Coevolution of Deception and Preferences: Darwin and Nash Meet Machiavelli

Heller, Yuval and Mohlin, Erik

Univeristy of Oxford

30 August 2014

Coevolution of Deception and Preferences: Darwin and Nash Meet Machiavelli*

Yuval Heller[†] and Erik Mohlin[‡]

24/02/15

Abstract

We develop a framework in which individuals' preferences co-evolve with their abilities to deceive others regarding their preferences and intentions. Specifically, individuals are characterized by a level of cognitive sophistication and by a subjective utility function. Increased cognition is costly, but higher level individuals are able to completely deceive lower level opponents about their preferences and intentions. Only individuals who are of the same cognitive level can observe each others preferences. Our main results show that, despite the limited possibility to observe preferences, and despite the strong form of deception, only efficient population states can be stable. We extend our model to study preferences that depend also on the opponent's type.

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

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 two decades this has changed dramatically. In particular, there is now a large literature on the evolutionary foundations of preferences (for an overview, see Robson & Samuelson, 2011). A prominent strand of this literature is the so-called "indirect evolutionary approach", pioneered by Güth & Yaari (1992) (term coined by Güth, 1995). This approach has been used to explain the existence of a variety of "non-standard" preferences that do not coincide with material payoffs, e.g. altruism, spite, and reciprocal preferences.[1] Typically, the non-materialistic preferences in question convey some form of commitment advantage that induces opponents to behave in a way that benefits individuals with non-materialistic preferences, as described by Schelling (1960) and Frank (1987). Indeed, Heifetz *et al.* (2007) show that this kind of result is generic.

*Valuable comments were provided by Vince Crawford, Itzhak Gilboa, Larry Samuelson, and Jörgen Weibull, as well as participants at presentations in Oxford, at G.I.R.L.13 in Lund, the Toulouse Economics and Biology Workshop, DGL13 in Stockholm, and the 25th International Conference on Game Theory at Stony Brook. Erik Mohlin was supported in part by the European Research Council, grant no. 230251.

[†]Affiliation: Queen's College and Department of Economics, University of Oxford. Address: Queen's College, High Street, Oxford, OX1 1AW, United Kingdom. E-mail: yuval.heller@queens.ox.ac.uk.

[‡]Affiliation: Nuffield College and Department of Economics, University of Oxford. Address: Nuffield College, New Road, Oxford OX1 1NF, United Kingdom. E-mail: erik.mohlin@nuffield.ox.ac.uk.

[1]For example, Bester & Güth (1998), Bolle (2000), and Possajennikov (2000) study combinations of altruism, spite, and selfishness. Ellingsen (1997) finds that preferences that induce aggressive bargaining can survive in a Nash demand game. Fershtman & Weiss (1998) study evolution of concerns for social status. Sethi & Somanthan (2001) study the evolution of reciprocity in the form of preferences that are conditional on the opponent's preference type. In the context of the finitely repeated Prisoner's Dilemma, Guttman (2003) explores the stability of conditional cooperation. Dufwenberg & Güth (1999) study firm's preferences for large sales. Güth & Napel (2006) study preference evolution when players use the same preferences in both ultimatum and dictator games. Koçkcesen & Ok (2000) investigate survival of more general interdependent preferences in aggregative games. Friedman & Singh (2009) show that vengefulness may survive if observation has some degree of informativeness.

A crucial feature of the indirect evolutionary approach is that preferences are explicitly or implicitly assumed to be at least partially observable.[2] Consequently the results are vulnerable to the existence of mimics who signal that they have, say, a preference for cooperation, but actually defect on cooperators, thereby earning the benefits of having the non-standard preference without having to pay the cost (Samuelson, 2001). The effect of varying the degree to which preferences can be observed has been investigated by Ok & Vega-Redondo (2001), Ely & Yilankaya (2001), Dekel *et al.* (2007), and Herold & 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 them to be the product of an evolutionary process.[3] *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. Mutual observation of preferences only occurs whether players of the same cognitive level meet. Our main result is that such a "grain" of perfect observability is sufficient to imply efficient play in any stable population state.*

Overview of the Model. As in standard evolutionary game theory we assume an infinite population of individuals who are uniformly randomly matched to play a symmetric normal form game.[4] Each individual has a type, which is a tuple, consisting of a *preference component* and a *cognitive component*. The preference component is identified with a utility function over the set of outcomes (i.e. action profiles). In an extension we allow for *type-interdependent preferences*, which are represented by utility functions that are defined over both action profiles and the opponent's type. The cognitive component is simply a natural number, representing the level of cognitive sophistication of the individual.[5] The cost of increased cognition is strictly positive.

When the individuals in a match are of *different* cognitive levels, the one with the higher level is assumed to be able to deceive the one with the lower level. For tractability reasons, and in order to "stack the cards" against our main result, we model a strong form of deception. The deceiver observes the opponent's preferences perfectly, and is allowed to choose whatever she wants the deceived party to believe about the deceiver's intended action choice. A strategy profile that is consistent with this form of deception is called a *deception equilibrium*. When both individuals are of the *same* cognitive level, we assume that each player observes the opponent's preferences, and as a result, the individuals play a Nash equilibrium of the complete information game induced by their preferences.[6]

[2]Gamba (2013) is an interesting exception. She assumes play of a self-confirming equilibrium, rather than a Nash equilibrium, in an extensive form game. This allows for evolution of non-materialistic preferences even when they are completely unobservable. An alternative is to allow for a dynamic that is not strictly payoff monotonic. This approach is pursued by Frenkel *et al.* (2014), who show that multiple biases (inducing non-materialistic preferences) can survive in plausible non-monotonic evolutionary dynamics even if they are unobservable, because each approximately compensates for the errors of the others.

[3]On this topic, Robson & 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 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)

[4]It is known that positive assortative matching is conducive to the evolution of altruistic behaviour (Hines & Maynard Smith, 1979) and non-materialistic preferences even when preferences are perfectly unobservable (Alger & Weibull, 2013). It is also known that finite populations allow for the evolution of spiteful behaviours (Schaffer, 1988) and non-materialistic preferences (Huck & Oechssler, 1999). By assuming that individuals are uniformly randomly matched in an infinite population, we avoid confounding these effects with the effect of endogenising the degree of observability.

[5]The one-dimensional representation of cognitive ability reflects the idea that if one is good at deceiving others, then one is more likely to be good also at reading others and avoiding to be deceived by them. In this paper we simplify this relation by assuming a perfect correlation between the two abilities, and leave the study of more general relations for future research.

[6]Our assumption that players of equal levels play Nash equilibrium relies on the, so-called, "folk-theorem of evolutionary game theory", which implies that any stable population state must be a Nash equilibrium of the underlying complete information game

The state of a population is described by a *configuration*, consisting of a type distribution and a behaviour policy. The *type distribution* is simply a finite support distribution on the set of types. The *behaviour policy* specifies a Nash equilibrium for each match between cognitive equals, and a deception equilibrium for each match between types of different cognitive levels. In a *neutrally stable configuration* all incumbents earn the same, and if a small group of mutants enter they earn weakly (strictly) less than the incumbents in any *focal* post-entry state. A focal post-entry state is one in which the incumbents behave against each other in the same way as before the mutants entered.

Results. We first show that the behavior of type $\bar{\theta}$ with the highest cognitive level in the incumbent population (henceforth, "highest type") in a stable configuration satisfies three properties (Theorem 1). First, type $\bar{\theta}$ plays efficiently (i.e. maximizing the fitness) when meeting itself. The intuition is that otherwise a highest-type mutant who mimics the play of $\bar{\theta}$ against all incumbents while playing efficiently against itself, would outperform type $\bar{\theta}$ (an application of a "secret-handshake" argument, Robson, 1990). The second result is that $\bar{\theta}$ maximizes its own fitness against all "lower" types (i.e. types with lower cognitive levels). The intuition is that otherwise a highest-type mutant who mimics the play of $\bar{\theta}$ against most incumbents, except that he maximizes the fitness against one of the lower types, would outperform type $\bar{\theta}$. Finally we show that a lower type cannot achieve more than the efficient level of fitness when being matched against $\bar{\theta}$. The intuition is that otherwise a highest-type mutant who mimics this lower type when being matched with $\bar{\theta}$, while mimicking $\bar{\theta}$ against all other opponents, would outperform type $\bar{\theta}$.

Next we restrict attention to generic games (i.e., each fitness payoff is independently drawn from a continuous distribution) and obtain our main result: *any stable configuration must induce efficient play* in all matches between all types. The proof can be brief sketched as follows. We first show that any type θ in a stable configuration must play efficiently when meeting *itself*. This is because otherwise a mutant could enter, which has the same level as θ and has the same utility function as θ, except that the efficient action, $\bar{a}$ is a subjective best-reply to itself. We show that there exists a focal post-entry state in which the mutant plays the same as the incumbent θ except playing the efficient profile$(\bar{a}, \bar{a})$ against itself and possibly against some of the higher types. Next, we show that *any* two types must play efficiently. The intuition is that otherwise the mean within-groups fitness is higher than the between-groups fitness, which implies instability to small perturbations in the frequency of the types (a type who becomes slightly more frequent will be have higher fitness relative to the other incumbents, and this will take the population away).

The existing literature (e.g., Dekel *et al.*, 2007) shows that if players perfectly observe the opponent's preferences (or do so with sufficiently high probability), then only efficient outcomes are stable. *Our key contribution is showing that a "grain" of perfect observability is enough to imply efficiency.* While, the existing models assume that any player perfectly observe the preferences of any other player, our model only assumes perfect observability as a "tie-breaking" rule: players with equal-levels s to perfectly observe each other's preferences, but by acquiring a higher cognitive level (which may incur an arbitrarily low cognitive cost), a player can completely deceive the opponent about his preferences.

Next we make a few observations, which allow us to fully characterize stable configurations in generic games. First, we observe that any generic game admits at most one efficient action, and that this action is either strict equilibrium or not a best-reply to itself. In the former case, we show the stability of an homogenous population players with the lowest cognitive level and with preferences such that the efficient action is dominant. Moreover, if the cognitive cost required for achieving the second cognitive level is sufficiently low, then this stable configuration is essentially unique. In any other case (i.e. the underlying game admits no efficient action, or the efficient action is not a best-reply to itself), there is no stable configuration.

for many evolutionary dynamics (see Nachbar, 1990 for a formal statement and proof).

Finally, we note that non-generic games may admit different kinds of stable configurations. One particular interesting family of non-generic games is the family of zero-sum games, such as the Rock-Paper-Scissors game (Example **??**). We analyse this game and characterize an heterogeneous stable population in which different cognitive levels co-exist, players with equal levels play the Nash equilibrium, and players with higher levels beat their opponents (but this gain is compensated by the higher cognitive costs). The construction is similar to that of Conlisk (2001).

Variants and Extensions. As mentioned above, our main result relies on having full observability as the "tie-breaking" rule. In Section 5 we consider the opposite assumption, namely, that players with equal level do not observe each others preferences. We show that this variant yields very different results. Specifically, we show that (1) any pure strict equilibrium (possibly, non-efficient) of the underlying game is stable, and (2) the highest type always play a Nash equilibrium (again, not necessarily efficient) when being matched against themselves.

Thus, whether stability implies efficiency or Nash equilibrium behavior, crucially depends on "grain" of exogenous observability assumption in the matches between equals in our model. The existing literature (e.g., Ok & Vega-Redondo (2001); Ely & Yilankaya (2001); Dekel *et al.* (2007); Norman (2012)) show that if players cannot observe each other's preferences, then only Nash equilibria can be stable. *Our results show that a "grain" of non-observability among players with equal levels is enough to imply strong relations between stability and Nash equilibria*, even though our setup allows a player to spend (possibly arbitrarily low) additional cognitive cost, obtains a higher level than his opponent, and fully observe his preferences.

In most of the paper we deal only with "type-neutral" preferences that depends only over action profiles. Section 6 extends the analysis to interdependent preferences that may also depend on the opponent's type. Herold & Kuzmics (2009) study a similar setup while assuming perfect observability of types among all players. Their key result shows that any mixed action that yields each player a payoff above his maxmin payoff, can be the outcome of a stable configuration. We show a slightly weaker extension of this result while assuming only perfect observability among players with equal levels. Specifically, we show that pure stable population states (i.e. populations in which everyone play the same pure action) are essentially Nash equilibria that yield each player a payoff above the maxmin value. We conclude by characterizing stable configurations in the "Hawk-Dove" game (Section 6.4). We show that such games admit heterogeneous stable configurations in which players with different levels co-exist, each type has discriminating preferences that induce cooperation only against itself, and higher types "exploit" lower types (and this is compensated by their higher cognitive levels.)

In Appendix B we briefly present another variant of the model in which the deceiver is *not* able to tailor the attempted deception to the current opponent's type. Instead, an individual has to use the same attempted deception against all opponents. One can show that qualitatively similar results also hold with this less flexible form of deception.

Further related literature. "Ok & Vega-Redondo (2001) and Ely & Yilankaya (2001) investigate the case in which preferences are unobservable, and all preferences defined over outcomes are allowed. They show that only Nash equilibria of the game with material/fitness payoffs can be implemented by evolutionarily stable preferences. More generally, Dekel *et al.* (2007) study environments in which there is a fixed probability that a player observes the preferences of the opponent. They confirm the previous results for unobservable preferences. Furthermore, they show that if preferences are perfectly, or almost perfectly, observable, then only efficient outcomes can be supported by neutrally stable preferences.[7] Our results indicate that when deception is introduced and observation is endogenised, then a pure profile has to be *both* Nash and efficient in order to be the sole outcome supported by neutrally stable preferences. Herold & Kuzmics (2009) expand the framework of Dekel *et al.* (2007)

[7]See Norman (2012) for related results in a dynamic model.

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 evolutionarily stable. In our setting a pure profile also has to be a Nash equilibrium in order to be the sole outcome supported by evolutionarily stable preferences. Herold & Kuzmics (2009) find that non-discriminating preferences (including selfish materialistic preferences) are typically not evolutionarily stable on their own. In contrast, certain preferences that exhibit discrimination are evolutionarily stable. Similarly, evolutionary stability requires the presence of discriminating preferences also in our setup."

There is a large literature in biology and evolutionary psychology on the evolution of 'theory of mind' (Premack & Woodruff 1979). According to the "Machiavellian intelligence" hypothesis (Humphrey, 1976), and "social brain" hypothesis (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). Our model incorporates these features.

There is a smaller literature on the evolution of strategic sophistication within game theory; see, e.g. Stahl (1993), Banerjee & Weibull (1995), Stennek (2000), Conlisk (2001), Abreu & Sethi (2003), Mohlin (2012), Rtischev (2012), and Heller (2015). As in these papers, we provide results to the effect that different degrees of cognitive sophistication may co-exist.

Kimborough *et al.* (2014) construct a model to demonstrate the advantage of having a theory of mind (understood as an ability to ascribe stable preferences to other players) over learning by reinforcement. In novel games the ascribed preferences allow the agents with a theory of mind to draw on past experience whereas a reinforcement learner without such a model has to start over again. Hopkins (2014) explains why costly signaling of altruism may be especially valuable for those agents who have a theory of mind.

Robson (1990) initiated a literature on evolution in cheap talk games by formulating the secret handshake effect: evolution selects an efficient stable state if mutants can send messages that the incumbents either do not see or not benefit from seeing. Against the incumbents a mutant plays the same action as the incumbents do, but against other mutants the mutant plays an action that is a component of the efficient equilibrium. Thus the mutants are able to invade unless the incumbents are already playing efficiently. See also Matsui (1991). As pointed out by Wärneryd (1991), and Schlag (1993), among others, problems arise if either the incumbents use all available messages (so that there is no message left for the incumbents to coordinate on) or the incumbents follow a strategy that induces the mutants to play an action that lowers the mutants' payoffs below those of the incumbents. Kim & Sobel (1995) use stochastic stability arguments, and Wärneryd (1998) uses complexity costs, to circumvent this problem. Similarly, evolution selects efficient outcome in our model, where the preferences also serve the function of messages

Structure. The rest of the paper is organized as follows. Section 2 presents the model. In Section 3 we define our stability notions. The results for the main model are presented in Section 4. Section 5 deals with a variant in which a player cannot observe the preferences of an opponent with the same cognitive level. 6 extends the model to include type-interdependent preferences. In Section 1 we discuss related literature. Section 7 concludes. Appendix A contains proofs not in the main text. Appendix B presents a variant of the model with uniform deception. Appendix C formally constructs heterogeneous stable populations in Rock-Scissor-Paper and Hawk-Dove games.

