta(x,\omega(x))\,x_{\theta''} > \kappa^{\max}(\psi+1-\psi')$ then the mutant θ^* outperforms the incumbent θ'. ■

Similarly, for type-neutral preferences we may show that, for any given configuration, if we push the marginal cost low enough then we ensure that if the configuration is NSC then in all outcomes between individuals of the same type, one of the symmetric Nash equilibria with the highest payoff is played.

[8] Alternatively a group of mutant may enter such that at the post-entry state the fraction $x_{\theta''}$ is unaltered.

Proposition 8 *Consider* ***type-neutral*** *preferences. For any* x*, if* $\kappa^{\max} < \bar{\kappa}(x, \omega(x))$*, and if* $(x, \omega(x))$ *is an NSC, then in every match between two individuals of the same type, a symmetric Nash equilibrium (in fitness payoffs) is played, and there is no symmetric Nash equilibrium (in fitness payoffs) that gives a higher payoff.*

Proof. Suppose $\kappa^{\max} < \bar{\kappa}$ and $(x, \omega(x))$ is an NSC. It follows from proposition 6 that the outcome of each match is a Nash equilibrium in fitness payoffs, and in every match between two individuals of the same type, a symmetric Nash equilibrium (in fitness payoffs) is played. Suppose that there is a type $\theta \in C(x)$ which plays a symmetric Nash equilibrium σ when it meets someone of the same type. To obtain a contradiction assume that there is another symmetric Nash equilibrium $\tilde{\sigma} = (\tilde{\sigma}^*, \tilde{\sigma}^*)$ such that $\pi(\sigma) = \pi(\sigma^*, \sigma^*) < \pi(\tilde{\sigma}^*, \tilde{\sigma}^*) = \pi(\tilde{\sigma})$. Consider a mutant θ' with $\psi' = \psi$ and with materialistic preferences. Since it has materialistic preferences it is able to behave exactly like type θ' against all incumbent types (because each outcome in x is a Nash equilibrium in fitness payoffs). Thus we have $w(\theta', \theta'') = w(\theta, \theta'')$ for all $\theta'' \in C(x)$. If the mutants play $\tilde{\sigma}$ against each other then $\pi(\tilde{\sigma}) = w(\theta', \theta') > w(\theta, \theta') = \pi(\sigma)$. Hence for a fraction ε of mutants we have

$$\begin{aligned}\Pi_{\theta'} - \Pi_{\theta} &= (w(\theta', \theta) - w(\theta, \theta)) x_{\theta} + (w(\theta', \theta') - w(\theta, \theta')) x_{\theta'} \\ &= (\pi(\sigma) - \pi(\sigma)) x_{\theta} + (\pi(\tilde{\sigma}) - \pi(\sigma)) x_{\theta'} \\ &= (\pi(\tilde{\sigma}) - \pi(\sigma)) x_{\theta'} > 0.\end{aligned}$$

■

If the underlying game is a coordination game such that the highest payoff a player can obtain is obtained in a symmetric Nash equilibrium, then such an efficient symmetric nash equilibrium is played in every match:

Proposition 9 *Consider* ***type-neutral*** *preferences. Suppose that there is a symmetric Nash profile* σ *such that* $\sigma \in \arg\max_{\sigma' \in \Delta} \pi(\sigma')$*. For any* x*, if* $\kappa^{\max} < \bar{\kappa}(x, \omega(x))$*, and if* $(x, \omega(x))$ *is an NSC, then in every match between two individuals of the same type, both players earn* $w^e = \arg\max_{\sigma' \in \Delta} \pi(\sigma')$*.*

Proof. Consider two types θ and θ'. Let w^e denote the highest payoff that can be obtained by a player, i.e. $w^e = \arg\max_{\sigma' \in \Delta} \pi(\sigma')$. By the assumption of the proposition this payoff can be obtained in a symmetric profile. Suppose we are in an ESC $(x, \omega(x))$. By definition $\Pi_{\theta'}(x) = \Pi_{\theta}(x)$. By the above lemma we know that $w(\theta, \theta) = w(\theta', \theta') = w^e$.

