Evolution of Perceptions and Play*

Daron Acemoglu
MIT

Muhamet Yildiz
MIT

June 2001

Abstract

An agent with a misperception may have an evolutionary advantage when his misperception and its behavioral implications are recognized by others. In such situations evolutionary forces can lead to misperceptions, yielding "irrational behavior," such as the play of strictly dominated strategies. We point out that this reasoning relies on the assumption of *subjective rationality*—agents are assumed to choose the behavior that maximizes their perceived payoffs. However, subjective rationality does not have solid evolutionary foundations: in the presence of misperceptions, agents who do not maximize their perceived payoffs may have greater fitness than those who do. We show that relaxing the subjective rationality requirement, somewhat paradoxically, leads to effectively rational behavior: although agents may have systematic misperceptions, they will develop other biases to undo these misperceptions, and will act *as if* they are rational. As a result, systematic biases in experimental settings may not necessarily translate into irrational behavior. We also demonstrate that the same evolutionary forces lead agents to play *as if* they have *a common prior*, even though each agent will have different and possibly incorrect perceptions of payoffs and the rules of the game.

Keywords: Common prior, evolution, neutral stability, misperceptions, perceptions, rationality.

JEL Classification: B40, C72, D84.

*We thank Abhijit Banerjee, Eric Van den Steen, and Jorgen Weibull for useful comments.

1 Introduction

Much of economics is built on the notion that agents will take actions that maximize their payoff or fitness. Alchian (1950), Friedman (1953), and Becker (1962) have all emphasized how "evolutionary" competition is likely to eliminate agents who deviate substantially from maximization. The evolutionary game theory literature formalizes these insights. Under replicator dynamics, evolutionary forces will typically eliminate all dominated strategies, and any rest point of an evolutionary process will correspond to a Nash equilibrium of an underlying game (see, e.g., the recent textbooks on evolutionary game theory, Weibull, 1997, or Samuelson, 1997).[1]

The psychology literature and recent research in behavioral economics, instead, emphasize a variety of biases and misperceptions (e.g., overconfidence, wishful thinking, law of small numbers). Although some of these "nonrational" types of behavior may result from cognitive limitations, many scholars argue that nonrational behavior can arise because it confers no payoff disadvantage and may even have payoff benefits. First, there is an informal folk theorem that given the complexity of the economic situations that agents are involved in and the speed with which evolution works, there will not be strong enough evolutionary forces to wipe out nonrational behavior (See Mullainathan and Thaler, 2000, for a very useful discussion and justification for this view. See also DeLong et al. 1990, Blume and Easley, 1992, for models where nonrational behavior generates higher "payoff" than rational behavior). Second, many economists have argued that certain types of behavior that appear "irrational" may be useful commitment devices. Robert Frank (1988), for example, points out how the irrational tendency to seek revenge may be useful by acting as a commitment to punish those who break their promises. Robson (1990), Banerjee and Weibull (1993), Dekel, Ely and Yilankaya (1998), and Kockesen et al (1999) similarly show how, when agents' preferences (types) are observable, evolution may select preferences or types that act "irrationally" to gain commitment advantage.

The argument that irrational behavior may be a useful commitment device requires agents' preferences to be (to some degree) observable. In fact, a number of contributions, including Dekel, Ely and Yilankaya (1998), Ely and Yilankaya (1999) and Ok

[1] Other (non-monotonic) evolutionary dynamics may lead to strategies that are not Nash equilibria (and are even dominated) in the underlying game. See Bendor et. al. (2001) and Stenneck (2000).

and Vega-Rodondo (2000) show that when preferences are not observed at all, evolution takes us back to Nash equilibria. All of these papers, irrespective of whether they allow preferences to be observed or unobserved, rely on a strong notion of *subjective rationality*: agents are assumed to always maximize their *perceived* utility or payoff. Other contributions in game theory also start from different perceptions and use a notion of subjective rationality to derive tight predictions (e.g., Harrison and Kreps, 1978, Morris, 1996, Van den Steen, 2001, Yildiz, 2000).

In this paper, we argue that subjective rationality does not have strong evolutionary foundations (see also Blume and Easley, 1992). Subjectively irrational agents—who systematically choose strategies that do not maximize their perceived payoffs—may obtain higher payoffs than subjectively rational agents. Somewhat paradoxically, once we allow mutations (deviations) that relax subjective rationality, there are strong forces pushing agents towards *effectively rational behavior* (with limited information).[2] Our main results can be summarized as: allowing subjective irrationality leads to effectively rational behavior (with limited information) in any *neutrally stable* outcome of evolution.[3]

Nevertheless, we find that there is no immediate link between rational behavior and accurate perceptions. Evolution will select agents who will act "effectively rationally", but these agents will have very different perceptions, and sometimes will make systematic mistakes (yet their mistakes will cancel each other). In this sense, our analysis provides a microfoundation for the common premise that agents will typically behave *as if* they are rational (e.g., Friedman, 1953). This reasoning suggests that systematic misperceptions in experimental settings or in situations with little relevance to long-run fitness do not necessarily translate into widespread "irrational" behavior.

