Evolution of Theories of Mind in Strategic Interactions[*]

Erik Mohlin[†]

October 27, 2009

Abstract

This paper explores the evolution of player's models of how other people think – their 'theories of mind'. There is considerable experimental evidence that when people face an unfamiliar game they do not play a Nash equilibrium, but rather they behave in accordance with the cognitive hierarchy (Camerer, Ho & Chong 2004 QJE) and level-k (Stahl & Wilson 1995, GEB, Nagel 1995 AER) models. The evolution of different theories of mind is formalized as the evolution of different cognitive types within the cognitive hierarchy and level-k models. The models are also extended to allow for partial observation of the opponent's types. It is found that evolution of types does not in general lead to Nash behavior in unfamiliar games. Under plausible assumptions evolution leads to states where different, relatively unsophisticated, types co-exist, in line with the experimental evidence. This result holds both with and without partial observation of types.

Keywords: Cognitive Hierarchy; Level-k; Theory of Mind; Evolution.
JEL codes: C73.

[*] Thanks to Jörgen Weibull for advice and guidance. Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged.

[†] E-mail: erik.mohlin@hhs.se. Mail: Department of Economics, Stockholm School of Economics, P.O. Box 6501, SE-113 83 Stockholm, Sweden.

1 Introduction

This paper studies the evolution of players' models of other players' beliefs – what psychologists and cognitive scientists call a *theory of mind*.[1] In particular the paper provides evolutionary foundations for the cognitive hierarchy (Camerer et al. (2004)) and level-k models (Stahl and Wilson (1995) and Nagel (1995)).

On the one hand it has been found empirically, that people often do not play a Nash equilibrium of a game, unless they have played it many times with sufficiently rich feedback. Cognitive hierarchy and level-k models have been shown to outperform Nash equilibrium as prediction in many one-shot experimental games (Camerer (2003)).[2] On the other hand, when players are allowed to gain experience they often eventually adapt to play a Nash equilibrium (for an overview see Camerer (2003)). The formalization of such equilibration processes has been the focus of a large literature on learning and evolution (e.g. Weibull (1995), Fudenberg and Levine (1998), and Sandholm (2007)).

Since the cognitive hierarchy and level-k models focus on explaining behavior of inexperienced individuals they do not conflict with the use of Nash equilibrium as a prediction for behavior in the long run. However it should be pointed out that many decisions and games with important consequences, such as choosing a career or a mate, do not allow for much learning to take place. Hence, the strategic ability displayed in games where there is little or no opportunity to learn, should be highly relevant for the success of a person, in economic as well as biological terms. Thus it is natural to investigate the evolutionary foundations of the behavior that is observed when people face new games. In particular, *why do people not play a Nash equilibrium strategy initially*, given that they often eventually adapt to do so? The need to investigate the CH and level-k models from an evolutionary perspective is accentuated by the extensive recent use of these models in applied work.

According to both the cognitive hierarchy model and the level-k model each individual belongs to a cognitive type $k \in \{0, 1, 2, ...\}$. The cognitive type of an individual is not defined in terms of her payoffs, as in the standard approach to incomplete information games, but rather in terms of her model of how other individuals think. The level-k *model* assumes that all individuals of type $k \geq 1$ play a best response given their beliefs. Furthermore type $k \geq 1$ believes that everyone else belongs to type $k-1$, and knows that type $k-1$ believes that everyone is of type $k-2$, and so on. Type 0 randomizes uniformly over the strategy space. Thus type 1 plays a best response to the uniform distribution, type 2 plays a best response to the what type 1 does, and so on. Similarly,

[1] The term 'theory of mind' was introduced in this context by Premack and Wodruff (1979).

[2] In order to claim that Nash equilibrium predictions are refuted by experimental data it is of course crucial to estimate preferences. This is not done in the mentioned studies, but the preferences needed to make the observed behavior conform to Nash equilibrium seem like a much less reasonable explanation than attributing the behavior to some form of incorrect expectations. See the discussion in Costa-Gomes and Crawford (2006).

the *cognitive hierarchy (CH) model* also assumes that all individuals of type $k \geq 1$ play a best response given their beliefs, and that type 0 randomizes uniformly over the strategy space. However, in contrast to the level-k model, it posits that an individual of type $k \geq 1$ believes that everyone else belongs to type 0 through $k-1$, and has correct beliefs about the relative fractions of these types. Type k knows how lower types form their beliefs, and knows that they best respond given their beliefs. Thus type k plays a best response against a weighted average of what lower types do.

When estimating the fractions of individuals in the population that behave in accordance with the different types in these models, it is commonly found that individuals of different sophistication coexist.[3] Most people are estimated to belong to type $1-2$ and individuals of type 3 and higher are very rare (see e.g. Costa-Gomes and Crawford (2006)).[4] It is natural to ask why evolution has produced players whose behavior in initial rounds of a game conforms to the level-k and CH models.[5] In particular, *why is the strategic sophistication heterogeneously distributed in the population*, and why is it so limited relative to the ideals of rationalistic game theory?

In the model of this paper each individual belongs to a cognitive type $k \in K = \{0, 1, ..., \kappa\}$. In the main model the types are taken from the cognitive hierarchy (CH) framework, as described above. The CH framework is only designed to capture situations where players lack specific information about the cognitive type of their opponent; the beliefs of a player only depends on that player's own type and not on the opponent's type. This might be a plausible assumption for anonymous experimental games (as well as for some real life interactions). But in many situations people face new interactions with people they have already met, and they know something about how their opponents usually reason in games. Therefore I extend the CH framework to allow for the case of partially observed types. This is done by assuming that an individual of a higher type can observe the type of an individual of a lower type, but not vice versa. Consequently the lower type individual behaves as in the standard CH model, whereas the higher type individual plays a best response to what the lower type individual does. When two individuals of the same type meet I assume that they understand that they are of the same type and play a Nash equilibrium. (Two alternatives to this assumption are considered in the appendix, section 8, and the main contrasts between the cases of observed and

[3] Further empirical evidence on the limitations of the theories of mind employed by humans is provided by psychological research on so called theory of mind tasks, e.g. Kinderman et al. (1998) and Apperly et al. (2007).

[4] Ohtsubo and Rapoport (2006) claim that these and similar studies underestimate the depth of reasoning since they can not differentiate between depth of reasoning and difficulties with computing best responses. Still, their own experiment also reveals limits and heterogeneity with respect to steps of thinking, with the mode being at the fourth step.

[5] There is reason to believe that strategic reasoning is a domain specific ability which is implemented by specialized modules; Cosmides and Tooby (1992). According to Penke et al. (2007) heterogeneity in domain specific abilities is best explained with frequency dependent selection (their term is balancing selection), and not with random variation.

unobserved types are preserved.)

For the evolutionary analysis consider a large population of individuals, each belonging to a cognitive type in K. A state is a probability distribution over types. Individuals are randomly matched to play a symmetric two-player game that is drawn from a class of games $\mathcal{G}$. The types of the drawn individuals, together with the state, determine the individuals' beliefs, and hence their actions and payoffs, in the drawn game. The individuals cannot condition their type on the particular game being drawn. Therefore the average payoff of a type is equal to a weighted sum of what that type earns on average in each of the games in $\mathcal{G}$ – with the weights are equal to the probabilities of drawing the different games. The fractions of types evolve in proportion to the average payoffs earned by individuals of the different types. Formally this is represented by the replicator dynamics. Separate analyzes are carried out for each different special cases when all probability is put on one game in $\mathcal{G}$. From the results in these special cases one can draw conclusions about which mixes of games in $\mathcal{G}$ that are conducive to the evolution of the distribution of types that we observe in the data. Formally, for each game in $\mathcal{G}$, one I use the map from types and states to payoffs, to define a *cognitive game*, and study evolution in that game.

The analyzed games include all 2 × 2-games; coordination games like the Stag Hunt, games with a unique ESS, like the Hawk-Dove game, and dominance solvable games, i.e. Prisoners' Dilemmas. It also includes dominance solvable games with an arbitrary number of strategies. In particular I define a class of 'Machiavellian games' in order to capture the kind of interactions that shaped the evolution of human theories of mind (and human cognitive abilities in general), according to the "social brain" or "Machiavellian intelligence" hypothesis (Humphrey (1976), Alexander (1990), and Byrne and Whiten (1998)). In Machiavellian games strategies are ordered from the most unsophisticated and naive to the most sophisticated and Machiavellian. This class of games includes the Travelers' Dilemma (Basu (1994)). In an appendix I also consider dominance solvable games with infinite (compact) strategy spaces.

It might seem obvious that the more sophisticated a theory of mind an individual uses, the more successful she will be. However, brief reflection reveals that a more sophisticated theory of mind is not always beneficial. Consider the CH model with unobserved types. In a coordination game each pure strategy is the unique best response to itself. In such a game all types, except type 0, will play the same strategy, namely the best response to the uniform distribution. All types except type 0 earn the same payoff and type 0 earns less than the other types. Thus there is no advantage associated with being of a higher type (as long as you are above type 0) and evolution can lead to any state where type 0 is extinct. In other games there are states where higher types earn strictly less than lower types so that lower types always survive: I show that in 2 × 2-games with a unique evolutionarily stable strategy (ESS), the asymptotically stable set of states include interior states where all types co-exist, and does not include any state where only one type exists. I provide sufficient conditions for the asymptotic stability of a set of states with a positive fraction of type 0, and this condition even applies to some dominance solvable games. In other dominance solvable games – the Machiavellian games – evolution leads to the state where everyone belongs to one of the highest types. But when types are observed this is not generally true: In the Traveler's Dilemma it might be the case that all states some type $k > 0$ does not exist, are unstable. Furthermore, in a 2×2-game with a unique ESS evolution from any interior initial condition leads to a state where all types $k > 0$ co-exist. The contrast between the results for unobserved and observed types also carries over to infinite dominance solvable games.

Taken together, these results explain how evolution could lead to a state with a heterogeneous population, most of which is constituted by relatively low types, in accordance with the empirical findings. From an evolutionary perspective the potential advantage of increased strategic sophistication, or a better theory of mind, has to be weighted against the cost of increased reasoning capacity. In order to avoid speculative assumptions about such costs I abstract from them in the formal analysis. However, when interpreting the results one should bear in mind that the existence of cognitive costs strengthens the case for lower types. To the best of my knowledge the present paper is the first one to study the evolutionary foundations of the CH model. It is also the first paper to extend the CH model to the case of partially observed types.

Evolutionary reasoning about rationality and maximizing behavior have a long tradition in economics, going back at least to the classical evolutionary motivation for the "as if" approach to economics, associated with Alchian (1950) and Friedman (1953). However, there are only a few studies of evolution of cognitive types.[6] A pioneering paper is Stahl (1993). In his model there is a set of types $n \in \{0, 1, 2, ...\}$ and all individuals are perfectly informed about the actual distribution of types in the population. Type 0 is divided into subtypes, each preprogrammed to different pure strategy. Type n believes that everyone else is of a lower type and is able to deduce what lower types will do. She chooses among strategies that are n^{th} order rationalizable conditional on the actual distribution of types. An individual of type n does not form any belief about what the opponent will choose among the set of n^{th} order rationalizable strategies. In order to choose among strategies in this set, each individual has a secondary strict preference ordering over strategies. Banerjee and Weibull (1995) study the interaction between individuals that are preprogrammed to different strategies and individuals that optimize given a correct belief about the strategy of the opponent (full information case) or the population distribution of strategies (incomplete information case). Another related paper is Stennek (2000) who studies the evolution of ascribing different degrees of rationality to one's opponent. An individual of type $d \in \{0, 1, 2, ...\}$ believes that everyone else is of type $d - 1$ and chooses some d-iterations undominated action (because she is rational

[6] There has also been some study of learning in this context e.g. in Nagel (1995), Ho et al. (1998), Stahl (2000), and Haruvy and Stahl (2008).

and also assumes that the opponents choose some $d-1$-iterations undominated action). Like in Stahl's model the choice among the d-iterations undominated actions is made in accordance with some preference over the pure strategies rather than on the basis of a belief about which strategy (in the set) that the opponent will chose.[7]

There are several differences between these papers and the present one: First, all these papers assume that a fixed game is played recurrently and that some individuals are preprogrammed to pure strategies, like in the conventional evolutionary game theory literature. Since the payoff to different strategies varies from one game to another these models can not address the question of evolution of behavior in unfamiliar games (or games where there is little scope for learning). Secondly, these papers also build on behavioral models that lack the kind of empirical support that the CH and level-k models have (see Costa-Gomes and Crawford (2006) and Camerer (2003)). Third, in all of these models there are many types whose behavior is not fully determined by best response given beliefs. Instead additional preference orderings are added in Stahl's and Stennek's studies, and several types are preprogrammed in the studies by Stahl and Bannerjee & Weibull. Such an approach potentially confounds the question of how theories of mind have evolved, and the question of how optimizing behavior has evolved. In contrast, in the CH and level-k models all types except level 0 best respond given their beliefs.

The results derived for the CH model carry over to the *level-k model* and (with one exception). In order to check the robustness of the results the model is extended to include a *Nash equilibrium (NE) type*, which is preprogrammed to a Nash equilibrium strategy. As another robustness check I investigate the impact of adding a *rational expectations (RE) type*, which somehow knows how all individuals form their beliefs. This type acts like it had the perfect theory of mind of a conventional *homo oeconomicus*, and best responds to the population distribution of strategies. (Note that if all individuals are of this type the model reduces to the standard set up and everyone plays the same Nash equilibrium.)

The paper is organized as follows: The next section presents the model; Section 2.1 introduces notation and defines the concept of a cognitive game. Section 2.2 describes the cognitive types according to the CH model. The evolutionary set up is presented in section 2.3.1 and the class of underlying games is described in section 2.4. Section 3 contains the main results both for the case of observed types and the case of unobserved types. Section 4 discusses the level-k model, as well as two extensions – the NE and RE types. Section 5 concludes. All proofs of results in the main text are in appendix A, section 6. The model is extended to games with infinite strategy spaces in appendix B, section 7. Appendix C, section 8, discusses alternative specifications of the model for observed types.

[7] Other, more distantly related, studies include Robson (2003) and Samuelson (2001*a*).

2 Model

2.1 Games

2.1.1 Underlying Game

Consider a symmetric two-player normal form game G with a finite pure strategy set S and mixed strategy set $\Delta(S)$.[8] Let $\sigma^s \in \Delta(S)$ denote the degenerate mixed strategy that puts all weight on pure strategy s. Payoffs are given by $\pi : S \times S \to \mathbb{R}$, where $\pi(s_i, s_j)$ is the payoff to player i when player i plays s_i and player j plays s_j. For mixed strategies the payoffs are given by $\tilde{\pi} : \Delta(S) \times \Delta(S) \to \mathbb{R}$ with

$$\tilde{\pi}(\sigma_i, \sigma_j) = \sum_{s \in S} \sum_{t \in S} \pi(s,t)\, \sigma_{i,s} \sigma_{j,t},$$

where $\sigma_{i,s}$ is the weight put on pure strategy s by player i's mixed strategy σ_i and $\sigma_{j,t}$ is the weight put on pure strategy t by player j's mixed strategy σ_j.

The games that will be studied in this paper are described later in sections 2.4 and 3. Here I only introduce one game, the Travelers' Dilemma, which will be used to illustrate the model. The game was introduced by Basu (1994). The more general form that I use here is taken from Capra et al. (1999). (In Basu's version of the game $a = 0$, $b = 200$ and $R = 2$.)

Definition 1 *The Travelers' Dilemma is a symmetric two-player normal form game with strategy space $S = \{a, a+1, a+2, ..., b\}$, for $a, b \in \mathbb{N}$, satisfying $a < b$. The payoff to player i (choosing strategy s_i) when facing player j (choosing strategy s_j) is*

$$\pi(s_i, s_j) = \begin{cases} s_i & \text{if } s_i = s_i \\ s_i + R & \text{if } s_i < s_i \\ s_j - R & \text{if } s_i > s_i \end{cases},$$

for some real number $R > 1$.

If the opponent plays a mixed strategy $\sigma_j \in \Delta(S)$ then expected payoff to a pure strategy $s_i \in S$ is

$$E[\pi(s_i, s_j) | \sigma_j] = \sum_{t \in S} \sigma_{j,t} \pi(s_i, t).$$

[8] The restriction to symmetric games is not essential as long as there is a single population playing the game in question. When analyzing evolutionary stability in asymmetric games played by a single population it is natural to assume that agents are uniformly randomly allocated to player positions, and that strategies specify actions to be taken conditional on received player position. The resulting game is symmetric.

Define the *pure best reply correspondence* $\beta : \Delta(S) \twoheadrightarrow S$ by

$$\beta(\sigma_j) = \arg\max_{s_i \in S} E\left[\pi(s_i, s_j) | \sigma_j\right].$$

If the best response is unique I will write $\beta(\sigma) = s$ rather than $\beta(\sigma) = \{s\}$. The *mixed best reply correspondence* $\tilde{\beta} : \Delta(S) \twoheadrightarrow \Delta(S)$ is defined by maximizing over $\Delta(S)$ instead of S. The uniform randomization over the set of pure best responses to σ, is denoted $\bar{\beta}(\sigma)$.

