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†
University of Oxford.

August 30, 2014

Abstract

We develop a framework in which individuals' preferences co-evolve with their abilities to deceive others regarding their preferences and intentions. We show that a pure outcome is stable, essentially if and only if it is an efficient Nash equilibrium. All individuals have the same deception ability in such a stable state. In contrast, there are non-pure outcomes in which non-Nash outcomes are played, and different deception abilities co-exist. 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.

*Valuable comments were provided by Vince Crawford, 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.

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

1 Introduction

For a long time economists took preferences as given. The study of their origin and formation was considered a question outside the scope of economics. Over the past 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 and Samuelson 2011). A prominent strand of this literature is the so-called "indirect evolutionary approach", pioneered by Güth and Yaari (1992).[1] 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.[2] 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, Shannon, and Spiegel (2007) show that this kind of result is generic.

A crucial feature of the indirect evolutionary approach is that preferences are explicitly or implicitly assumed to be at least partially observable.[3] 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 and Vega-Redondo (2001), Ely and Yilankaya (2001), Dekel, Ely, and Yilankaya (2007), and Herold and Kuzmics (2009). They confirm that the degree to which preferences are observed decisively influences the outcome of preference evolution. However, the degree to which preferences are observed is still exogenous in these models. In reality we would expect both the preferences, and the ability to observe or conceal them to be the product of an evolutionary process.[4]

[1] The term was coined in Güth (1995).

[2] For example, Bester and 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 and Weiss (1998) study evolution of concerns for social status. Sethi and 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 and Güth (1999) study firm's preferences for large sales. Güth and Napel (2006) study preference evolution when players use the same preferences in both ultimatum and dictator games. Koçkcesen and Ok (2000) investigate survival of more general interdependent preferences in aggregative games. Friedman and Singh (2009) shows that vengefulness may survive if observation has some degree of informativeness.

[3] 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.

[4] On this topic, Robson and Samuelson (2011) write: "The standard argument is that we can observe preferences because people give signals – a tightening of the lips or flash of the eyes – that 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

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. The main model deals with preferences defined over action profiles. We later extend it to interdependent preferences that depend on the opponent's preferences.

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.[5] 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. In the main model we restrict attention to the standard case of utility functions that are defined over action profiles, which we refer to as *type-neutral preferences*. 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. The cost of increased cognition is strictly positive. The cognitive levels of the individuals in a match determine the probability that one individual observes the opponent's preferences and is able to deceive the opponent.

When both individuals are of the same cognitive level they are assumed to play a Nash equilibrium of the complete information game induced by their preferences, just as in the standard indirect evolutionary approach. However, when the individuals in a match are of different cognitive levels, the one with the higher level is able to deceive the one with the lower level. In the main model (with type-neutral preferences), 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. In the extension with type-interdependent preferences, the deceiver is also allowed to choose whatever she wants the deceived party to believe about the deceiver's type. A strategy profile that is consistent with this form of deception is called a *deception equilibrium*.

We also analyse an extension of our 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. The same results as in the main model hold for this less flexible form of deception.

---

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)

[5] It is known that positive assortative matching is conducive to the evolution of altruistic behaviours (Hines and Maynard Smith 1979) and non-materialistic preferences even when preferences are perfectly unobservable (Alger and Weibull 2013). It is also known that finite populations allow for the evolution of spiteful behaviours (Schaffer 1988) and non-materialistic preferences (Huck and 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.

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 (evolutionarily) stable configuration* (NSC or ESC) 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.

Consider a society that has settled upon a convention, represented by a configuration that induces play of the same pure outcome in all matches. For this case we are essentially able to provide a characterisation of neutral (evolutionary) stability. For the main model, with type-neutral preferences, we find that if the configuration is an NSC, then everyone is of the lowest cognitive type, and the induced outcome is efficient. Moreover, if the marginal cost of cognition is low enough then the outcome is also a Nash equilibrium of the underlying game, in fitness payoffs. Conversely, any efficient strict pure Nash equilibrium of the underlying game can be implemented in an NSC. For the extension with type-interdependent preferences, we find that if the configuration is an NSC, then everyone is of the lowest cognitive type, and the induced outcome gives all players at least their pure maxmin payoff. Moreover, if the marginal cost of cognition is low enough then the outcome is also a Nash equilibrium. Conversely, any pure Nash equilibrium in which all players get more than the pure minmax payoff, can be implemented in an ESC. (Recall that the pure minmax and pure maxmin payoffs need not coincide.)

When we consider configurations that induce play of more than one pure outcome, the above characterisation breaks down in an interesting way. For both type-neutral and type-interdependent preferences we are able to construct NSCs (even ESCs in the case of type-interdependent preferences) in which some matches result in non-Nash outcomes, and in which different cognitive levels co-exist. Still, even in this context the highest types have to play efficiently among themselves, and the cognitive cost determines the size of deviations from Nash behaviour that can persist.