2 Model

We consider a large population of agents, each of which is endowed with a type that determines her subjective preferences and her cognitive level. The agents are randomly matched to play a symmetric two-player game. A dynamic evolutionary process of cultural learning, or biological inheritance, increases the frequency of more successful types. In the next section, we present a static solution concept to capture stable population states in such environments.

2.1 Underlying Game

Consider a symmetric two-player normal form game G with a finite set A of pure actions and a set $\Delta(A)$ of mixed actions (or strategies). We use the letter a (σ) to describe a typical pure action (mixed action). Payoffs are given by $\pi : A \times A \to \mathbb{R}$, where $\pi(a, a')$ is the payoff to a player using action a against action a'. The payoff function is extended to mixed actions in the standard way, where $\pi(\sigma, \sigma')$ denotes the material payoff to a player using strategy σ, against an opponent using strategy σ'. With a slight abuse of notation let a denote the degenerate mixed strategy that puts all weight on pure strategy a. We adopt this convention for probability distributions throughout the paper.

Remark 1. The restriction to symmetric games is without loss of generality when dealing with interactions in a single population. In cases in which the interaction is asymmetric, it can be captured in our setup (as is standard in the literature; see, e.g. Selten, 1980 and Samuelson, 1991) by embedding the asymmetric interaction in a larger, symmetric game in which nature first randomly assigns the players to roles in the asymmetric interaction.

2.2 Types

We imagine a large (technically infinite) population of individuals who are uniformly randomly matched to play the game G. Each individual i in the population is endowed with a *type*

$$\theta = (u, n) \in \Theta = U \times \mathbb{N},$$

consisting of (von Neumann-Morgenstern) *preferences*, identified with a utility function, $u \in U$ and a *cognitive level* $n \in \mathbb{N}$. Let $\Delta(\Theta)$ be the set of all finite support probability distributions on Θ. A population is represented by a finite support *type distribution* $\mu \in \Delta(\Theta)$. Let $C(\mu)$ denote the support (carrier) of type distribution $\mu \in \Delta(\Theta)$. Elements of $C(\mu)$ will be called incumbents. Given a type θ, we use u_θ and n_θ to refer to its preferences and cognitive level, respectively.

In the main model we assume that the preferences are defined over action profiles, as in Dekel *et al.* (2007). This means that any preferences can be represented by a utility function of the form

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

The set of all possible (modulo affine transformations) utility functions on $A \times A$ is $U = [0, 1]^{|A|^2}$. Let $BR_u(\sigma')$ denote the set of best replies to strategy σ' given preferences u, i.e. $BR_u(\sigma') = \arg\max_{\sigma \in \Delta(A)} u(\sigma, \sigma')$.

Later, in Section 6, we analyse *type-interdependent* preferences, which depend also on the opponent's type, as in Herold & Kuzmics (2009). In contrast preferences defined solely over action profiles will be referred to as *type-neutral* preferences.

There is a fitness cost to increased cognition, represented by a positive and strictly increasing cognitive 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 k_n denote the cost of having cognitive level n. Hence $k_\theta = k_{n_\theta}$ denotes the cost of having type θ. Without loss of generality, we assume that $k_1 = 0$. In some of our results we will make the additional assumption that k_2 is sufficiently small.

2.3 Configurations

A complete description of a state of the population is constituted by a type distribution and a behaviour policy for each type in the support of the type distribution. An individual's behaviour is assumed to be (subjectively) rational in the sense that it maximizes her subjective preferences given the belief she has about the opponent's expected behaviour. However, her beliefs may be incorrect, if she is deceived by her opponent. An individual is deceived if and only if her opponent is of a higher cognitive level.

If two individuals of the same cognitive level are matched to play, then they play a Nash equilibrium of the game induced by their preferences. Given two preferences $u, u' \in U$, let $NE(u,u') \subseteq \Delta(A) \times \Delta(A)$ be the set of mixed equilibria of the game induced by the preferences u and u', i.e.

$$NE(u,u') = \{(\sigma,\sigma') \in \Delta(A) \times \Delta(A) : \sigma \in BR_u(\sigma') \text{ and } \sigma' \in BR_{u'}(\sigma)\}.$$

Remark 2. Similar to most of the existing literature of the indirect evolutionary approach (e.g., Güth & Yaari, 1992, Dekel *et al.*, 2007, Section 3), we assume full observability of the opponent's preferences. However, while the existing literature assumes this for all interactions, we only assume this when two incumbents with the same cognitive level interact. When two agents with different levels meet, the observability is endogenously determined by the deception efforts of the player with the higher level, and the observability assumption is only used as a "tie breaking rule". In Section XXX we analyse the opposite "tie breaking rule", according to when agents with the same cognitive level interact, they do not observe the opponent's preferences.

If two individuals of different cognitive levels are matched to play, then the individual with the higher cognitive level (henceforth, the *higher type*) observes the opponent's preferences perfectly, and is able to deceive the opponent (henceforth, the *lower type*). The deceiver is allowed to choose whatever she wants the deceived party to believe about the deceiver's intended action choice. The deceived party best responds given her possibly incorrect belief.

For simplicity, we assume that if the deceived party has multiple best replies, then the deceiver is allowed to break indifference, and choose which of the best replies she wants the deceived party to play. Consequently the deceiver is able to induce the deceived party to play any strategy that is a best reply to some belief about the opponent's mixed action, given the deceived party's preferences.

Given preferences $u \in U$, let $\Sigma(u)$ denote the set of *undominated strategies*. By the minmax theorem, $\Sigma(u)$ is also the set of actions that are best replies to at least one strategy of the opponent (given the preferences u). Formally, we define

$$\Sigma(u) = \{\sigma \in \Delta(A) : \text{there exists } \sigma' \in \Delta(A) \text{ such that } \sigma \in BR_u(\sigma')\}.$$

We say that a strategy profile is a *deception equilibrium* if the strategy profile is optimal from the point of view of player i under the constraint that player j has to play an undominated strategy. Formally:

Definition 1. Given two types θ, θ' with $n_\theta > n_{\theta'}$, a strategy profile $(\tilde{\sigma}, \tilde{\sigma}')$ is a *deception equilibrium* if

$$(\tilde{\sigma}, \tilde{\sigma}') \in \arg\max_{\sigma \in \Delta(A), \sigma' \in \Sigma(u_{\theta'})} u_\theta(\sigma, \sigma').$$

Let $DE(\theta,\theta')$ be the set of all such deception equilibria.

We are now in a position to define our key notion of a configuration, by combining a type distribution with a behaviour policy, as represented by Nash equilibria and deception equilibria.

Definition 2. A *configuration* is a pair (μ, b) where $\mu \in \Delta(U)$ is a type distribution, and $b : C(\mu) \times C(\mu) \longrightarrow \Delta(A)$ is a *behaviour policy* such that for each $\theta, \theta' \in C(\mu)$:

$$n_\theta = n_{\theta'} \Longrightarrow (b_\theta(\theta'), b_{\theta'}(\theta)) \in NE(\theta,\theta'), \text{ and}$$

$$n_\theta > n_{\theta'} \Longrightarrow (b_\theta(\theta'), b_{\theta'}(\theta)) \in DE(\theta,\theta').$$

We interpret $b_\theta(\theta') = b(\theta,\theta')$ as the strategy of type θ when being matched with type θ'.

Given a configuration (μ, b) we call the types in its support the *incumbents*.

Note that standard arguments imply that for any type distribution μ there exists a mapping $b : C(\mu) \times C(\mu) \longrightarrow \Delta(A)$ such that (μ, b) is a configuration.

The expected fitness to an individual of type θ in configuration (μ, b) is:

$$\Pi_\theta((\mu,b)) = \sum_{\theta' \in C(\mu)} \mu(\theta') \cdot \pi(b_\theta(\theta'), b_{\theta'}(\theta)) - k_\theta.$$

When all incumbent types have the same expected fitness, we say that the configuration is *balanced*, and denote this uniform expected payoff by $\Pi((\mu,b))$.

Remark 3. Our model assumes that a player may use different deceptions against different types with lower cognitive levels. We note that all our results remain the same (with minor changes to the proofs) in an alternative setup in which individuals have to use the *same* mixed action in their deception efforts towards all opponents with lower cognitive levels. We refer to this as *uniform deception*. The formal changes in the model that are required to implement this variant are described in Appendix A.

3 Evolutionary Stability

3.1 Definitions

Recall that a neutrally stable strategy (Maynard Smith & Price, 1973 and Maynard Smith, 1982) is a strategy that, if played by most of the population, weakly outperforms any other strategy. Similarly, an evolutionarily stable strategy is a strategy that, if played by most of the population, strictly outperforms any other strategy.

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

We extend the notions of neutral and evolutionary stability, from strategies to configurations. We begin by defining the type game that is induced by a configuration.

Definition 4. For any configuration (μ, b) the corresponding *type game* $\Gamma_{(\mu,b)}$ is the symmetric two-player game where each player's strategy space is $C(\mu)$, and the payoff to strategy θ, against strategy θ', is $\pi(b_\theta(\theta'), b_{\theta'}(\theta)) - k_\theta$.

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. A similar methodology was used in Mohlin (2012). Note that each type distribution with support in $C(\mu)$ is represented by a mixed strategy in $\Gamma_{(\mu,b)}$.

We want to capture robustness with respect to small groups of individuals, henceforth called *mutants*, which introduce new types and new behaviours into the population. 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 $\mu' \in \Delta(\Theta)$. Consequently the post-entry type distribution is $\tilde{\mu} = (1-\varepsilon)\cdot\mu + \varepsilon\cdot\mu'$. That is, for each type $\theta \in C(\mu) \cup C(\mu')$, $\tilde{\mu}(\theta) = (1-\varepsilon)\cdot\mu(\theta) + \varepsilon\cdot\mu'(\theta)$. In line with most of the literature on the indirect evolutionary approach we assume that adjustment of behaviour is infinitely faster than the adjustment of the type distribution.[8] Thus we assume that the post-entry type distribution quickly stabilizes into a configuration $\left(\tilde{\mu},\tilde{b}\right)$. There may exist many such post-entry type configurations, all with the same type distribution, but with different behaviour policies. We note that incumbents do not have to adjust their behaviour against other incumbents in order to continue playing Nash equilibria, and deception equilibria, among themselves. For this reason, we assume that the incumbents maintain the same pre-entry behaviour among themselves. In doing so we also follow Dekel *et al.* (2007). Formally:

Definition 5. Let (μ, b) and $\left(\tilde{\mu},\tilde{b}\right)$ be two configurations such that $C(\mu) \subseteq C(\tilde{\mu})$. We say that $\left(\tilde{\mu},\tilde{b}\right)$ is *focal* (with respect to (μ,b)) if $\theta,\theta' \in C(\mu)$ implies that $\tilde{b}_\theta(\theta') = b_\theta(\theta')$.

Standard fixed point arguments imply that for every configuration (μ,b) and every type distribution $\tilde{\mu}$ satisfying $C(\mu) \subseteq C(\tilde{\mu})$, there exists a behaviour policy $\tilde{b}$ such that $\left(\tilde{\mu},\tilde{b}\right)$ is a focal configuration.

Our stability notion requires that the incumbents outperform all mutants in all configurations that are focal relative to the initial configuration.

Definition 6. A configuration (μ,b) is a *neutrally stable configuration (NSC)***,** if for every $\mu' \in \Delta(\Theta)$, there is some $\bar{\varepsilon} \in (0,1)$ such that for all $\varepsilon \in (0,\bar{\varepsilon})$, it holds that if $\left(\tilde{\mu},\tilde{b}\right)$, where $\tilde{\mu} = (1-\varepsilon)\cdot\mu + \varepsilon\cdot\mu'$, is a focal configuration, then μ is an NSS in the type game $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$. The configuration (μ,b) is an *evolutionarily stable configuration (ESC)* if the same conditions imply that μ is an ESS in the type game $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$ for each $\mu' \neq \mu$.

3.2 Remarks

We discuss four issues related to our notion of stability.

1. The main stability notion that we use in the paper is NSC. The stronger notion of ESC is not useful in our main model because there always exist equivalent types that have slightly different preferences (as the set of preferences is a continuum) and induce the same behaviour as the incumbents. Such mutants would always achieve the same fitness as the incumbents in post-entry configurations, and thus ESCs will never exist. Note that the stability notions in Dekel *et al.* (2007) and Alger & Weibull (2013) are also based on neutral stability.[9] In Section 6 we study a variant of the model in which the preferences may depend also on the opponent's types. This will allow for the existence of ESCs.

2. Observe that Definition 6 implies internal stability with respect to small perturbations in the frequencies of the incumbent types (because when $\mu' = \mu$, then μ is required to be an NSS in $\Gamma_{(\mu,b)}$). By standard arguments, internal stability implies that any NSC is "balanced": all incumbent types obtain the same fitness.

[8] Sandholm (2001) and Mohlin (2010) are exceptions.

[9] In their stability analysis of *homo hamiltonensis* preferences Alger & Weibull (2013) disregard mutants who are behaviourally indistinguishable from *homo hamiltonensis* upon entry.

3. By simple adaptations of existing results in the literature, one can show that NSCs and ESCs are dynamically stable. NSCs are Lyapunov stable: no small change in the population composition can lead it away from μ in the type game $\Gamma_{(\tilde{\mu},\tilde{b})}$, if types evolve according to the replicator dynamic (Thomas, 1985, Bomze & Weibull, 1995). ESCs are also asymptotically stable: populations starting close enough to μ eventually converge to μ in $\Gamma_{(\tilde{\mu},\tilde{b})}$ if types evolve according to a smooth payoff-monotonic selection dynamic (Taylor & Jonker, 1978, Cressman, 1997, Sandholm, 2010).

4. The stability notions of Dekel *et al.* (2007) and Alger & Weibull (2013) only consider monomorphic groups of mutants (i.e. all mutants having the same type). We also consider stability against polymorphic groups of mutants (as do Herold & Kuzmics, 2009). One advantage of our approach is that it allows us to use an adaptation of the well-known notion of ESS, which immediately implies dynamic stability and internal stability, whereas Dekel *et al.* (2007) have to introduce a novel notion of stability without these properties. We note that our results remain similar with an analogous notion of stability that deals only with monomorphic mutants, except that in this case stability of pure outcomes would imply only a weaker notion of efficiency that compares the fitness only to symmetric profiles, as discussed in Remark 4.2 below.

4 Results

4.1 Preliminaries

We say that a strategy profile is *efficient* if it maximizes the sum of fitness payoffs. Formally:

Definition 7. A strategy profile (σ, σ') is *efficient in the game* $G = (A, \pi)$ if $\pi(\sigma, \sigma') + \pi(\sigma', \sigma) \geq \pi(a, a') + \pi(a', a)$, for each action profile (a, a').

Let $\bar{\pi} = max_{a,a' \in A}(0.5 \cdot (\pi(a, a') + \pi(a, a')))$ denote the efficient payoff - the average payoff achieved by players who play an efficient profile. If a symmetric strategy profile (σ^*, σ^*) is efficient, then we say that the strategy σ^* is efficient.

A pure Nash equilibrium (a, a) is strict if $\pi(a, a) > \pi(a', a)$ for all $a' \in A$. If a symmetric action profile $(a, a)$is a (strict) Nash equilibrium, of the fitness game, then we say that the action a is a (strict) Nash equilibrium, of the fitness game.

Given a configuration (μ, b) let $\bar{n} = \max_{\theta \in C(\mu)} n_\theta$ denote the the maximal cognitive level of the incumbents. We refer to incumbents with this cognitive level as the *highest types*.

A deception equilibrium is *fitness maximizing* if it maximizes the fitness of the type with the higher level of the types in the match (under the restriction that the type with the lower level plays an action that is not dominated given her preferences). Formally:

Definition 8. Let θ, θ' be two types with $n_\theta > n_{\theta'}$. A deception equilibrium $(\tilde{\sigma}, \tilde{\sigma}') \in DE(\theta, \theta')$ is *fitness maximizing* if:

$$(\tilde{\sigma}, \tilde{\sigma}') \in argmax_{\sigma \in \Delta(A),\, \sigma' \in \Sigma(u_{\theta'})} \pi(\sigma, \sigma').$$

Let $FMDE(\theta, \theta') \subseteq DE(\theta, \theta')$ denote the set of all such fitness-maximizing deception equilibria of two types θ, θ' with $n_\theta > n_{\theta'}$. Note that $FMDE(\theta, \theta')$ might be an empty set (if there is no action profile that maximizes both the fitness and the subjective utility of the higher type).