I will allow myself to abuse the notation in the following way: I will use s as a short hand for the mixed strategy σ^s that puts all weight on the pure strategy s. Thus $\beta(s)$ (and likewise for $\tilde{\beta}$ and $\bar{\beta}$) stands for the (pure) best response to the mixed strategy σ^s, and $\tilde{\pi}(s_i, s_j)$ stands for the payoff of to i playing σ^{s_i} against j playing σ^{s_j}.

2.1.2 Cognitive Game

Consider a single population consisting of a finite set of cognitive types. Unless otherwise noted the set will be $K = \{0, 1, 2, ..., \kappa\}$. The set of probability distributions over K is $\Delta(K)$ so a *population state* is a point

$$x = (x_0, x_1, ...x_\kappa) \in \Delta(K).$$

Suppose that individuals from this population are randomly matched to play a fixed symmetric two-player normal form game G. All individuals of the same type play the same strategy. At state $x \in \Delta(K)$ the distribution of strategies among individuals of type k is $z_k(x) \in \Delta(S)$. The weight put on pure strategy s by type k at state x is $z_{k,s}(x)$. The *population distribution of strategies* is $q(x) \in \Delta(S)$, where the weight put on strategy s in state x is

$$q_s(x) = \sum_{j=0}^{\kappa} z_{j,s}(x)\, x_j.$$

For a given game G, the expected payoff of type k at state x is given by the function $\Pi_k^G : \Delta(K) \to \mathbb{R}$, with

$$\Pi_k^G(x) = \sum_{s \in S} z_{k,s}(x) \left(\sum_{t \in S} q_t(x)\, \pi(s, t) \right).$$

Let $\Pi^G(x) = \left(\Pi_0^G(x), ..., \Pi_\kappa^G(x)\right)'$. The function $\Pi^G : \Delta(K) \to \mathbb{R}^{|K|}$, together with the set of types K and the set of population states $\Delta(K)$ defines a *cognitive population game* (see Sandholm (2007) about population games), or a cognitive game for short.

2.2 Cognitive Types

2.2.1 Unobserved Types

Consider the CH model, as it is usually formulated, with *unobserved* types. Type 0 randomizes uniformly over the strategy space S, independently of the population state x. For any set of strategies X let $U(X)$ denote the uniform distribution over X. If $X = S$ and there is no risk of confusion I will write U for short. Type 1 believes that everyone else is of type 0, i.e. believes that the population state is $\hat{x}^1 = (1, 0, ..., 0)'$ so type 1 expects to meet strategy $\hat{q}^1 = z_0$. This induces type 1 to play the (mixed or pure) strategy $z_1 = \bar{\beta}\left(\hat{q}^1(x)\right)$. Individuals of type $k > 1$ believe that all other individuals belong to type 0 through $k-1$, and that the population state is

$$\hat{x}^k = \frac{1}{\sum_{i=0}^{k-1} x_i}\left(x_0, x_1, ..., x_{k-1}, 0, ..., 0\right)'.$$

Note that in states where $x_0 = x_1 = 0$ the beliefs and behavior of type 2 is not well-defined according to the model. Hence the beliefs of type $k > 1$ are also not well-defined. Since we are generally interested in interior initial conditions, and since the system will never move from the interior when starting there, this is not an important limitation. In all other states the beliefs and the behavior of all types are well-defined.

Since type k has a correct belief about the relative fractions of lower types, she knows how the lower types will behave. Hence type k has the following belief about the population distribution of strategies.

$$\hat{q}^k(x) = \frac{1}{\sum_{i=0}^{k-1} x_i}\sum_{j=0}^{k-1} z_j(x)\, x_j.$$

Type k best replies given these beliefs. The set of pure best replies is $\beta\left(\hat{q}^k(x)\right)$. If $\beta\left(\hat{q}^k(x)\right)$ is not a singleton, then the individual is assumed to randomize uniformly among the elements in $\beta\left(\hat{q}^k(x)\right)$. Let $\bar{\beta}$ denote the uniform distribution over the pure strategies in β. So behavior of type k at state x is given by

$$z_k(x) = \bar{\beta}\left(\hat{q}^k(x)\right).$$

In order to apply the CH model to the Travelers' Dilemma game we need the following lemma.

Lemma 1 *In the Travelers' Dilemma, the best reply to the uniform randomization over a set of strategies* $\{a, a+1, ..., b\}$ *is* $s = \max\{t \in S : b - 2R \geq t\}$.

Note that if $2R$ is an integer, then the best reply is $s = b - 2R$. Using this lemma we have:

Example 2 *Consider the Travelers' Dilemma with $a = 0$, $b = 200$ and $R = 2$, and let $\kappa = 2$. Type 0 randomizes uniformly over $S = \{2, 3, ..., 200\}$. By the above lemma, type 1 plays $\beta(U(S)) = 196$. Thus type 2 believes that a fraction $x_0 / (x_0 + x_1)$ plays $U(S)$ and that a fraction x_1 plays 196. If $x_1 / (x_0 + x_1)$ is close to 0 then type 2 plays 196 and if $x_1 / (x_0 + x_1)$ is close to 1 then type 2 plays 195.*

2.2.2 Observed Types

Now consider the case of (partially) *observed* types. In the CH model it is implicitly assumed that individuals do not observe each other's type. Relaxing the assumption is not straightforward. Suppose two individuals A and B of different types k_A and k_B with $k_A < k_B$ meet to play a game. Since B is a higher type than A it seems reasonable to assume that B *can* observe A's type. Since the CH model postulates that every type is aware only of lower types, it seems reasonable to assume that A can *not* observe B's type. The most conservative way to extend the CH model here seems to be to assume that A believes that B belongs to one of the types $k < k_A$, with probabilities equal to the relative fractions of these types. Note that if $x_0 > 0$ then A will assign positive probability to B being of type 0. So whatever action B takes, A will not be surprised.

As before, type 0 randomizes uniformly against all opponents. For higher types things get more complicated. If an individual A of type k_A meets and opponent B of type $k_B < k_A$ then A detects B's type and therefore plays a best response $\beta(z_{k_B})$, to what B does. If $k_B > k_A$ then A cannot identify which type B belongs to, so she forms beliefs like in the case of unobserved types. That is, she forms the expectation $\hat{q}^{k_A}(x)$, as defined above, and best responds to this, i.e. plays $\bar{\beta}\left(\hat{q}^{k_A}(x)\right)$.

If $k_A = k_B = k$ then each individual understands that the opponent is of the same type. In this case I will assume that they play a Nash equilibrium. In case there are many Nash equilibria I assume that they will choose a symmetric Nash equilibrium whose component strategies are ESS. This way of modeling the encounter between two individuals of the same type might be considered arbitrary. However, it does capture essential parts of the CH model; each individual is overconfident and reasons iteratively. Moreover, in the appendix, section 8, I consider two alternative specifications of behavior in the case of partially observable types, which both yield similar results.

Example 3 *Consider the Travelers' Dilemma with $a = 0$, $b = 200$ and $R = 2$, and let $\kappa = 2$. Type 0 randomizes uniformly. Type 1 plays $\beta(U) = 196$ against type 0 and type 2. Type 1 plays the Nash equilibrium strategy 2 against other individuals of type 1. Type 2 plays 196 against type 0, and 195 against type 1. Type 2 plays 2 against another individual of the same type.*

2.3 Evolution

2.3.1 Dynamics

Above a cognitive game was defined relative to a fixed underlying game. But we are interested in the question of what types that survive when individuals face several different games. To formalize this idea it is assumed that every time two individuals are drawn to play, a game G is randomly drawn from a finite set of games $\mathcal{G}$ according to a measure μ. Individuals do not condition their type on the drawn game, so the average payoff to type k is a μ-weighted sum of the average payoffs that type k individuals earn in each of the games in $\mathcal{G}$. Formally $\Pi_k^G(x)$ is the payoff to type k in game G in state x, and the average payoff of type k in state x is

$$\Pi_k^{\mathcal{G}}(x) = \sum_{G \in \mathcal{G}} \mu(G) \Pi_k^G(x).$$

The average payoff in the population is

$$\bar{\Pi}^{\mathcal{G}}(x) = \sum_{k=0}^{\kappa} x_k \Pi_k^{\mathcal{G}}(x).$$

Evolution of types is determined by the *replicator dynamics*

$$\dot{x}_k = [\Pi_k^{\mathcal{G}}(x) - \bar{\Pi}^{\mathcal{G}}(x)] x_k.$$

I will not analyze this dynamic explicitly. Instead I analyze each of the special cases where $\mu(G) = 1$ for some game $G \in \mathcal{G}$. From these extreme cases one can draw conclusions about what weights on the underlying games in $\mathcal{G}$ that may support the evolution of different type distributions.

Since behavior of types $k \geq 2$ generally is not continuous in the population state we can not assume that payoff is continuous in the population state. This poses a difficulty for the existence and uniqueness of solutions to the replicator dynamics. However, in applications below we will find ways to handle this potential problem.

2.3.2 Stability

We are interested in Lyapunov stable and asymptotically stable states of the replicator dynamics. In the CH model the dynamics is not well-defined for states where $x_0 = x_1 = 0$ so such states can not be stable, but they can still be attracting. For reasons that will become clear, asymptotically stable sets are also of importance. Hence we need the following definitions:

Definition 4 *A closed set $A \subset \Delta(K)$ is Lyapunov stable if every neighborhood B of A contains a neighborhood B^0 of A such that if the system starts in $B^0 \cap \Delta(K)$ at t_0 then the system remains in B at all $t \geq t_0$.*

A closed set $A \subset \Delta(K)$ is asymptotically stable if it is Lyapunov stable and if there exists a neighborhood B^* of A such that if the system starts in B^* at t_0 then as $t \to \infty$ the system goes asymptotically to A.

The basin of attraction of a closed set $A \subset \Delta(K)$ is the set of states such that starting from such a state the system goes to A as $t \to \infty$.

A set $A \subset \Delta(K)$ is an attractor if its basin of attraction is a neighborhood of A.

Stability of a point is defined as the stability of the singleton $\{x\}$. Note that a Lyapunov stable set is asymptotically stable if and only if it is an attractor. For more on these concepts see Weibull (1995). A state is polymorphic if it contains positive fractions of more than one type. Otherwise the state is monomorphic. Finally the concept of an evolutionarily stable strategy (ESS) will be used:

Definition 5 *A strategy $\sigma \in \Delta(S)$ is an evolutionarily stable strategy (ESS) if (i) $\tilde{\pi}(\sigma', \sigma) \leq \tilde{\pi}(\sigma, \sigma)$ for all $\sigma' \in \Delta(S)$, and (ii) $\tilde{\pi}(\sigma', \sigma) = \tilde{\pi}(\sigma, \sigma)$ implies $\tilde{\pi}(\sigma', \sigma') < \tilde{\pi}(\sigma, \sigma')$ for all $\sigma' \neq \sigma$.*

2.4 The Class of Underlying Games

This section goes through the types of games that will be analyzed below. All symmetric two-player, two-strategy games fall into one of three categories of games which share the same best reply properties (see Weibull (1995)). Consider a game

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

I will assume that payoffs are generic, in the sense that $a_{11} - a_{21} \neq a_{22} - a_{12}$. Three different cases can be discerned. (I) If $a_{11} - a_{21} > 0$ and $a_{22} - a_{12} > 0$, then there are two symmetric pure strategy equilibria, both of which correspond to evolutionarily stable strategies (ESS), and a symmetric mixed strategy equilibrium, which is not ESS. This is the class of 2×2 *coordination games*. The Stag Hunt Game falls into this category. (II) If $a_{11} - a_{21} < 0$ and $a_{22} - a_{12} < 0$, then there are two asymmetric pure strategy equilibria and one symmetric mixed strategy equilibrium, where only the latter corresponds to an ESS. This is the class of 2×2 *games with a unique interior ESS*. The Hawk Dove Game falls into this category.[9] (III) If $a_{11} - a_{21} < 0$ and $a_{22} - a_{12} > 0$, then the game is *dominance solvable*. The Prisoners' Dilemma Game falls into this category. Games with $a_{11} - a_{21} > 0$ and $a_{22} - a_{12} < 0$ have the same properties provided that the strategies are relabeled.

[9] The Battle of Sexes also fall into this category, provided that strategies are renamed.

Since these games only have two strategies, different types will often not distinguish themselves behaviorally. Thus it might be trivial that higher types do not earn strictly more than lower types. One could generalize the 2×2 coordination games to games with any finite number of strategies where each strategy is a best reply to itself, and one could generalize 2×2 games with a unique interior ESS to games with any finite number of strategies where there is a unique interior ESS. However, as will become clear below, neither of these kinds of games provide any fertile ground for breeding higher types. The games where higher types have the strongest advantage are dominance solvable games. Therefore I will consider dominance solvable games with more than two strategies.

Humans have extraordinary cognitive abilities compared to other animals (Roth and Dicke (2005)). According to the prominent "social brain" or "Machiavellian intelligence" hypothesis these abilities developed as a result of the demands of social competition and interaction rather than the demands of the natural environment (Jolly (1966), Humphrey (1976), Alexander (1990), Byrne and Whiten (1998), Dunbar (1998), Dunbar (2003) and Flinn et al. (2005)). In a single person decision problem there is a fixed benefit of being smart but in a strategic situation there is a potential advantage of relative intelligence; it may be important to be smarter than the opponent. For a concrete example of what might have been an interaction relevant for the evolution of theory of mind, consider the following (much cited) story from an experiment by Menzel (1974). A subordinate chimpanzee named Belle attempted to prevent a dominant chimpanzee named Rock from finding food that she had seen the experimenters hide:

> If Rock was not present, Belle invariably led the group to food and nearly everybody got some. In tests conducted when Rock was present, however, Belle became increasingly slower in her approach to the food. The reason was not hard to detect. As soon as Belle uncovered the food, Rock raced over, kicked or bit her, and took it all.
>
> Belle accordingly stopped uncovering the food if Rock was close. She sat on it until Rock left. Rock, however, soon learned this, and when she sat on one place for more than a few seconds, he came over, shoved her aside, searched her sitting place and got the food.
>
> Belle next stopped going all the way. Rock, however, countered by steadily expanding the area of his search through the grass near where Belle sat. Eventually, Belle sat farther and farther away, waiting until Rock looked in the opposite direction before she moved toward the food at all – and Rock in turn seemed to look away until Belle started to move somewhere. On some occasions Rock started to wander off, only to wheel around suddenly precisely as Belle was about to uncover the food.
>
> In other trials when we hid an extra piece of food about 10 feet away from the large pile, Belle led Rock to the single piece, and while he took it she raced for the pile. When Rock started to ignore the single piece of food to keep his

watch on Belle, Belle had temper tantrums. (Menzel (1974) pp. 134-5.)

We can order the strategies of Belle and Rock according to complexity and deceptiveness from very simple and naive strategies to very sophisticated and Machiavellian strategies. An important feature of this interaction is that for each strategy chosen by Rock, Belle has incentives to choose a somewhat more sophisticated strategy in order to save some of the food from Rock. Likewise Rock has an incentive to choose a more sophisticated strategy than Belle. The following symmetric game is in line with the above story: The strategy space has a linear order, $S = \{1, 2, ... |S|\}$, such that increasing numbers represent increasing deceptiveness. A strategy s is primarily targeted at outsmarting the strategy slightly below s. Formally, for a given strategy s_j of the opponent, player i maximizes her payoff by some strategy above s_j, i.e. $\beta(s_j) > s_j$. Moreover, suppose that the farther away s_i is from the strategy directly above s_j the lower is the payoff to player i. This feature ensures that there is not one single strategy that is optimal against all lower strategies. Formally we have:

Definition 6 *A symmetric two-player game G is a Machiavellian game, if the following holds: S has a linear order, so that one can write $S = \{1, 2, ... |S|\}$. For each strategy $t < |S|$ of the opponent, the payoff $\pi(s,t)$, as a function of strategy s, is single peaked with its maximum at some $s > t$. If $t = |S|$, then $\pi(s,t)$ is maximized at $s = t$.*

The Travelers' Dilemma is a Machiavellian game.

3 Results

3.1 Unobserved Types

First consider coordination games. In all states all types $k > 0$ earn the same and earn more than type 0, so have the following simple result.