In the next subsection we discuss related literature. The rest of the paper is organized as follows. Section 2 presents the main model, with type-neutral preferences and deception tailored to each opponent. In Section 3 we define our stability notions. Results for the main model are presented in Sections 4 and 5. Section 6 extends the model to include type-interdependent preferences. Section 7 concludes. Appendix A analyses the alternative assumption that each individual uses the same deception for all opponents. Appendix B contains proofs not in the main text.

1.1 Related Literature

Ok and Vega-Redondo (2001) and Ely and 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, Ely, and Yilankaya (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.[6] 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 and Kuzmics (2009) expand the framework of Dekel, Ely, and Yilankaya (2007) to include interdependent preferences, i.e. preferences that depend on the opponent's preference type. Under perfect or almost perfect observability, if all preferences that depend on the opponent's type are considered, then any symmetric outcome above the minmax material payoff is 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 and 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.

There is a large literature in biology and evolutionary psychology on the evolution of 'theory of mind' (Premack and 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, Dunbar, and Bentall 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 and Weibull (1995), Stennek (2000), Conlisk (2001), Abreu and Sethi (2003), Mohlin (2012), and Heller (2014). As in these papers, we provide results to the effect that different degrees of cognitive sophistication may co-exist. The model of Conlisk (2001) is very similar to our analysis of the Rock-Paper-

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

Scissors game in Section 5.3, below.

Rtischev (2012) provides a model where agents differ with respect to their ability to detect and conceal their strategies. He focuses on a leader-follower game where the follower benefits from making a credible and visible commitment. The follower may pay a cost to be transparent. If the follower is transparent, then the leader can observe the follower's strategy if and only if the follower has paid to be transparent and the leader pays a cost to obtain "mindsight". There is an equilibrium in which committed transparent followers, uncommitted non-transparent followers, leaders mindsight, and leaders without mindsight co-exist. Kimborough, Robalino, and Robson (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 signalling 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 ESS 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. 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 and Sobel (1995) use stochastic stability arguments, and Wärneryd (1998) uses complexity costs, to circumvent this problem. Similarly, efficient Nash equilibria are selected in our model too. Preferences serve the function of messages and, since the set of preferences is uncountable, there are always unused "messages".

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, Ely, and Yilankaya (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 and 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 many 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.

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

If two individuals of different cognitive levels are matched to play, then the individual with the higher cognitive level observes the opponent's preferences perfectly, and is able to deceive the opponent. 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. Dispensing with this assumption comes at additional notational cost, but the results are qualitatively similar.

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'.

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 2 *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 and 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* $\tilde{\pi}(\sigma', (1-\varepsilon)\sigma + \varepsilon\sigma') \leq \tilde{\pi}(\sigma, (1-\varepsilon)\sigma + \varepsilon\sigma')$*. 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.[7] 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, Ely, and Yilankaya (2007). Formally:

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

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, Ely, and Yilankaya (2007) and Alger and Weibull (2013) are also based on neutral stability.[8] 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.

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 $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$, if types evolve according to the replicator dynamic (Thomas 1985, Bomze and Weibull 1995). ESCs are also asymptotically stable: populations starting close enough to

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

μ eventually converge to μ in $\Gamma_{(\tilde{\mu},\tilde{b})}$ if types evolve according to a smooth payoff-monotonic selection dynamic (Taylor and Jonker 1978, Cressman 1997, Sandholm 2010).

4. The stability notions of Dekel, Ely, and Yilankaya (2007) and Alger and 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 and 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, Ely, and Yilankaya (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 below.

4 Characterisation of Stable Pure Configurations

In this section we consider configurations in which everyone plays the same pure action. We interpret such configurations as representing a state of a population that has settled on a convention, which is shared by everyone. We show that such configurations are stable essentially if and only if the action profile is both efficient and a Nash equilibrium of the fitness game.

4.1 Definitions

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')*.*

If a symmetric strategy profile (σ^*, σ^*) is efficient, then we say that the strategy σ^* is efficient. Similarly 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. A Nash equilibrium is strict if $\pi(a, a) > \pi(a', a)$ for all $a' \in A$.

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

Definition 8 *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.

Preferences $u \in U$ are *completely indifferent* if they induce indifference between all action profiles, i.e. if $u(a, a') = u(a'', a''')$ for all combinations of a, a', a'', and a'''. Preferences $u \in U$ are said to be *strategically indifferent* if they induce a player to be indifferent between all action profiles in which the opponent's action is fixed; i.e. it holds that $u(a, a') = u(a'', a')$, for all actions $a, a', a'' \in A$. Note that a utility function is strategically indifferent if and only if it is strategically equivalent (Moulin and Vial 1978) to the completely indifferent utility.