We also show that in any neutrally stable outcome of evolution, agents may have different perceptions, but the play will always correspond to a Nash equilibrium outcome

[2]We call an agent as *effectively rational with limited information* iff he maximizes his expected fitness with respect to his information. Such an agent may lack some information, but will never have any bias nor play dominated strategies given his information set. (A stronger notion of rationality will correspond to *effectively rational with full information*.) We refer to this type of behavior as "effectively rational" since agents, despite their misperceptions, act *as if* they are rational. Alternatively, this type of behavior could be referred to as "objectively rational" to contrast it with subjectively rational behavior. To minimize terminology, we stick to the term "effectively rational".

[3]Neutral stability is a weaker form of evolutionary stability, where stable strategies do not need to perform strictly better than all mutations in pot-entry population.

of a game where players have a common prior.[4] This result is of interest since there is disagreement about the plausibility of the common prior assumption in game theory (see for example, Gul, 1998, and Aumann, 1998). Since the common-prior assumption is a crucial epistemological foundation for equilibrium behavior in most games, the criticisms regarding the common prior assumption have important implications for all equilibrium analysis (see Aumann, 1987, and Aumann and Brandenburger, 1995, for the importance of the common prior assumption, and Dekel and Gul, 1997, for the related criticism of equilibrium analysis). Our results suggest that equilibrium analysis with the common-prior assumption may have good evolutionary foundations—even when players' *expressed* beliefs are incompatible with each other.

It is important to note that the neutrally stable behavior will be effectively rational only with limited information; some misperceptions may remain in the form of "ignorance", so agents will not necessarily play a Nash equilibrium outcome of the underlying full-information game. When perceptions are observable, agents may develop misperceptions in order to suppress information that would otherwise reduce the payoff to all parties.[5]

The rest of the paper is organized as follows. In the next section, we give a number of examples that illustrate our results, and also discuss related literature in more detail. Section 3 analyzes evolution in a single-agent decision problem (without strategic interactions) more formally. Section 4 analyzes a multi-agent/game-theoretic situation and contains the main results of the paper. Since our main point is that effectively rational behavior emerges once we allow agents to be subjectively irrational, in this section, we allow perceptions to be observed, creating the most favorable environment for deviations from rational behavior (see, e.g., Dekel, Ely, and Yilankaya, 1998). Section 5 analyzes evolution when perceptions are not observed. Section 6 concludes.

[4]But they may not be able to distinguish some underlying games from each other. This common prior will simply correspond to the frequencies with which these underlying games are played.

[5]This result is obtained because perceptions are more than "cheap talk": they also determine how individuals process the world. A player who perceives two different situations in exactly the same way has to play the same strategy in both situations. We show below that when perceptions are simply "cheap talk" systematic misperceptions that affect behavior cannot arise.

2 Examples, Outline and Related Literature

2.1 Examples and Outline

Consider the following 2x2 game and interpret the payoffs as fitness levels, so that evolutionary dynamics will favor strategies that have higher payoffs:

Example of underlying game (Game 1):

1\2	L	R
l	0,0	2,2
r	1,1	3,0

This game is dominance solvable and hence has a unique Nash equilibrium in which player 1 chooses r and player 2 plays L. Now imagine a meta-game in which players have different perceptions of their payoffs—they may see the payoff matrix differently— and an evolutionary process determining the frequency of different perceptions. The fitness of different perceptions in this evolutionary meta-game are still given by the payoffs in the underlying game. Moreover, let us make two important assumptions:

1. *observability:* agents' perceptions are observable to others in the game (for example, the perception of the payoff matrix that one holds is engraved on their forehead).[6]

2. *subjective rationality:* players continue choose whichever action maximizes their perceived payoffs.

Then, consider a type for player 1 who perceives this game instead as

Misperceived game (Game 2):

1\2	L	R
l	2,0	4,2
r	1,1	3,0

This type mistakenly perceives that l is a dominant strategy, and given subjective rationality, will choose l over r. This will encourage his opponent to play R, and so a player with this type of misperception will receive the "true payoff" of 2. This misperception benefits player 1 because it enables him to commit to playing his dominated strategy in the underlying game, thus enticing player 2 to change his play. Without the

[6]See Frank (1988) for a defense of the view that perceptions and preferences are, at least to some degree, observable to others.

misperception, player 1 could not have committed to playing his dominated strategy, and player 2 would have preferred to respond by L.