Proposition 1 *In any cognitive game based on an underlying coordination game, evolution from any interior initial state converges to a state where $x_0 = 0$ and all other types exist in the same relative fractions as in the initial state.*

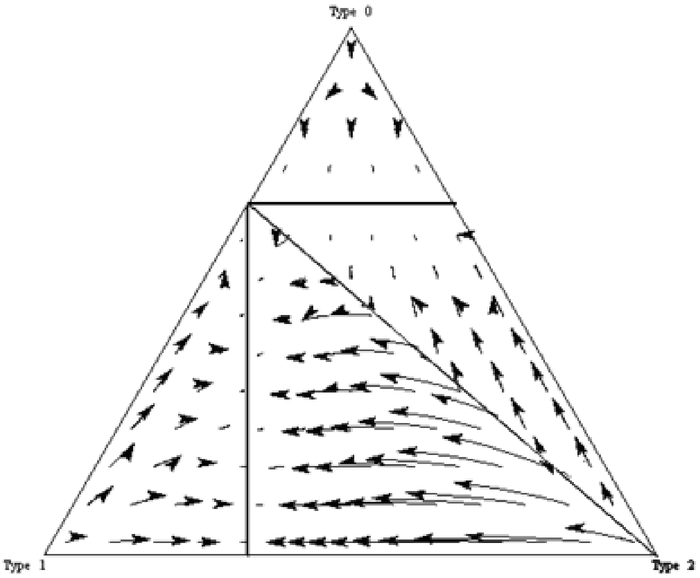
Thus in coordination games there is no evolutionary advantage of belonging to a higher type as long as one does not belong to type 0. In cognitive games based on underlying games with a unique interior ESS we get a different result:

Proposition 2 *In any cognitive game based on an underlying 2×2-game with a unique interior ESS, let*

$$X^{ESS} = \left\{x \in \Delta(K) : z(x) = \sigma^{ESS}\right\}.$$

No monomorphic states are stable and X^{ESS} is the unique asymptotically stable set, with the whole interior as its basin of attraction.

For the case of $K = \{0, 1, 2\}$ the following figure illustrates the proposition.[10] The vertices of the simplex (the edges of the triangle) represent the states where everyone is of the same type – as labeled in the figure. The figure was generated by assuming that the general payoff matrix for 2×2-games satisfies $a_{11} - a_{21} = -2$ and $a_{22} - a_{12} = -1$. The unique ESS of this game is $(1/3, 2/3)$. The thick (vertical and horizontal) lines represent X^{ESS}, the set of states where aggregate behavior corresponds to the ESS. In the area below (southwest of) the thin diagonal line type 2 plays H and above that line type 2 plays D. Starting from any point that is not monomorphic, evolution leads to some point in X^{ESS}. In all states in X^{ESS} two or more types co-exist. In all but three states in X^{ESS}, all types co-exist.



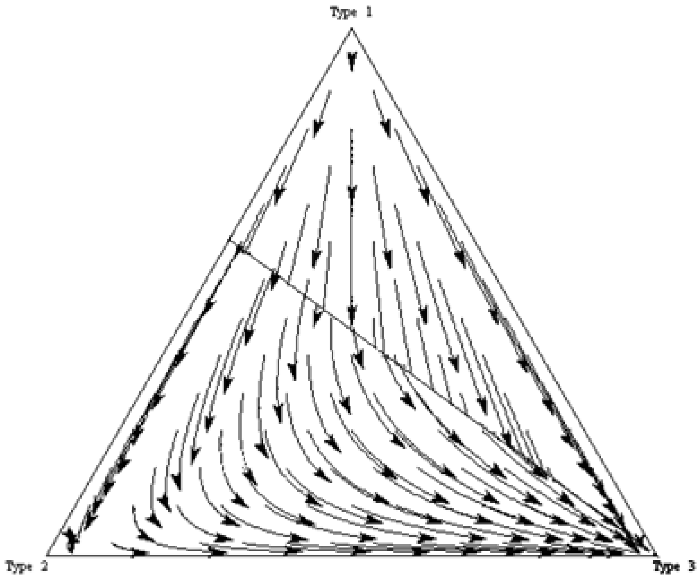
Machiavellian games are a less friendly environment for types with a less sophisticated theory of mind. In line with the social brain hypothesis we find that higher types have a strict advantage over lower types.

Proposition 3 *In a cognitive game based on an underlying Machiavellian game: Let $\tilde{k}$ be the minimum number of iterated best responses to the uniform distribution that are required to reach the Nash equilibrium strategy. Evolution from any interior initial condition leads to states where asymptotically only types $k \geq \tilde{k}$ exist.*

For the case of $\kappa = 3$, this is illustrated in the figure below (using $R = 3/2$, $a = 0$, and $b = 7$). Type 0 is extinct so the figure only depicts what happens when type 1, 2,

[10] All figures created with the help of the software Dynamo (Sandholm and Dokumaci (2007)), with some additional editing to obtain diagrams capturing discontinuous behavior.

and 3 are present.



Above the (thin) diagonal line type 3 plays the best response to what type 1 does and below that line type 3 plays the best response to what type 2 does.

It might be thought that the result for Machiavellian games – that being more sophisticated is strictly advantageous – carries over to any dominance solvable game, but this is not the case. Recall the definition of strict dominance:

Definition 7 *A pure strategy $s_i \in S_i$ of player i is strictly dominated if there is a mixed strategy $\sigma_i \in \Delta(S_i)$ of player i such that $\pi(s_i, s_{-i}) < \pi(\sigma_i, s_{-i})$ for all $s_{-i} \in S_{-i}$.*

A game is dominance solvable (in mixed and pure strategies) if iterated elimination of strictly dominated strategies leads to one remaining strategy profile. This profile constitutes a Nash equilibrium. Similarly a game is dominance solvable in *pure* strategies if iterated elimination of strategies that are strictly dominated by pure strategies. The following is an example of a game that is dominance solvable in *mixed*, but not in pure strategies.

Example 8 *The game G^{MDS} is defined by the following payoff matrix (MDS stands for mixed dominance solvable):*

$$\begin{array}{cccc} & A & B & C \\ A & 3 & 2 & 0 \\ B & 8 & 0 & 0 \\ C & 3 & 3 & 1 \end{array}.$$

Strategy A is strictly dominated by e.g. the mixed strategy that puts probability $1/4$ on B and probability $3/4$ on C. After deletion of strategy A, strategy C strictly dominates strategy B. Thus the remaining strategy profile (C,C) is the unique Nash equilibrium.

It has been demonstrated (Samuelson and Zhang (1992)) that in a population of individuals which are preprogrammed to different pure strategies, evolution under the replicator dynamic wipes out any strategy that does not survive iterated elimination of strictly dominated strategies. However a very different result is obtained in the case of evolution of cognitive types. The game G^{MDS} shows that it might be the case that cognitive evolution leads to stable states where aggregate behavior does not correspond to a Nash equilibrium and where a positive fraction of type 0 survives:

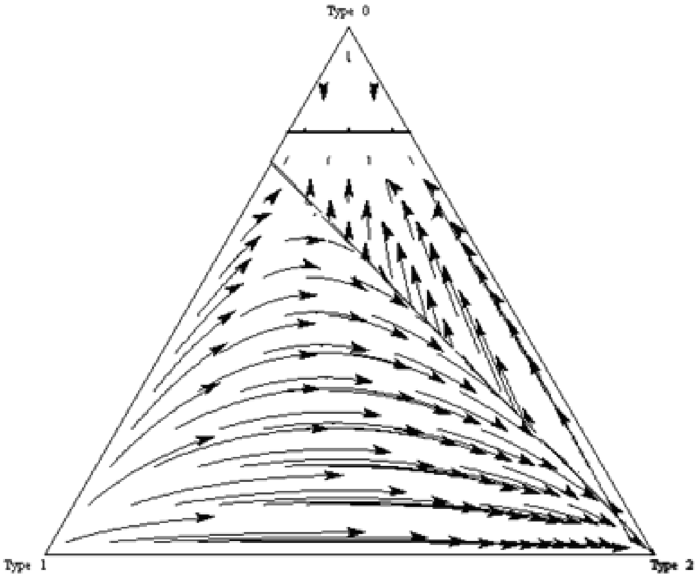
Proposition 4 *In the cognitive game based on game G^{MDS}: **(a)** A state is Lyapunov stable if and only if it belongs to*

$$R = \{x \in \Delta(K) : x_0 = 15/19\}.$$

***(b)** No state is asymptotically stable and the set R is the unique asymptotically stable set.*
***(c)** The basin of attraction of R includes states where x_0 is arbitrarily small.*

There are no states with well defined beliefs where aggregate behavior corresponds to the Nash equilibrium. Nash equilibrium-behavior is the limit of behavior as one moves (from some sets of states) towards the states where beliefs are not defined. These limit states are not attracting, since they do no belong to R.

For the case of $\kappa = 2$ the figure below illustrates the dynamics. The thick (horizontal) line denotes the set R. Below the thin diagonal line type 2 plays C and above the thin line type 2 plays B. If type 0 and 1 is present in the population then aggregate behavior does not correspond to the Nash equilibrium, and if they are not present then beliefs of type 2 are not well-defined. However, Nash equilibrium behavior is the limit of behavior as one moves, from certain states below the diagonal, towards the state where everyone is of type 2 (where beliefs are not defined). Still such this state is not attracting since evolution from states arbitrarily close to the southwest corner (where everyone is of type 2) leads to the set R.



To get an intuition for these results note that there is a region where everyone plays the best reply to the uniform distribution. Furthermore note that the payoff that type 0 earns against an individual playing a best response to the uniform distribution, is higher than the payoff that an individual playing a best response to the uniform distribution earns when meeting another individual doing the same thing. That is

$$\tilde{\pi}(U, \beta(U)) > \tilde{\pi}(\beta(U), \beta(U)).$$

So when the fraction of type 0 is small and everyone except type 0 plays $\beta(U)$ then type 0 earns more than all other types. In general we can formulate the following sufficient condition for a set of states with $x_0 > 0$ to be asymptotically stable, which holds for all games, not only for dominance solvable games.

Proposition 5 *Consider a finite symmetric two-player normal form game. Denote*

$$\begin{aligned} \tilde{\pi}(U, \beta(U)) - \tilde{\pi}(\beta(U), \beta(U)) &= A, \\ \tilde{\pi}(\beta(U), U) - \tilde{\pi}(U, U) &= B. \end{aligned}$$

Suppose that the best reply to U is strict. There is some $a \in (0,1)$ such that if $A/(A+B) > a$, then the set of states where $x_0 = A/(A+B)$ is an asymptotically stable set.

There are also cognitive games with asymptotically stable sets containing a positive fraction of type 1. One example can be constructed by adding a strictly dominated strategy to the Hawk-Dove game in the following way:

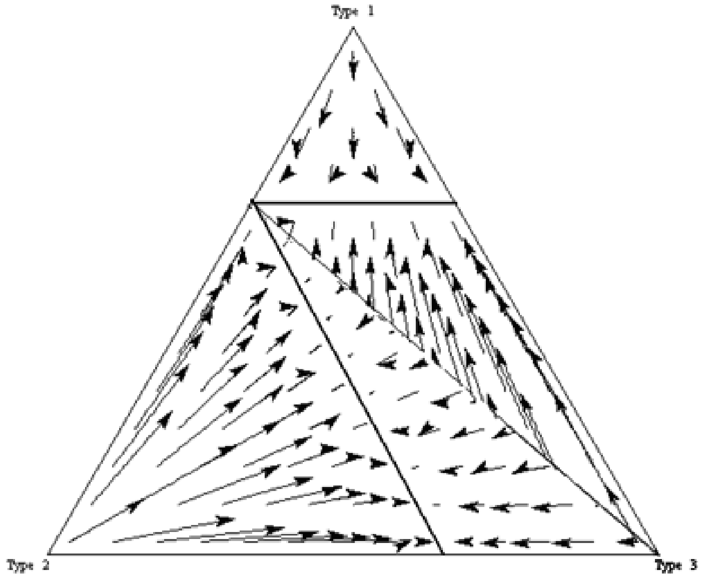
Example 9 *The game G^{HD+} is defined by the following payoff matrix, with $a > b$. Strategy A is strictly dominated by both strategy B and C. After deletion of strategy A, the remaining strategies constitute a Hawk-Dove like game where the ESS puts weight $b/(a+b)$ on strategy B and weight $a/(a+b)$ on strategy C.*

$$\begin{array}{cccc} & A & B & C \\ A & -1 & -a-1 & -b-1 \\ B & 0 & -a & 0 \\ C & 0 & 0 & -b \end{array}.$$

In the game G^{HD+} we have the following result:

Proposition 6 ***(a)*** *For any κ, evolution from any interior initial state converges to the set X^{ESS}.* ***(b)*** *If $\kappa = 3$ then $X^{ESS} = \{x_1 = a/(a+b)\} \cup \{x_2 = b/(a+b)\}$. Evolution from any interior initial state where $x_2 > bx_1/a$ converges to a state where $x_2 = b/(a+b)$, and evolution from any interior initial state where $x_2 \leq bx_1/a$ converges to a state where $x_1 = a/(a+b)$.*

Part (b) of this proposition is illustrated in the figure below (generated by letting $a = 2$ and $b = 1$). The thick diagonal line represents the set of states where $x_2 = 1/3$, and the thick horizontal line represents the set of states where $x_1 = 2/3$. The thin (diagonal) line represented the states where $x_2 = x_1/2$. Below this line type 3 plays strategy B, and evolution from any state in this region leads to some state where $x_2 = 1/3$. Starting from above this line, where type 3 plays strategy C, evolution leads to some state where $x_1 = 2/3$.

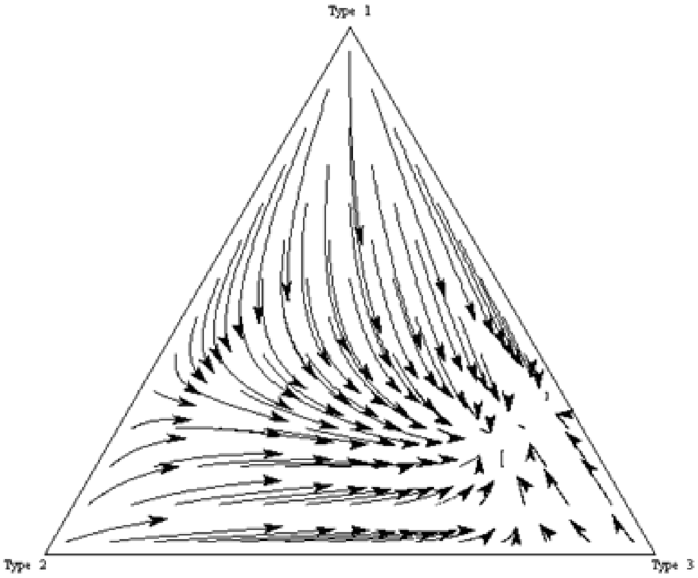


3.2 Observed Types

In coordination games things do not change at all when observability is introduced. In games with a unique interior ESS the analysis is changed somewhat but it is still the case that the state where everyone is of the highest type, is unstable. In Machiavellian games things change when types are observed – it need no longer be the case that higher types earn more than lower types.

Proposition 7 *Consider the cognitive game based on the Travelers' Dilemma, with observable types.* ***(a)*** *Suppose $2R \in \mathbb{N}$. If $\kappa < b - a - 3R + 1$ and $1 + a + 2R \leq b$ then type 0 is extinct and every state where $x_k = 0$ for some $k \in K \backslash \{0\}$, is unstable*[11]*.* ***(b)*** *Suppose $\kappa = 3$, and $\kappa < b - a - 3R + 1$. Evolution from any interior initial state converges to a uniqe interior state.*

The figure below illustrates part (b) of the above proposition – using $\kappa = 3$, $R = 3/2$, $a = 0$, and $b = 7$, like the figure illustrating evolution in the cognitive game based on the Travelers' Dilemma without observability. Evolution from any interior initial condition converges to the unique asymptotically stable state $x = (13/55, 7/55, 7/11)$.



The intuition for this result is that in this game the lower, more Machiavellian, strategies are more destructive so that higher types, when meeting each other, earn less than what lower types earn when meeting the higher types. Another way to put it is to say

[11] In fact when we restrict attention to $K/\{0\}$ the cognitive game is an anti-coordination game in the sense of Kojima and Takahashi (2007).

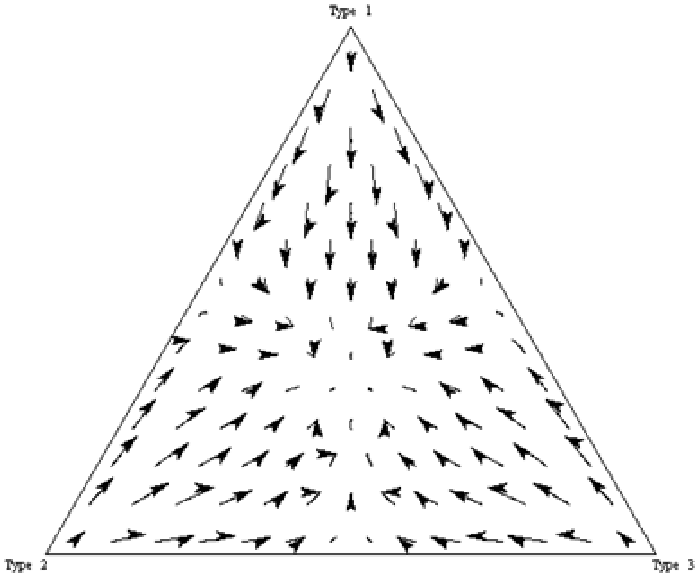
that lower types have a commitment advantage relative to higher types; a lower type may be committed to less destructive strategy, and may thereby induce higher types to take a less destructive strategy, something that might benefit both types. When there is a high fraction of the high type this favors the growth of the low type.It should be noted that proposition 16 demonstrates a possible mechanism for the evolution of cooperative behavior. The crucial element in this mechanism is the partial observability of types. It is well known that under complete information about preferences there might be a commitment advantage of not having preferences exclusively for what conveys fitness. Evolution might then lead to states where people have preferences that induce behavior that is not a Nash equilibrium (e.g. Güth and Yaari (1992), Dekel et al. (2007)). A similar advantage can accrue to bounded rationality (e.g. Banerjee and Weibull (1995)), as long as this boundedness of rationality is observed by other players. However, as pointed out by e.g. Samuelson (2001*b*) it is important that assumptions of observability are convincingly motivated and not ad hoc. Otherwise one can create evolutionary models for the survival of almost any deviation from the standard paradigm of rational maximization of preferences that reflect fitness. Since a theory of mind is precisely about how good one is at understanding how other people think, employing a sophisticated theory of mind means that one is generally good at detecting how others form their beliefs, i.e. what their theory of mind is. Hence it is very reasonable to assume that those who use a sophisticated theory of mind are able to observe the theory of mind of those who use a less sophisticated theory of mind. Thus in the context of the model of this paper the assumption of observability of types is not ad hoc but follows naturally from first principles. Consequently the result on cooperation in the above proposition does not rely on ad hoc assumption of observability.