4.2 Stability Implies Nash and Efficiency

We will show that if (μ, a^*) is stable then a^* must be both a Nash equilibrium (of the underlying fitness game) and an efficient action. We begin by presenting a simple lemma that shows that if a configuration is pure, then all incumbents must have the minimal cognitive level, since having a higher ability does not yield any advantage when everyone plays the same action.

Lemma 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*

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

If (μ, a^*) is an NSC, then 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. ■

Next, we show that any outcome of a pure NSC is efficient. The intuition is that if the incumbents play inefficiently among themselves, then they can be invaded by a heterogeneous group of mutants (also of cognitive level 1). That is, since mutants observe each other's preferences they use their preferences as a "secret-handshake" to achieve efficiency; see Robson (1990).

When facing incumbents, mutants play the same as the incumbents, but they play more efficient action profiles among themselves. If these efficient action profiles are asymmetric, the mutants use their heterogeneity as a correlation device to induce such asymmetric behaviour. Formally:

Proposition 2 *If (μ, a^*) is an NSC, then a^* is efficient.*

Proof. To obtain a contradiction assume that (μ, a^*) is an NSC but a^* is not efficient. The inefficiency implies the existence of actions a, a' such that $\pi(a^*, a^*) < 0.5 \cdot (\pi(a, a') + \pi(a, a'))$. Let $u_1, u_2, u_3 \in U$ be three different mutant preferences, all of which are strategically indifferent, and let $\mu' \in \Delta(U)$ be the distribution of mutants that assign a mass of $\frac{1}{3}$ to each of the following three types: $\theta_1 = (u_1, 1), \theta_2 = (u_2, 1)$, and $\theta_3 = (u_3, 1)$.

For each $\varepsilon \in (0, 1)$ we have a post-entry type distribution $\tilde{\mu} = (1 - \varepsilon) \cdot \mu + \varepsilon \cdot \mu'$. Let $\left(\tilde{\mu}, \tilde{b}\right)$ be the focal configuration in which (1) $\tilde{b}_\theta(\theta') = a^*$ if $\theta \in C(\mu)$ or $\theta' \in C(\mu)$, and (2) for each $i, j \in \{1, 2, 3\}$,

$$b_{\theta_i}(\theta_j) = \begin{cases} a & j = i + 1 \bmod 3 \\ a' & j = i - 1 \bmod 3 \\ a^* & i = j \end{cases}.$$

Since Proposition 1 implies that all the incumbent types in $C(\mu)$ have cognitive level 1, it is immediate that $\left(\tilde{\mu}, \tilde{b}\right)$ is indeed a configuration. Moreover, each incumbent type earns

a payoff of $\pi(a^*, a^*)$, and each mutant type $\theta_i \in C(\mu')$ earns a strictly higher payoff:

$$\Pi_{\theta_i}\left(\left(\tilde{\mu}, \tilde{b}\right)\right) = \left(1 - \frac{2 \cdot \varepsilon}{3}\right) \cdot \pi(a^*, a^*) + \frac{2 \cdot \varepsilon}{3} \cdot 0.5 \cdot (\pi(a, a') + \pi(a, a')) > \pi(a^*, a^*).$$

This implies that μ is not an NSS in $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$, and thus (μ, a^*) is not an NSC. ■

Remark 3 *Note that our proof above shows that a configuration that induces a pure inefficient outcome is unstable in a strong sense: (1) all mutant types in μ' strictly outperform all incumbent types, and (2) this holds for any $\varepsilon \in (0, 1)$, and not only for small ε.*

Remark 4 *Dekel, Ely, and Yilankaya (2007) work with a framework in which there are no cognitive levels and no deception, and there is an exogenous probability p for each player to privately observe her opponent's preferences. For $p = 1$, they show (Proposition 2 in their paper) a result that is similar to our Proposition 2. Still, there is one key difference: in their setup stability of a pure outcome is characterised by a weaker notion of efficiency. An action is efficient in the sense of Dekel, Ely, and Yilankaya (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, Ely, and Yilankaya (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, as established in Proposition 2 above.*

4.3 Strict Nash and Efficiency Implies Stability

The following 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. 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 3 *If a^* is both efficient and a strict Nash equilibrium (in fitness payoffs), then there exists a type distribution μ such that (μ, a^*) is an NSC.*

Proof. Let a^* be an efficient action that is also a strict Nash equilibrium. Consider a monomorphic configuration (μ, 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 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 a^* (since a^* is strictly dominant for θ^*). Since the incumbent is always playing a^*, and (a^*, a^*) is a strict Nash equilibrium of G, mutants that do not play 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 a^* whenever they are matched with θ^*, the incumbents' average fitness is given by $\pi(a^*, a^*)$, and since mutants cannot obtain an average fitness strictly higher than this when they are matched among themselves (since (a^*, a^*) is efficient), they cannot obtain a strictly higher average fitness either. We conclude that (μ, a^*) is an NSC. This argument is similar to the one used to prove Proposition 6 of Dekel, Ely, and Yilankaya (2007). ■