To obtain a contradiction suppose $w(\theta,\theta') < w^e$ and $w(\theta',\theta) < w^e$. If x_θ is increased at the expense of $x_{\theta'}$, and the fraction of all other types is kept constant, then $\Pi_{\theta'} < \Pi_\theta$. This implies that $(x,\omega(x))$ is not an ESC. We have established that $w(\theta,\theta') = w^e$ or $w(\theta',\theta) = w^e$.

To obtain a contradiction suppose $w(\theta,\theta') < w^e$ and $w(\theta',\theta) = w^e$. If $x_{\theta'}$ is increased at the expense of x_θ, and the fraction of all other types is kept constant, then $\Pi_{\theta'} > \Pi_\theta$. This implies that $(x,\omega(x))$ is not an ESC. We have established that $w(\theta,\theta') = w^e$ and $w(\theta',\theta) = w^e$. ■

We can say the following for any cost level:

Proposition 10 *Consider* ***type-neutral*** *preferences.*

With one kind of mutant at a time: For any x, if $(x,\omega(x))$ is an NSC, and σ is the (symmetric) outcome when two individuals of type $\theta \in C(x)$ meet each other, then there is no symmetric profile σ' with a higher payoff, i.e. it holds that $\pi(\sigma) \geq \pi(\sigma')$ for all symmetric σ'.

[With more than one kind of mutant at a time: For any x, if $(x,\omega(x))$ is an NSC, and σ is the (symmetric) outcome when two individuals of type $\theta \in C(x)$ meet each other, then there is no profile with a higher average payoff per player.]

Proof. Suppose only *a single type of mutant at a time* is allowed. Suppose x is an NSC in which type θ plays the symmetric profile σ. If there is a symmetric profile $\sigma' = (\sigma'_i, \sigma'_j)$ such that $\pi(\sigma) < \pi(\sigma')$ then a mutant θ' can invade, which is of the same cognitive type θ, is indifferent between all actions, and behaves exactly like θ, except when meeting someone of its own type θ', in which case it would play σ'.

If *more than one kind of mutant* is allowed to enter at the same time then the condition needs to be strenghthened to rule out asymmetric profiles $\sigma' = (\sigma'_i, \sigma'_j)$ such that $\pi(\sigma) < \pi(\sigma')$. To see this consider two mutants both of the same cognitve level as θ, and both indifferent between all actions. One of them plays σ'_i against σ'_j, and the other plays σ'_j against σ'_i (which is not the same thing since the profile is asymmetric). ■

4 Discussion

To be added.

References

Alger, I. and Weibull, J. W. (2013), 'Homo moralis, preference evolution under incomplete information and assortative matching', *Econometrica* **81**(6), 2269–2302.

Banerjee, A. and Weibull, J. W. (1995), Evolutionary selection and rational behavior, *in* A. Kirman and M. Salmon, eds, 'Learning and Rationality in Economics', Blackwell, Oxford, UK, chapter 12, pp. 343–363.

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

Bolle, F. (2000), 'Is altruism evolutionarily stable? and envy and malevolence? remarks on bester and güth', *Journal of Economic Behavior and Organization* **42**, 131–133.

Dekel, E., Ely, J. C. and Yilankaya, O. (2007), 'Evolution of preferences', *Review of Economic Studies* **74**, 685–704.

Dunbar, R. I. M. (1998), 'The social brain hypothesis', *Evolutionary Anthropology* **6**, 178–190.

Gul, F. and Pesendorfer, W. (2010), 'Interdependent preference models as a theory of intentions'. Manuscript.

Güth, W. and Napel, S. (2006), 'Inequality aversion in a variety of games - an indirect evolutionary analysis', *The Economic Journal* **116**, 1037–1056.

Güth, W. and Yaari, M. E. (1992), Explaining reciprocal behavior in simple strategic games: An evolutionary approach, *in* U. Witt, ed., 'Explaining process and change:', University of Michigan, Ann Arbor, pp. 22–34.