The introduction of observability also affects the results in other classes of games. In the class of 2 × 2-games with a unique interior ESS we have the following result:

Proposition 8 *Consider the cognitive game based on a 2×2-game with a unique interior ESS, with observable types.* ***(a)*** *If $\kappa > 2a/(a+b)$ then evolution from any interior initial state converges to the state where $x_0 = 0$ and $x_k = 1/\kappa$ for all $k > 0$.* ***(b)*** *If $\kappa \leq 2a/(a+b)$ then every state where $x_k = 0$ for some $k \in K$, is unstable.*

Part (a) of the above proposition is illustrated below, for the case of $\kappa = 3$. Evolution from any interior initial condition converges to the unique asymptotically stable state

$x = (1/3, 1/3, 1/3)$.



4 Extensions to Other Types

4.1 The Level-K Model

As mentioned above, according to the level-k model, type 0 plays the uniform randomization over the strategy space and type 1 best responds to this, like in the CH model. An individual of type $k > 0$ believes that everyone else is of type $k - 1$ and plays a best response to what type $k - 1$ does. The framework developed above can be used for the level-k model like we have used it for the CH model so far. The model becomes much simpler, since the beliefs and behavior of different types in the level-k model do not depend on the distribution of types in the population. It is straightforward to go through the proofs in the appendix, and verify that all the propositions stated above carry over to the level-k model with the only exception being proposition 4. Some of the propositions could be somewhat strenghtened but I will not provide such an analysis here.

4.2 A Nash equilibrium type

The CH (and level-k) model constitutes an alternative to Nash equilibrium as a prediction of behavior in situations where individuals lack experience with sufficiently similar games, or when individuals receive insufficient feedback to allow for adaptation and equilibration. Taken together the results presented above can explain why lower types are not driven to extinction and why individuals of different types may co-exist. As a result of the co-existence of several types, some of which are quite unsophisticated behavior will generally not correspond to a Nash equilibrium.

In this section I study a Nash equilibrium (NE) type that is preprogrammed to play a Nash equilibrium strategy. In the case of multiple equilibria I will assume that the NE type plays a symmetric equilibrium. Among many such equilibria I assume that the NE type plays chooses one that corresponds to an ESS. If there are many such equilibria (as is the case e.g. in coordination games) I assume that the NE type chooses the ESS that is the best reply against the uniform distribution over the strategy space.

First consider the case when *types are not observed*: The question arises as to if and how the CH-types take the existence of the NE type into account. One could reason that the NE type is on the same level of sophistication as type 0, since both play a strategy regardless of what others do. In line with this one could assume that all types except type 0 and type NE take the existence of type NE into account when forming their beliefs about the population distribution of strategies. However, I will not make this assumption, but instead I will assume that the CH types form beliefs as described above, without taking the NE type into account. I do this for three reasons: First, letting the CH types adjust their behavior to what the NE type does seems to bias the model in favor of the CH types. Second, since the CH and level-k models do not include the NE type, any assumption about how these types form beliefs about the NE type would be arbitrary. Third, from an evolutionary point of view the question of whether a mutant NE type could take over a population of CH types should be well approximated by the case when the CH types do not know about the NE type.

In coordination games the NE type plays $\beta(U)$, like type $k > 0$, and thus the fraction $x_{NE}/(1 - x_0 - x_{NE})$ will stay constant. In games with a unique interior ESS the NE type plays the ESS so may states, with $x_{NE} > 0$, including the state with $x_{NE} = 1$ belong to the set X^{ESS}. However, in Machiavellian games the results are more interesting, as described by the following proposition:

Proposition 9 *Consider the cognitive game based on an underlying Machiavellian game with unobserved types $K \cup \{NE\}$. Let $\beta^{\kappa}(U)$ be the κ times iterated best against the uniform distribution.* ***(a)*** *If $\pi(\beta^{\kappa}(U), \beta^{\kappa}(U)) < \pi(s^{NE}, \beta^{\kappa}(U))$ then the unique asymptotically stable state is the state where $x_{NE} = 1$.* ***(b)*** *If $\pi(\beta^{\kappa}(U), \beta^{\kappa}(U)) > \pi(s^{NE}, \beta^{\kappa}(U))$ then both the state where $x_{NE} = 1$ and the state where $x_{\kappa} = 1$ are asymptotically stable.*

Recall that $\tilde{k}$ was defined to be the minimum number of iterations against the uniform distribution needed to reach the Nash equilibrium strategy. Note that if $\kappa \geq \tilde{k} - 1$ then it always holds that $\pi(\beta^{\kappa}(U), \beta^{\kappa}(U)) < \pi(s^{NE}, \beta^{\kappa}(U))$. Equivalently, a necessary condition for $\pi(\beta^{\kappa}(U), \beta^{\kappa}(U)) > \pi(s^{NE}, \beta^{\kappa}(U))$ to hold is that $\kappa < \tilde{k} - 1$. This if κ is large enough then case (a) applies but if the highest type currently in the population is not high enough then case (b) applies and evolution might lead to a a state where type

NE is extinct. It is relevant to consider the case of a low κ, since it is plausible to assume that a NE type mutant will not emerge until relatively sophisticated other types already are abundant.

Now consider the case when *types are observed*: Since type NE, like type 0, is pre-programmed to a strategy, I will assume that type $k > 0$ best respond to it by playing the Nash equilibrium strategy too, while type 0 plays U against type NE. Under these assumptions it follows straightforwardly from the proof of proposition 7 that in a cognitive game based on the Travelers' Dilemma with observed types, evolution from any interior initial condition leads to states where $x_{NE} = 0$.

Proposition 10 *Consider the cognitive game based on the Travelers' Dilemma, with observable types $K \cup \{NE\}$. If $\kappa < b - a - 3R + 1$ and $1 + a + 2R \leq b$ then types 0 and NE are extinct and every state where $x_k = 0$ for some $k \in K \backslash \{0\}$, is unstable.*

In the class of 2×2-games with a unique interior ESS we also have the result that type NE is extinct:

Proposition 11 *Consider the cognitive game based on a 2×2-game with a unique interior ESS, with observable types $K \cup \{NE\}$.* ***(a)*** *If $\kappa > 2a/(a+b)$ then evolution from any interior initial state converges to the state where $x_0 = 0$, $x_{NE} = 0$, and $x_k = 1/\kappa$ for all $k > 0$.* ***(b)*** *If $\kappa \leq 2a/(a+b)$ then evolution from any interior initial state leads to states where $x_{NE} = 0$, and every state where $x_k = 0$ for some $k \in K$, is unstable.*

4.3 A Rational Expectations (RE) Type

4.3.1 Introducing the RE Type

In the CH model players believe that they are the only one of their type and that everyone else is less sophisticated. Hence they reason iteratively, in a finite number of steps, about what other individuals will do, and then best respond given their beliefs. In contrast, traditional game theory assumes that players are aware of the fact that there are other individuals that form beliefs in the same way as they do. They behave as if they solved a fix-point problem, and play a best reply to the population distribution of strategies. In this section such a *rational expectations (RE)* type is introduced into the CH model.[12] This is done by changing the beliefs and behavior of type κ only. It is now assumed that type κ has a correct belief about the composition and behavior of the population and plays a best response.[13] Formally type κ-individuals have a correct belief about the

[12] This is the term used e.g. by Stahl and Wilson (1995). Other authors, such as Costa-Gomes and Crawford (2006), refer to this type as 'sophisticated'.

[13] One could have modified the CH model by letting all individuals of all types $k > 0$ be aware of the existence of other individuals of the same type. However, since evolution works piece-wise and not in leaps we should expect the ultra-sophisticated to develop out of a high type, such as type κ, rather than from a lower type type.

population state, i.e. $\hat{x}^\kappa = x$. (As before, superscript refers to the type that has the belief and subscript refers to what type the belief is about.). They have a true belief about the play of lower types, and a belief $\hat{z}_k^\kappa$ about their own type, so

$$\hat{z}^\kappa = \left(z_0, z_1, ..., \hat{z}_k^\kappa\right).$$

Consequently they believe that the aggregate behavior is

$$\hat{q}^\kappa = \sum_{i=0}^{\kappa-1} z_i x_i + \hat{z}_\kappa^\kappa x_\kappa.$$

When types are not observed they best respond to this belief, i.e. choose an element in

$$\tilde{\beta}\left(\hat{q}^\kappa\right) = \arg\max_{\sigma\in\Delta(S)} \tilde{\pi}\left(\sigma, \hat{q}^\kappa\right).$$

The behavior of type κ individuals satisfies the following fixpoint equation

$$\hat{z}_\kappa^\kappa \in \tilde{\beta}\left(\sum_{i=0}^{\kappa-1} z_i x_i + \hat{z}_\kappa^\kappa x_\kappa\right).$$

The set of solutions to this equation is

$$Z^\kappa = \left\{\sigma \in \Delta\left(S\right) : \sigma \in \tilde{\beta}\left(\sum_{i=0}^{\kappa-1} z_i x_i + \sigma x_\kappa\right)\right\},$$

and we have the following result.

Lemma 2 *Z^κ is non-empty.*[14]

It should be noted that if one sets $\kappa = 1$ and $x_0 = 0$ then everyone has plays a best response to the actual distribution of strategies so that one Nash equilibrium is played by the whole population.

Now consider the case of *observed types.* When an RE individual faces an opponent of a lower type then the RE individuals knows the opponents' type and knows how that type will behave. If two *RE* individuals meet then they play a Nash equilibrium.

[14] The set Z^κ need not be a singleton but in the games considered in this paper this will not cause any problems in the analysis.

4.3.2 Evolution with RE Type

When *types are unobserved* we have the following result, regardless of whether Z^{κ} is a singleton or not:

Proposition 12 *In the CH model with unobserved types and where κ is an RE type: For any underlying game, evolution in the cognitive game converges, from any interior initial state, to a state where aggregate behavior corresponds to a Nash equilibrium of the underlying game.*

This result implies that all types are wiped out that do not play a strategy that is in the support of a Nash equilibrium strategy. So unless the game has a completely mixed Nash equilibrium the fraction of type zero individuals goes to zero. Moreover, if the strategy $\beta(U)$ is not in the support of a Nash equilibrium the fraction of type 1 also goes to zero. This implies that in the game G^{MDS1} no state with $x_0 > 0$ or $x_1 > 0$ is stable. (In contrast it was found above, in the CH model without the *RE* type, that the asymptotically stable set of states have $x_0 > 0$, and that all but one state in this set has $x_1 > 0$.) In the games with a unique interior ESS the asymptotically stable set of states X^{ESS} has expanded and now includes the state where everyone is of the highest type, now the RE type. In Machiavellian games we still have the result that only the types that are high enough to play the Nash equilibrium survive. In conclusion the addition of the RE type to the CH framework seems to worsen the survival chances of lower types.

However, moving to *observable types* we can find the same kind of commitment advantage of lower types as we found in the absence of the *RE* type. For example it is clear that one can prove an analog of proposition 7, according to which every state where $x_k = 0$ for some $k \in K \backslash \{0\}$, is unstable. In addition to this there are undoubtedly costs of increased sophistication with respect to theory of mind abilities (Holloway (1996), Dunbar (2003), and Roth and Dicke (2005)). More reasoning power requires a larger brain and brain tissue is metabolically very costly.[15] In this paper cognition costs were not included in the formal analysis since any assumption on the cost function would seem arbitrary. Still such costs must be taken into account when interpreting the results. Costs of cognition might inhibit the evolution of the hyper-sophisticated *RE* type in two ways: First, given that an *RE* mutation emerges, the cognitive costs of this creature may bee too high to be compensated by the increased payoff resulting from improved decisions (solving fixpoint problems is difficult). Second, such a mutation might never emerge. The reason is that evolution occurs in small steps and not in big leaps. An *RE* mutant should therefore not be expected to arise before there already exists sufficiently sophisticated types who do not solve fixpoint problems. The benefits of increased sophistication, below the *RE* type, may not outweigh the costs.

[15] The brain stands for approximately two percent of the body's weight but utilizes about twenty percent of the total body metabolism at rest; Holloway (1996).

5 Conclusion

This paper makes two main novel contributions. (1) It is the first paper to perform an evolutionary analysis of the empirically successful, and widely applied, cognitive hierarchy (CH) model. (2) It extends the CH model to the case of partially observed types, and performs an evolutionary analysis of these types too.

The purpose of this paper was to provide foundations for the existence of bounded and heterogeneous theory of mind abilities. For the case of unobserved cognitive type the occurrence of cognitive arms races was verified in the class of 'Machiavellian games'. But there are also games where higher types do not have an advantage, even when types are not observed. In 2×2-games with a unique interior ESS the unique asymptotically stable set of states includes states where all types coexist, and does not include any monomorphic states. Sufficient conditions were also given for when a positive fraction of type 0 are asymptotically stable, even in dominance solvable games. The survival prospects for unsophisticated individuals are increased further when we move from the incomplete information scenario and allow higher types to observe lower types. In this case it was found that evolution in a cognitive game based on an underlying 2×2-game with a unique interior ESS always leads to a unique asymptotically stable state where all types, except type 0, co-exist. There are even Machiavellian games where evolution leads to an asymptotically stable state where all types (except type 0) co-exist. It was noted that this amounts to a mechanism for the evolution of cooperative behavior. Taken together, an evolutionary process based on payoffs earned in these different games may plausibly lead to a polymorphic population where most individuals belong to relatively low types.

6 Appendix A: Proofs

6.1 Preliminaries

Proof of Lemma 1. The expected payoff against the uniform randomization over a set $\{a, a+1, ..., b\}$ is

$$\begin{aligned}
&E\left[\pi\left(s_i, \sigma_j\right) | \sigma_j \sim U\left(\{a, a+1, ..., b\}\right)\right] \\
&= \frac{1}{b-a+1}\left(s_i + \sum_{s_j=a}^{s_i-1}\left(s_j - R\right) + \sum_{s_j=s_i+1}^{b}\left(s_i + R\right)\right) \\
&= \frac{1}{b-a+1}\left(s_i + \sum_{s_j=a}^{s_i-1} s_j - \left(s_i - a\right)R + \left(b - s_i\right)\left(s_i + R\right)\right) \\
&= \frac{1}{b-a+1}\left(\sum_{s_j=a}^{s_i-1} s_j + \left(b - 2R + 1\right)s_i - s_i^2 + R\left(a+b\right)\right).
\end{aligned}$$

The increase from s_i from s to $s+1$ results in a change of payoff by

$$\begin{aligned}
&\frac{1}{b-a+1}\left(\sum_{s_j=a}^{s} s_j - \sum_{s_j=a}^{s-1} s_j + \left(b - 2R + 1\right) - \left(s+1\right)^2 + s^2\right) \\
&= \frac{1}{b-a+1}\left(s + \left(b - 2R + 1\right) - 2s - 1\right) \\
&= \frac{1}{b-a+1}\left(b - 2R - s\right).
\end{aligned}$$

This is positive if and only if $b - 2R > s$, so if $2R$ is an integer then the best reply is $s = b - 2R$. Generally the best reply is $s = \max\{t \in S : b - 2R \geq t\}$. ■

6.2 The CH model

6.2.1 Unobserved Types

Proof of Proposition 1. Since all types $k \geq 1$ play the same strategy they earn the same in all states, and they earn more than type 0. ■

Proof of Proposition 2. By the assumption about generic payoffs we have $\sigma^{ESS} \neq U$. Note the following property of 2×2-game with a unique interior ESS: If $z_i\left(x\right) > \sigma_i^{ESS}$ then strategy $i \in \{H, D\}$ earns more than strategy $j \neq i$.