Remark 5 *Observe that the stability of a^* in the proof above is strict with respect to any mutant type who either introduces different behaviour (plays an action $a' \neq a^*$) or is of a different cognitive level (larger than 1). Mutants can achieve the same fitness as the incumbent only if they are "outcome equivalent" to the incumbents: they have the same minimal cognitive level as the incumbents and always plays action a^* like the incumbents.*

The results of this section implies as a corollary that being Nash and efficient is essentially a necessary and sufficient condition for an action to be the pure outcome of an NSC. Formally:

Corollary 1

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

Example 1 *For coordination games, like the following Stag Hunt game*

$$\begin{array}{cc} & \begin{array}{cc} S & H \end{array} \\ \begin{array}{c} S \\ H \end{array} & \begin{array}{cc} 3,3 & 0,1 \\ 1,0 & 2,2 \end{array} \end{array},$$

the above propositions imply that there is an NSC in which (S, S) is the outcome of every match, and no other pure profile can be the unique outcome in an NSC.

5 Multiple Outcome Configurations

Move beyond the focus on populations that have settled on a convention, in the form of a configuration that induces a unique pure outcome, we now allow for more diverse

populations, as represented by configurations inducing many different pure or mixed outcomes. In this setting general results are much harder to come by. However, we show that stability still implies a limited form of efficiency: the types with the highest cognitive level in the population have to play efficiently in any NSC. In contrast a NSC no longer implies Nash equilibrium play. We demonstrate this by a counterexample, based on Rock-Paper-Scissors.

5.1 Efficiency among the Highest-Level Types

The following result states that any type that has the highest level of cognition in the population must play an efficient action when meeting itself, provided that there is at least one action that is never played in the current configuration.

Proposition 4 *Let (μ^*, b^*) be an NSC in which at least one action is never played. Define $\bar{n} = \max_{\theta \in C(\mu^*)} n_\theta$. If $\theta_0 = (u_{\theta_0}, \bar{n})$ then $b^*_{\theta_0}(\theta_0)$ is efficient.*

Proof. Assume that $\sigma^* = b^*_{\theta_0}(\theta_0)$ is not efficient. Thus there are actions a', a'' such that $0.5 \cdot (\pi(a', a'') + \pi(a'', a')) > \pi(\sigma, \sigma)$. Let A^+ and A^- be the set of actions that are sometimes and never played in (μ^*, b^*), respectively. We consider three mutually exclusive and jointly exhaustive cases.

Case 1: Suppose that there is an efficient profile (a', a'') consisting of unused actions $a', a'' \in A^-$. (Note that we allow for $a' = a''$.) Let θ_1, θ_2, and θ_3 be three mutant types, all of cognitive level $\bar{n}$, but with three different kinds of preferences, u_1, u_2, and u_3, such that for some $\alpha \in \mathbb{R}_{++}$,

$$\begin{aligned}
&u_i\left(a^+, \hat{a}^+\right) = 0 \text{ for all } a^+, \hat{a}^+ \in A^+,\\
&u_i\left(a^-, \hat{a}^-\right) = 0 \text{ for all } a^-, \hat{a}^- \in A^-,\\
&u_i\left(a^-, a^+\right) = -1 \text{ for all } a^- \in A^-, \text{ and } a^+ \in A^+,\\
&u_i\left(a^+, a^-\right) = -\alpha i \text{ for all } a^+ \in A^+ \text{ and } a^- \in A^-.
\end{aligned}$$

Note that u_1, u_2, and u_3 cannot be obtained from each other as affine transformations. Moreover, we can always find an α such that these three types are not among the incumbents.

Let the type distribution of the mutants be $\mu' = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Fix an incumbent of the highest level $\theta_0 = (u_{\theta_0}, \bar{n})$. For any $\varepsilon > 0$ there is a focal post entry-configuration $\left(\tilde{\mu}, \tilde{b}\right)$, where $\tilde{\mu} = (1-\varepsilon) \cdot \mu^* + \varepsilon \cdot \mu'$, such that (1) in any match with an incumbent $\theta \in C(\mu^*)$ each of the types θ_1, θ_2, and θ_3 play the same profile as θ_0 does when facing the incumbent

$\theta \in C(\mu^*)$, and (2) in matches between mutants, it holds that for each $i, j \in \{1,2,3\}$,

$$b_{\theta_i}(\theta_j) = \begin{cases} a' & j = i+1 \bmod 3 \\ a'' & j = i-1 \bmod 3 \\ \sigma^* & i = j \end{cases}.$$

Thus, against an incumbent $\theta \in C(\mu^*)$ the mutants earn exactly the same as θ_0. Against any of the mutants the type θ_0 earns $\pi(\sigma^*, \sigma^*)$, while the mutants earn

$$\Pi_{\theta_i}\left(\left(\tilde{\mu}, \tilde{b}\right)\right) = \left(1 - \frac{2 \cdot \varepsilon}{3}\right) \cdot \pi(\sigma^*, \sigma^*) + \frac{2 \cdot \varepsilon}{3} \cdot 0.5 \cdot (\pi(a, a') + \pi(a, a')) > \pi(\sigma^*, \sigma^*),$$

against θ_0. This implies that μ is not an NSS in $\Gamma_{(\tilde{\mu}, \tilde{b})}$, and thus (μ^*, b^*) is not an NSC.