Heller, Y. (2013), Three steps ahead. Manuscript.

Herold, F. and Kuzmics, C. (2009), 'Evolutionary stability of discrimination under observability', *Games and Economic Behavior* **67**, 542–551.

Hines, W. G. S. and Smith, J. M. (1979), 'Games between relatives', *Journal of Theoretical Biology* **79**(1), 19–30.

Holloway, R. (1996), Evolution of the human brain, *in* A. Lock and C. R. Peters, eds, 'Handbook of Human Symbolic Evolution', Clarendon, Oxford, pp. 74–125.

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

Humphrey, N. K. (1976), The social function of intellect, *in* P. P. G. Bateson and R. A. Hinde, eds, 'Growing Points in Ethology', Cambridge University Press, Cambridge, pp. 303–317.

Kinderman, P., Dunbar, R. I. M. and Bentall, R. P. (1998), 'Theory-of-mind deficits and causal attributions', *British Journal of Psychology* **89**, 191–204.

Koçkcesen, L. and Ok, E. A. (2000), 'Evolution of interdependent preferences in aggregative games', *Games and Economic Behavior* **31**, 303–310.

Maynard Smith, J. and Price, G. R. (1973), 'The logic of animal conflict', *Nature* **246**(5427), 15–18.

Mohlin, E. (2010), 'Internalized social norms in conflicts: An evolutionary approach', *Economics of Governance* **11**(2), 169–181.

Mohlin, E. (2012), 'Evolution of theories of mind', *Games and Economic Behavior* **75**(1), 299–312.

Ok, E. A. and Vega-Redondo, F. (2001), 'On the evolution of individualistic preferences: An incomplete information scenario', *Journal of Economic Theory* **97**, 231–254.

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

Premack, D. and Wodruff, G. (1979), 'Does the chimpanzee have a theory of mind', *Behavioral and Brain Sciences* **1**, 515–526.

Robson, A. J. and Samuelson, L. (2011), The evolutionary foundations of preferences, *in* J. Benhabib, A. Bisin and M. Jackson, eds, 'The Social Economics Handbook', North Holland, pp. 221–310.

Schaffer, M. E. (1988), 'Evolutionarily stable strategies for finite population and a varable contest size', *Journal of Theoretical Biology* **132**, 469–478.

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

Stahl, D. O. (1993), 'Evolution of smart$_n$ players', *Games and Economic Behavior* **5**(4), 604–617.

Stennek, J. (2000), 'The survival value of assuming others to be rational', *International Journal of Game Theory* **29**, 147–163.

Taylor, P. D. and Jonker, L. B. (1978), 'Evolutionary stable strategies and game dynamics', *Mathematical Biosciences* **40**(1-2), 145 – 156.

5 Appendix

5.1 Example in which not all ESC outcomes are Nash (Type-Interdependent Preferences)

We define the *discriminating materialists*:

Definition 10 *The set of DM-types, Θ^{DM}, is the set of types such that for $\theta_i = (\phi_i, \psi_i) \in \Theta^{DM}$:*

(a) If $\phi_i \neq \phi_j$ then $u^{\phi_i}(a^m, a_j, \phi_j) > u^{\phi_i}(a_i, a_j, \phi_j)$ for all a_i and a_j.

(b) If $\phi_i = \phi_j$ then $u^{\phi_i}(a_i, a_j, \phi_j) = \pi(a_i, a_j)$ for all a_i and a_j.