(i) First we show that a state $x \in \Delta(K)$ is a rest point of the replicator dynamics if and only if it is monomorphic or belongs to X^{ESS}:

(i.i) To see that all points in X^{ESS} are rest points, note that all strategies in the support of a Nash equilibrium earn the same payoff against the Nash equilibrium strategy. Since the unique ESS is interior, all strategies earn the same against σ^{ESS}. Hence if $x \in X^{ESS}$ then all types earn the same, so x is a rest point. Furthermore, it is trivial that monomorphic states are rest points.

(i.ii) To see the converse, that all rest points are either monomorphic or in X^{ESS}, consider a polymorphic state x where $z(x) \neq \sigma^{ESS}$. Behavior cannot correspond to any of the two asymmetric Nash equilibria so $z(x) \neq \sigma^{ESS}$ implies $z(x) \neq \sigma^{NE}$, which means that one strategy earns more than the other. If the lowest type mixes then the second lowest type plays a pure strategy – since we have assumed $\sigma^{ESS} \neq U$. These two types earn different payoffs so the state is unstable.

(ii) Now we show that no monomorphic states are stable.

(ii.i) To see that a state with $x_k = 1$, $k \in \{1, ..., \kappa - 1\}$, is unstable, note that there is some $\varepsilon > 0$ such that in any state with $x_k \in (1-\varepsilon, 1)$, type k plays $\bar{\beta}\left(\hat{q}^k(x)\right)$ and all types $k' > k$ play $\bar{\beta}\left(\bar{\beta}\left(\hat{q}^k(x)\right)\right)$. When $\varepsilon \to 0$ we have that type $k' > k$ earns

$$\Pi_{k'>k} \to \tilde{\pi}\left(\bar{\beta}\left(\bar{\beta}\left(\hat{q}^k(x)\right)\right), \bar{\beta}\left(\hat{q}^k(x)\right)\right),$$

whereas type k earns

$$\Pi_k \to \tilde{\pi}\left(\bar{\beta}\left(\hat{q}^k(x)\right), \bar{\beta}\left(\hat{q}^k(x)\right)\right).$$

Since $\sigma^{ESS} \neq U$ we have

$$\tilde{\pi}\left(\bar{\beta}\left(\bar{\beta}\left(\hat{q}^k(x)\right)\right), \bar{\beta}\left(\hat{q}^k(x)\right)\right) > \tilde{\pi}\left(\bar{\beta}\left(\hat{q}^k(x)\right), \bar{\beta}\left(\hat{q}^k(x)\right)\right).$$

Thus there is some $\delta < \varepsilon$ such that if $x_{k'} \in (0, \delta)$, then $\Pi_{k'>k} > \Pi_k$.

(ii.ii) A similar argument shows that the state with $x_0 = 1$ is unstable.

(ii.iii) To see that the state with $x_\kappa = 1$ is unstable let $\delta = \min_{i \in \{H,D\}} \sigma_i^{ESS}$. Since σ^{ESS} is interior we have $\delta > 0$. Assume, without loss of generality, that $\beta(U) = D$. If $x_\kappa = 1 - \varepsilon$, and $x_0 = \varepsilon$, then type κ plays $\beta(U) = D$. In this state $z_H(x) = 1 - \varepsilon/2$ and $z_D(x) = \varepsilon/2$. If $\varepsilon < \delta$ then $z_H(x) = \varepsilon/2 < \varepsilon < \delta = \min_{i \in \{H,D\}} \sigma_i^{ESS} \leq \sigma_H^{ESS}$. Thus H earns a higher payoff than D against $z(x)$. It follows that type 0 earns more than type κ in all states where $x_0 < \delta$.

(iii) Now we show that X^{ESS} is the unique asymptotically stable set, with the whole interior as its basin of attraction. Suppose $K = \{0, 1, 2, .., \kappa\}$. If the system starts in X^{ESS} then the system remains in this set, so assume $x^0 \notin X^{ESS}$. Let X^I denote the set of states where one or more types $k \geq 1$ are indifferent between the strategies;

$$X^I = \left\{x \in \Delta(K) : \exists k \text{ s.t. } \hat{z}^k(x) = \sigma^{ESS}\right\}.$$

The set X^I is closed, since a type is indifferent betwen strategies only when they yild the exact same expected payoff. Let $int(X)$ denote the interior of X.

(iii.i) Suppose that $x^0 \in int\,(X)$, but $x^0 \notin X^I$. Since X^I is closed, there is a nbd B of x^0 such that in every state $x \in B$ all types use the same strategy as in x^0. Since $x^0 \notin X^{ESS}$ one strategy i is overweighted relative to its weight in the ESS, i.e. $z_i\left(x^0\right) > \sigma_i^{ESS}$. This implies that strategy i earns less than strategy $j \neq i$. Thus the fractions of the types that play strategy i decrease as the system moves away from x^0. A type k that initially plays strategy i does so because it mistakenly believes that strategy i is underweighted relative to to its weight in the ESS, i.e. $\hat{z}_i^k\left(x^0\right) < \sigma_i^{ESS}$. As the fractions of all types playing strategy i decrease, it continues to hold that $\hat{z}_i^k\left(x\right) < \sigma_i^{ESS}$, so no type that plays i switches to j. There may be some types that start out by playing j which eventually come to believe that strategy i is underrepresented relative to the ESS (since the fraction that plays i decreases). Thus either (1) the fraction of types playing pure strategy i decreases until a state in X^{ESS} is reached, or (2) the fraction of types playing pure strategy i goes to zero, and the fraction of type 0 (which put probability 1/2 on strategy i) decreases until a state in X^{ESS} is reached.

(iii.ii) Suppose that $x^0 \in int\,(X) \cap X^I$. Since $x^0 \notin X^{ESS}$ and $x^0 \in int\,(X)$ (i) implies that evolution will lead away from x^0, and that the fractions of all types will change (inclucing type 0). This will change the beliefs of all types $k \geq 2$, so that they are no longer indifferent between strategies. Thus the system moves away from X^I, and the rest follows from (iii.i). ■

Proof of Proposition 3. Suppose type k plays strategy s. The only source of difference between the beliefs of type k and type $k+1$ is that type k is aware of the existence of type k. Because of the single peak property type $k+1$ will either think that s is a best response to the population or that some $t > s$, is a best response. Hence type $k+1$ will play a weakly higher strategy than type k. Type κ will always play a weakly higher strategy than all types $k < \kappa$.

If $x_k/x_{k+1} \to 0$ for all k, then type 1 plays $\beta\left(U\right)$, type 2 plays $\beta\left(\beta\left(U\right)\right)$, and so on. (Here I use $\beta\left(U\right)$ as short hand for the mixed strategy $\sigma^{\beta(U)}$ that puts all probability on the pure strategy $\beta\left(U\right)$.) Type k plays the k times iterated best response to the uniform distribution. Clearly this is the highest strategy that k will play in any state $x \in X$. Thus type k never plays the Nash equilibrium strategy (i.e. the highest strategy $|S|$) unless $k \geq \tilde{k}$.

Suppose that s is the best response according to the beliefs of type κ. If x_κ is small enough then s is indeed the best response. If x_κ is not small enough then the best response will be $t > s$. No type will play that strategy. Because of the single peak property strategy s will be among the best strategies used in the population. Thus no type plays a better response to the population than type κ does.

Not everyone will play the same strategy since at least type 0 will play U. Hence, type κ will earn above the average and increase in fraction. If k and κ play different strategies then type κ will grow at a higher rate than type k. If k and κ play the same strategy $s < |S|$ then both types will grow at the same rate, and no other type will grow faster.

Eventually when x_k is large enough type κ will change to playing a higher strategy $t > s$. Then κ grows at a higher rate than k. ■

Proof of Proposition 4. Type 0 randomizes. Type 1 plays B always. Type 2 plays B if $x_0 > 3x_1$ and type 3 plays B if $x_0 > 3(x_1 + x_2)$. In general type $k \geq 2$ plays B if $x_0 > 3\sum_{i=1}^{k-1} x_i$.

(a) Consider the points where $x_0, x_1 > 0$. In any such state there is a $k^* \in \{1, ..., \kappa\}$ such that $3\sum_{i=1}^{k^*} x_i > x_0 \geq 3\sum_{i=1}^{k^*-1} x_i$.

Case I: Suppose first that $3\sum_{i=1}^{k^*} x_i > x_0 > 3\sum_{i=1}^{k^*-1} x_i$. In this case everyone up to, and including, type k^* plays B (except type 0), and everyone above type k^* plays C. Payoff of type 0 is

$$\begin{aligned}\Pi_0 &= \frac{1}{3}\left(x_0 + 2\left(\frac{1}{3}x_0 + \sum_{i=1}^{k^*} x_i\right)\right) + \frac{1}{3}\left(\frac{8}{3}x_0\right) \\ &+ \frac{1}{3}\left(x_0 + 3\left(\frac{1}{3}x_0 + \sum_{i=1}^{k^*} x_i\right) + \left(\frac{1}{3}x_0 + 1 - \sum_{i=1}^{k^*} x_i - x_0\right)\right) \\ &= \frac{1}{3}\left(1 + \frac{17}{3}x_0 + 4\sum_{i=1}^{k^*} x_i\right)\end{aligned}$$

Payoff of type $k \in \{1, .., k^*\}$ is $\Pi_{k\in\{1,..,k^*\}} = 8x_0/3$. Payoff of type $k \in \{k^* + 1, ...\kappa\}$ is

$$\Pi_{k\in\{k^*+1,...\kappa\}} = x_0 + 3\left(\frac{1}{3}x_0 + \sum_{i=1}^{k^*} x_i\right) + \left(\frac{1}{3}x_0 + 1 - \sum_{i=1}^{k^*} x_i - x_0\right) = \frac{4}{3}x_0 + 2\sum_{i=1}^{k^*} x_i + 1.$$

I.I Suppose $x_0 > 0$, $\sum_{i=1}^{k^*} x_i > 0$, and $\sum_{i=k^*+1}^{\kappa} x_i > 0$ (i.e. some randomize uniformly, some play B, and some play C). In any stable such state we have $\Pi_0 = \Pi_{k\in\{1,..,k^*\}} = \Pi_{k\in\{k^*+1,...\kappa\}}$. This requires $\Pi_0 = \Pi_{k\in\{1,..,k^*\}}$ and $\Pi_{k\in\{1,..,k^*\}} = \Pi_{k\in\{k^*+1,...\kappa\}}$. Putting this together yields $\sum_{i=1}^{k^*} x_i = 3/2$, which is impossible. So no state with $x_0 > 0$, $\sum_{i=1}^{k^*} x_i > 0$, and $\sum_{i=k^*+1}^{\kappa} x_i > 0$ is stable.

I.II Suppose $x_0 > 0$, $\sum_{i=1}^{k^*} x_i > 0$, and $\sum_{i=k^*+1}^{\kappa} x_i = 0$ (i.e. some randomize uniformly, some play B, and no one plays C). Then $\sum_{i=1}^{k^*} x_i = 1 - x_0$ and we only need to have $\Pi_0 = \Pi_{k\in\{1,..,k^*\}}$, which reduces to $x_0 = 15/19$. Hence states with $x_0 = 15/19$ are Lyapunov stable. Also note that $x_0 > 15/19$ if and only if $\pi^0 < \pi^{k\in\{1,..,k^*\}}$, so neither $x_0 = 1$ nor $x_1 = 1$ is asymptotically stable.

I.III Suppose $x_0 > 0$, $\sum_{i=1}^{k^*} x_i = 0$, and $\sum_{i=k^*+1}^{\kappa} x_i > 0$ (i.e. some randomize uniformly, no one plays B, and some play C), implying $x_1 = 0$ and $3\sum_{i=1}^{k^*} x_i > x_0$. Then we only need to have $\Pi_0 = \Pi_{k\in\{k^*+1,...\kappa\}}$, but

$$\Pi_0 - \Pi_{k\in\{k^*+1,...\kappa\}} = \frac{5}{9}x_0 - \frac{2}{3} < 0,$$

so no point with $x_0 > 0$, $\sum_{i=1}^{k^*} x_i = 0$, and $\sum_{i=k^*+1}^{\kappa} x_i > 0$ is stable.

I.IV If $x_0 > 0$ and $x_1 \geq 0$ then we are in the same situation as when $x_0 > 0$, $\sum_{i=1}^{k^*} x_i > 0$, and $\sum_{i=k^*+1}^{\kappa} x_i \geq 0$.

I.V If $x_0 = 0$ and $x_1 > 0$ then type 1 will play B and everyone else will play C. In this case

$$\Pi_{k\in\{2,\dots\kappa\}} - \Pi_2 = 2\sum_{i=1}^{k^*} x_i + 1 > 0,$$

so no point where $x_1 > 0$, and $\sum_{i=2}^{\kappa} x_i > 0$ is stable.

Case II: Now suppose $3\sum_{i=1}^{k^*} x_i > x_0 = 3\sum_{i=1}^{k^*-1} x_i$. Then everyone up to type $k^* - 1$ plays B (except type 0), and everyone above type k^* plays C. Type k^* randomizes between B and C. In any such state we must have $x_0 > 0$. Payoff of type 0 is (using $\sum_{i=1}^{k^*-1} x_i = x_0/3$)

$$\begin{aligned}\Pi_0 &= \frac{1}{3}\left(x_0 + 2\left(\frac{1}{3}x_0 + \frac{1}{3}x_0 + \frac{1}{2}x_{k^*}\right)\right) + \frac{1}{3}\left(\frac{8}{3}x_0\right)\\ &+ \frac{1}{3}\left(x_0 + 3\left(\frac{1}{3}x_0 + \frac{1}{3}x_0 + \frac{1}{2}x_{k^*}\right) + \left(\frac{1}{3}x_0 + 1 - \frac{1}{3}x_0 - \frac{1}{2}x_{k^*} - x_0\right)\right)\\ &= \frac{7}{3}x_0 + \frac{2}{3}x_{k^*} + \frac{1}{3}\end{aligned}$$

Payoff of type k^* is

$$\begin{aligned}\Pi_{k^*} &= \frac{1}{2}\left(\frac{8}{3}x_0 + x_0 + 3\left(\frac{1}{3}x_0 + \frac{1}{3}x_0 + \frac{1}{2}x_{k^*}\right) + \left(\frac{1}{3}x_0 + 1 - \frac{1}{3}x_0 - \frac{1}{2}x_{k^*} - x_0\right)\right)\\ &= \frac{7}{3}x_0 + \frac{1}{2}x_{k^*} + \frac{1}{2}.\end{aligned}$$

In any stable state we have $\Pi_0 = \Pi_{k^*}$ or

$$\frac{7}{3}x_0 + \frac{2}{3}x_{k^*} + \frac{1}{3} = \frac{7}{3}x_0 + \frac{1}{2}x_{k^*} + \frac{1}{2},$$

which reduces to $x_{k^*} = 1$. This contradicts our assumption that $x_0 > 1$, so no state with $x_0 = 3\sum_{i=1}^{k^*-1} x_i$ is stable.

Case III: Recall that if $x_0 = x_1 = 0$ then the beliefs of any type $k \geq 2$ with $x_k > 0$ are undefined. However we can still say something about the movement of the system close to such a point: In any neighborhood of a state where $x_0 = ... = x_{k'-1} = 0$, and $x_{k'} > 0$ there are states where

$$x_0 = \varepsilon > x_1 = ... = x_{k'-1} = 0.$$

In such a state at least type k' plays B. If $x_0 = \varepsilon$ is small enough then the best reply to the population distribution of play is C.

Suppose $x_0 = \varepsilon > 0$ and $x_k = 0$ for all $k > 0$ such that $k \neq k'$. Then since $\pi\left(U, \sigma^B\right)$ we have that x_0 grows and $x_{k'}$ declines and the system moves away from the initial state. ($\sigma^B \in \Delta(S)$ denotes the strategy that puts all weight on pure strategy B.) This process continues until the payoff of these types is equal. Hence no monomorphic state with $x_0 = x_1 = 0$ is is attracting.

Suppose $x_0 = \varepsilon > 0$, $x_{k'}, x_{k''} > 0$ for some $k^{''} > k^{'}$, and $x_k = 0$ for

$$k \in \{1, ..., k^{'} - 1\} \cup \{k' + 1, ..., k'' - 1\}.$$

Then type k'' plays C. Since x_0 is small we have $\bar{\Pi} \in (\Pi_{k'}, \Pi_{k''})$ so $x_{k'}$ declines and the system moves away from the initial state. Hence no state with $x_0 = x_1 = 0$ is attracting

(b) We have shown above that the set of polymorphic stable states is identical to R. Furthermore we know that no monomorphic state is an attractor. Also no point in R is an attractor. In order to prove (b) we now proceed to show that R is an asymptotically stable set. Suppose $x_0 \in (3/4, 1]$. Then

$$3\sum_{i=1}^{\kappa} x_i = 3(1 - x_0) < 3/4 < x_0$$

so all types except type 0 play B. In this case $\sum_{i=1}^{k^*} x_i = 1 - x_0$, so payoff of type 0 is

$$\Pi_0 = \frac{1}{3}\left(1 + \frac{17}{3}x_0 + 4\sum_{i=1}^{k^*} x_i\right) = \frac{5}{9}x_0 + \frac{5}{3}.$$

Payoff of type 1 and 2 is $\Pi_1 = \Pi_2 = 8x_0/3$, so $\Pi_0 > \Pi_1 = \Pi_2$ if and only if

$$\frac{5}{9}x_0 + \frac{5}{3} - \frac{8}{3}x_0 = \frac{15}{9} - \frac{19}{9}x_0 > 0,$$

or equivalently $x_0 < 15/19$. So for $x_0 < 15/19$ we have $\Pi_0 > \Pi_1 = \Pi_2$ and for $x_0 > 15/19$ we have $\Pi_0 < \Pi_1 = \Pi_2$. And for $x_0 = 15/19$ we have $\Pi_0 = \Pi_1 = \Pi_2$. This shows that from any initial state satisfying $x_0 \in (3/4, 1]$ evolution leads to the set R. This concludes the proof of (b).