To see the forces shaping the evolution of perceptions, we simply apply standard evolutionary reasoning to a meta-game where we allow player 1 to either have the correct perception of Game 1 or misperceive it as Game 2 (in other words, we allow mutations over accurate and inaccurate perceptions). For simplicity, we allow no misperceptions for player 2. To highlight the choices facing player 2, we also write out the four strategies for player 2, even though there is no evolutionary selection over these actions (instead, player 2 chooses his action optimally after observing the perceptions of player 1). Since players are subjectively rational given their perceptions, the payoffs in this meta-game are

Meta-game
of the evolution of perceptions:

1\2	LL	LR	RL	RR
accurate perception	1,1	1,1	3,0	3,0
misperception	0,0	**2,2**	0,0	2,2

Strategy LL for player 2 corresponds to choosing L when player 1 is observed to have accurate perceptions and also when he is observed to have misperceptions. Strategy LR corresponds to choosing L when player 1 has accurate perceptions, but R when he has misperceptions. The payoffs then follow immediately from Game 1, incorporating the fact that when he has accurate perceptions, player 1 will choose r, and when he misperceives the payoffs, he will choose l.

The best play for player 1 is clearly to develop a misperception, inducing him to play a *dominated strategy,* l, in the underlying game. The best response of player 2 is to choose LR. This generates the payoff (2,2), which is obviously not a Nash equilibrium outcome of the underlying game (and it contains the play of a dominated strategy along the equilibrium path). Replicator or similar monotonic dynamics will therefore encourage the development of misperceptions.

Intuitively, when player 2 sees an opponent with accurate perceptions, he will choose L, anticipating that his opponent will choose r, giving both players (1,1). In contrast, when faced with a player with a misperception, he will correctly reason that his opponent will play l, and he will choose R, leading to (2,2). This example therefore suggests that evolutionary pressures, far from wiping out misperceptions, may favor apparently "nonrational" behavior. Specifically, two types of nonrational behavior arise as an outcome of the evolution of perceptions in this case: first, agents may have systematic

misperceptions; second, these misperceptions lead to the play of strictly dominated (and hence non-equilibrium) strategies in the underlying game.

The role of the *observability* assumption above is clear: if the misperception of player 1 were not observed, it could not act as a commitment device. Specifically, a mutant (deviation) without misperception would choose r and receive the higher payoff of 3 at the expense of player 2.

The *subjective rationality* assumption is equally important: it guarantees that the misperception of player 1 will translate into a specific action, encouraging player 2 to change his behavior.[7] Is subjective rationality reasonable in an evolutionary setting? Consider the choice between two options, A and B, where A has much higher payoff (in the extreme case, agents who choose B die). But, suppose that all agents have developed a misperception and think that B is preferable to A. Now imagine selection over two different types, one that is subjectively rational and the other subjectively irrational (in the sense that he will not necessarily play the actions that maximize his perceived payoffs—in fact, here we take the subjectively irrational type to play the action that minimizes his payoff). The subjectively rational type will choose B since he perceives it to be preferable to A, while the subjectively irrational type will choose A despite the fact that he perceives B to yield a higher payoff. Evolution will favor the subjectively irrational type, choosing A. In fact, a subjectively irrational type with this misperception will do as well as a subjectively rational agent with accurate perceptions, who perceives A to yield higher payoff and chooses it. This simple example demonstrates that, in the presence of misperceptions, there are no compelling reasons to expect evolution to favor subjectively rational agents.

Let us now return to the analysis of Game 1 and allow players not only to misperceive payoffs, but also to deviate from subjective rationality (for example, mutations that change how a given perception or plan is mapped into behavior). Our major result (Proposition 5) states that in this case where we allow subjective irrationality, evolutionary forces will take us towards "effectively rational play"—i.e., towards behavior that is

[7] To many economists and game theorists, subjective rationality is natural, almost tautological: preferences, through the axiom of revealed preferences, simply represent choices. But in an evolutionary model where preferences (perceptions) can be observable at the beginning of the game, this revealed-preference notion does not apply, since we cannot possibly observe the representation of the choice before the choice has been made.

indistinguishable from that resulting from rational play and does not feature dominated strategies given the information set.

To better understand this result, let us now return to the above meta-game and allow 4 different "strategies" for player 1: accurate perception and subjective rationality; misperception and subjective rationality; accurate perception and subjective irrationality; and misperception and subjective irrationality. Here by misperception we mean that player 1 perceives Game 2 when the true game is Game 1, while by subjective irrationality we simply mean player 1 playing the dominated strategy in the perceived game. We allow no misperceptions for player 2.