A configuration is pure if everyone plays the same action. Formally:

Definition 9. A configuration (μ, b) is *pure* if there exists $a^* \in A$ such that $b_\theta(\theta') = a^*$ for each $\theta, \theta' \in C(\mu)$.

With a slight abuse of notation we denote such a pure configuration by (μ, a^*), and we refer to a^* as the *outcome* of the configuration.

In order to simplify the notation and the arguments in the proofs we assume throughout this section that underlying game admits at least 3 actions (i.e. $|A| \geq 3$). The results can be extended to games with two actions but this makes the notation more cumbersome and the proofs less instructive.

4.2 Characterization of the Highest Types' Behavior

In this section we characterize the behavior of an incumbent type, $\bar{\theta} = (u, \bar{n})$, which has the highest level of cognition in the population. We show that the behavior satisfies the following three conditions:

1. Type $\bar{\theta}$ plays an efficient action when meeting itself.
2. Type $\bar{\theta}$ maximizes its fitness in all interactions with types with lower cognitive levels.
3. Any opponent with a lower cognitive level achieves at most $\bar{\pi}$ when being matched with $\bar{\theta}$.

Theorem 1. *Let (μ^*, b^*)be an NSC, and $\underline{\theta}, \bar{\theta} \in C(\mu^*)$.*

1. *If $n_{\bar{\theta}} = \bar{n}$ then$\pi\left(b_{\bar{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\bar{\theta}\right)\right) = \bar{\pi}$.*
2. *If $n_{\underline{\theta}} < n_{\bar{\theta}} = \bar{n}$ then $\left(\left(b_{\bar{\theta}}\left(\underline{\theta}\right), b_{\underline{\theta}}\left(\bar{\theta}\right)\right)\right) \in FMDE\left(\bar{\theta}, \underline{\theta}\right)$.*
3. *If $n_{\underline{\theta}} < n_{\bar{\theta}} = \bar{n}$ then $\pi\left(b_{\underline{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\underline{\theta}\right)\right) \leq \bar{\pi}$.*

The formal proof can be found in the appendix. here we provide a brief sketch of it. The proof utilises mutants (denoted θ_1, θ_2, and $\hat{\theta}$ below) with the highest cognitive level $\bar{n}$ and with a specific kind of utility functions, called *indifferent and pro-generous*, that make a player indifferent between all their own actions, but may make the player prefer the opponent to take an action that allows the player to obtain the highest possible fitness payoff.

1. To prove part 1 of the Theorem, assume to the contrary that $\pi\left(b_{\bar{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\bar{\theta}\right)\right) < \bar{\pi}$. Let $a_1, a_2 \in A$ be any two actions such that (a_1, a_2) is an efficient action profile (i.e. $0.5 \cdot (\pi(a_1, a_2) + \pi(a_1, a_2)) = \bar{\pi}$). Consider two mutant types θ_1 and θ_2, $(\theta_1 \neq \theta_2)$ which are of the highest cognitive level that is present in the population, and have indifferent and pro-generous utility functions. Suppose equal fractions of these two mutant types enter. There is a post-entry population, in which; the incumbents keep their pre-entry play among each other, the mutants play a fitness-maximizing deception equilibria against lower types, the mutants mimic the play of $\bar{\theta}$ against all incumbents of level $\bar{n}$, the mutants mimic $\bar{\theta}$ when facing a mutant of the same type, and when mutant type θ_1 meets mutant type θ_2 they play the efficient profile (a_1, a_2). In such a focal post-entry configuration a mutant earns weakly higher fitness than $\bar{\theta}$ against the incumbents, and the mutants on average earn strictly more than $\bar{\theta}$ against the mutants. This implies that (μ^*, b^*) cannot be an NSC.

2. To prove part 2 , assume to the contrary that $\left(\left(b_{\bar{\theta}}\left(\underline{\theta}\right), b_{\underline{\theta}}\left(\bar{\theta}\right)\right)\right) \notin FMDE\left(\bar{\theta}, \underline{\theta}\right)$. Suppose mutants of type $\hat{\theta}$ enter. Consider a post-entry configuration in which the incumbents keep their pre-entry play among each other, and the mutants mimic the play of $\bar{\theta}$, except that they play a fitness-maximizing deception equilibria against all lower types. The mutants obtains a weakly higher payoff than $\bar{\theta}$ against all types, and a strictly higher payoff than $\bar{\theta}$ against some lower type. Thus(μ^*, b^*) cannot be an NSC.

3. To prove part 3 , assume to the contrary that $\pi\left(b_{\underline{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\underline{\theta}\right)\right) > \bar{\pi}$. Suppose mutants of type $\hat{\theta}$ enter. Consider a post-entry configuration in which the incumbents keep their pre-entry play among each other, while the mutants; (i) play fitness-maximizing deception equilibria against lower types, (ii) mimic type $\underline{\theta}$

against type $\bar{\theta}$, and (iii) mimic the play of $\bar{\theta}$ against all other types. The mutant $\hat{\theta}$ earns strictly more than $\bar{\theta}$ against both $\hat{\theta}$ and $\bar{\theta}$. The mutant earns weakly more than $\bar{\theta}$ against all other types. This implies that (μ^*, b^*) cannot be an NSC.

Remark. The first part of Theorem 1 (a highest types must play an efficient strategy when meeting itself) is similar to Dekel *et al.*'s (2007) Proposition 2, which shows that only efficient outcomes can be stable in in a setup with perfect observability and no deception. We should note that Dekel *et al.* (2007) use a weaker notion of efficiency. An action is efficient in the sense of Dekel *et al.* (2007) (DEY-efficient) if its fitness is highest among the symmetric strategy profiles (i.e. action a is DEY-efficient if $\pi(a,a) \geq \pi(\sigma,\sigma)$ for all strategies $\sigma \in \Delta(A)$). Observe that our notion of efficiency (Definition 7) implies DEY-efficiency, but the converse is not necessarily true. The weaker notion of DEY-efficiency is the one relevant in the set up of Dekel *et al.* (2007), because they consider only monomorphic groups mutants; i.e. all mutants that enter at the same time are of the same type. A similar result would hold also in our setup, if we imposed a similar limitation on the set of feasible mutants. However, without such a limitation, heterogeneous mutants can correlate their play, and our stronger notion of efficiency is required to characterise stability.

An immediate corollary of Theorem 1 is that a game without efficient actions (i.e., a game that has only non-symmetric efficient profiles) do not admit any NSCs.

Corollary 1. *If G does not admit an efficient action (i.e. if $\pi(a,a) < \bar{\pi}$ for each $a \in A$), then the game does not admit any NSC.*

4.3 Characterization of Pure NSCs

We now present three results which, together with Theorem 1, allow us to completely characterise NSCs with pure outcomes. The first proposition shows that in a pure NSC all incumbents have the minimal cognitive level, since having a higher ability does not yield any advantage when everyone plays the same action.

Claim 1. If (μ, a^*) is an NSC, and $(u,n) \in C(\mu)$, then $n = 1$.

Proof. Since all players earn the same game payoff of $\pi(a^*, a^*)$, they must also incur the same cognitive cost, or else the fitness of the different incumbent types would not be balanced (which contradicts (μ, a^*) being an NSC). Moreover, this uniform cognitive level must be level 1. Otherwise a mutant of a lower level, who strictly prefers to play a^* against all actions, would strictly outperform the incumbents in nearby post-entry focal configurations. □

The following proposition shows that if k_2 (the cost of having cognitive level 2) is sufficiently small, then any outcome of a pure NSC must be a Nash equilibrium of the underlying game. The reason is that if the pure outcome is not a Nash equilibrium, then the population can be invaded by mutants with cognitive level 2, who deceive the incumbents into thinking they face other incumbents, and best reply to the incumbents' play.

Proposition 1. *Suppose*

$$k_2 < \delta := \min_{a,a',a'' \ s.t.\ \pi(a,a'') \neq \pi(a',a'')} \left| \pi(a,a'') - \pi(a',a'') \right|. \tag{1}$$

If (μ, a^*) is an NSC, then (a^*, a^*) is a symmetric Nash equilibrium, in fitness payoffs.

Proof. Assume to the contrary that (μ, a^*) is a pure NSC and a^* is not a best response to itself; i.e. there exist $a' \in A$ such that $\pi(a', a^*) > \pi(a^*, a^*)$. Assume without loss of generality that a' is a best reply against a^* (in fitness terms). By Proposition 1, all incumbents have cognitive level 1. Consider a mutant $\theta' = (\pi, 2)$ with cognitive level 2 and materialistic preferences. There is a focal post-entry configuration in which mutants play the deception equilibrium (a', a^*) against the incumbents. Observe that the mutants obtain a strictly higher payoff when facing an incumbent, than what two incumbents earn against each other:

$$\pi(a', a^*) - k_2 > \pi(a', a^*) - \delta \geq \pi(a^*, a^*).$$

This implies that if the mutants are sufficiently rare, they outperform the incumbents in the post-entry focal configuration. □

The last proposition shows that any action that is both efficient and a strict Nash equilibrium, can be induced as the outcome of an NSC. The intuition is as follows (and is similar to Dekel *et al.*, 2007, Proposition 6). Consider a monomorphic population in which all individuals have cognitive level 1 and the efficient strict Nash action is a dominant action. The action being strict Nash and efficient, implies that any group of mutants is weakly outperformed.

Proposition 2. *If $\bar{a}$ is an efficient action and a strict Nash equilibrium (in fitness payoffs), then there exists a type distribution μ such that $(\mu, \bar{a})$ is an NSC.*

Proof. Consider a monomorphic configuration $(\mu, \bar{a})$ consisting of type $(\theta^*, 1)$ where all incumbents are of cognitive level 1 and of the same preference type θ^*, which strictly prefers to play $\bar{a}$ regardless of what the opponent plays. Observe, that after any mutant's entry, in all focal post-entry configurations the incumbent θ^* will always play $\bar{a}$ (since $\bar{a}$ is strictly dominant for θ^*). Since the incumbent is always playing $\bar{a}$, and $(\bar{a}, \bar{a})$ is a strict Nash equilibrium of G, mutants that do not play $\bar{a}$ when they are matched with θ^* will obtain strictly less fitness than the incumbents if their population share is sufficiently small. But for mutants that play $\bar{a}$ whenever they are matched with θ^*, the incumbents' average fitness is given by $\pi(\bar{a}, \bar{a})$, and since mutants cannot obtain an average fitness strictly higher than this when they are matched among themselves (since $(\bar{a}, \bar{a})$ is efficient), they cannot obtain a strictly higher average fitness either. We conclude that $(\mu, \bar{a})$ is an NSC. □

Part 1 of Theorem 1, together with Propositions 1-2 imply that pure NSCs outcomes are essentially efficient Nash equilibria. Formally:

Corollary 2 (Characterization of pure NSCs)**.**

1. *If (a^*, a^*) is both efficient and a strict Nash equilibrium in fitness payoffs, then it is the outcome of a pure NSC.*

2. *If (a^*, a^*) is the outcome of a pure NSC and $k_2 < \delta$, then it is both efficient and a Nash equilibrium in fitness payoffs.*

4.4 Characterization of NSCs in Generic Games

In this section we characterize NSCs in generic games, by which we mean games in which any two different action profiles both gives an individual player different payoffs, and yields different total payoffs.

Definition 10. A (symmetric) game is generic if for each $a, a', b, b' \in A$, $\{a, a'\} \neq \{b, b'\}$ implies:

$$\pi(a, a') \neq \pi(b, b'), \text{ and } \pi(a, a') + \pi(a', a) \neq \pi(b, b') + \pi(b', b).$$

For example, if the entries of the payoff matrix π are drawn from a continuous distribution on an open subset of the real numbers, then the induced game is generic with probability one.

Note that a generic game admits at most one efficient action profile. Due to Corollary 1, we can restrict attention to games with a single efficient action. Let $\bar{a}$ denote this unique efficient action.

Our first result shows that each incumbent type plays efficiently when meeting itself.

Theorem 2. *Let (μ^*, b^*)be an NSC of a generic game and let $\theta \in C(\mu^*)$. Then $\pi(b_\theta(\theta), b_\theta(\theta)) = (\bar{a}, \bar{a})$.*

The outline of the proof is as follows (the formal proof is in Appendix A.3). In order to obtain a contradiction assume that there exists an incumbent that does not play $\bar{a}$. Let $\bar{\theta} \in C(\mu^*)$ be the type with the highest cognitive level, among the types that do not play $\bar{a}$ against themselves. Let $\hat{u}$ be a utility function which is identical to $u_{\bar{\theta}}$ except that: (1) the payoff of the outcome $(\bar{a}, \bar{a})$ is increased by the minimal amount required to make it a best reply to itself, and (2) the payoffs of some other outcome is altered slightly (to ensure $\hat{u}$ is not already an incumbent) in a way that does not force $\hat{\theta}$ to behave differently from $\bar{\theta}$. Suppose a mutant of type $\hat{\theta} = \left(\hat{\theta}, n_{\bar{\theta}}\right)$ enters. Consider a focal post-entry configuration in which the mutants mimic the play of the incumbent $\bar{\theta}$ in all matches except: (1) they play the efficient profile $(\bar{a}, \bar{a})$ among themselves (which yields a higher payoff than what $\bar{\theta}$ achieves when being matched against θ'), and (2) when the mutant faces a higher type they either play $(\bar{a}, \bar{a})$ or the same deception equilibrium that the higher type plays against $\bar{\theta}$. It follows that the mutant $\hat{\theta}$ earns a strictly higher payoff than $\bar{\theta}$ against $\hat{\theta}$, and a weakly higher fitness than type $\bar{\theta}$ against all other types. Thus the the mutants strictly outperform the incumbents, which contradicts (μ^*, b^*)being an NSC.

The following Theorem shows that any two types play efficiently when meeting each other. The intuition is that populations in which within-type matchings yields higher payoff ($\bar{\pi}$ according to Theorem 2) than out-group matchings (an average payoff of less than $\bar{\pi}$) are dynamically unstable: a slight increase in a group's frequency would increase its fitness and would take the population away from the initial state.

Theorem 3. *Let (μ^*, b^*) be an NSC of a generic game, and let $\theta, \theta' \in C(\mu^*)$. Then $b_\theta(\theta'), b_{\theta'}(\theta) = \bar{a}$.*

Proof. Assume to the contrary that there exist two types $\theta_j, \theta_k \in C(\mu^*)$, $\theta_j \neq \theta_k$, such that $b_\theta(\theta') \neq \bar{a} \neq b_{\theta'}(\theta)$. This implies that $0.5 \cdot \left(\pi\left(b_{\theta_j}(\theta_k), b_{\theta_k}(\theta_j)\right) + \pi\left(b_{\theta_k}(\theta_j), b_{\theta_j}(\theta_k)\right)\right) < \bar{\pi}$. The fact that (μ^*, b^*) is an NSC implies that μ^* is an NSS in the type game $\Gamma_{(\mu^*, b^*)}$. Let B be the payoff matrix of the type game $\Gamma_{(\mu^*, b^*)}$ and let $n = |C(\mu^*)|$. It is well known (e.g., Hofbauer & Sigmund, 1988, Exercise 6.4.3 and Hofbauer (2011, pp. 1-2)) that an interior strategy profile of a normal form game can be an NSS only if the payoff matrix is negative semi-definite with respect to the tangent space, i.e. only if $x \cdot Bx \leq 0$ for each $x \in \mathbb{R}^n$ such that $\sum_i x_i = 0$.

Assume without loss of generality that type θ_j (θ_k) is represented by the j^{th} (k^{th}) row of the matrix B. Let the column vector x be defined as follows:

$$x(i) = \begin{cases} 1 & i = j \\ -1 & i = k \\ 0 & i \neq j, k \end{cases}.$$

That is the vector x has all entries equal to zero, except the j^{th} entry which is equal to 1, and the k^{th} entry which is equal to -1. We have

$$\begin{aligned} x \cdot Bx &= B_{jj} - B_{jk} - B_{jk} + B_{kk} \\ &= \bar{a} - k_{n_{\theta_j}} + \bar{a} - k_{n_{\theta_k}} - \left(\pi\left(b_{\theta_j}(\theta_k), b_{\theta_k}(\theta_j)\right) - k_{n_{\theta_j}} + \pi\left(b_{\theta_k}(\theta_j), b_{\theta_j}(\theta_k)\right) - k_{n_{\theta_k}}\right) \\ &= 2 \cdot \bar{a} - \left(\pi\left(b_{\theta_j}(\theta_k), b_{\theta_k}(\theta_j)\right) + \pi\left(b_{\theta_k}(\theta_j), b_{\theta_j}(\theta_k)\right)\right) > 0. \end{aligned}$$

Thus B is not negative semi-definite. $\square$

Note that that in a generic game any pure action is either a strict equilibrium, or not a best reply to itself. Combining the results of this section with the above charachterisation of pure NSCs yields the following corollary, which fully characterises the NSCs of generic games.