The type which combines ϕ-preferences with cognitive type ψ, is denoted $\phi\psi$. When there is no risk of confusion we will simply write $\phi\psi$. We assume:

$$\{M\psi\}_{\psi\in\Psi} \cup \{DM\psi\}_{\psi\in\Psi} \cup \{MMDa\psi\}_{\psi\in\Psi, a\in A\times A} \subseteq \Theta.$$

To prove the result we want for type-interdependent preferences we first need to prove a lemma regarding ESCs involving multiple types. Consider a set of types $\{(T,i)\}_{i=1}^{J}$, for arbitrarily large (but finite) J. Let

$$w(T\psi, T\psi') = \begin{cases} t & \text{if } \psi > \psi' \\ w & \text{if } \psi = \psi' \\ s & \text{if } \psi < \psi' \end{cases}.$$

Thus t is the payoff that a player of type T earns when deceiving an opponent of type T, and s is the payoff earned by the deceived party. When two individuals of the same type meet they earn w. The following result can be proved:

Lemma 1 *Suppose*

$$t - w > k_{i+1} - k_i \text{ for all } i. \tag{3}$$

(i) If $2w < s + t$ there is an ESC $(x^*, \omega^*(x^*))$, such that x^* is mixed, with $\sum_{i=1}^{I} x^*_{Ti} = 1$.

(ii) If $2w = s + t$ there is an NSC $(x^*, \omega^*(x^*))$, such that x^* is mixed, with $\sum_{i=1}^{I} x^*_{Ti} = 1$.

(iii) If $2w > s + t$ there is no NSC and hence no ESC. The replicator dynamic converges to the boundary of the face where $\sum_{i=1}^{I} x^*_{Ti} = 1$.

Proof. Available request. ■

We begin constructing an ESC by limiting attention to a set of discriminating materialist types $\{(DM,\psi)\}_{\psi=1}^{J}$, for some finite J. For brevity we denote the type (DM,i) by DMi. Let

$$t^m = \max_{a',a''} \pi\left(a', \arg\max_{a} \pi\left(a,a''\right)\right),$$
$$s^m = \pi\left(\arg\max_{a} \pi\left(a,\hat{a}\right),\bar{a}\right), \text{ for } (\bar{a},\hat{a}) = \arg\max_{a',a''} \pi\left(a', \arg\max_{a} \pi\left(a,a''\right)\right).$$

or, since fitness payoffs are assumed to be generic,

$$t^m = \max_{a'} \pi\left(\beta\left(\beta\left(a'\right)\right),\beta\left(a'\right)\right),$$
$$s^m = \pi\left(\beta\left(\hat{a}\right),\beta\left(\beta\left(\hat{a}\right)\right)\right), \text{ for } \hat{a} = \arg\max_{a'} \pi\left(\beta\left(\beta\left(a'\right)\right),\beta\left(a'\right)\right).$$

The payoff t^m is the highest fitness that a player can achieve against a materialist player. In this case the deceived party earns fitness s^m. Moreover, let w_{mm} be the payoff the symmetric Nash equilibrium that two individuals of type DM play in the case they both observe each other's preferences. Thus

$$w\left(DM\psi, DM\psi'\right) = \begin{cases} t^m & \text{if } \psi > \psi' \\ w_{mm} & \text{if } \psi = \psi' \\ s^m & \text{if } \psi < \psi' \end{cases}.$$

The payoffs are

	$DM1$	$DM2$	$DM3$	$\ldots$	$DMI-1$	DMI
$DM1$	$w_{mm}-k_1$	s^m-k_1	s^m-k_1	$\ldots$	s^m-k_1	s^m-k_1
$DM2$	t^m-k_2	$w_{mm}-k_2$	s^m-k_2	$\ldots$	s^m-k_2	s^m-k_2
$DM3$	t^m-k_3	t^m-k_3	$w_{mm}-k_3$	$\ldots$	s^m-k_3	s^m-k_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$DMI-1$	t^m-k_{I-1}	t^m-k_{I-1}	t^m-k_{I-1}	$\ldots$	$w_{mm}-k_{I-1}$	s^m-k_{I-1}
DMI	t^m-k_I	t^m-k_I	t^m-k_I	$\ldots$	t^m-k_I	$w_{mm}-k_I$

If the type set is $\Theta = \{(DM,\psi)\}_{\psi=1}^{J}$ then lemma 1 gives conditions under which there is an ESC involving multiple types.

The following lemma provides a condition for when such an ESC obtains also with a larger type set.