Case 2: Suppose that there is an efficient profile (a', a'') consisting of used actions $a', a'' \in A^+$. (Note that we allow for $a' = a''$.) We can use exactly the same construction as in case 1.

Case 3: Suppose that there is an efficient profile (a', a'') consisting of one used action $a' \in A^+$, and one unused action $a'' \in A^-$. (Hence $a' \neq a''$.) We need to modify the construction of preferences used above. Let θ_1, θ_2, and θ_3 be three mutant types, all of cognitive level $\bar{n}$, but with three different kinds of preferences, u_1, u_2, and u_3, such that for some $\alpha \in \mathbb{R}_{++}$,

$$\begin{aligned}
&u_i\left(a^+, \hat{a}^+\right) = 0 \text{ for all } a^+, \hat{a}^+ \in A^+, \\
&u_i\left(a^-, \hat{a}^-\right) = 0 \text{ for all } a^-, \hat{a}^- \in A^-, \\
&u_i(a', a'') = u_i(a'', a') = 0, \\
&u_i\left(a^-, a^+\right) = -1 \text{ for all } a^- \in A^-, \text{ and } a^+ \in A^+, \text{ such that } \left(a^-, a^+\right) \notin \{(a', a''), (a'', a')\}, \\
&u_i\left(a^+, a^-\right) = -\alpha i \text{ for all } a^+ \in A^+ \text{ and } a^- \in A^-, \text{ such that } \left(a^-, a^+\right) \notin \{(a', a''), (a'', a')\}.
\end{aligned}$$

If there are at least two actions being used in (μ^*, b^*), so that $|A^+| \geq 2$, then the subjective payoff matrix for preferences u_i contain all of the values 0, -1, and $-\alpha i$. In this case u_1, u_2, and u_3 cannot be obtained from each other as affine transformations. Moreover, we can always find an α such that these three types are not among the incumbents. If there is only one action that is being used in (μ^*, b^*), so that $A^+ = \{a'\}$, then we have a pure outcome configuration. Hence it is without loss of generality to assume that there are at least two actions being used in (μ^*, b^*), so that $|A^+| \geq 2$.

Let the type distribution of the mutants be $\mu' = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. The rest of the proof is as in case 1. ■

The condition that there must be an action that is not played in the stable configuration in order for the highest types to play efficiently is reminiscent of the condition, in the evolutionary cheap talk literature (Wärneryd 1991, and Schlag 1993) that there

must be a message that is not sent in an ESS in order for ESS to imply efficiency. The similarity is not a coincidence, since the preferences are effectively used as messages in our setting.

5.2 Cognitive Cost and Nash Behaviour

We can say something about how the cognitive cost function relates to the size of deviations from Nash equilibrium.

Observation 1 *Let* $\bar{n} = \max_{\theta \in C(\mu)} n_\theta$. *If* (μ^*, b^*) *is an NSC then, for all* $\theta' \in C(\mu^*)$,

$$\sum_{\theta'' \in C(\mu^*)} \mu^*_{\theta''} \max_a \left[\pi\left(a, b^*_{\theta''}(\theta')\right) - \pi\left(b^*_{\theta'}(\theta''), b^*_{\theta''}(\theta')\right)\right] \leq k_{\bar{n}+1} - k_{n(\theta')}. \tag{2}$$

Proof. The payoff to type θ' is

$$\Pi_{\theta'}(\mu^*, b^*) = \sum_{\theta'' \in C(\mu^*)} \mu^*_{\theta''} \pi\left(b^*_{\theta'}(\theta''), b^*_{\theta''}(\theta')\right) - k_{n(\theta')}.$$

Let u^π denote preferences that coincide with material fitness, and consider a mutant $\tilde{\theta} = (u^\pi, \bar{n}+1)$. The payoff to type $\tilde{\theta}$ is at least

$$\sum_{\theta'' \in C(\mu^*)} \mu^*_{\theta''} \max_a \left[\pi\left(a, b^*_{\theta''}(\theta')\right) - \pi\left(b^*_{\theta'}(\theta''), b^*_{\theta''}(\theta')\right)\right] - k_{\bar{n}+1}.$$

To ensure that $\tilde{\theta}$ is unable to invade, (2) must hold. ■

To interpret this observation note that the left-hand side of (2) measures the average distance between the actual behaviour of type θ' and the behaviour that would constitute a best response for θ', in terms of fitness. The distance is measured in terms of the loss of fitness payoff. If type θ' plays a best response against all opponents then the left-hand side is zero. Thus we see that if there are large deviations from playing a Nash equilibrium, then the highest types must pay a large cognitive cost for a configuration to be neutrally stable.