Corollary 3. *Let G be a generic game.*

1. *If G admits an efficient action $\bar{a}$ that is also a Nash equilibrium, then there is a pure NSC in which all types (i) have cognitive level one, and (ii) always play $\bar{a}$. Converesely, if $k_2 < \delta$, then this NSC is essentially unique: any NSC must satisfy both properties (i) and (ii).*
2. *Otherwise, the game does not admit any NSC.*

4.5 Non-pure NSCs in non-Generic Games

The previous two subsections fully characterized (i) pure NSCs and (ii) NSCs in generic games. In this section we analyse non-pure NCSs in non-generic games. Non-generic games may be of interest in various setups, such as: (1) normal-form representation of generic extensive-form games (the induced matrix is typically non-generic), and (2) special interesting families of games, such as zero-sum games. Unlike generic games, non-generic games can admit (non-pure) NSCs with multiple cognitive level and non-Nash behavior. To demonstrate this we consider the Rock-Paper-Scissors game, with the payoff matrix

$$\begin{array}{cccc} & R & P & S \\ R & 0,0 & -1,1 & 1,-1 \\ P & 1,-1 & 0,0 & -1,1 \\ S & -1,1 & 1,-1 & 0,0 \end{array}.$$

The following result shows that, under mild assumptions on the cognitive costs function, this game admits an NSC in which all players have the same materialistic preferences, but players of different cognitive levels co-exist, and non-Nash profiles are played in all matches of two individuals of different cognitive levels. More precisely, when individuals of different cognitive levels meet, the higher level individual deceives the lower level individual to take a pure action that the higher level individual then best responds to. Thus the higher level individual earns 1 and her opponent earns -1. Individuals of the same cognitive level play the unique Nash equilibrium. This means that higher-level types will obtain the payoff 1 more often than lower-level types, and lower-level types will obtain the payoff -1 more often than higher-level types. In the NSC this payoff difference is offset exactly by the higher cognitive cost paid by higher types. Moreover, the cognitive cost is increasing so that at some point the cost of cognition outweighs any payoff differences that may arise from the underlying game. This implies that there is an upper bound on the cognitive sophistication in the population.

Proposition 3. *Let G be a Rock-Paper-Scissors game. Let u^{π} denote the (materialistic) preference such that $u^{\pi}(a,a') = \pi(a,a')$ for all profiles (a,a'). Suppose that the marginal cognitive cost is small but non-vanishing, so that (a) there is an N such that $k_N \leq 2 < k_{N+1}$, and (b) it holds that $1 > k_{n+1} - k_n$ for all $n \leq N$.*[10] *There exists an NSC (μ^*, b^*), such that $C(\mu^*) \subseteq \{(u^{\pi}, n)\}_{n=1}^{N}$, and μ^* is mixed (i.e. $|C(\mu^*)| > 1$). The behaviour of the incumbent types is as follows: if the individuals in a match are of different cognitive level, then the higher level plays Paper and the lower level plays Rock; if both individuals in a match are of the same cognitive level, then they both play the unique nash equilibrium (i.e. randomise uniformly over the three actions).*

Appendix C.2 contains a formal proof of this result an relates it to a similar construction in Conlisk (2001).

[10] If we define δ as in (1) then we have $\delta = 1$, in Proposition 3. Thus the condition that $\delta > k_{n+1} - k_n$ for all n in Proposition 3, may be viewed as an extension of the condition $k_2 < \delta$ in Proposition 1.

5 Non-Observability Among Cognitive Equals

The main model assumes that types with equal levels can observe each others preferences. In this section we present an alternative "tie-breaking rule" according to which players are unable to observe the preferences of an opponent with the same cognitive level.

5.1 Changes to the Baseline Model

Play Between Types with Equal Cognitive Levels. If two individuals of the same cognitive level are matched to play, then they play a Bayes-Nash equilibrium of the Bayesian game in which each player only knows that his opponent has the same cognitive level (but he cannot observe his opponent's preferences). Given a distribution of types μ, we say that $n \in \mathrm{proj}_N C(\mu)$ if there exists a type $(u,n) \in C(\mu)$. Given a distribution μ and $n \in \mathrm{proj}_N C(\mu)$, let Θ_n denote the set of types with level n, and let μ_n denote the distribution of types conditional on having cognitive level n:

$$\mu_n(u,n) = \frac{\mu(u,n)}{\sum_{(u',n)\in C(\mu)} \mu(u',n)}.$$

Given distribution μ and $n \in \mathrm{proj}_N C(\mu)$, a vector of action profiles $(\sigma_\theta)_{\theta\in\Theta_n}$ is a Bayes-Nash equilibrium of the game between types with level n, if each player best-reply to the aggregate behavior of players with the same cognitive level. Let $BNE(n) \subseteq (\Delta A)^{|\Theta_n|}$ be the set of Bayes-Nash equilibria of the game induced by the interactions between players with cognitive level n, i.e.

$$BNE(n) = \left\{ (\sigma_\theta)_{\theta\in\Theta_n} \in (\Delta A)^{|\Theta_n|} : \forall \sigma_{(u,n)} \in \Theta_n,\ \sigma_{(u,n)} \in argmax_{\sigma\in\Delta(A)} \sum_{\theta\in\Theta_n} \mu_n(\theta)\cdot u(\sigma,\sigma_\theta) \right\}.$$

The play among types with equal levels is analogous to the play of all types in the unobservable case of Dekel *et al.* (2007, Section 4). Players with different cognitive levels play deception equilibria as in the baseline model.

Redefining Configurations. We redefine the notion of a configuration as follows:

Definition 11. A *configuration* is a pair (μ, b) where $\mu \in \Delta(U)$ is a type distribution, and $b : C(\mu) \times C(\mu) \longrightarrow \Delta(A)$ is a *behaviour policy* such that:

1. Players play the same against all opponents with the same level, i.e. for each $n \in \mathrm{proj}_N C(\mu)$, if $\theta, \theta', \theta'' \in \Theta_n$ then $b_\theta(\theta') = b_\theta(\theta'')$.
2. Types with the same level play a Bayes-Nash equilibrium, i.e. for each $n \in \mathrm{proj}_N C(\mu)$

$$\left((b_\theta(\theta))_{\theta\in\Theta_n}\right) \in BNE(n).$$

3. Types with different levels play deception equilibria, i.e. for each $\theta, \theta' \in C(\mu)$:

$$n_\theta > n_{\theta'} \Longrightarrow (b_\theta(\theta'), b_{\theta'}(\theta)) \in DE(\theta,\theta').$$

Redefining NSC. Unlike in the baseline model, focal configurations (in which the incumbents keep their pre-entry play) do not always exist. Accordingly, we modify the definition NSC slightly to require the existence of a focal configuration for any given distribution of mutants.[11] Formally:

[11] All the results remain the same if one only requires δ-focality à la Dekel *et al.* (2007)

Definition 12. A configuration (μ, b) is a *neutrally stable configuration (NSC)***,** if for every $\mu' \in \Delta(\Theta)$, there is some $\bar{\varepsilon} \in (0,1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$, it holds that; (1) there exists a focal configuration $\left(\tilde{\mu}, \tilde{b}\right)$, where $\tilde{\mu} = (1-\varepsilon) \cdot \mu + \varepsilon \cdot \mu'$, and (2) for each such focal configuration $\left(\tilde{\mu}, \tilde{b}\right)$, the distribution μ is an NSS in the type game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$.

5.2 Results

The following simple results show that with non-observable equals, it is no longer the case that stable outcomes must be efficient. The first result shows that any strict Nash equilibrium (even non-efficient) is the outcome of a pure NSC. The argument is similar to Dekel *et al.* (2007, Proposition 5-b) and is presented briefly.

Proposition 4. *Let $a^* \in A$ be a strict Nash equilibrium. Then there exists a pure NSC (μ, a^*) where μ contains a single type $(1, u^*)$ such that a is a strictly dominant action given the utility function u^*.*

Proof. Observe that the incumbents always play a^* in any focal configuration (and that such focal configuration always exist). The fact that a^* is a strict equilibrium implies that any mutant who does not always play a^* against the incumbents is strictly outperform if the mutants are sufficiently rare. If the mutant type has level one, it implies that he must always play a^* also against mutants, and that it achieves the same payoff as the incumbents. If the mutant has higher cognitive level, then if the mutants are sufficiently rare the higher cognitive cost implies that it is strictly outperformed. □

The next result shows that players with the highest type must play a Nash equilibrium (of the underlying game) among themselves.The argument is similar to Dekel *et al.* (2007, Proposition 5-a) and is presented briefly.

Claim 2. Let (μ^*, b^*) be an NSC and let $\bar{n}$ be the highest cognitive level. The action profile $(b_\theta(\theta))_{\theta \in \Theta_{\bar{n}}}$ is a Bayes-Nash equilibrium of the fitness game. That is:

$$\forall \theta^* \in \Theta_{\bar{n}},\ b_{\theta^*}(\theta^*) \in argmax_{\sigma \in \Delta(A)} \sum_{\theta \in \Theta_n} \mu_n(\theta) \cdot \pi(\sigma, b_\theta(\theta)).$$

Proof. Assume to the contrary the the highest types do not play a Bayes-Nash equilibrium of the fitness game. Let $\bar{\theta} \in \Theta_{\bar{n}}$ be one such type. That is

$$b_{\theta^*}(\theta^*) \notin argmax_{\sigma \in \Delta(A)} \sum_{\theta \in \Theta_n} \mu_n(\theta) \cdot \pi(\sigma, b_\theta(\theta)).$$

Let $\theta' = (u', \bar{n})$ be a mutant type which is indifferent between any two outcomes, and which mimics the play of θ^* against lower types, while best-replying in the induced fitness game between the highest types. Such a mutant type would strictly outperform the incumbent $\bar{\theta}$ in any focal configuration. □

6 Type-Interdependent Preferences

In this section we describe an extension of our baseline model, such that the preferences may depend not only on action profiles, but also on the opponent's type.

6.1 Changes to the Baseline Model

We briefly describe how to amend the model to handle type-interdependent preferences. Our construction is similar to that of Herold & Kuzmics (2009).

When the preferences of a type depend on the opponent's type, we can no longer work with the set of all possible preferences, because it would create problems of circularity and cardinality.[12] Instead, we must restrict attention to a pre-specified set of feasible preferences. We begin by defining Θ_{ID} as an arbitrary set of labels. Each label is a pair $\theta = (u, n) \in \Theta_{ID}$, where $n \in \mathbb{N}$ and u is a type-interdependent utility function that depends on the played action profile as well as the opponent's label,

$$u : A \times A \times \Theta_{ID} \to \mathbb{R}.$$

Each label $\theta = (u, n)$ may now be interpreted as a type. The definition of u extends to mixed actions in the obvious way. We use the label u also to describe its associated utility function u. Thus $u(\sigma, \sigma', \theta')$ denotes the subjective payoff that a player with preferences u earns when she plays strategy σ against an opponent with type θ' who plays strategy σ'.

Let U_{ID} denote the set of all preferences that are part of some type in Θ_{ID}, i.e. $U_{ID} = \{u : \exists n \in \mathbb{N}$ s.t. $(u, n) \in \Theta_{ID}\}$. For each type-neutral preference $u \in U$ we can define an equivalent type-interdependent preference $u \in U_{ID}$, which is independent of the opponent's type; that is, $u'(\sigma, \sigma', \theta') = u''(\sigma, \sigma', \theta'')$ for each $u', u'' \in U_{ID}$. Let U_N denote the set of all such type-interdependent versions of the type-neutral preferences of the baseline model. All of our results allow, but do not require, that $U_N \subseteq U_{ID}$.

Next, we amend the definitions of Nash equilibrium, undominated strategies, and deception equilibrium. The best-reply correspondence now takes both strategies and types as arguments: $BR_u(\sigma', \theta') = \arg\max_{\sigma \in \Delta(A)} u(\sigma, \sigma', \theta')$. Accordingly we adjust the definition of the set of Nash equilibria,

$$NE(\theta, \theta') = \{(\sigma, \sigma') \in \Delta(A) \times \Delta(A) : \sigma \in BR_u(\sigma', \theta') \text{ and } \sigma' \in BR_{u'}(\sigma, \theta)\},$$

and the set of *undominated strategies*

$$\Sigma(\theta) = \{\sigma \in \Delta(A) : \text{there exists } \sigma' \in \Delta(A) \text{ and } \theta' \in \Theta_{ID} \text{ such that } \sigma \in BR_u(\sigma', \theta')\}.$$

Finally, we adapt the definition of deception equilibrium. Given two types θ, θ' with $n_\theta > n_{\theta'}$, a strategy profile $(\tilde{\sigma}, \tilde{\sigma}')$ is a *deception equilibrium* if

$$(\tilde{\sigma}, \tilde{\sigma}') \in \arg\max_{\sigma \in \Delta(A), \sigma' \in \Sigma(\theta')} u_\theta(\sigma, \sigma', \theta').$$

Let $DE(\theta, \theta')$ be the set of all such deception equilibria. The rest of our model remains unchanged.

6.2 Pure Maxmin and Minimal Fitness

The pure maxmin and minmax values give a minimal bound to the fitness of an NSC. Given a game $G = (A, \pi)$, define $\underline{M}(\bar{M})$ as its pure maxmin (minmax) value:

$$\underline{M} = \max_{a_1 \in A} \min_{a_2 \in A} \pi(a_1, a_2),$$

$$\overline{M} = \min_{a_2 \in A} \max_{a_1 \in A} \pi(a_1, a_2).$$

[12]The circularity comes from the fact that each type contains a preferences component, which is identified with a utility function defined over types (and action profiles). To see that this creates a problem if the set of types is unrestricted, let Θ_* be the set of types and suppose that the corresponding set of preferences, U_*, contains all mappings $u : A \times A \times \Theta_* \to \mathbb{R}$. The cardinality of this set is $|U| \cdot |\Theta_*|$, but if U_* is indeed the set of *all* mappings $u : A \times A \times \Theta_* \to \mathbb{R}$, then we must have $|U_*| = |U| \cdot |\Theta_*|$. Since $|\Theta_*| \geq |U_*|$ this is a contradiction. See also footnote 10 in Herold & Kuzmics (2009).

The pure maxmin value $\underline{M}$ is the minimal fitness payoff a player can guarantee herself in the sequential game in which she plays first, and the opponent replies in an arbitrary way (i.e. not necessarily in a way that maximizes the opponent's fitness.) The pure minmax value $\overline{M}$ is the minimal fitness payoff a player can guarantee herself in the sequential game in which her opponent plays first an arbitrary action, and she best-replies to the opponent's pure action. It is immediate that $\underline{M} \leq \overline{M}$, and that the minmax value in mixed actions is between these two values.

Let $a_{\underline{M}}$ be a maxmin action of a player; an action $a_{\underline{M}}$ guarantees that the player's payoff is at least $\underline{M}$,

$$a_{\underline{M}} \in \arg\max_{a_1 \in A} \min_{a_2 \in A} \pi\left(a_1, a_2\right).$$

The following simple lemma (which holds also in the baseline model with type-neutral preferences) shows that the maxmin value is a lower bound on the fitness payoff obtained in an NSC. The intuition is that if the payoff is lower, then a mutant of cognitive level 1, with preferences such that the maxmin action $a_{\underline{M}}$ is dominant, will outperform the incumbents.

Definition 13. Given a pure action $a^* \in A$, let $u^{a*} \in U_N$ be the (type-neutral) preferences in which the player obtains a payoff of 1 if she plays a^* and a payoff of 0 otherwise (i.e. a^* is a dominant action regardless of the opponent's preferences).