Lemma 2 *If $2w_{mm} > t^m + m$ and $m \leq s^m$ then any sufficiently small group of mutants not belonging to $\{(DM,i)\}_{i=1}^{I}$ that enter a population consisting of types $\{(DM,i)\}_{i=1}^{I}$ is strictly outperformed.*

Proof. Suppose a type $\theta = (\phi, \psi) \notin \{(DM, i)\}_{i=1}^{I}$ with $\psi \leq I$, enters a population consisting of types $\{(DM, i)\}_{i=1}^{I}$. Against $DM\psi'$ with $\psi' < \psi$, the highest fitness that θ can earn is the same as a materialist could earn, i.e. t^m. Against $DM\psi'$ with $\psi' > \psi$, the highest fitness that θ can earn is the minmax payoff m, since $DM\psi'$ minmaxes anyone who is not in $\{(DM, i)\}_{i=1}^{I}$. If $m \leq s^m$ then this is weakly less than what $DM\psi$ earns against $DM\psi'$, when $\psi' > \psi$. Against $DM\psi'$ with $\psi' = \psi$, the highest fitness that θ can earn is $\frac{1}{2}(t^m + m)$. If $w_{mm} > \frac{1}{2}(t^m + m)$ then this is strictly less than what $DM\psi$ earns against $DM\psi'$, when $\psi' = \psi$. Thus in any game with $w_{mm} > \frac{1}{2}(t^m + m)$ and $m \leq s^m$ the type $\theta = (\phi, \psi)$ is strictly outperformed by the $DM\psi$ type. Furthermore if a mutant with $\psi > I$ enters then by the definition of I she earns less than $M1$. ■

Using lemmas 1, and 2, we have:

Proposition 11 *Consider a set of types including $\{(DM, \psi)\}_{\psi=1}^{J}$. Suppose (3) and $2w_{mm} > t^m + m$ and $m \leq s^m$ hold.*

(i) If $2w_{mm} < s^m + t^m$ (i.e. if $s^m + t^m > 2w_{mm} > t^m + m$) there is an ESC $(x^*, \omega^*(x^*))$, such that x^* is mixed, with $\sum_{i=1}^{I} x^*_{DMi} = 1$.

(ii) If $2w_{mm} = s^m + t^m$ (i.e. if $s^m + t^m = 2w_{mm} > t^m + m$) there is an NSC $(x^*, \omega^*(x^*))$, such that x^* is mixed, with $\sum_{i=1}^{I} x^*_{DMi} = 1$.

(iii) If $2w_{mm} > s^m + t^m$ (i.e. if $s^m + t^m < 2w_{mm} > t^m + m$) there is no NSC and hence no ESC. The replicator dynamic converges to the boundary of the face where $\sum_{i=1}^{I} x^*_{DMi} = 1$.

Proof. For the type set $\{(DM, \psi)\}_{\psi=1}^{J}$ this follows from lemma 1. By lemma 2 all other types will earn strictly less than the incumbents. ■

Example 11 *The **Prisoners' Dilemma with outside option** (PDO) has payoff matrix*

$$\begin{array}{ccccc} & C & D & O \\ C & b, b & s, t & 0, 0 \\ D & t, s & f, f & 0, 0 \\ O & 0, 0 & 0, 0 & 0, 0 \end{array}, \tag{4}$$

with $t > b > 0, f > s \geq 0$. Consider the types $\{(DM, \psi)\}_{\psi=1}^{J}$. In the case of mutual observation the Nash equilibrium (D, D) is played. Thus $w_{mm} = f$, $t^m = t$, and $s^m = s \geq m = 0$. If $s + t > 2f > t$ then $s^m + t^m > 2w_{mm} > t^m + m$. Thus we are in case (i) of the above proposition. There is an ESC x^*, such that x^* is mixed, with $\sum_{i=1}^{I} x^*_{DMi} = 1$. The only outcomes that will be observed in matches with individuals of different types are (D, C), and (C, D).