5.3 Application: Rock-Paper-Scissors

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 results shows that for any cognitive cost function the environment admits a heterogeneous NSC in which players of different cognitive levels co-exist, and non-Nash profiles are played in all matches of two individuals of different types: types with a higher cognitive level deceive, and defeat, those with a lower cognitive level. 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 5 *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 there is an N such that*

$$k_N \leq 2 < k_{N+1},$$

and suppose that[9]

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

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:

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

Proof. 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 3 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 3 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

[9] If we define δ as in (1) then we have $\delta = 1$, in Proposition 5. Thus the condition that $\delta > k_{n+1} - k_n$ for all n in Proposition 5, may be viewed as an extension of the condition $k_2 < \delta$ in Proposition 1.

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 6 *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.*

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 and 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.[10] 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

[10] 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 and Kuzmics (2009).

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. $U_{ID} = \{u : \exists n \in \mathbb{N} \text{ s.t. } (u, n) \in \Theta_{ID}\}$. 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:

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

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

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 $\bar{M}$ 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 $\underline{M} \leq \bar{M}$, 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 $\underline{M}$,

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

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 9 *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).*

Lemma 2 *Assume that $(u^{a_{\underline{M}}}, 1) \in \Theta_{ID}$. Let (μ, b) be an NSC. Then $\Pi(\mu, b) \geq \underline{M}$.*

Proof. Assume to the contrary that $\Pi(\mu, b) < \underline{M}$. 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 type distribution. Furthermore, due to continuity it holds that $\Pi_\theta\left(\tilde{\mu}, \tilde{b}\right) < \underline{M}$ for any $\theta \in C(\mu)$ 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 observe that the proofs of Lemma 1 and Proposition 1 hold with minor adaptations also in the type-interdependent setup. Thus, if (μ, a^*) is a pure NSC, then:

(1) all incumbents have cognitive level 1, and (2) a^* is a symmetric Nash equilibrium, provided that $k_2 < \delta$.

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 10 *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 6 *Assume that $\left(u^{a^*}_{a_{\bar{M}}}, 1\right) \in \Theta_{ID}$. If action a^* is a symmetric Nash equilibrium and $\pi(a^*, a^*) > \bar{M}$, 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 $\pi(a^*, a^*) > \bar{M}$ 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 $\pi(a^*, a^*) > \bar{M}$ 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 2

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 $k_2 < \delta$, 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 $l \in (0, 1)$.[11]

$$\begin{array}{cc} & \begin{array}{cc} H & \quad D \end{array} \\ \begin{array}{c} H \\ D \end{array} & \begin{array}{cc} 0,0 & \quad 1+g, 1-l \\ 1-l, 1+g & \quad 1,1 \end{array} \end{array} . \tag{3}$$

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 11 *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 $u^n(H, D, \theta') = 2$, $u^n(H, H, \theta') = 1$, and $u^n(D, D, \theta') = u^n(D, H, \theta') = 0$.

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 natural assumptions on the cognitive cost function we can 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.

[11] If $g = l < 1$ then we may interpret the game as a mini-version of the Nash demand game, representing a simple bargaining interaction: the players have to agree how to divide a resource that is worth two fitness points. If both play D, they divide the resource equally. Playing H (bargaining aggressively) against someone playing D, yields a gain of $g > 0$, and imposes an equal loss on the opponent. If both players are aggressive they do not reach an agreement, and so the resource is lost.

When individuals of the same type meet, they play mutual cooperation (D, D). When individuals of different types meet, they play (H, D) or (D, H).

Proposition 7 *Let G be the game represented in (3), where $g > 0$ and $l \in (0, 1)$. Suppose that there is an N such that $k_N \leq l + g < k_{N+1}$, and suppose 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. $C(\mu^*) > 1$), and the behaviour of an incumbent $\theta \in C(\mu^*)$ facing another incumbent $\theta' \in C(\mu^*)$ is given by

$$b_\theta^*(\theta') = \begin{cases} D & \text{if } n_\theta \geq n_{\theta'} \\ H & \text{if } n_\theta < n_{\theta'} \end{cases}. \tag{4}$$

(ii) If $g = l$ then there exists an NSC (μ^*, b^*), such that $C(\mu^*) \subseteq \{(u^n, n)\}_{n=1}^N$, and μ^* is mixed (i.e. $C(\mu^*) > 1$), and the behaviour among incumbents is given by (4).

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

Proof. The proof is similar to the proof of Proposition 5. Under the described behavioural policy we have, for $\theta, \theta' \in \{(u^n, n)\}_{n=1}^\infty$,

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

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 3 in Appendix B. For example, to see that Lemma 3 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 3. 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 7 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 (4), 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 (u^n, n') is not among the incumbents, then by the definition of an ESC, it must earn strictly 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 type (u^n, n') would earn.