Proposition 5. *Assume that $(u^{a_{\underline{M}}}, 1) \in \Theta_{ID}$. Let (μ, b) be an NSC. Then $\Pi\left(\mu, b\right) \geq \underline{M}$.*

Proof. Assume to the contrary that $\Pi\left(\mu, b\right) < \underline{M}$. Consider a monomorphic group of mutants with type $(u^{a_{\underline{M}}}, 1)$. The fact that $a_{\underline{M}}$ is a maxmin action implies that

$$\pi_{\left(u^{a_{\underline{M}}},1\right)}\left(\left(\tilde{\mu}, \tilde{b}\right)\right) \geq \underline{M}$$

in any post-entry configuration. Furthermore, due to continuity it holds that $\Pi_\theta\left(\tilde{\mu}, \tilde{b}\right) < \underline{M}$ for any $\theta \in C\left(\mu\right)$ in all sufficiently close focal post-entry configuration. This contradicts μ being an NSS in $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$, and thus it contradicts (μ, b) being an NSC. ☐

6.3 Characterisation of Pure Stable Configurations

In this subsection we show that, essentially, a pure action can be an outcome of an ESC if and only if it is a Nash equilibrium that yields each player a payoff above her minmax/maxmin value.

We first adapt Propositions 1-1 to the current setup. Specifically, we show that if (μ^*, a^*) is a pure NSC, then: (1) all incumbents have the same cognitive level, and (2) a^* is a symmetric Nash equilibrium, provided that the marginal cognitive costs are sufficiently small (smaller than δ, as defined in (1)).

Proposition 6. *Let (μ^*, a^*) be a pure NSC. Then:*

1. *$\theta, \theta' \in C\left(\mu^*\right) \Rightarrow n_\theta = n_{\theta'}$.*

2. *If for each n $k_{n+1} - k_n < \delta$, then a^* is a symmetric Nash equilibrium (in fitness payoffs).*

Proof.

1. Since all players earn the same game payoff of $\pi\left(a^*, a^*\right)$, they must also incur the same cognitive cost, or else the fitness of the different incumbent types would not be balanced (which contradicts (μ, a^*) being an NSC).

2. Assume to the contrary that there exist $a' \in A$ such that $\pi(a', a^*) > \pi(a^*, a^*)$. Assume without loss of generality that a' is a best reply against a^* (in fitness terms). By part (i) all the incumbents have cognitive level n. Consider a monomorphic group of mutants $\theta' = (\pi, n+1)$. There is a focal post-entry configuration in which the mutants play the deception equilibrium (a', a^*) against the incumbents. Observe that the mutants obtain a strictly higher payoff when facing an incumbent, than what two incumbents earn against each other:

$$\pi(a', a^*) - (k_{n+1} - k_n) > \pi(a', a^*) - \delta \geq \pi(a^*, a^*).$$

This implies that if the mutants are sufficiently rare, they outperform the incumbents in the post-entry focal configuration.

□

Let $a_{\bar{M}}$ be a minmax action, i.e. an action that guarantees that the opponent's payoff is at most $\bar{M}$;

$$a_{\bar{M}} \in \arg\min_{a_2 \in A} \max_{a_1 \in A} \pi(a_1, a_2).$$

Definition 14. Given any two actions $\tilde{a}, \tilde{a}' \in A$, let $u^{\tilde{a}}_{\tilde{a}'}$ be the discriminating preferences defined by the following utility function: For all a',

$$u^{\tilde{a}}_{\tilde{a}'}(a, a', \theta') = \begin{cases} 1 & \text{if } u_{\theta'} = u^{\tilde{a}}_{\tilde{a}'} \text{ and } a = \tilde{a} \\ 1 & \text{if } u_{\theta'} \neq u^{\tilde{a}}_{\tilde{a}'} \text{ and } a = \tilde{a}' \\ 0 & \text{otherwise} \end{cases}.$$

In words, the preferences $u^{\tilde{a}}_{\tilde{a}'}$ are such that $\tilde{a}$ is a dominant action against an opponent with the same preferences, and $\tilde{a}'$ is the dominant action against all other opponents.

The following result shows that any action a^* that is both a symmetric Nash equilibrium and yields a payoff above the minmax value can be implemented as the unique pure outcome of an ESC. (Recall that θ is used to denote that probability distribution μ puts all weight on θ, i.e. $\mu(\theta) = 1$.)

Proposition 7. *Assume that* $\left(u^{a^*}_{a_{\bar{M}}}, 1\right) \in \Theta_{ID}$. *If action* a^* *is a symmetric Nash equilibrium and* $\pi(a^*, a^*) > \overline{M}$, *then* $\left(\left(u^{a^*}_{a_{\bar{M}}}, 1\right), a^*\right)$ *is an ESC.*

Proof. Suppose that all incumbents are of type $\left(u^{a^*}_{a_{\bar{M}}}, 1\right)$. Note that in all focal post-entry configurations the incumbent $\left(u^{a^*}_{a_{\bar{M}}}, 1\right)$ always plays either a^* or $a_{\bar{M}}$. Against a mutant $(\theta, 1)$ with cognitive level 1, an incumbent plays a^* if and only if $u(\theta) = u^{a^*}_{a_{\bar{M}}}$. The fact that $\pi(a^*, a^*) > \overline{M}$ implies that any mutant $\theta \neq \left(u^{a^*}_{a_{\bar{M}}}, 1\right)$ earns a strictly lower payoff against the incumbents in any post-entry configuration. As a result, if the frequency of mutants is sufficiently small, then they are strictly outperformed. Against a mutant (θ, n) with cognitive level $n > 1$, an incumbent may play either a^* or $a_{\bar{M}}$. Since a^* is a symmetric Nash equilibrium and $\pi(a^*, a^*) > \overline{M}$ the mutants earn at most $\pi(a^*, a^*)$ in matches against incumbents. Consequently, as the fraction of mutants vanishes the average fitness of mutants is weakly less than $\pi(a^*, a^*) - k_n$, and the average fitness of the incumbents is $\pi(a^*, a^*)$. Since k is strictly increasing this implies that $\left(\left(u^{a^*}_{a_{\bar{M}}}, 1\right), a^*\right)$ is an ESC. □

The results of this section imply the following corollary, which characterises pure outcomes of stable configurations in terms of being Nash equilibria that yield payoffs above the pure maxmin/minmax values.

Corollary 4.

1. *If action* a^* *is a Nash equilibrium and* $\pi(a^*, a^*) > \bar{M}$, *then it is the pure outcome of an ESC.*

2. If action a^* is a pure outcome of an NSC and $\forall n,\ k_{n+1} - k_n < \delta$, then a^* is a symmetric Nash equilibrium and $\pi(a^*, a^*) \geq \underline{M}$.

6.4 Application: In-Group Cooperation and Out-Group Exploitation

The following table represents a family of Hawk-Dove games. When both players play D (Dove) they earn 1 each and when they both play H (Hawk) they earn 0. When a player plays H against an opponent playing D, she obtains an additional gain of $g > 0$ and the opponent incurs a loss of $l \in (0, 1)$.

$$\begin{array}{ccc} & H & D \\ H & 0,0 & 1+g, 1-l \\ D & 1-l, 1+g & 1,1 \end{array}. \tag{2}$$

It is natural to think of mutual play of D as the cooperative outcome. We define preferences that induce players to cooperate with their own kind and to seek to exploit those who are not of their own kind.

Definition 15. Let u^n denote the preferences such that:

(1) If $u_{\theta'} = u^n$ and $n_{\theta'} = n$ then $u^n(D, a', \theta') = 1$ and $u^n(H, a', \theta') = 0$ for all a'.

(2) If $u_{\theta'} \neq u^n$ or $n_{\theta'} \neq n$ then $u^n(H, a', \theta') = 1$ and $u^n(D, a', \theta') = 0$ for all a'.

Thus, facing someone who is of the same type, an individual with u^n-preferences strictly prefers cooperation, in the sense of playing D. When facing someone who is not of the same type, an individual with u^n-preferences prefers the exploitative outcome (H, D), and after that she prefers the destructive outcome (H, H) over the remaining outcomes.

Under the assumption that $g > l$ and that the marginal cognitive costs are sufficiently small (but non-vanishing), we construct an ESC in which only individuals with preferences from $\{u^i\}_{i=1}^{\infty}$ are present. Individuals of different cognitive levels co-exist, and non-Nash profiles are played in all matches between equals. When individuals of the same level meet, they play mutual cooperation (D, D). When individuals of different levels meet, the higher level plays H and the lower level plays D. The gain from obtaining the high payoff of $1 + g$ against lower types is exactly compensated by the higher cognitive costs. In contrast, if $g < l$ then the game does not admit this kind of stable configuration.

Proposition 8. *Let G be the game represented in (2), where $g > 0$ and $l \in (0, 1)$. Suppose that the marginal cognitive cost is small but non-vanishing, so that (a) there is an N such that $k_N \leq l + g < k_{N+1}$, and (b) it holds that $g > k_{n+1} - k_n$ for all $n \leq N$.*

(i) If $g > l$ then there exists an ESC (μ^*, b^*), such that $C(\mu^*) \subseteq \{(u^n, n)\}_{n=1}^{N}$, and μ^* is mixed (i.e. $|C(\mu^*)| > 1$). The behaviour of the incumbents is as follows: if the individuals in a match are of different cognitive levels, then the higher level plays H and the lower level plays D; if both individuals in a match are of the same cognitive level, then they both play D.

(ii) If $g = l$ then there exists an NSC with the above properties.

(iii) If $g < l$ then there does not exist any NSC (μ^*, b^*), such that $C(\mu^*) \subseteq \{(u^n, n)\}_{n=1}^{\infty}$.

The formal proof is presented in Appendix C.3.

7 Conclusion and Directions for Future Research

We have developed a model in which preferences co-evolve with the ability to detect others' preferences and misrepresent one's own preferences. We do this by allowing for heterogeneity with respect to costly cognitive ability. The assumption of an exogenously-given level of observability of the opponent's preferences, which has characterised the indirect evolutionary approach so far, is replaced by a Machiavellian notion of deception equilibrium, which endogenously determines what each player observes.

Only one, seemingly small, aspect of the observability structure remains exogenous: the observability among players with equal cognitive levels. Our main results surprisingly shows that a "grain" of perfect observability among equals is enough to imply that only efficient configurations can be stable. However, we later show that this result crucially depend on the "tie-breaking" rule. If we assume that players with equal levels cannot observe each other's preferences, then stable configurations do not have to be efficient (instead they are closely related to Nash equilibria of the underlying games). We leave for future research a model that will endogenize also the observability among equals.

Our model assumes a very powerful form of deception. This allows us to derive sharp results that clearly demonstrate effects of endogenising observation, and introducing deception. We expect similar but weaker effects to be present when deception takes a weaker form. Specifically, we think that the "Bayesian" deception is an interesting model for future research: each incumbent type is associated with a signal, agents with high cognitive levels can mimic the signals of types with lower cognitive levels, and agents maximise their preferences given the received signals and the correct Bayesian inference about the opponent's type.

In a companion paper (Heller & Mohlin, 2014) we study environments in which players are randomly matched, and make inferences about the opponent's type by observing her past behaviour (rather than observing the type directly as is standard in the "indirect evolutionary approach"). In future research, it would be interesting to combine both approaches and allow the observation of the past behaviour to be influenced by deception.

Most papers taking the indirect evolutionary approach study the stability of preferences defined over material outcomes. Moreover, it is common to restrict attention to some parameterised class of such preferences. Since we study preferences defined on the more abstract level of action profiles (or the joint set of action profiles and opponent's types in the case of type-interdependent preferences) we do not make predictions about whether some particular kind of preferences over material outcomes, from a particular family of utility functions, will be stable or not. It would be interesting to extend our model to such classes of preferences. Furthermore, with preferences defined over material outcomes it would be possible to study co-evolution of preferences and deception not only in isolated games, but also when individuals play many different games using the same preferences. We hope to come back to these questions and we invite others to employ and modify our framework in these directions.

A Proofs

A.1 Preliminaries

This subsection section contains notation and definitions that will be used in the following proofs.

A generous action is an action such that if played by the opponent, it allows a player to achieve the maximal fitness payoff. Formally:

Definition 16. Action $a_g \in A$ is *generous*, if there exists an $a \in A$ such that $\pi(a, a_g) \geq \pi(a', a'')$ for all $a', a'' \in A$.

Fix a generous action $a_g \in A$ of the game G. A second-best generous action is an action such that if played by the opponent, it allows a player to achieve the maximal fitness payoff under the constraint that the opponent is not allowed to play the generous action a_g. Formally:

Definition 17. Action $a_{g_2} \in A$ is *second-best generous*, conditional on $a_g \in A$ being first best generous, if there exists $a \in A$ such that $\pi(a, a_{g_2}) \geq \pi(a', a'')$ for alll$a', a'' \in A$ such that $a'' \neq a_g$.

Fix a generous action $a_g \in A$, and fix a second-best generous action $a_{g_2} \in A$, conditional on $a_g \in A$ being first best generous. For each $\alpha \geq \beta \geq 0$, let $u_{\alpha,\beta}$ be the following utility function:

$$u_{\alpha,\beta}(a,a') = \begin{cases} \alpha & a' = a_g \\ \beta & a' = a_{g_2} \\ 0 & \text{otherwise} \end{cases}.$$

Observe that such a utility function$u_{\alpha,\beta}$ satisfies:

1. *Indifference* - the utility functiononly depends on the opponent's action; i.e., the player is indifferent between any two of his own actions.

2. *Pro-generosity* - the utility is highest (second-highest) if the opponent plays the (second-best) generous action, and lowest otherwise.

Let $U_{GI} = \{u_{\alpha,\beta} | \alpha \geq \beta \geq 0\}$ be the family of all such preferences, called *pro-generosity indifferent preferences*. Note that U_g includes a continuum of different utilities (under the assumption that G includes at least three actions). Thus for any set of incumbent types we can always find a utility function in U_g which is does not belong to any of the current incumbents.

A.2 Proof of Theorem 1 (Behaviour of the Highest Types)

A.2.1 Proof of Theorem 1.1

Assume to the contrary that $\pi\left(b_{\bar{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\bar{\theta}\right)\right) < \bar{\pi}$. (Note that the definition of $\bar{\pi}$ implies that the opposite inequality is impossible). Let $a_1, a_2 \in A$ be any two actions such that (a_1, a_2) is an efficient action profile, i.e. $0.5 \cdot (\pi(a_1,a_2) + \pi(a_1,a_2)) = \bar{\pi}$. Let θ_1, θ_2 be two types that satisfy: (1) the types are not incumbents; $\theta_1, \theta_2 \notin C(\mu*)$, (2) both types have the highest incumbent cognitive level; $n_{\theta_1} = n_{\theta_2} = \bar{n}$, and (3) both types have different pro-generosity indifferent preferences; $u_{\theta_1}, u_{\theta_2} \in U_{GI}$ and $u_{\theta_1} \neq u_{\theta_2}$. Let μ' be the distribution that assigns mass 0.5 to each of these types. The post-entry type distribution is $\tilde{\mu} = (1-\epsilon)\cdot\mu + \epsilon\cdot\mu'$. Let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality; $\tilde{b}_\theta(\theta') = b_\theta(\theta')$ for each incumbent pair $\theta, \theta' \in C(\mu^*)$.

2. In matches between mutants and incumbents of lower types, behaviour is such that the mutants maximize their fitness; $\left(\tilde{b}_{\theta_i}(\theta'), \tilde{b}_{\theta'}(\theta_i)\right) \in FMDE(\theta_i, \theta')$ for each $i \in \{1,2\}$ and $\theta' \in C(\mu^*)$ with $n_{\theta'} < \bar{n}$.

3. In matches between mutants and incumbents of the highest type, the mutants mimick $\bar{\theta}$ and the higher types play the same way they play against $\bar{\theta}$; $\left(\tilde{b}_{\theta_i}(\theta'), \tilde{b}_{\theta'}(\theta_i)\right) = \left(b_{\bar{\theta}}(\theta'), b_{\theta'}\left(\bar{\theta}\right)\right)$, for each $i \in \{1,2\}$ and $\theta' \in C(\mu^*)$ with $n_{\theta'} = \bar{n}$.

4. Two mutants of *different* types play efficiently when meeting each other; $\tilde{b}_{\theta_i}(\theta_j) = a_i$ for each $i \neq j \in \{1,2\}$.

5. Two mutants of the *same* type play like $\bar{\theta}$ plays against itself, when meeting each other;$\tilde{b}_{\theta_i}(\theta_i) = b_{\bar{\theta}}\left(\bar{\theta}\right)$ for each $i \in \{1,2\}$.