(c) We saw above that in any neighborhood of a monomorphic state $x_k = 1$, $k \neq 0$, there is a point with $x_0 = \varepsilon > 0$ from which evolution leads to a state where only type k and 0 exist and where $\Pi_k = \Pi_0$. Hence there are points where x_0 is arbitrarily small from which evolution leads to R. ■

Proof of Proposition 5. Note that $B \geq 0$ so $A/(A + B) > 0$ implies $A > 0$. Since the best reply to U is strict, there exists an $a \in (0, 1)$ such that if $x_0 \geq a$ then

$$\beta(U) = \arg\max_{\sigma \in \Delta(S)} \sigma \cdot A\left(\hat{q}^k(x)\right),$$

for all k. Thus if $x_0 \geq a$ then type $k \geq 1$ play $\beta(U)$ so that all types $k \geq 1$ earn the same payoff. We have

$$\begin{aligned}\Pi_0 - \Pi_{k>0} &= x_0 \tilde{\pi}(U,U) + (1-x_0)\tilde{\pi}(U,\beta(U)) - (x_0 \tilde{\pi}(\beta(U),U) + (1-x_0)\tilde{\pi}(\beta(U),\beta(U))) \\ &= x_0 (\tilde{\pi}(U,U) - \tilde{\pi}(\beta(U),U)) + (1-x_0)(\tilde{\pi}(U,\beta(U)) - \tilde{\pi}(\beta(U),\beta(U))) \\ &= -Bx_0 + (1-x_0)A.\end{aligned}$$

This is positive if and only if $A/(A+B) > x_0$ (implying $A > 0$). Hence if $A/(A+B) > x_0 > a$ then $\Pi_0 > \Pi_{k>0}$, and if $x_0 > A/(A+B) > a$ then $\Pi_0 < \Pi_{k>0}$. ■

Proof of Proposition 6. (a) All types $k \geq 1$ play strategy B or C. Since B and C strictly dominate A type 0 will always earn less than the other types and be extinct. After removing strategy A we can apply proposition 2.

(b) The ESS puts weight $b/(a+b)$ on strategy B and weight $a/(a+b)$ on strategy C. Let $\kappa = 3$. Type 1 plaus $\beta(U) = C$, so type 2 plays $\beta(\beta(U)) = B$. From the point of view of type 3, the expected payoff to strategies B and C are

$$\mathbb{E}\left[\Pi_B(x) \,|\hat{q}^3(x)\right] = \frac{1}{x_1+x_2}(x_1\pi(B,C) + x_2\pi(B,B)) = \frac{-ax_2}{x_1+x_2},$$

and

$$\mathbb{E}\left[\Pi_C(x) \,|\hat{q}^3(x)\right] = \frac{1}{x_1+x_2}(x_1\pi(C,C) + x_2\pi(C,B)) = \frac{-bx_1}{x_1+x_2}.$$

Thus 3 chooses B instead of C if $-ax_2 > -bx_1$, or equivalently $x_2 < bx_1/a$. Thus there are three different cases to consider.

(i) Suppose that $x_2 < bx_1/a$, so that type 3 plays B. Then only type 1 plays C, so the set $X^{ESS} \cap \{x \in \Delta(K) : x_2 < bx_1/a\}$ is constituted by states such that $x_1 = a/(a+b)$.

(ii) Suppose instead that $x_2 > bx_1/a$, so that type 3 plays C. Then only type 2 plays B, so the set $X^{ESS} \cap \{x \in \Delta(K) : x_2 > bx_1/a\}$ is constituted by states such that $x_2 = b/(a+b)$.

(iii) Suppose that $x_2 = bx_1/a$, so that type 3 randomizes uniformly between B and C. Then the set $X^{ESS} \cap \{x \in \Delta(K) : x_2 > bx_1/a\}$ is constituted by states such that $x_1 + x_3/2 = a/(a+b)$, or equivalently, using $x_2 = bx_1/a$ and $x_3 = 1 - x_1 - x_2$, we have $x_1 = a/(a+b)$. ■

6.2.2 Observed Types

Proof of Proposition 7. (a) (i) *Behavior*: Since $2R \in \mathbb{N}$ it holds that $\beta(U) = b - 2R$. Type 2 plays $\beta(U)$ if x_0/x_1 is sufficiently large and plays $s_2 = \beta(\beta(U)) = b - 2R - 1$ if x_0/x_1 is sufficiently small. Iterating on this, using $\beta(s) = s - 1$ for $s < a$, we get

$$\beta\left(\hat{q}^k(x)\right) \in \{b-2R, b-2R-1, ..., b-2R-(k-1)\},$$

and which strategy that is chosen from this set depends on the belief of type k. Since $\beta\left(U\right) = b - 2R$ the smallest number of iterated best responses to the uniform distribution that are required to reach the Nash equilibrium strategy a is $\tilde{k} = b - 2R - a + 1$. Thus type $k \geq \tilde{k}$ plays

$$\beta\left(\hat{q}^{k}\left(x\right)\right) \in \left\{b - 2R, b - 2R - 1, ..., a\right\}.$$

When type $k > 0$ encounters type $k' > k$, then k plays $\beta\left(\hat{q}^{k}\left(x\right)\right)$, and k' plays $\beta\left(\beta\left(\hat{q}^{k}\left(x\right)\right)\right) = \beta\left(\hat{q}^{k}\left(x\right)\right) - 1$, unless $\beta\left(\hat{q}^{k}\left(x\right)\right) = a$ in which case k' plays a. Since the only difference between the beliefs of type k and type $k-1$ lies in the fact that type k acknowledges the existence ot type $k-1$ we have $\beta\left(\hat{q}^{k-1}\left(x\right)\right) \geq \beta\left(\hat{q}^{k}\left(x\right)\right)$. When type k encounters type $k' < k$, with $k' > 0$, then type k plays $\beta\left(\hat{q}^{k'}\left(x\right)\right) - 1$, unless $\beta\left(\hat{q}^{k'}\left(x\right)\right) = a$ in which case type k plays a. Since type k acknowledges the existence of type $k-1$ we have that type k' plays a weakly lower strategy against k than against $k-1$.

(ii) *Payoffs*: Let $w\left(i, j, x\right)$ denote the payoff of type $i \in K$ when meeting type $j \in K$ in state $x \in \Delta\left(K\right)$. First, compare the payoffs of types $k > 1$ and $k - 1 > 0$ when they encounter type $k' > k$: Since we have $\beta\left(\hat{q}^{k-1}\left(x\right)\right) \geq \beta\left(\hat{q}^{k}\left(x\right)\right)$, and since type k' best responds to this, it holds that $w\left(k-1, k', x\right) \geq w\left(k, k', x\right)$ for all x. Second, since $k > k'$ and type $k+1 > k'$ use the same strategy against type k' if follows that $w\left(k, k', x\right) = w\left(k+1, k', x\right)$. Third, in order to have $w\left(k-1, k, x\right) > w\left(k, k, x\right)$ for all k and all x, we need to have $w\left(\kappa - 1, \kappa, x\right) > w\left(\kappa, \kappa, x\right) = a$ for all x. This requires that

$$\pi\left(\beta\left(\hat{q}^{\kappa-1}\left(x\right)\right), \beta\left(\beta\left(\hat{q}^{\kappa-1}\left(x\right)\right)\right)\right) = \beta\left(\hat{q}^{\kappa-1}\left(x\right)\right) - 1 - R > a.$$

Since $\beta\left(\hat{q}^{\kappa-1}\left(x\right)\right) \leq b - 2R - \left(\kappa - 2\right)$ for all x, we need to have $b - 2R - \left(\kappa - 2\right) - 1 - R > a$, or equivalently $\kappa < b - a - 3R + 1$. Fourth, suppose $k > k'$. Note

$$w\left(k, k', x\right) = \pi\left(\beta\left(\hat{q}^{k'}\left(x\right)\right) - 1, \beta\left(\hat{q}^{k'}\left(x\right)\right)\right) \geq b - 2R - \left(k' - 1\right) - 1 + R = b - R - k'.$$

In order to have $w\left(k, k', x\right) > w\left(k', k', x\right)$ we need $b - R - k' > a$ or equivalently $k' < b - a - R$. This is implied by $\kappa < b - a - 3R + 1$ (since $R > 1$). Thus, summing up so far, if $\kappa < b - a - 3R + 1$ then for every k and every x we have

$$w\left(1, k, x\right) \geq ... \geq w\left(k-1, k, x\right) > w\left(k, k, x\right) < w\left(k+1, k, x\right) = ... = w\left(\kappa, k, x\right).$$

We also need to consider the payoffs involving type 0. Since $\beta\left(U\right)$ is a pure strategy we have $\tilde{\pi}(\beta(U), U) > \tilde{\pi}(U, U)$. Furthermore

$$\begin{aligned}
\tilde{\pi}(U, \beta(U)) &= \frac{1}{b-a+1}\left(\sum_{s=a}^{\beta(U)-1}(s+R) + \beta(U) + \sum_{s=\beta(U)+1}^{b}(\beta(U)-R)\right) \\
&= \frac{1}{b-a+1}\left(\sum_{s=a}^{b-2R-1}(s+R) + b - 2R + \sum_{s=b-2R+1}^{b}(b-3R)\right) \\
&= \frac{1}{b-a+1}\left((b-2R-a)\left(\frac{b-2R-1+a}{2}+R\right) + b - 2R + 2R(b-3R)\right) \\
&= \frac{1}{b-a+1}\left((b-2R-a)\left(\frac{b-1+a}{2}\right) + b - 2R + 2R(b-3R)\right) \\
&= \frac{(b-a-2R)}{b-a+1}\left(\frac{b-1+a}{2}\right) + \frac{b-2R}{b-a+1} + 2R\frac{b-3R}{b-a+1} \\
&< \left(\frac{b-1+a}{2}\right) + 1 + 2R \\
&= \frac{1}{2}(b+1+a+4R)
\end{aligned}$$

Now I show that type 0 in the cognitive game is strictly dominated by a uniform randomizations over the remaining types. The average (weighted uniformly) payoff to type k against type k' is

$$\begin{aligned}
&\frac{1}{\kappa-1}\sum_{k=1}^{\kappa} w(k, k', x) \\
&\geq \frac{1}{\kappa-1}\left((b-3R-1) + ... + (b-3R-(k'-1)) + a + (\kappa-k')(b-R-(k'-1))\right) \\
&= \frac{1}{\kappa-1}\left((k'-1)(b-3R) - (1+2+...+(k'-1)) + a + (\kappa-k')(b-R-(k'-1))\right) \\
&= \frac{1}{\kappa-1}\left((k'-1)(b-3R) - \frac{1}{2}k'(k'-1) + a + (\kappa-k')(b-R-(k'-1))\right).
\end{aligned}$$

This is increasing in k' since

$$\frac{\partial}{\partial k}\left((k-1)(b-3R) - \frac{1}{2}k(k-1) + a + (\kappa-k)(b-R-(k-1))\right) = k - 2R - \kappa - \frac{1}{2},$$

so

$$\begin{aligned}\frac{1}{\kappa-1}\sum_{k=1}^{\kappa} w\left(k,k',x\right) &\geq \frac{1}{\kappa-1}\left(R+a-b-R\kappa+b\kappa\right)\\ &=\frac{1}{\kappa-1}\left(\left(\kappa-1\right)b+\left(1-\kappa\right)R+a\right)\\ &=b+R+\frac{1}{\kappa-1}a\\ &\geq b+R.\end{aligned}$$

Now we have

$$\tilde{\pi}\left(U,\beta\left(U\right)\right)<\frac{1}{2}\left(b+1+a+4R\right)\leq b+R<\frac{1}{\kappa-1}\sum_{k=1}^{\kappa} w\left(k,k',x\right)$$

if and only if $1+a+2R\leq b$. Thus type 0 in the cognitive game is strictly dominated by a uniform randomizations over the remaining types.

(iii) *Dynamics*: Since type 0 is strictly dominated in the cognitive game we can disregard this type. The payoff matrix for the remaining types in state x is

$$\mathbf{A}=\begin{pmatrix} a & w\left(1,k>1,x\right) & w\left(1,k>1,x\right) & . & w\left(1,k>1,x\right)\\ w\left(k>1,1,x\right) & a & w\left(2,k>2,x\right) & . & w\left(2,k>2,x\right)\\ w\left(k>1,1,x\right) & w\left(k>2,2,x\right) & a & . & w\left(3,k>3,x\right)\\ . & . & . & . & .\\ w\left(k>1,1,x\right) & w\left(k>2,2,x\right) & w\left(k>3,3,x\right) & . & a \end{pmatrix}.$$

Consider a state x where $x_i=0$ for at least one type $i>0$. Then either (I) there is at least one type k such that $x_k=0$ and $x_{k+1}>0$, or (II) there is some k such that $x_{k'}=0$ for all $k'>k$, and $x_{k-1}>0$. In case (I) the average payoff to types k and $k+1$ are, using $x_k=0$,

$$\Pi_k\left(x\right)=\sum_{i=1}^{k-1} w\left(k>i,i,x\right)x_i+\sum_{i=k+1}^{\kappa} w\left(k,k'>k,x\right)x_i,$$

and

$$\Pi_{k+1}\left(x\right)=\sum_{i=1}^{k-1} w\left(k+1>i,i,x\right)x_i+ax_{k+1}+\sum_{i=k+2}^{\kappa} w\left(k+1,k'>k+1,x\right)x_i.$$

Using $x_k = 0$ we get

$$\begin{aligned}\Pi_k(x) - \Pi_{k+1}(x) &= \sum_{i=1}^{k-1} \left(w(k > i, i, x) - w(k+1 > i, i, x)\right) x_i \\ &+ \sum_{i=k+2}^{\kappa} \left(w(k, k' > k, x) - w(k+1, k' > k+1, x)\right) x_i \\ &+ \left(w(k, k' > k, x) - a\right) x_{k+1}.\end{aligned}$$

Since $w(k > i, i, x) = w(k+1 > i, i, x)$, $w(k, k' > k, x) \geq w(k+1, k' > k+1, x)$, and $w(k, k' > k, x) > a$, this is strictly positive. Thus a mutant of type k entering the population will earn more than type $k+1$. Hence the state x is not stable.

In case (II) the average payoff to type k is, using the fact that $x_{k'} = 0$ for all $k' \geq k$;

$$\Pi_k(x) = \sum_{i=1}^{k-1} w(k > i, i, x) x_i,$$

and the payoff to type $k-1$ is

$$\Pi_{k-1}(x) = \sum_{i=1}^{k-2} w(k-1 > i, i, x) x_i + a x_{k-1} + \sum_{i=k}^{\kappa} w(k-1, k' > k-1, x) x_i,$$

so

$$\begin{aligned}\Pi_k(x) - \Pi_{k-1}(x) &= \sum_{i=1}^{k-2} \left(w(k > i, i, x) - w(k-1 > i, i, x)\right) x_i \\ &+ \left(w(k > k-1, k-1, x) - a\right) x_{k-1}.\end{aligned}$$

Since $w(k > i, i, x) = w(k-1 > i, i, x)$, and $w(k > k-1, k-1, x) > a$, this is strictly positive. Thus mutant of type k entering the population will earn more than type $k-1$. Hence the state x is not stable.

(b) *Behavior and Payoffs*: Let $\kappa = 3$. Suppose $\kappa < b - a - 3R + 1$, i.e. $2 + a + 3R < b$. This implies $1 + a + 2R \leq b$, so type 0 will be extinct. Type 1 plays $\beta(U) = b - 2R$ against types 2 and 3. Since $x_0 = 0$ type 2 plays $\beta(\beta(U)) = b - 2R - 1$, against type 3. The payoffs of the cognitive game are

$$\mathbf{A} = \begin{pmatrix} a & b - 3R - 1 & b - 3R - 1 \\ b - R - 1 & a & b - 3R - 2 \\ b - R - 1 & b - R - 2 & a \end{pmatrix}.$$

(Note that the assumption that $\kappa < b - a - 3R + 1$, implies $a < b - 3R - 2 < b - 3R - 1$ and $a < b - R - 2 < b - R - 1$, as was to be expected from the general analysis in (a)).