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

7 Conclusion

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 Nash assumption that has characterised the indirect evolutionary approach is complemented by a more Machiavellian notion of deception equilibrium.

We obtain particularly clean results for populations represented by pure configurations, in in which the same pure outcome is played in all matches. For type-neutral preferences, a pure configuration is essentially a neutrally stable configuration (NSC), if and only if the induced action profile is both efficient and a Nash equilibrium. For type-interdependent preferences, being a Nash equilibrium is still a necessary and sufficient condition for stability, but instead of efficiency the critical condition is that payoffs are above the pure minmax and maxmin payoffs. The results are different in the case of configurations that induce play of a non-pure outcome. Regardless of whether preferences are type-neutral or type-interdependent, we are able to construct stable configurations in which some matches result in non-Nash outcomes, and in which different cognitive levels coexist.

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 and 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 Appendix: 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 12 *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* $\theta, \theta' \in C(\mu)$:

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

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

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 13 *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* $\left(\tilde{\mu}, \tilde{b}\right)$, *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* $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$.

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.*

B Appendix: Result 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

$$\pi\left(b_\theta\left(\theta'\right), b_{\theta'}\left(\theta\right)\right) = \begin{cases} t & \text{if } n_\theta > n_{\theta'} \\ w & \text{if } n_\theta = n_{\theta'} \\ s & \text{if } n_\theta < n_{\theta'} \end{cases}. \tag{5}$$

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 (5). 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 3 *Suppose $t \geq w \geq s$. Suppose that there is an N such that*

$$k_N - k_1 \leq t - s < k_{N+1} - k_1, \tag{6}$$

and suppose that

$$t - w > k_{n+1} - k_n \text{ for all } n \leq N. \tag{7}$$

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 (5).

(i) 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. $C(\mu^*) \subseteq \{(u,n)\}_{n=1}^{N}$.

(ii) 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. $C(\mu^*) \subseteq \{(u,n)\}_{n=1}^{N}$.

(iii) If $2w > s + t$ then $\Gamma_{(\mu,b)}$, admits no NSS and hence no ESS.

We now prove this result, starting with the following lemma:

Lemma 4 *$(u, N+1)$ 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.*

Proof. Since $s \leq w \leq t$ this follows from $t - k_{N+1} < s - k_1$ 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

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

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:

Lemma 5 *Consider the game with payoff matrix* $\mathbf{A}$*. Suppose (7) holds.*

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

Proof. Condition (7) 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, (7) implies that $(u,n+1)$ is the unique best response to (u,n) for all $n \in \{1,..,N-2\}$. ■

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. If $\mathbf{A}$ is negative 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 $\mathbf{A}$ has a unique ESS, which is also the unique Nash equilibrium of the game; see Hofbauer and Sigmund (1988), p. 72. If $\mathbf{A}$ is negative semi-definite with respect to the tangent space, i.e. if $v \cdot \mathbf{A}v \leq 0$ for all $v \in \mathbb{R}_0^d$, then $\mathbf{A}$ has an NSS, which need not be unique. Moreover, the set of Nash equilibria coincides with the set of NSS and constitutes a nonempty convex subset of the simplex (Hofbauer and Sandholm 2009, Theorem 3.2).

One can show:

Lemma 6 *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 $2w \geq (s+t)$.

■

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 3.

Proof of Lemma 3. (i) If $2w < s + t$ then by Lemma 6 $\mathbf{A}$ is negative definite w.r.t. the tangent space, implying that it has a unique ESS. Lemma 5 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.

By Lemma 4, type $(u, N + 1)$, and higher types of higher levels earn strictly less than $(\theta, 1)$. Thus regardless of whether $(\theta, 1)$ is in the support of the ESS of $\mathbf{A}$, type $(u, N + 1)$ and types of higher levels earn strictly less than the strategies in the support of the ESS of $\mathbf{A}$.

(ii) 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 Lemma 4 rules out higher-level types.

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

References

Abreu, D., and R. Sethi (2003): "Evolutionary Stability in a Reputational Model of Bargaining," *Games and Economic Behavior*, 44(2), 195–216.

Alger, I., and J. W. Weibull (2013): "Homo Moralis, Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81(6), 2269–2302.

Banerjee, A., and J. W. Weibull (1995): "Evolutionary Selection and Rational Behavior," in *Learning and Rationality in Economics*, ed. by A. Kirman, and M. Salmon, pp. 343–363. Blackwell, Oxford.

Bester, H., and W. Güth (1998): "Is Altruism Evolutionarily Stable?," *Journal of Economic Behavior and Organization*, 34, 193–209.

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

Bomze, I. M., and J. W. Weibull (1995): "Does Neutral Stability imply Lyapunov Stability?," *Games and Economic Behavior*, 11(2), 173–192.