In virtue of point 1 the construction $\left(\tilde{\mu}, \tilde{b}\right)$ is a focal configuration (with respect to (μ, b)). By point 2 the mutants θ_1 and θ_2 earn weakly more than $\bar{\theta}$ against lower types. By point 3 the mutants earn exactly the same as $\bar{\theta}$ aginst all the highest incumbent types (including $\bar{\theta}$). By 4 and 5 the mutants on average earn strictly more than $\bar{\theta}$ against the mutants. In total average fitness earned by θ_1 and θ_2 is strictly higher than that of $\bar{\theta}$, against a population who follows $\left(\tilde{\mu}, \tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^* in the population game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$. Thus, μ^* is not a symmetric Nash equilibrium, and therefore it is not an NSS, in $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$, which implies that that μ^* is not an NSC.

Remark 4. In this proof the mutants only outperform the incumbents on average. This allows for the possibility that one mutant earns strictly less than $\bar{\theta}$, while the other mutant earns strictly more than $\bar{\theta}$. That is enough to prove the desired result. However, we could also have proved our result with a construction involving three different mutants $\theta_1, \theta_2, \theta_3$, each of which earns strictly more than $\bar{\theta}$ in the post-entry configuration. This would be achieved if the three mutant types are equally frequent, and any two mutants of the *same* type play like $\bar{\theta}$ plays against itself (point 5 above), but play as follows when facing another mutant: (i) mutant θ_1 plays a_1 against θ_2 and a_2 against θ_3, (ii) mutant θ_2 plays a_1 against θ_3 and a_2 against θ_1, and (iii) mutant θ_3 plays a_1 against θ_1 and a_2 against θ_2.

A.2.2 Proof of Theorem 1.2

Assume to the contrary that $\left(\left(b_{\bar{\theta}}\left(\underline{\theta}\right), b_{\underline{\theta}}\left(\bar{\theta}\right)\right)\right) \notin FMDE\left(\bar{\theta}, \underline{\theta}\right)$. Let $\hat{\theta}$ be a type that satisfies: (1) not being an incumbent; $\hat{\theta} \notin C\left(\mu*\right)$, (2) having the highest incumbent cognitive level; $n_{\hat{\theta}} = \bar{n}$, and (3) having pro-generosity indifferent preferences; $u_{\hat{\theta}} \in U_{GI}$. Let μ' be the distribution that assigns mass one to type $\hat{\theta}$. The post-entry type distribution is $\tilde{\mu} = (1-\epsilon) \cdot \mu + \epsilon \cdot \mu'$. Let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality; $\tilde{b}_\theta\left(\theta'\right) = b_\theta\left(\theta'\right)$ for each $\theta, \theta' \in C\left(\mu^*\right)$.

2. In matches between mutants and incumbents of lower types, behaviour is such that the mutants maximize their fitness; $\left(\tilde{b}_{\hat{\theta}}\left(\theta'\right), \tilde{b}_{\theta'}\left(\hat{\theta}\right)\right) \in FMDE\left(\hat{\theta}, \theta'\right)$ for each $\theta' \in C\left(\mu^*\right)$ with $n_{\theta'} < \bar{n}$.

3. In matches between mutants and incumbents of the highest type, the mutant mimicks $\bar{\theta}$ and the higher types play the same way they play against $\bar{\theta}$; $\left(\tilde{b}_{\hat{\theta}}\left(\theta'\right), \tilde{b}_{\theta'}\left(\hat{\theta}\right)\right) = \left(b_{\bar{\theta}}\left(\theta'\right), b_{\theta'}\left(\bar{\theta}\right)\right)$, for each $\theta' \in C\left(\mu^*\right)$ with $n_{\theta'} = \bar{n}$.

Note that $\left(\tilde{\mu}, \tilde{b}\right)$ is a focal configuration (with respect to (μ, b)), and that $\hat{\theta}$ obtains a strictly higher fitness than $\bar{\theta}$ against a population who follows $\left(\tilde{\mu}, \tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^* in the population game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$. Thus, μ^* is not a symmetric Nash equilibrium, and therefore it is not an NSS, in $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$, which implies that that μ^* is not an NSC.

A.2.3 Proof of Theorem 1.3

Assume to the contrary that $\pi\left(b_{\underline{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\underline{\theta}\right)\right) > \bar{\pi}$, which immediately implies that $\pi\left(b_{\bar{\theta}}\left(\underline{\theta}\right), b_{\underline{\theta}}\left(\bar{\theta}\right)\right) < \bar{\pi}$. Let $\hat{\theta}$ be a type that satisfies: (1) not being an incumbent; $\hat{\theta} \notin C\left(\mu*\right)$, (2) having the highest incumbent cognitive level; $n_{\hat{\theta}} = \bar{n}$, and (3) having pro-generosity indifferent preferences; $u_{\hat{\theta}} \in U_{GI}$. Let μ' be the distribution that assigns mass one to type $\hat{\theta}$. The post-entry type distribution is $\tilde{\mu} = (1-\epsilon) \cdot \mu + \epsilon \cdot \mu'$. Let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality; $\tilde{b}_\theta\left(\theta'\right) = b_\theta\left(\theta'\right)$ for each $\theta, \theta' \in C\left(\mu^*\right)$.

2. In matches between mutants and incumbents of lower types, behaviour is such that the mutants maximize their fitness; $\left(\tilde{b}_{\hat{\theta}}\left(\theta'\right), \tilde{b}_{\theta'}\left(\hat{\theta}\right)\right) \in FMDE\left(\hat{\theta}, \theta'\right)$ for each $\theta' \in C\left(\mu^*\right)$ with $n_{\theta'} < \bar{n}$.

3. In a match between a mutant $\hat{\theta}$ and the incumbent $\bar{\theta}$, the mutant mimicks $\underline{\theta}$, and the incumbent $\bar{\theta}$ plays the same way it plays against $\underline{\theta}$; $\left(\tilde{b}_{\hat{\theta}}\left(\bar{\theta}\right), \tilde{b}_{\bar{\theta}}\left(\hat{\theta}\right)\right) = \left(b_{\underline{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\underline{\theta}\right)\right)$.

4. $\hat{\theta}$ plays against itself the same way $\bar{\theta}$ plays against itself; $\left(\tilde{b}_{\hat{\theta}}\left(\hat{\theta}\right),\tilde{b}_{\hat{\theta}}\left(\hat{\theta}\right)\right)=\left(\tilde{b}_{\bar{\theta}}\left(\bar{\theta}\right),\tilde{b}_{\bar{\theta}}\left(\bar{\theta}\right)\right)$, and

5. $\hat{\theta}$ mimics $\bar{\theta}$ against all other highest types, and these higher types play against $\hat{\theta}$ the same as they play against $\bar{\theta}$; $\left(\tilde{b}_{\hat{\theta}}\left(\theta'\right),\tilde{b}_{\theta'}\left(\hat{\theta}\right)\right)=\left(b_{\bar{\theta}}\left(\theta'\right),b_{\theta'}\left(\bar{\theta}\right)\right)$ for each $\theta'\neq\bar{\theta}$ with $n_{\theta'}=\bar{n}$.

Note that $\left(\tilde{\mu},\tilde{b}\right)$ is a focal configuration (with respect to (μ,b)). By point 2 the mutant $\hat{\theta}$ earns weakly more than $\bar{\theta}$ against lower types. By point 3 (and Theorem 1.1) the mutants earn strictly more than $\bar{\theta}$ against type $\bar{\theta}$. By 3 and 4 (and Theorem 1.1 and) the mutant earns strictly more than $\bar{\theta}$ against the mutant. By 5 earns the same as $\bar{\theta}$ against all other types. In total average fitness earned by $\hat{\theta}$ is strictly higher than that of $\bar{\theta}$, against a population who follows $\left(\tilde{\mu},\tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^* in the population game $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$. Thus, μ^* is not a symmetric Nash equilibrium, and therefore it is not an NSS, in $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$, which implies that that μ^* is not an NSC.

A.3 Proof of Theorem 2 (Each Type Plays Efficiently Against Itself)

We begin by proving a lemma.

Lemma 1. *If $(\sigma_1,\sigma_2)\in DE\left(\theta_1,\theta_2\right)$ then there exist actions $a_1,a_1'\in supp\left(\sigma_1\right)$ and $a_2,a_2'\in supp\left(\sigma_2\right)$ such that:$(a_1,a_2)\in DE\left(\theta_1,\theta_2\right)$, and $(a_1',a_2')\in DE\left(\theta_1,\theta_2\right)$, with $\pi\left(a_1,a_2\right)\geq\pi\left(\sigma_1,\sigma_2\right)$, and $\pi\left(a_1',a_2'\right)\leq\pi\left(\sigma_1,\sigma_2\right)$.*

Proof. Note that for any mixed deception equilibrium (σ_1,σ_2), and any action $a\in supp\left(\sigma_2\right)$, the profile (σ_1,a) is also a deception equilibrium (because othewise the deceiver would not induce the decieved party to take a mixed action that puts positive weight on a). It follows that there are actions actions $a_2,a_2'\in supp\left(\sigma_2\right)$ such that (σ_1,a_2) and (σ_1,a_2') are deception equilibria, with $\pi\left(\sigma_1,a_2\right)\geq\pi\left(\sigma_1,\sigma_2\right)$ and $\pi\left(\sigma_1,a_2'\right)\leq\pi\left(\sigma_1,\sigma_2\right)$. Furthermore, if (σ_1,a_2) and (σ_1,a_2') are deception equilibria then, for any action $a\in supp\left(\sigma_1\right)$, the profiles (a,a_2) and (a,a_2') are also deception equilibria, with $\pi\left(\sigma_1,a_2\right)=\pi\left(a,a_2\right)$ and $\pi\left(\sigma_1,a_2'\right)=\pi\left(a,a_2'\right)$. Hence there are actions $a_1,a_1'\in supp\left(\sigma_1\right)$ such that (a_1,a_2) and (a_1',a_2') are deception equilibria, with $\pi\left(a_1,a_2\right)=\pi\left(\sigma_1,a_2\right)\geq\pi\left(\sigma_1,\sigma_2\right)$, and $\pi\left(a_1,a_2'\right)=\pi\left(\sigma_1,a_2'\right)\leq\pi\left(\sigma_1,\sigma_2\right)$. □

We are now able to prove Theorem 2. Assume to the contrary that there exists an incumbent that does not play $\bar{a}$ (the unique efficient action) against itself. (From Theorem 1.1 we know that $\mathring{\theta}$ cannot be of the highest type in the population, denoted $\bar{n}$ above.) Let $\mathring{\theta}\in C\left(\mu^*\right)$ be the type with the highest cognitive level, among those types who do not play $\bar{a}$ against themselves. That is, $\left(b_{\mathring{\theta}}\left(\mathring{\theta}\right),b_{\mathring{\theta}}\left(\mathring{\theta}\right)\right)\neq\left(\bar{a},\bar{a}\right)$, and if $n_\theta>n_{\mathring{\theta}}$, then $\left(b_\theta\left(\theta\right),b_\theta\left(\theta\right)\right)=\left(\bar{a},\bar{a}\right)$.

We consider a mutant $\hat{\theta}=\left(n_{\mathring{\theta}},\hat{u}\right)\notin C\left(\mu^*\right)$. If $\Sigma\left(u_{\mathring{\theta}}\right)=\Delta$, then we let $\hat{u}\in U_{GI}$ be such that $\hat{\theta}=\left(n_{\mathring{\theta}},\hat{u}\right)\notin C\left(\mu^*\right)$. If $\Sigma\left(u_{\mathring{\theta}}\right)\neq\Delta$, then we fix a dominated action $\underline{a}\in A\backslash\Sigma\left(u_{\mathring{\theta}}\right)$, and let $\hat{u}$ be defined as follows:

$$\hat{u}\left(a,a'\right)=\begin{cases}max_{a\in A}\left(u_{\mathring{\theta}}\left(a,\bar{a}\right)\right) & a=a'=\bar{a}\\ u_{\mathring{\theta}}\left(\underline{a},a'\right)-\beta_{a'} & a=\underline{a}\text{ and }a'\neq\bar{a}\\ u_{\mathring{\theta}}\left(a,a'\right) & \text{otherwise}\end{cases},$$

where each $\beta_{a'}\geq0$ is chosen such that $\hat{\theta}=\left(n_{\mathring{\theta}},\hat{u}\right)\notin C\left(\mu^*\right)$. That is, if $\Sigma\left(u_{\mathring{\theta}}\right)\neq\Delta$, then the utility function $\hat{u}$ is constructed from the utility function $u_{\mathring{\theta}}$ by arbitrarily lowering the payoff of some of the outcomes associated with the (already) dominated action $\underline{a}$ and which do not involve action $\bar{a}$, while increasing the payoff of the outcome $(\bar{a},\bar{a})$ by the minimal amount that make it a best reply to itself. Note that this definition of $\hat{u}$ is valid also for the case of $\bar{a}=\underline{a}$. It follows that $a\in\Sigma\left(u_{\mathring{\theta}}\right)\cup\{\bar{a}\}$ if and only if $a\in\Sigma\left(\hat{u}\right)$. (To see this, note that if $\Sigma\left(u_{\mathring{\theta}}\right)\neq\Delta$ and $\underline{a}=\bar{a}$, then $\Sigma\left(\hat{u}\right)=\Sigma\left(u_{\mathring{\theta}}\right)\cup\{\bar{a}\}$. Otherwise $\Sigma\left(\hat{u}\right)=\Sigma\left(u_{\mathring{\theta}}\right)$.) Thus, $\hat{\theta}$ can be induced to play exactly the same pure actions as $\mathring{\theta}$, unless $\bar{a} = \underline{a}$, in which case $\hat{\theta}$ can be induced to play $\bar{a}$ in addition to all actions that $\mathring{\theta}$ can be induced to play.)

Let μ' be the distribution that assigns mass one to type $\left(n_{\mathring{\theta}}, \hat{u}\right)$. Let the post-entry type distribution be $\tilde{\mu} = (1-\epsilon) \cdot \mu + \epsilon \cdot \mu'$, and let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality; $\tilde{b}_{\theta}(\theta') = b_{\theta}(\theta')$ for each $\theta, \theta' \in C(\mu^*)$.

2. In matches between the mutant type $\hat{\theta}$ and any lower type $\theta' \in C(\mu^*)$ (with $n_{\theta'} < n_{\hat{\theta}}$), we distinguish two cases.

    (a) Suppose that $\Sigma\left(u_{\mathring{\theta}}\right) = \Delta$. In this case let $\left(\tilde{b}_{\hat{\theta}}(\theta'), \tilde{b}_{\theta'}\left(\hat{\theta}\right)\right) \in FMDE\left(\hat{\theta}, \theta'\right)$.

    (b) Suppose that $\Sigma\left(u_{\mathring{\theta}}\right) \neq \Delta$. In this case let $\left(\tilde{b}_{\hat{\theta}}(\theta'), \tilde{b}_{\theta'}\left(\hat{\theta}\right)\right) = (a_1, a_2)$, for some $(a_1, a_2) \in DE\left(\mathring{\theta}, \theta'\right)$ such that $\pi(a_1, a_2) \geq \pi\left(b_{\mathring{\theta}}(\theta'), b_{\theta'}\left(\mathring{\theta}\right)\right)$. By lemma **XXX insert ref to lemma above XXX** above such a profile (a_1, a_2) exists.

3. If $(\sigma_1, \sigma_2) \in DE(\theta_1, \theta_2)$ then there exist actions $a_1 \in supp(\sigma_1)$ and $a_2 \in supp(\sigma_2)$ such that $(a_1, a_2) \in DE(\theta_1, \theta_2)$ and$\pi(a_1, a_2) \geq \pi(\sigma_1, \sigma_2)$

4. In matches between the mutant type $\hat{\theta}$ and any incumbent type θ' with same level, the mutant $\hat{\theta}$ mimics $\mathring{\theta}$, and the incumbent θ' treats the mutant $\hat{\theta}$ like the incumbent $\mathring{\theta}$; $\left(\tilde{b}_{\hat{\theta}}(\theta'), \tilde{b}_{\theta'}\left(\hat{\theta}\right)\right) = \left(b_{\mathring{\theta}}(\theta'), b_{\theta'}\left(\mathring{\theta}\right)\right)$ if $n_{\theta'} = n_{\hat{\theta}}$ and $\theta' \neq \hat{\theta}$.