Dynamics: In order to show that evolution from any interior initial state converges to a unique interior state it is sufficient to show that the game is stable (Sandholm (2007)). Define the tangent space

$$T\left(K\backslash\{0\}\right)=\left\{v\in\mathbb{R}^{|\kappa|}:\sum_{i=1}^{|\kappa|}v_i=0\right\}.$$

A normal form game is stable if and only if the payoff matrix is negative definite with respect to the tangent space. The payoff matrix $\mathbf{A}$ is negative definite with respect to the tangent space if $v\cdot\mathbf{A}v<0$, for all $v\in T\left(K\backslash\{0\}\right)$, $v\neq\mathbf{0}$. We have

$$\frac{1}{2}v\cdot\left(\mathbf{A}+\mathbf{A}'\right)v=2\left(1+2R+a-b\right)v_1^2+2\left(2+2R-b+a\right)v_2^2+2\left(2+2R+a-b\right)v_1v_2.$$

This is a function of v that equals zero at $v=\mathbf{0}$ so it is sufficient to show that the function is concave. Computing the second order derivatives gives the Hessian

$$\left(\begin{array}{cc}4\left(1+2R+a-b\right) & 2\left(2+2R+a-b\right)\\ 2\left(2+2R+a-b\right) & 4\left(2+2R+a-b\right)\end{array}\right).$$

The first principal minor is $4\left(1+2R+a-b\right)<0$ and the second principal minor is the determinant of the Hessian, which is

$$\begin{aligned}&4\left(1+2R+a-b\right)4\left(2+2R+a-b\right)-2\left(2+2R+a-b\right)2\left(2+2R+a-b\right)\\&=4\left(2+2R+a-b\right)\left(4\left(1+2R+a-b\right)-\left(2+2R+a-b\right)\right)\\&=4\left(2R+a-b+2\right)\left(6R+3a-3b+2\right)\\&=12\left(a-\left(b-2R-2\right)\right)\left(a-\left(b-2R-\frac{2}{3}\right)\right)>0\end{aligned}$$

Thus the matrix is negative definite so $\frac{1}{2}v\cdot\left(\mathbf{A}+\mathbf{A}'\right)v$ is indeed concave in v. ■

Proof of Proposition 8. (a) All 2×2-games with a unique interior ESS are strategically equivalent to the game

$$\left(\begin{array}{cc}-a & 0\\ 0 & -b\end{array}\right),$$

for some some $a,b>0$. Let the first strategy be H and the second strategy be D. The ESS puts weight $b/\left(a+b\right)$ on strategy H and weight $a/\left(a+b\right)$ on strategy D, so the ESS payoff is $-ab/\left(a+b\right)$. Suppose without loss of generality that $a>b$. Then type 1 plays $\beta\left(U\right)=D$ against all opponents that are not of type 1. Level k plays D or H against a higher type $k'>k$, and best responds to what a lower type $k'<k$ does. Thus in each encounter between two different types $k>0$ and $k'>0$ $(k\neq k')$ the payoff is zero. Finally we have $\tilde{\pi}(U,U) = -(b+a)/4$, and $\tilde{\pi}(\beta(U), U) = \tilde{\pi}(U, \beta(U)) = -b/2$, so the payff matrix of the cognitive game is

$$\begin{pmatrix} -(b+a)/4 & -b/2 & -b/2 & . & -b/2 \\ -b/2 & -ab/(a+b) & 0 & . & 0 \\ -b/2 & 0 & -ab/(a+b) & . & 0 \\ . & . & . & . & . \\ -b/2 & 0 & 0 & . & -ab/(a+b) \end{pmatrix}.$$

A mix between the types above 0 strictly dominates type 0 in the cognitive game if and only if

$$\frac{-ab}{\kappa(a+b)} > \frac{-b}{2}$$

or equivalently $\kappa > 2a/(a+b)$. After deletion of type 0 from the above matrix, what remains is $-ab/(a+b)\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. From above we know that we need to show that the payoff matrix $-ab/(a+b)\mathbf{I}$ is negative definite with respect to the tangent space. One can transform the problem to one of checking negative definiteness with respect to the space $\mathbb{R}^{\kappa-1}$ rather than the tangent space. This is done with the following transformation matrix P (Weissing (1991)):

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & -1 \end{pmatrix}.$$

Now check whether $-ab/(a+b)(\mathbf{P}\cdot\mathbf{IP})$ is negative definite with respect to $\mathbb{R}^{\kappa-1}$. We have

$$\mathbf{P}\cdot\mathbf{IP} = \begin{pmatrix} 2 & 1 & . & 1 \\ 1 & 2 & . & 1 \\ . & . & . & . \\ 1 & 1 & . & 2 \end{pmatrix}.$$

This matrix is positive definite so $-ab/(a+b)\mathbf{P}\cdot\mathbf{IP}$ is negative definite. Thus evolution from any interior initial state converges to the unique interior ESS, which clearly puts equal weight on all types.

(b) Suppose $\kappa < 2a/(a+b)$. In this case type 0 is not dominated. Each type is the worst reply to itself. An argument similar to that in the proof of proposition 7(a) shows that all states where $x_k = 0$ for some $k \leq \kappa$ are unstable. ■

6.3 Extensions to Other Types

6.3.1 The NE Type

Proof of Proposition 9. (a) Assume $\kappa \geq \tilde{k}$. Suppose type k plays strategy s. The only source of difference between the beliefs of type k and type $k+1$ is that type k is aware of the existence of type k. Because of the single peak property type $k+1$ will either think that s is a best response to the population or that some $t > s$, is a best response. Hence type $k+1$ will play a weakly higher strategy than type k. Type κ will always play a weakly higher strategy than all types $k < \kappa$.

Suppose that s is the best response according to the beliefs of type κ. If x_κ and x_{NE} are small enough then s is indeed the best response. If x_κ and x_{NE} are not small enough then the best response will be $t > s$. No type $k < \kappa$ will play that strategy. Because of the single peak property strategy s earns more than all strategies $s' < s$ which may be used by types $k < \kappa$. Thus type κ earns weakly more than type $k < \kappa$. Furthermore not everyone will play the same strategy since at least type 0 will play U. Hence, type κ will earn above the average of types $\{0, 1, ..., \kappa\}$. This means that the relative fraction $x_\kappa / \sum_{i=0}^{\kappa} x_i$ will increase (even though x_κ may decrease, if x_{NE} earns sufficiently much more that type κ). There are three cases to consider: (1) If type $k < \kappa$ plays a different strategy than type κ, then $x_\kappa / \sum_{i=0}^{\kappa} x_i$ will grow at a higher rate than $x_k / \sum_{i=0}^{\kappa} x_i$. (2) If type k and type κ play the same strategy $s < |S|$ then both types will grow or decrease at the same rate, and no other relative fraction $x_{k'} / \sum_{i=0}^{\kappa} x_i$, $k' < \kappa$, will grow faster. Eventually when $x_k / \sum_{i=0}^{\kappa} x_i$ is large enough type κ will change to playing a higher strategy $t > s$. Then $x_\kappa / \sum_{i=0}^{\kappa} x_i$ grows at a higher rate than $x_k / \sum_{i=0}^{\kappa} x_i$. (3) Since $\kappa \geq \tilde{k}$ it may be that type k and type κ play the same strategy $s = |S|$. In this case they earn the same.

From this we can conclude that if $\kappa \geq \tilde{k}$ then asymptotically only types $k \geq \tilde{k}$ and type NE may exist, i.e. only states with $\sum_{i=\tilde{k}}^{\kappa} x_i + x_{NE} = 1$ may be attracting. To see that only states with $x_{NE} = 1$ are asymptotically stabe consider a state with $\sum_{i=\tilde{k}}^{\kappa} x_i + x_{NE} = 1$, and let $k^* = \min_{k \in \{\tilde{k}, \tilde{k}+1, ... \kappa\}} x_k > 0$. Every nbd of such a state contains states where $\sum_{i=0}^{\tilde{k}-1} x_i = \varepsilon$ and where $x_0, x_1, ..., x_{\tilde{k}-1}$ are such that k^* plays a strategy s such that $\beta(s) = s^{NE}$. In such a state type k^* earns below the average and type NE earns above the average. Thus only states with $x_{NE} = 1$ are asymptotically stable.

If $\kappa < \tilde{k}$ then asymptotically only types κ and NE may exist. An argument similar to that in the previous paragraph establishes that only $x_{NE} = 1$ is asymptotically stable, givne the asumption that $\pi(\beta^\kappa(U), \beta^\kappa(U)) < \pi(s^{NE}, \beta^\kappa(U))$.

(b) Assume $\pi(\beta^\kappa(U), \beta^\kappa(U)) > \pi(s^{NE}, \beta^\kappa(U))$. Now in every nbd of the state with $x_\kappa = 1$ type κ will play a strategy $s \leq \beta^\kappa(U)$ and for any such strategy s it holds that $\pi(s, s) > \pi(s^{NE}, s)$. ■

Proof of Proposition 11. All types, including type NE earn $-ab/(a+b)$ against type NE. Type NE earns $-ab/(a+b)$ against all types. Since $a > b$ implies $-ab/(a+b) <$

$-b/2$ it holds that type NE earns less than type $k > 0$ against type 0. Since $-ab/(a+b) < 0$ type NE earns less than type $k > 0$ against type $k > 0$. Thus in any interior inital state type NE earns less than type $k > 0$. The rest follows from the proof of proposition 8. ■

6.3.2 The RE Type

Proof of Lemma 2. It is a standard result that $\tilde{\beta}$ is compact valued, convex valued and upper hemi-continuous. For a fixed population state $x \in \Delta(K)$, and given behavior of other types $(z_0, ..., z_{\kappa-1})$, the function $\hat{q}^{\kappa}(x, z_0, ..., z_{\kappa-1}, \hat{z}^{\kappa}_{\kappa}) : \Delta(S) \to \Delta(S)$ is continuous in $\hat{z}^{\kappa}_{\kappa}$. Hence the composite correspondence $\tilde{\beta}(\hat{q}^{\kappa}(\hat{z}^{\kappa}_{\kappa}))$ is compact-valued, convex-valued and upper hemi-continuous. By Kakutani's theorem it has a fixpoint. ■

Proof of Proposition 12. If we are in a state x where $z(x) \in \tilde{\beta}(z(x))$ then $z(x)$ is a Nash equilibrium strategy. If $z(x) \notin \tilde{\beta}(z(x))$ then there is at least some $i \in K$ such that $z_i(x) \notin \tilde{\beta}(z(x))$ so $\Pi_i < \bar{\Pi} < \Pi_{\kappa}$. Hence $\dot{x}_{\kappa} > 0$ and $\dot{x}_{\kappa} \geq \dot{x}_j$ for all types j which are present in the population. It follows that eventually either we end up in a state x with $z(x) \in \tilde{\beta}(z(x))$ or $x_{\kappa} = 1$ in which case we also end up with $z(x) \in \tilde{\beta}(z(x))$. ■

7 Appendix B: Extension to Infinite Games

Proofs of results in this appendix are available from the author upon request.

7.1 A Class of Infinite Dominance Solvable Games

In this section I extend the model of the paper to the case of games with infinite, compact, strategy spaces. It is straightforward to extend the notation for finite games to games with infinite, compact, strategy spaces but one thing requires rephrasing: Type k believes that the aggregate weight put on pure strategy s is

$$\hat{q}^k_s(x) = \frac{1}{\sum_{i=0}^{k-1} x_i} \sum_{j=0}^{k-1} \hat{z}^k_{j,s}(x) x_j.$$

I define a class of dominance solvable games with infinite strategy spaces. The interaction between Belle and Rock, mentioned in the introduction, can be viewed as a special case of this class. Formally the following assumptions define the class of games, which is is essentially the same class of games as in Heifetz et al. (2007), except for assumption 4 (see also Moulin (1984)).

Definition 10 *A game belongs to the class $\mathcal{G}^{DSI}$ of dominance solvable infinite games if the strategy space S is a compact interval $[a,b] \subset \mathbb{R}$, and the payoff $\pi : S \times S \to \mathbb{R}$ is twice continuously differentiable, satisfying (1)-(4):*

1. *Payoff is strictly concave in the own strategy;*

$$\frac{\partial^2 \pi\left(s_i, s_j\right)}{\partial s_i^2} < 0.$$

2. *The game has strategic complements, i.e.*

$$\frac{\partial^2 \pi\left(s_i, s_j\right)}{\partial s_j \partial s_i} > 0.$$

3. *The own strategy has a larger effect on the marginal payoff than the other's strategy has;*

$$\left|\frac{\partial^2 \pi\left(s_i, s_j\right)}{\partial s_i^2}\right| > \left|\frac{\partial^2 \pi\left(s_i, s_j\right)}{\partial s_j \partial s_i}\right|.$$

4. *The marginal benefit of increasing the own strategy at (a, a) is weakly positive*

$$\left.\frac{\partial \pi\left(s_i, s_j\right)}{\partial s_i}\right|_{s_i = s_j = a} \geq 0,$$

   and the marginal benefit of increasing the own strategy at (b, b) is weakly negative

$$\left.\frac{\partial \pi\left(s_i, s_j\right)}{\partial s_i}\right|_{s_i = s_j = b} \leq 0.$$

Condition (1) says that the marginal benefit of increasing one's strategy is decreasing in one's own strategy. Condition (2) incorporates the observation, from the game between the chimpanzees, that the marginal benefit of increasing one's strategy is increasing in the opponent's strategy.[16] Condition (3) says that a change in the own strategy has a larger effect on the marginal benefit than an equally large change in the other's strategy. Finally, (4) assures that the best reply function will always be defined by the first order condition. More importantly this implies that the Nash equilibrium will not be reached in a finite number of iterations of the best reply function, unless one starts at the Nash equilibrium. In this way we achieve maximal separation of behavior of different types.

The class of games $\mathcal{G}^{DSI}$ includes Bertrand duopoly with imperfect substitutes, production with positive externality, arms races, and rent seeking. Since there are only two players the case of strategic substitutes can be handled by simply inverting the strategy space of one of the players. (Note that this does not affect (3)). These games include

[16] Note also that in the context of games where the strategy space S is a closed interval in $\mathbb{R}$, and the with a payoff that is twice continuously differentiable and satisfies property (1), the property (2) characterizes strictly supermodular games (or plainly supermodular if the inequality is weak).

Cournot duopoly and public goods provision. For an interesting discussion of these families of games see Eaton (2004).

In this class if games the best reply is single valued. For convenience define the *pure best reply function* $b : S \to S$, that maps a pure strategy s_j to its pure best reply $b(s_j)$

$$b(s_j) = \arg\max_{s_i \in S} \pi(s_i, s_j).$$

We have the following preliminary results.

Lemma 3 *Games in class $\mathcal{G}^{DSI}$ have the following properties:*

(a) *The best reply correspondence $\beta : \Delta(S) \twoheadrightarrow S$ is single-valued, i.e for all types $k \in K$ and all $x \in \Delta(K)$ we have that $z_k(x) = \beta\left(\hat{q}^k(x)\right)$ puts all weight on one pure strategy, and $b : S \to S$ is well-defined.*

(b) *The game is dominance solvable, so there is a unique Nash equilibrium, denoted $\left(s^{NE}, s^{NE}\right)$.*

(c) *For any pure strategy $s \in S$, if $s \gtrless s^{NE}$ then $b(s) \gtrless s^{NE}$.*

(d) *Let $s \in S$ be a pure strategy. If $s \gtreqqless s^{NE}$ then*

$$\left.\frac{\partial \pi(s_i, s_j)}{\partial s_i}\right|_{s_i = s_j = s} \lesseqgtr 0,$$

with equality only if $s = s^{NE}$.

7.2 Evolution of Types in Infinite Games

7.2.1 Unobserved Types

We now analyze a cognitive game based on an underlying game from class $\mathcal{G}^{DSI}$. It turns out that if the best reply against the uniform distribution is a different strategy than the Nash equilibrium strategy, then each type $k > 1$ plays a strategy that is strictly between the strategy of type $k-1$ and the Nash equilibrium strategy. This means that higher types earn strictly more than lower types. Formally:

Proposition 13 *Let s_k be the pure strategy played by type k. i.e. the pure strategy that is given all mass by z_k. If $s_1 < s^{NE}$ then $s_1 < s_2 < s_3 < ... < s^{NE}$ and if $s_1 > s^{NE}$ then $s_1 > s_2 > s_3 > ... > s^{NE}$. In either case $\Pi_0(x) < \Pi_1(x) < ... < \Pi_\kappa(x)$ for all $x \in \Delta(K)$.*

The following result is needed for the evolutionary analysis. The reason that it is a non-trivial result is that behavior changes with the population state, and has to do so in a way that makes $\Pi_k(x)$ Lipschitz continuous.

Lemma 4 $\Pi_k(x)$ *is Lipschitz continuous.*

Since $\Pi_k(x)$ is Lipschitz continuous the vector field $[\Pi_k(x) - \bar{\Pi}(x)]x_k$ is also Lipschitz continuous. By the Picard-Lindelöf theorem the replicator dynamics has a unique solution $\xi(\cdot, x^0) : T \to \Delta(K)$. through any initial condition x_0. Moreover the solution is continuous in t and x^0. This facilitates the analysis of the evolution of the system. Since higher types earn more than lower types evolution will lead to ever more sophisticated individuals. Also, as the following proposition establishes, behavior approaches the Nash equilibrium in the limit.