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

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

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

Dufwenberg, M., and W. Güth (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, R. I. M. (1998): "The Social Brain Hypothesis," *Evolutionary Anthropology*, 6, 178–190.

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

Ely, J. C., and O. Yilankaya (2001): "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97, 255–272.

Fershtman, C., and Y. Weiss (1998): "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics*, 70(1), 53–73.

Frank, R. 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.

Friedman, D., and N. Singh (2009): "Equilibrium Vengeance," *Games and Economic Behavior*, 66(2), 813–829.

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

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

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

Güth, W., and M. E. Yaari (1992): "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in *Explaining Process and Change*, ed. by U. Witt, pp. 22–34. University of Michigan Press, Ann Arbor, MI.

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

Heifetz, A., C. Shannon, and Y. Spiegel (2007): "What to Maximize if You Must," *Journal of Economic Theory*, 133(1), 31–57.

Heller, Y. (2014): "Three Steps Ahead," Forthcoming in Theoretical Economics.

Heller, Y., and E. Mohlin (2014): "Stable Observable Behavior," Working paper.

Herold, F., and C. Kuzmics (2009): "Evolutionary Stability of Discrimination under Observability," *Games and Economic Behavior*, 67, 542–551.

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

Hofbauer, J., and W. H. Sandholm (2009): "Stable Games and Their Dynamics," *Journal of Economic Theory*, 144(4), 1665–1693.

Hofbauer, J., and K. Sigmund (1988): *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, Cambridge.

Holloway, R. (1996): "Evolution of the Human Brain," in *Handbook of Human Symbolic Evolution*, ed. by A. Lock, and C. R. Peters, pp. 74–125. Clarendon Press, Oxford.

Hopkins, E. (2014): "Competitive Altruism, Mentalizing and Signalling," Forthcoming in American Economic Journal: Microeconomics.

Huck, S., and J. Oechssler (1999): "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior*, 28, 13–24.

Humphrey, N. K. (1976): "The Social Function of Intellect," in *Growing Points in Ethology*, ed. by P. P. G. Bateson, and R. A. Hinde, pp. 303–317. Cambridge University Press, Cambridge.

Kim, Y.-G., and J. Sobel (1995): "An Evolutionary Approach to Pre-Play Communication," *Econometrica*, 63(5), 1181–1193.

Kimborough, E. O., N. Robalino, and A. J. Robson (2014): "The Evolution of "Theory of Mind": Theory and Experiments," Cowles Foundation Discussion Paper No. 1907R, Yale University.

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

Koçkcesen, L., and E. A. Ok (2000): "Evolution of Interdependent Preferences in Aggregative Games," *Games and Economic Behavior*, 31, 303–310.

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

Maynard Smith, J., and G. R. Price (1973): "The Logic of Animal Conflict," *Nature*, 246(5427), 15–18.

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

——— (2012): "Evolution of Theories of Mind," *Games and Economic Behavior*, 75(1), 299–312.

Moulin, H., and J. Vial (1978): "Strategically Zero-Sum Games: The Class of Games whose Completely Mixed Equilibria Cannot be Improved Upon," *International Journal of Game Theory*, 7(3), 201–221.

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

Ok, E. A., and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, 97, 231–254.

Possajennikov, A. (2000): "On the Evolutionary Stability of Altruistic and Spiteful Preferences," *Journal of Economic Behavior and Organization*, 42, 125–129.

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

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

Robson, A. J., and L. Samuelson (2011): "The Evolutionary Foundations of Preferences," in *The Social Economics Handbook*, ed. by J. Benhabib, A. Bisin, and M. Jackson, pp. 221–310. North Holland.

Rtischev, D. (2012): "Evolution of Mindsight, Transparency and Rule-Rationality," Working Paper, MPRA.

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

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

Sandholm, W. H. (2001): "Preference Evolution, Two-Speed Dynamics, and Rapid Social Change," *Review of Economic Dynamics*, 4, 637–679.

——— (2010): "Local Stability under Evolutionary Game Dynamics," *Theoretical Economics*, 5(1), 27–50.

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

Schelling, T. C. (1960): *The Strategy of Conflict*. Harvard University Press.

Schlag, K. H. (1993): "Cheap Talk and Evolutionary Dynamics," Bonn Department of Economics Discussion Paper B-242.

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

Sethi, R., and E. Somanthan (2001): "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273–297.

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

Stennek, J. (2000): "The Survival Value of Assuming Others to be Rational," *International Journal of Game Theory*, 29, 147–163.

Taylor, P. D., and L. B. Jonker (1978): "Evolutionary Stable Strategies and Game dynamics," *Mathematical Biosciences*, 40(1–2), 145–156.

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

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

——— (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," in *Game Equilibrium Models I. Evolution and Game Dynamics*, ed. by R. Selten, pp. 29–97. Springer.