5. The mutant plays efficiently when meeting itself; $\tilde{b}_{\hat{\theta}}\left(\hat{\theta}\right) = \bar{a}$.

6. In matches between the mutant type $\hat{\theta}$ and any higher type $\theta' \in C(\mu^*)$ (with $n_{\theta'} > n_{\hat{\theta}}$), we distinguish two cases. Pick a profile $(a_1, a_2) \in DE\left(\theta', \mathring{\theta}\right)$, such that $\pi(a_2, a_1) \geq \pi\left(b_{\mathring{\theta}}(\theta'), b_{\theta'}\left(\mathring{\theta}\right)\right)$. By lemma **XXX insert ref to lemma above XXX** above such a profile (a_1, a_2) exists. Moreover, (since $a \in \Sigma\left(u_{\mathring{\theta}}\right) \cup \{\bar{a}\}$ if and only if $a \in \Sigma(\hat{u})$) it is either the case that $(a_1, a_2) \in DE\left(\theta', \hat{\theta}\right)$ or that $a_2 = \bar{a}$. In the latter case, if $a_2 = \bar{a}$, then $(\bar{a}, \bar{a}) \in DE\left(\theta', \hat{\theta}\right)$. This is due to the fact that$(b_{\theta'}(\theta'), b_{\theta'}(\theta')) = (\bar{a}, \bar{a})$ implies that $\bar{a}$ is a best reply to $\bar{a}$ for type θ'. Note that $\pi(\bar{a}, \bar{a}) \geq \pi\left(b_{\mathring{\theta}}(\theta'), b_{\theta'}\left(\mathring{\theta}\right)\right)$ by Theorem 1.3.

    (a) If $u_{\theta'}(a_1, a_2) > u_{\theta'}(\bar{a}, \bar{a})$ let $\left(\tilde{b}_{\theta'}\left(\hat{\theta}\right), \tilde{b}_{\hat{\theta}}(\theta')\right) = (a_1, a_2)$.

    (b) If $u_{\theta'}(a_1, a_2) \leq u_{\theta'}(\bar{a}, \bar{a})$ let $\left(\tilde{b}_{\theta'}\left(\hat{\theta}\right), \tilde{b}_{\hat{\theta}}(\theta')\right) = (\bar{a}, \bar{a})$.

By point 1, $\left(\tilde{\mu}, \tilde{b}\right)$ is a focal configuration (with respect to (μ, b)). By point 2 the mutant $\hat{\theta}$ earns weakly more than $\mathring{\theta}$ against lower types. By point 3 the mutant $\hat{\theta}$ earns the same as $\mathring{\theta}$ against all incumbents of level $n_{\mathring{\theta}}$. By 3 and 4, (and the assumption that $\mathring{\theta}$ deos not play efficiently against itself) the mutant $\hat{\theta}$ earns strictly more than $\mathring{\theta}$ against $\hat{\theta}$. By 5 the mutant $\hat{\theta}$ earns weakly more than $\mathring{\theta}$ against all incumbents of higher cognitive level. In total average fitness earned by $\hat{\theta}$ is strictly higher than that of $\mathring{\theta}$, against a population who follows $\left(\tilde{\mu}, \tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^* in the population game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$. Thus, μ^* is not a symmetric Nash equilibrium, and therefore it is not an NSS of $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$, which implies that that μ^* is not an NSC.

B Variant with Uniform Deception

In this section we describe how to adapt our model in a way that requires players to use the *same* mixed action in their deception efforts towards all opponents with lower cognitive levels. We implement this change by replacing the definition of configuration with a new notion of *configuration with uniform deception*.

Definition 18. A *configuration with uniform deception* is a pair (μ, b) where $\mu \in \Delta(U)$ is a type distribution, and $b : C(\mu) \times C(\mu) \longrightarrow \Delta(A)$ is a behavioural policy such that

1. For each type $\theta \in C(\mu)$, there exists $\tilde{\sigma}(\theta)$ that satisfies

$$\tilde{\sigma}(\theta) \in \arg\max_{\sigma \in \Delta(A)} \left(\sum_{\theta' \in C(\mu),\ n_{\theta'} < n_{\theta}} \mu(\theta') \cdot \max_{\sigma' \in BR_u(\sigma)} u_{\theta}(\sigma, \sigma') \right), \text{ and}$$

2. For each $\theta, \theta' \in C(\mu)$; $b_{\theta}(\theta') = \tilde{\sigma}(\theta)$ and

$$n_{\theta} = n_{\theta'} \Longrightarrow (b_{\theta}(\theta'), b'_{\theta}(\theta)) \in NE(\theta, \theta'), \text{ and}$$

$$n_{\theta} > n_{\theta'} \Longrightarrow b_{\theta'}(\theta) \in BR_{u_{\theta'}}(\tilde{\sigma}(\theta)).$$

We interpret $\tilde{\sigma}(\theta)$ as the strategy that lower levels are deceived into believing is being played by type θ, and we interpret $b_{\theta}(\theta')$ as the strategy of type θ when being matched with type θ'.

We restrict our definition of a neutrally stable configuration to a configuration with uniform deceptions:

Definition 19. A configuration (μ, b) is a *neutrally stable configuration (NSC) with uniform deception***,** if for every $\mu' \in \Delta(\Theta)$, there is some $\bar{\varepsilon} \in (0,1)$ such that if $(\tilde{\mu}, \tilde{b})$, where $\tilde{\mu} = (1-\varepsilon) \cdot \mu + \varepsilon \cdot \mu'$, is a focal configuration with uniform deceptions, then μ is an NSS in the type game $\Gamma_{(\tilde{\mu}, \tilde{b})}$.

An analogous change can be made to the setup of interdependent preferences. All other details of the model are unchanged. It is relatively straightforward to see that *all our results hold also in this setup of uniform deceptions, with minor adaptations to the proofs.*

C Constructions of Heterogeneous NSCs in Examples

The first subsection presents a lemma on stable heterogeneous populations, which will later be used to construct NSCs in the Rock-Paper-Scissors and Hawk-Dove games, with type-neutral and type-interdependent preferences.

C.1 A Useful Lemma on Stable Heterogeneous Populations

Consider a configuration (μ, b), consisting of a type distribution with (finite) support $C(\mu) \subseteq \{(u, n)\}_{n=1}^{\infty}$, and behaviour policies such that

$$\pi(b_{\theta}(\theta'), b_{\theta'}(\theta)) = \begin{cases} t & \text{if } n_{\theta} > n_{\theta'} \\ w & \text{if } n_{\theta} = n_{\theta'} \\ s & \text{if } n_{\theta} < n_{\theta'} \end{cases}. \tag{3}$$

Thus t is the payoff that a player of type θ earns when deceiving an opponent of type θ', and s is the payoff earned by the deceived party. When two individuals of the same type meet they earn w. Our first lemma concerns the type game $\Gamma_{(\mu,b)}$ that is induced by a configuration (μ, b), such that $C(\mu) \subseteq \{(u, n)\}_{n=1}^{\infty}$ and with behaviour policies given by (3). Although we have normalised $k_1 = 0$ in the main text, we do not omit reference to k_1 in what follows. This is done to simplify the proofs.

Lemma 2. *Suppose $t \geq w \geq s$. Suppose that there is an N such that*

$$k_N - k_1 \leq t - s < k_{N+1} - k_1, \tag{4}$$

and suppose that

$$t - w > k_{n+1} - k_n \text{ for all } n \leq N. \tag{5}$$

Consider the type game $\Gamma_{(\mu,b)}$ induced by a configuration (μ, b) with a type distribution such that $C(\mu) \subseteq \{(u,n)\}_{n=1}^{\infty}$, and with behaviour policies given by (3).

1. *If $2w < s + t$ then $\Gamma_{(\mu,b)}$ has a unique ESS $\mu^* \in \Delta(C(\mu))$, which is mixed, i.e. $C(\mu^*) > 1$, and in which no type above N is present, i.e. $C(\mu^*) \subseteq \{(u,n)\}_{n=1}^{N}$.*
2. *If $2w = s + t$ then $\Gamma_{(\mu,b)}$ has an NSS $\mu^* \in \Delta(C(\mu))$, which is mixed, i.e. $C(\mu^*) > 1$, and in which no type above N is present, i.e. $C(\mu^*) \subseteq \{(u,n)\}_{n=1}^{N}$.*
3. *If $2w > s + t$ then $\Gamma_{(\mu,b)}$, admits no NSS and hence no ESS.*

The rest of this subsection is devoted to proving this result.

First note that that type $(u, N+1)$ earns strictly less than $(u,1)$ at all population states, and (u, N) earns at least as much as $(u,1)$ at least at some population state. This immediately follows from $s \leq w \leq t$ and $t - k_{N+1} < s - k_1$ and $s - k_1 \leq t - k_N$. For this reason it is sufficient to consider the type distributions with support in $\{(u,n)\}_{n=1}^{N}$. The payoffs for a type game with all these types present are

$$\begin{array}{ccccccc}
 & (u,1) & (u,2) & (u,3) & \dots & (u,N-1) & (u,N) \\
(u,1) & w-k_1 & s-k_1 & s-k_1 & \dots & s-k_1 & s-k_1 \\
(u,2) & t-k_2 & w-k_2 & s-k_2 & \dots & s-k_2 & s-k_2 \\
(u,3) & t-k_3 & t-k_3 & w-k_3 & \dots & s-k_3 & s-k_3 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
(u,N-1) & t-k_{N-1} & t-k_{N-1} & t-k_{N-1} & \dots & w-k_{N-1} & s-k_{N-1} \\
(u,N) & t-k_N & t-k_N & t-k_N & \dots & t-k_N & w-k_N
\end{array},$$

or in matrix form

$$\mathbf{A} = \begin{pmatrix}
w-k_1 & s-k_1 & s-k_1 & \dots & s-k_1 & s-k_1 \\
t-k_2 & w-k_2 & s-k_2 & \dots & s-k_2 & s-k_2 \\
t-k_3 & t-k_3 & w-k_3 & \dots & s-k_3 & s-k_3 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
t-k_{N-1} & t-k_{N-1} & t-k_{N-1} & \dots & w-k_{N-1} & s-k_{N-1} \\
t-k_N & t-k_N & t-k_N & \dots & t-k_N & w-k_N
\end{pmatrix}.$$

Inspecting the matrix $\mathbf{A}$ we make the following observation:

Claim 3. Consider the game with payoff matrix $\mathbf{A}$. Suppose (5) holds.

1. $(u, n+1)$ is the unique best response to n for all $n \in \{1, .., N-2\}$.
   (a) If $t - k_N > s - k_1$ then (u, N) is the unique best reply to $(u, N-1)$.
   (b) If $t - k_N = s - k_1$ then (u, N) and $(u, 1)$ are the only two best replies to $(u, N-1)$.
   (c) $(u, 1)$ is the unique best response to (u, N).

Proof. Condition (5) implies that $t - k_{N+1} > w - k_N$, and the definition of N implies $t - k_{N+1} < s - k_1$. Taken together this implies that $w - k_N < s - k_1$, which means that $(u, 1)$ is the unique best response to (u, N).

The definition of N entails $t - k_N \geq s - k_1$. If $t - k_N > s - k_1$ then (u, N) is the unique best reply to $(u, N-1)$. If $t - k_N = s - k_1$ then (u, N) and $(u, 1)$ are the only two best replies to $(u, N-1)$. Furthermore, (5) implies that $(u, n+1)$ is the unique best response to (u, n) for all $n \in \{1, .., N-2\}$. $\square$

It is an immediate consequence of the above lemma that all Nash equilibria of $\mathbf{A}$ are mixed; i.e. that they have more than one type in their support. Next, we examine the stability properties of such equilibria. As discussed in the proof of Theorem 3, it is well-known that if $\mathbf{A}$ is negative definite (semi-definite) with respect to the tangent space, i.e. if $v \cdot \mathbf{A}v < 0$ for all $v \in \mathbb{R}_0^d = \{v \in \mathbb{R}^d : \sum_{i=1}^d v_i = 0\}$, $v \neq \mathbf{0}$, then $\mathbf{A}$ admits a unique ESS (but not necessarily a unique NSS). Moreover, the set of Nash equilibria coincides with the set of NSS and constitutes a nonempty convex subset of the simplex (Hofbauer & Sandholm 2009, Theorem 3.2).

One can show:

Claim 4. If $2w \geq (\leq)\, s + t$ then $\mathbf{A}$ is positive (negative) semi-definite w.r.t. the tangent space.

Proof. Let

$$\mathbf{K} = \begin{pmatrix} -k_1 & -k_1 & \dots & -k_1 \\ -k_2 & -k_2 & \dots & -k_2 \\ \vdots & \vdots & \ddots & \vdots \\ -k_N & -k_N & \dots & -k_N \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} w & s & \dots & s \\ t & w & \dots & s \\ \vdots & \vdots & \ddots & \vdots \\ t & t & \dots & w \end{pmatrix},$$

so that

$$\mathbf{A} = \mathbf{B} + \mathbf{K}.$$

Note that $v'\mathbf{K}v = 0$ for all $v \in \mathbb{R}_0^N$, $v \neq \mathbf{0}$, so that $v'\mathbf{A}v < 0$ for all $v \in \mathbb{R}_0^N$, $v \neq \mathbf{0}$, if and only if $v'\mathbf{B}v < 0$ for all $v \in \mathbb{R}_0^N$, $v \neq \mathbf{0}$. Moreover, note that $v'\mathbf{B}v < 0$ for all $v \in \mathbb{R}_0^N$, $v \neq \mathbf{0}$, if and only if $v'\bar{\mathbf{B}}v < 0$ for all $v \in \mathbb{R}_0^N$, $v \neq \mathbf{0}$, where

$$\bar{\mathbf{B}} = \frac{1}{2}\left(\mathbf{B} + \mathbf{B}^T\right).$$

One can transform the problem to one of checking negative definiteness with respect to $\mathbb{R}^{N-1}$ rather than the tangent space $\mathbb{R}_0^N$; see, e.g. Weissing (1991). This is done with the $N \times (N-1)$ matrix $\mathbf{P}$ defined by

$$p_{ij} = \begin{cases} 1 & \text{if } n = j \text{and } n, j < N \\ 0 & \text{if } n \neq j \text{and } n, j < N \\ -1 & \text{if } n = N \end{cases}.$$

We have

$$\mathbf{P}'\bar{\mathbf{B}}\mathbf{P} = \left(w - \frac{1}{2}(s+t)\right)(\mathbf{I} + \mathbf{1}\mathbf{1}'),$$

where $\mathbf{1}$ is an $N-1$-dimensional vector with all entries equal to 1, and I is the identity matrix. The matrix $\mathbf{P}'\bar{\mathbf{B}}\mathbf{P}$ has one eigenvalue (of multiplicity $N-1$) that is equal to $2w - (s+t)$. Finally, note that this eigenvalue is non-negative if and only if $2w \geq (s+t)$. $\square$

It follows that if $2w \leq s+t$ then the game with payoff matrix $\mathbf{A}$ admits an NSS. If $2w > s+t$ then the game does not have a mixed NSS. We are now able to prove Lemma 2.

1. If $2w < s+t$ then by Lemma 4 $\mathbf{A}$ is negative definite w.r.t. the tangent space, implying that it has a unique ESS. Lemma 3 implies that there can be no pure Nash equilibria (and hence no pure ESS). Thus $\mathbf{A}$ has a unique Nash equilibrium, which is mixed. As observed earlier, type $(u, N+1)$ (and higher types) earn strictly less than $(\theta, 1)$ for all population states, which implies that that this unique equilibrium remains an ESS also when they are included in the set of feasible types.

2. If $2w = s + t$ then $\mathbf{A}$ is both positive and negative semi-definite w.r.t. the tangent space. In this case $\mathbf{A}$ does not have an ESS but it does have a set of NSSs, all of which are Nash equilibria. Moreover, we know that $\mathbf{A}$ has no pure NE, and so all NSS are mixed. Again, type $(u, N+1)$ (and higher types) can be ignored because they always earn strictly less than $(\theta, 1)$.