Proposition 14 *Starting from any interior initial condition evolution leads to the state where everyone is of the highest type: As $t \to \infty$, $x_\kappa \to 1$. Also, in the limit as the number of types goes to infinity behavior corresponds to a Nash equilibrium: As $t \to \infty$, we have $s_k = b(s_{k-1})$ for all $k \geq 2$ and $\lim_{k\to\infty} s_k = s^{NE}$.*

The reason that behavior result asymptotically approaches the Nash equilibrium strategy is that in the limit type k believes that almost everyone is of type $k-1$. Hence type k plays a best reply to what type k does. Taking k to infinity means that we iterate the best reply operator infinitely many times, a process that converges to the Nash equilibrium strategy in this class of games. Note the similarity to the conjecture of Stahl (1993), p. 613, that there might be games where evolution leads to infinitely sophisticated types.

7.2.2 Observed Types

The analysis of infinite games with *observed* types verifies the finding from the finite case, that introducing observability may allow lower types to survive in a setting where they would be extinct if types were not observed. We show this by an example that should be relevant from the point of view of the social brain hypothesis (like the interaction between Belle and Rock).

The *continuous Machiavellian game* G^{CM}, has a strategy space $S = [0, 1]$ and the strategies are ranked according to how Machiavellian they are, with 1 being the most naive and simple strategy and 0 being the most elaborate and Machiavellian one. The incentive to outsmart and choose a more Machiavellian strategy than the opponent can be represented by a two player guessing game. Specifically it will be assumed that the person who guesses closest to one half of the opponent's guess earns the most. In order to allow for the possibility that not only the relative degree of Machiavellianism matters a term is added that captures positive or negative side effects of the strategies used. In total assume that payoff is

$$\pi(s_i, s_j) = -\left(s_i - \frac{1}{2}s_j\right)^2 + cs_is_j,$$

with $-1 < c < 1$. The first term in the payoff function is the guessing game component and the second term represents positive or negative effects of the sophistication/ deceitfulness.

If $c \in (-1, 1)$ then the first partial derivative of the payoff function w.r.t. s_i is negative for all $s_i \geq s_j$, so that one always has incentives to outsmart one's opponent by choosing a lower, more Machiavellian strategy.[17]

Lemma 5 *If $-1 < c < 1$ then the continuous Machiavellian game G^{CM} satisfies the definition of $\mathcal{G}^{DSI}$. The unique Nash equilibrium is $(0, 0)$. If $c > 1/4$ then total surplus is increased by increasing one players strategy unilaterally above a point (such as the Nash equilibrium) where both play the same strategy.*

The marginal benefit of increasing one's strategy unilaterally above a point where both play the same strategy is always negative. So if $c > 1/4$ then Lemma 5 says that one can increase total surplus at one's own expense, by increasing one's strategy unilaterally above a point where both play the same strategy. Hence it would be a cooperative act to increase one's strategy from the Nash equilibrium. Solving the model one finds.

Proposition 15 *If $c \in \left(4/\sqrt{7} - 1, 1\right) \approx (0.512, 1)$, then type 0 is extinct and every state where $x_k = 0$ for some $k \in K \backslash \{0\}$, is unstable.*

Proposition 15 says that if $c > 4/\sqrt{7} - 1$ then lower types, by committing to high strategies, can survive in the evolutionary process. Thus we get evolution of a kind of cooperation in this game, like in the Travelers' Dilemma according to proposition 7.

8 Appendix C: Alternative Models for the Case of Partially Observed Types

Above i extended the CH model to handle the case of partially observed types. In this section I consider two other ways of extending the CH model to the case of partially observed types. *Proofs of results in this appendix are available from the author upon request.*

8.1 Alternative 1

The first alternative is only marginally different from the specification used above. If an individual A of type k_A meets and opponent B of type $k_B < k_A$ then A detects B's type and therefore plays a best response $\beta\left(z_{k_B}\right)$, to what B does. If $k_B > k_A$ then A cannot identify which type B belongs to, so she forms beliefs like in the case of unobserved types. That is, she forms the expectation $\hat{q}^{k_A}(x)$, as defined above, and best responds to this, i.e. plays $\bar{\beta}\left(\hat{q}^{k_A}(x)\right)$. If $k_A = k_B = k$ then I will assume that both play $\bar{\beta}\left(\bar{\beta}\left(\hat{q}^{k}(x)\right)\right)$.

[17] If $c = 0$ we have a pure two-player guessing game. The game can also be interpreted as an arms (or patent) race where lower strategies represent higher arms (or R&D) expenditures.

The intended interpretation is that A understands that B is of type k, but that A falsely believes that B does not understand that A is of the same type. Instead, A thinks that B believes that A is of a type above k, i.e. A thinks that B forms the belief $\hat{q}^k(x)$ about what A will do. The same kind of reasoning is performed by B.

This model is more complicated to use but I how by example that in the Travellers' Dilemma it might be the case that higher types earn less than lower types in some states, so that evolution leads to states where some low types exist.

Proposition 16 *Consider the cognitive game based on the Travelers' Dilemma with* $R = 3/2$*, and let* $\kappa = 2$*. Evolution from any interior initial state converges to the state where* $x_0 = 0$*,* $x_1 = 1/4$*,* $x_2 = 3/4$*.*

8.2 Alternative 2

According to the CH model everyone is overconfident in the sense that everyone believes that they are the only one who knows how everyone else thinks and behaves. In the second alternative model this assumption is somewhat relaxed. Individuals understand that there are other types of equal or higher sophistication, but they still do not know how these types behave and reason in a finite number of steps.

Suppose two individuals A and B of different types k_A and k_B with $k_A < k_B$ meet to play a game. If individual B observes the type of individual A then it is reasonable to assume that B understands what A will do, so that B plays a best response to A's strategy. On the other hand if individual A understands that B is of some higher type than A then it is not clear what A should expect B to do. A natural assumption to make is that A understands that since B is of a higher type, B will not play a strategy that A would never play. (For the moment, disregard her beliefs about other individuals of her own type k.)

In order to make a sensible assumption about what it is that a player A would never do, one needs to think about why an individual forms beliefs in accordance with a certain type rather than another. I suggest that this is because it requires effort and reasoning power to entertain higher order beliefs – i.e. beliefs about beliefs, beliefs about beliefs about beliefs, and so on – and because this ability is heterogeneously distributed. Evidence on the limitations of the theory of mind employed by humans is provided by research on pure theory of mind tasks. In a typical experiment subjects are faced with short texts and then asked questions about the beliefs of the characters in the story. The more layers of beliefs about beliefs that the questions involve, the harder it is for people answer correctly. Let a first order belief be a belief about some non-mental state; e.g. a belief that it snows. An organism is said to be first order intentional (Dennett (1987)) if it is capable of forming beliefs about first order beliefs, e.g. able to form the belief that someone believes that is snows. An organism is second order intentional if it can form beliefs about beliefs about non-mental states, and so forth for higher order intentionality. (Note that this means that being i^{th} order intentional is the same as being able to form $(i+1)^{th}$ order beliefs.) Kinderman et al. (1998) show that that normal humans find tasks of greater than fourth-order intentionality very hard (see also Apperly et al. (2007)).[18]

Type 1 only needs to form a first order belief about what type 0 will do. It is therefore reasonable to assume that type 1 is also first-order intentional. This means that type 1 can form a belief about its own belief, e.g. form the belief "I believe that type 0 will play $U(S)$". Type 2 needs to form a belief about what type 1 thinks that type 0 will do. Accordingly it is reasonable to assume that type 2 is second-order intentional, being able to form beliefs about what she believes that type 1 thinks that type 0 will do. Generally a type k individual needs to be able to entertain beliefs up to the k^{th} order about what other types will believe and do. It is reasonable to assume that she is also able to reflect on the fact that she has these beliefs. Hence she is k^{th} order intentional.

Since a type 1 individual is able to form first order beliefs and best respond given these beliefs type 1 will never play a first order dominated strategy. Since type 1 is first order intentional she is able to reflect on the fact that she will never play a first order dominated strategy. Consequently a type 1 individual should expect that a higher type opponent also does not play a first order dominated strategy. Similarly, a type k individual should expect that an opponent of type $k' > k$ does not play a k^{th} order dominated strategy.

A type k individual does not know what a type $k' > k$ opponent chooses within the set of strategies that are not k^{th} order dominated. Therefore I will assume that type k follows the principle of insufficient reason and forms the belief that opponents of type $k' > k$ randomize uniformly over this set of strategies.

The above line of reasoning also has implications for what assumption to make about what an individual of type k believes that other individuals of type k will do. Since she is unable to form more than k^{th} order beliefs the best she can do is to form the same beliefs as she forms about higher type opponents.

Formally, define $R_0 = S$ and recursively define the set of strategies that are not i^{th} order dominated.

$$R_i = \{s \in S : s = b(t) \text{ for some } t \in R_{i-1}\}.$$

As before let $U(X)$ denote the uniform randomization over a set X. Type $k \geq 1$ plays

$$s_k = \begin{cases} \bar{\beta}(U) & \text{against } k' = 0 \\ \bar{\beta}\left(\bar{\beta}\left(U\left(R_{k'}\right)\right)\right) & \text{against } k' \in \{1, 2, ...k-1\} \\ \bar{\beta}\left(U\left(R_k\right)\right) & \text{against } k' \in \{k, k+1, ...\kappa\} \end{cases}.$$

In the games considered in this paper, the best replies will generally be unique. It should be noted that since

$$\bar{\beta}\left(\bar{\beta}\left(U\left(R_{k'}\right)\right)\right) \in U\left(R_{k'}\right),$$

[18] The difficulties that higher order beliefs pose also matter in strategic settings: Kübler and Weizsäcker (2004) estimate a quantal response model of beliefs in an information cascade experiment and find that the noise is increasing for higher order beliefs.

a lower type will never be surprised by what a higher type does.

The findings from the CH model with observed types are confirmed in the second alternative model:

Proposition 17 *Consider a cognitive game based on any Travelers' Dilemma with* $2R \in \mathbb{N}$. ***(a)*** *If* $R + 2 < \kappa$ *then the state* $x_\kappa = 1$ *is unstable.* ***(b)*** *All states with* $x_k = 1$ *for* $k < \kappa$ *are unstable.* ***(c)*** *If* $R = 3/2$ *and* $\kappa = 4$ *then evolution from any interior initial state converges to the state where* $x_0 = x_2 = x_3 = 0$ *and* $x_1 = x_4 = 1/2$.

Lower types have a commitment advantage relative to higher types. The results from the CH model also go through for the MCH model in coordination games. In games with a unique interior ESS things change slightly but the general principle holds; that everyone being of the highest type is unstable. In total, the CH and MCH models yield very similar results.

References

Alchian, A. A. (1950), 'Uncertainty, Evolution and Economic Theory', *Journal of Political Economy* **58**, 211–222.

Alexander, R. D. (1990), 'How did Humans Evolve? Reflections on the Uniquely Unique Species', University of Michigan Museum of Zoology Special Publication No 1.

Apperly, I. A., Back, E., Samson, D. and France, L. (2007), 'The Cost of Thinking about False Beliefs: Evidence from Adult's Performance on a Non-Inferential Theory of Mind Task', *Cognition* **106**, 1093–1108.

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

Basu, K. (1994), 'The Travellers' Dilemma: Paradoxes of Rationality in Game Theory', *American Economic Review (Papers and Proceedings)* **84**(2), 391–395.

Byrne, R. W. and Whiten, A. (1998), *Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans.*, Oxford University Press, Oxford.

Camerer, C. F. (2003), *Behavioral Game Theory*, Princeton University Press, Princeton.

Camerer, C. F., Ho, T.-H. and Chong, J.-K. (2004), 'A Cognitive Hierarchy Model of Games', *Quarterly Journal of Economics* **199**, pp. 861–898.

Capra, C. M., Goeree, J. K., Gomez, R. and A., H. C. (1999), 'Anomalous Behavior in a Traveller's Dilemma', *American Economic Review* **89**(3), 678–690.

Cosmides, L. and Tooby, J. (1992), Cognitive Adaptations for Social Exchange, *in* J. Barkow, L. Cosmides and J. Tooby, eds, 'The Adapted Mind: Evolutionary Psychology and the Generation of Culture.', Oxford University Press, New York.

Costa-Gomes, M. A. and Crawford, V. P. (2006), 'Cognition and Behavior in Two-Person Guessing Games: an Experimental Study', *American Economic Review* **96**, 1737–1768.

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

Dennett, D. C. (1987), *The Intentional Stance*, MIT Press, Cambridge, Massachusetts.

Dunbar, R. I. M. (1998), 'The Social Brain Hypothesis', *Evolutionary Anthropology* **6**, 178–190.

Dunbar, R. I. M. (2003), 'The Social Brain: Mind Language and Society in an evolutionary Perspective', *Annual Review of Anthropology* **32**, 163–181.

Eaton, B. (2004), 'The Elementary Economics of Social Dilemmas', *Canadian Journal of Economics* **37**, 805–29.

Flinn, M. V., Geary, D. C. and Ward, C. V. (2005), 'Ecological Dominance, Social Competition, and Coalitionary Arms Races: Why Humans Evolved Extraordinary Intelligence', *Evolution and Human Behavior* **26**, 10–46.

Friedman, M. (1953), *Essays in Positive Economics*, University of Chichago Press, chapter The Methodology of Positive Economics.

Fudenberg, D. and Levine, D. K. (1998), *The Theory of Learning in Games*, MIT Press, Camridge, MA.

Güth, W. and Yaari, M. E. (1992), Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach, *in* U. Witt, ed., 'Explaining process and change:', University of Michigan, Ann Arbor, pp. 22–34.

Haruvy, E. and Stahl, D. (2008), Learning Transference between Dissimilar Symmetric Normal-Form Games. mimeo.

Heifetz, A., Shannon, C. and Spiegel, Y. (2007), 'The Dynamic Evolution of Preferences', *Economic Theory* **32**(2), 251–286.

Ho, T.-H., Camerer, C. and Weigelt, K. (1998), 'Iterated Dominance and Iterated Best Response in Experimental "p-Beauty Contests"', *American Economic Review* **88**(4), 947–969.

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

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

Jolly, A. (1966), 'Lemur social behavior and primate intelligence', *Science,* **153**, 501–506.

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

Kojima, F. and Takahashi, S. (2007), 'Anti-Coordination Games and Dynamic Stability', *International Game Theory Review* **9**(4), 667–688.

Kübler, D. and Weizsäcker, G. (2004), 'Limited Depth of Reasoning and Failure of Cascade Formation in the Laboratory', *Review of Economic Studies* **71**, 425–441.

Menzel, E. W. (1974), A Group of Young Chimpanzees in a One-Acre Field, *in* R. J. Schusterman, J. A. Thomas and F. Wood, eds, 'Behavior of Non-Human Primates', Vol. 4, Academic Press, pp. 83–153.

Moulin, H. (1984), 'Dominance Solvability and Cournot Stability', *Mathematical Social Sciences* **7**, 83–102.

Nagel, R. (1995), 'Unraveling in Guessing Games: An Experimental Study', *American Economic Review* **85**, 1313–1326.

Ohtsubo, Y. and Rapoport, A. (2006), 'Depth of Reasoning in Strategic form Games', *The Journal of Socio-Economics* **35**, 31–47.

Penke, L., Denissen, J. J. A. and Miller, G. F. (2007), 'The Evolutionary Genetics of Personality', *European Journal of Personality* **21**, 549–587.

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

Robson, A. J. (2003), 'The Evolution of Rationality and the Red Queen', *Journal of Economic Theory* **111**, 1–22.

Roth, G. and Dicke, U. (2005), 'Evolution of the Brain and Intelligence', *TRENDS in Cognitive Sciences* **9**(5), 250–257.

Samuelson, L. (2001*a*), ‘Analogies, Adaptation, and Anomalies’, *Journal of Economic Theory* **97**(2), 320–366.

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

Samuelson, L. and Zhang, J. (1992), ‘Evolutionary Stability in Asymmetric Games’, *Journal of Economic Theory* **57**, 363–391.

Sandholm, W. H. (2007), Population Games and Evolutionary Dynamics. Book manuscript, to be published by MIT Press.

Sandholm, W. H. and Dokumaci, E. (2007), ‘Dynamo: Phase Diagrams for Evolutionary Dynamics (Software suite)’, http://www.ssc.wisc.edu/ whs/dynamo.

Stahl, D. O. (1993), ‘Evolution of Smartn Players’, *Games and Economic Behavior* **5**, 604–617.

Stahl, D. O. (2000), ‘Rule Learning in Symmetric Normal-Form Games: Theory and Evidence’, *Games and Economic Behavior* **32**, 105–138.

Stahl, D. O. and Wilson, P. W. (1995), ‘On Players’ Models of Other Players: Theory and Experimental Evidence’, *Games and Economic Behavior* **10**, 218–254.

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

Weibull, J. W. (1995), *Evolutionary Game Theory*, MIT Press, Cambridge Massachusetts.

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