3. If $2w < s + t$ then $\mathbf{A}$ is positive definite w.r.t. the tangent space, implying that it has no NSC.

C.2 Proof of Proposition 3: Construction of Equilibrium in Rock-Paper-Scissors Game

Formally the behaviour of the incumbent types is as follows:

$$b_\theta^*(\theta') = \begin{cases} (0,1,0) & \text{if } n_\theta > n_{\theta'} \\ \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) & \text{if } n_\theta = n_{\theta'} \\ (1,0,0) & \text{if } n_\theta < n_{\theta'} \end{cases}.$$

Under the described behavioural policy we have

$$\pi\left(b_\theta(\theta'), b_{\theta'}(\theta)\right) = \begin{cases} 1 & \text{if } n_\theta > n_{\theta'} \\ 0 & \text{if } n_\theta = n_{\theta'} \\ -1 & \text{if } n_\theta < n_{\theta'} \end{cases}.$$

Start by restricting attention to the set of types $\{(u^\pi, n)\}_{n=1}^\infty$. That is, for the moment we use $\{(u^\pi, n)\}_{n=1}^\infty$ instead of Θ as the set of all types. All definitions can be amended accordingly. Lemma 2 in Appendix B implies that there is an NSC (μ^*, b^*), such that $C(\mu^*) \subseteq \{(u^\pi, n)\}_{n=1}^N$, and μ^* is mixed. Lemma 2 establishes that the type game between the types $\{(u^\pi, n)\}_{n=1}^N$ behaves much like an N-player version of a Hawk-Dove game: it has a unique symmetric equilibrium that is in mixed strategies and that is neutrally or evolutionarily stable, depending on whether the payoff matrix of the type game is negative semi-definite, or negative definite, with respect to the tangent space.

It remains to show that types not in $\{(u^\pi, n)\}_{n=1}^\infty$ are unable to invade. Suppose a mutant of type (u', n') enters. Incumbents of level $n > n'$ will give the mutant a belief that induces the mutant to play some action a' and then play action $a' + 1 \bmod 3$, which is the incumbents' best response to a'. Thus, against incumbents of level $n > n'$ the mutant earns -1. Against incumbents of level $n < n'$, the mutant will earn at most 1. Against incumbents of level n' the mutant earns at most 0. Against itself the mutant (or a group of mutants for that matter) will earn 0. Thus any mutant of level n' earns weakly less than the incumbents of level n', in any focal post-entry configuration.

Remark 5. Our analysis is similar to that of Conlisk (2001). Like us, he works with a hierarchy of cognitive types (though in his case it is fixed and finite), where higher cognitive types carry higher cognitive costs. He stipulates that when a high type meets a low type the high type gets 1 and the low type gets -1. If two equals meet both get 0. He shows that there is a neutrally stable equilibrium of this game between types (using somewhat different arguments than we do), and explores comparative static effects of changing costs. However, unlike in our model, in Conlisk's model all individuals have the same materialistic preferences and the payoffs earned from deception are not derived from an underlying game.

C.3 Proof of Proposition 8: Construction of Equilibrium in Hawk-Dove and Type-Interdependent Preferences

Formally, the behaviour of an incumbent $\theta \in C(\mu^*)$ facing another incumbent $\theta' \in C(\mu^*)$ is given by

$$b^*_\theta(\theta') = \begin{cases} D & \text{if } n_\theta \geq n_{\theta'} \\ H & \text{if } n_\theta < n_{\theta'} \end{cases}. \tag{6}$$

Under the described behavioural policy we have, for $\theta, \theta' \in \{(u^n, n)\}_{n=1}^\infty$,

$$\pi(b_\theta(\theta'), b_{\theta'}(\theta)) = \begin{cases} 1+g & \text{if } n_\theta > n_{\theta'} \\ 1 & \text{if } n_\theta = n_{\theta'} \\ 1-l & \text{if } n_\theta < n_{\theta'} \end{cases}.$$

Start by restricting attention to the set of types $\{(u^n, n)\}_{n=1}^\infty$. That is, for the moment, let $\{(u^n, n)\}_{n=1}^\infty$, instead of Θ_{ID}, be the set of all types. All definitions can be amended accordingly. Under this restriction on the set of types, the desired results (i)–(iii) follow from Lemma 2 in Appendix B. For example, to see that Lemma 2 implies part (i) for the restricted type set, note that $g > l$ implies that $2w < t + s$, and $g > k_{n+1} - k_n$ implies that $t - w > k_{n+1} - k_n$, in the language of Lemma 2. The arguments for (ii) and (iii) are analogous.

Next, allow for a larger set of types Θ_{ID}, such that $\{(u^n, n)\}_{n=1}^\infty \subseteq \Theta_{ID}$. The fact that part (iii) of Proposition 8 holds for the restricted set of types implies that it also holds for any larger set of types. It remains to prove parts (i) and (ii) for the full set of types. We prove only part (i). The proof of part (ii) is very similar.

Consider a population consisting exclusively of types from the set $\{(u^n, n)\}_{n=1}^\infty$, and assume that the type distribution of these incumbents, together with the behaviour policy (6), would have constituted an ESC if the type set had been restricted to $\{(u^n, n)\}_{n=1}^\infty$. Suppose a mutant of type $(u', n') \notin \{(u^n, n)\}_{n=1}^\infty$ enters. If it is the case that type $\left(u^{n'}, n'\right)$ is not among the incumbents, then by the definition of an ESC, it must earn weakly less against the incumbents than what the incumbents earn against each other. Thus it is sufficient to show that the mutant of type (u', n') earns less than what a mutant or incumbent of the same cognitive level, i.e. type $\left(u^{n'}, n'\right)$, would earn.

Against an incumbent (u^n, n) of level $n > n'$ a mutant of type (u', n') earns at most $1 - l$, and type $\left(u^{n'}, n'\right)$ earns $1 - l$. Against an incumbent (u^n, n) of level $n = n'$ a mutant of type (u', n') earns at most $1 - l$, and type $\left(u^{n'}, n'\right)$ earns 1. Against incumbents (u^n, n) of level $n < n'$ a mutant of type (u', n') earns at most $1 + g$, and type $\left(u^{n'}, n'\right)$ earns $1 + g$. Thus in all cases, a mutant $(u', n') \notin \{(u^n, n)\}_{n=1}^\infty$ earns strictly less than what a mutant or incumbent of type $\left(u^{n'}, n'\right)$ earns. Hence if mutants are sufficiently rare they will earn strictly less than incumbents in any focal post-entry configuration.

References

Abreu, Dilip, & Sethi, Rajiv. 2003. Evolutionary Stability in a Reputational Model of Bargaining. *Games and Economic Behavior*, **44**(2), 195–216.

Alger, Ingela, & Weibull, Jorgen W. 2013. Homo Moralis, Preference Evolution under Incomplete Information and Assortative Matching. *Econometrica*, **81**(6), 2269–2302.

Banerjee, Abhijit, & Weibull, Jörgen W. 1995. Evolutionary Selection and Rational Behavior. *Pages 343–363 of:* Kirman, Alan, & Salmon, Mark (eds), *Learning and Rationality in Economics*. Oxford: Blackwell.

Bester, Helmut, & Güth, Werner. 1998. Is Altruism Evolutionarily Stable? *Journal of Economic Behavior and Organization*, **34**, 193–209.

Bolle, Friedel. 2000. Is Altruism Evolutionarily Stable? And Envy and Malevolence? Remarks on Bester and Güth. *Journal of Economic Behavior and Organization*, **42**, 131–133.

Bomze, Immanuel M, & Weibull, Jörgen W. 1995. Does Neutral Stability imply Lyapunov Stability? *Games and Economic Behavior*, **11**(2), 173–192.

Conlisk, John. 2001. Costly Predation and the Distribution of Competence. *American Economic Review*, **91**(3), 475–484.

Cressman, Ross. 1997. Local Stability of Smooth Selection Dynamics for Normal form Games. *Mathematical Social Sciences*, **34**(1), 1–19.

Dekel, Eddie, Ely, Jeffrey C., & Yilankaya, Okan. 2007. Evolution of Preferences. *Review of Economic Studies*, **74**, 685–704.

Dufwenberg, Martin, & Güth, Werner. 1999. Indirect Evolution vs. Strategic Delegation: A Comparison of Two Approaches to Explaining economic institutions. *European Journal of Political Economy*, **15**(2), 281–295.

Dunbar, Robin I. M. 1998. The Social Brain Hypothesis. *Evolutionary Anthropology*, **6**, 178–190.

Ellingsen, Tore. 1997. The Evolution of Bargaining Behavior. *The Quarterly Journal of Economics*, **112**(2), 581–602.

Ely, Jeffery C., & Yilankaya, Okan. 2001. Nash Equilibrium and the Evolution of Preferences. *Journal of Economic Theory*, **97**, 255–272.

Fershtman, Chaim, & Weiss, Yoram. 1998. Social Rewards, Externalities and Stable Preferences. *Journal of Public Economics*, **70**(1), 53–73.

Frank, Robert H. 1987. If Homo Economicus Could Choose his own Utility Function, Would He Want One with a Conscience? *The American Economic Review*, **77**(4), 593–604.

Frenkel, S., Heller, Y., & Teper, R. 2014. *The endowment effect as a blessing.* mimeo.

Friedman, Daniel, & Singh, Nirvikar. 2009. Equilibrium Vengeance. *Games and Economic Behavior*, **66**(2), 813–829.

Gamba, Astrid. 2013. Learning and Evolution of Altruistic Preferences in the Centipede Game. *Journal of Economic Behavior and Organization*, **85**(C), 112–117.

Güth, Werner. 1995. An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives. *International Journal of Game Theory*, **24**(4), 323–344.

Güth, Werner, & Napel, Stefan. 2006. Inequality Aversion in a Variety of Games: An Indirect Evolutionary Analysis. *The Economic Journal*, **116**(October), 1037–1056.

Güth, Werner, & Yaari, Menahem E. 1992. Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach. *Pages 22–34 of:* Witt, Ulrich (ed), *Explaining Process and Change*. Ann Arbor, MI: University of Michigan Press.

Guttman, Joel M. 2003. Repeated Interaction and the Evolution of Preferences for Reciprocity. *The Economic Journal*, **113**(489), 631–656.

Heifetz, Aviad, Shannon, Chris, & Spiegel, Yossi. 2007. What to Maximize if You Must. *Journal of Economic Theory*, **133**(1), 31–57.

Heller, Yuval. 2015. Three steps ahead. *Theoretical Economics*, **10**, 203–241.

Heller, Yuval, & Mohlin, Erik. 2014. *Stable Observable Behavior*. Unpublished.

Herold, Florian, & Kuzmics, Christoph. 2009. Evolutionary Stability of Discrimination under Observability. *Games and Economic Behavior*, **67**, 542–551.

Hines, W Gord S, & Maynard Smith, J. 1979. Games Between Relatives. *Journal of Theoretical Biology*, **79**(1), 19–30.

Hofbauer, Josef. 2011. Deterministic Evolutionary Game Dynamics. *In: Proceedings of Symposia in Applied Mathematics*, vol. 69.

Hofbauer, Josef, & Sandholm, William H. 2009. Stable Games and Their Dynamics. *Journal of Economic Theory*, **144**(4), 1665–1693.

Hofbauer, Josef, & Sigmund, Karl. 1988. *The Theory of Evolution and Dynamical Systems*. Cambridge: Cambridge University Press.

Holloway, R. 1996. Evolution of the Human Brain. *Pages 74–125 of:* Lock, A., & Peters, C. R. (eds), *Handbook of Human Symbolic Evolution*. Oxford: Clarendon Press.

Hopkins, Ed. 2014. Competitive Altruism, Mentalizing and Signalling. *American Economic Journal: Microeconomics*, **6**, 272–292.

Huck, Steffen, & Oechssler, Jörg. 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. *Pages 303–317 of:* Bateson, P. P. G., & Hinde, R. A. (eds), *Growing Points in Ethology*. Cambridge: Cambridge University Press.

Kim, Yong-Gwan, & Sobel, Joel. 1995. An Evolutionary Approach to Pre-Play Communication. *Econometrica*, **63**(5), 1181–1193.

Kimborough, Erik O., Robalino, Nikolaus, & Robson, Arthur J. 2014. *The Evolution of "Theory of Mind": Theory and Experiments*. Cowles Foundation Discussion Paper No. 1907R, Yale University.

Kinderman, P., Dunbar, R. I. M., & Bentall, R. P. 1998. Theory-of-Mind Deficits and Causal Attributions. *British Journal of Psychology*, **89**, 191–204.

Koçkcesen, Levent, & Ok, Efe A. 2000. Evolution of Interdependent Preferences in Aggregative Games. *Games and Economic Behavior*, **31**, 303–310.

Matsui, Akihiko. 1991. Cheap-talk and Cooperation in a Society. *Journal of Economic Theory*, **54**(2), 245–258.

Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, John, & Price, G. R. 1973. The Logic of Animal Conflict. *Nature*, **246**(5427), 15–18.

Mohlin, Erik. 2010. Internalized Social Norms in Conflicts: An Evolutionary Approach. *Economics of Governance*, **11**(2), 169–181.

Mohlin, Erik. 2012. Evolution of Theories of Mind. *Games and Economic Behavior*, **75**(1), 299–312.

Nachbar, John H. 1990. "Evolutionary" Selection Dynamics in Games: Convergence and Limit Properties. *International Journal of Game Theory*, **19**(1), 59–89.

Norman, Thomas W. L. 2012. Equilibrium Selection and the Dynamic Evolution of Preferences. *Games and Economic Behavior*, **74**(1), 311–320.

Ok, Efe A., & Vega-Redondo, Fernando. 2001. On the Evolution of Individualistic Preferences: An Incomplete Information Scenario. *Journal of Economic Theory*, **97**, 231–254.

Possajennikov, Alex. 2000. On the Evolutionary Stability of Altruistic and Spiteful Preferences. *Journal of Economic Behavior and Organization*, **42**, 125–129.

Premack, D.G., & Woodruff, G. 1979. Does the Chimpanzee have a Theory of Mind. *Behavioral and Brain Sciences*, **1**, 515–526.

Robson, Arthur J. 1990. Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake. *Journal of Theoretical Biology*, **144**(3), 379–396.

Robson, Arthur J., & Samuelson, Larry. 2011. The Evolutionary Foundations of Preferences. *Pages 221–310 of:* Benhabib, Jess, Bisin, Alberto, & Jackson, Matthew (eds), *The Social Economics Handbook*. North Holland.

Rtischev, Dimitry. 2012. *Evolution of Mindsight, Transparency and Rule-Rationality*. Unpublished.

Samuelson, Larry. 1991. Limit Evolutionarily Stable Strategies in Two-Player, Normal Form Games. *Games and Economic Behavior*, **3**(1), 110–128.

Samuelson, Larry. 2001. Introduction to the Evolution of Preferences. *Journal of Economic Theory*, **97**(2), 225–230.

Sandholm, William H. 2001. Preference Evolution, Two-Speed Dynamics, and Rapid Social Change. *Review of Economic Dynamics*, **4**, 637–679.

Sandholm, William H. 2010. Local Stability under Evolutionary Game Dynamics. *Theoretical Economics*, **5**(1), 27–50.

Schaffer, Mark E. 1988. Evolutionarily Stable Strategies for a Finite Population and a Variable Contest Size. *Journal of Theoretical Biology*, **132**, 469–478.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Harvard University Press.

Schlag, Karl H. 1993. *Cheap Talk and Evolutionary Dynamics*. Bonn Department of Economics Discussion Paper B-242.

Selten, Reinhard. 1980. A Note on Evolutionarily Stable Strategies in Asymmetric Animal Conflicts. *Journal of Theoretical Biology*, **84**(1), 93–101.

Sethi, Rajiv, & Somanthan, E. 2001. Preference Evolution and Reciprocity. *Journal of Economic Theory*, **97**, 273–297.

Stahl, Dale O. 1993. Evolution of Smart$_n$ Players. *Games and Economic Behavior*, **5**(4), 604–617.

Stennek, Johan. 2000. The Survival Value of Assuming Others to be Rational. *International Journal of Game Theory*, **29**, 147–163.

Taylor, Peter D., & Jonker, Leo B. 1978. Evolutionary Stable Strategies and Game dynamics. *Mathematical Biosciences*, **40**(1–2), 145–156.

Thomas, Bernhard. 1985. On Evolutionarily Stable Sets. *Journal of Mathematical Biology*, **22**(1), 105–115.

Wärneryd, Karl. 1991. Evolutionary Stability in Unanimity Games with Cheap Talk. *Economics Letters*, **36**(4), 375–378.

Wärneryd, Karl. 1998. Communication, Complexity, and Evolutionary Stability. *International Journal of Game Theory*, **27**(4), 599–609.

Weissing, Franz, J. 1991. Evolutionary Stability and Dynamic Stability in a Class of Evolutionary Normal Form Games. *Pages 29–97 of:* Selten, R. (ed), *Game Equilibrium Models I. Evolution and Game Dynamics*. Springer.