Meta-game of the evolution of perceptions and play:

1/2	LL	LR	RL	RR
accurate perception and subjective rationality	**1,1**	1,1	3,0	3,0
misperception and subjective rationality	0,0	2,2	0,0	2,2
accurate perception and subjective irrationality	0,0	0,0	2,2	2,2
misperception and subjective irrationality	**1,1**	3,0	1,1	3,0

Again the first letter for player 2 denotes his action when he observes player 1 with accurate perception, and the second letter gives his action following a misperception by player. For example, if player 1 has misperceptions and is subjectively rational (row 2), he will play l, so when player 2 chooses L, the outcome is (0,0) and when he chooses R, the outcome is (2,2). In contrast, if player 1 has a misperception and is subjectively irrational (row 4), he perceives l to give higher payoff, but still plays r. Now if player 2 chooses L, the outcome is (1,1), and if he chooses R, the outcome is (3,0).

In this meta-game, the outcome (2,2) is no longer (evolutionarily or neutrally) stable. To see this, first consider the case that player 1 misperceives the payoffs and is subjectively rational and player 2 plays LR (or RR), leading to (2,2). Imagine a mutation of player 1 that leads to subjective irrationality, while still keeping the misperception. Recall that whether player 1 has accurate perceptions is seen by player 2, but whether he is subjectively rational or not is not observed. Therefore, with both LR and RR, player 2 will continue to play R, but the mutant player 1 will respond with r and obtain the higher true payoff 3, taking the system away from (2,2). Similarly, the incumbent population in which player 1 has accurate perceptions but behaviorally irrational and player 2 plays RL (or RR) is invaded by the behaviorally rational mutants with accurate perceptions. It is therefore straightforward to verify that any rest point of our

evolutionary process will yield the outcome (1,1)—the Nash equilibrium outcome of the underlying game. This can be achieved as a combination of player 1 developing an accurate perception and subjective rationality and player 2 responding with LL, or player 1 developing a misperception and subjective irrationality and player 2 again responding with LL.

In essence, when agents are limited to subjective rationality, misperceptions provide commitment. However, once we allow the evolution of "play" (how perceptions are mapped into behavior) as well as perceptions, players can undo their misperceptions by modifying their degree of subjective rationality. This destroys the commitment value of misperceptions. The crucial difference here is between the observability of misperceptions and the non-observability of play. As is standard in game theory, we do not (and cannot) allow strategies to depend on the strategies of other players. So while which action a player chooses can depend on his and his opponents' perceptions and past actions, it cannot depend on the exact mapping that his opponents use to translate these perceptions into actions (e.g., it cannot depend on their actions in the future or the actions they could have taken in other situations).

The sense in which the resulting behavior is "rational" needs elaboration. Evolutionary forces take us back to equilibrium outcomes of the underlying game, but this could be supported either by accurate perceptions and subjective rationality, or by misperceptions and subjective irrationality. This explains our choice of terminology: "effectively" rational behavior may be a combination of inaccurate perceptions and subjective irrationality. So, in practice we may observe agents with systematic biases in how they perceive the world or how they make decisions. But these biases will be undone by other biases they develop and will not shape their important decisions (their decisions affecting long-run fitness). The notion of "rational behavior" that emerges is therefore weaker than the usual definition of rationality as it allows misperceptions.[8]

Yet agents are also rational in the stronger sense of effectively having a common prior. Although agents may appear to have systematic misperceptions, viewing the same situation in different ways, they will behave *as if* they all share the same beliefs. For example, suppose a certain behavior, say risk-taking, has a fitness benefit. Some agents

[8]It is also weaker than the usual definition because it requires behavior to be effectively rational with limited information—that is, rational given the information set of agents, which may not feature full information. See below.

may view this behavior as "too risky", while others view it as "fun" and well worth the risk. But at the end, both groups will undertake the activity *as if* they had the same beliefs (and preferences). In this sense, our analysis also provides a foundation for the "common prior" assumption: even though agents may have different perceptions, they will act as if they have effectively common priors.

Consider an example to illustrate this point further. Suppose that each of two agents have to take one of three actions, a, b and c, and there are three possible states of the world, A, B and C. For simplicity, ignore strategic interactions and assume that each agent's payoff is only a function of his own action and the state of the world. Each agent receives a payoff of 1 when the action matches the state of the world, and a payoff of 0 otherwise. Clearly, one possible evolutionary equilibrium involves each agent recognizing the underlying state of the world correctly, and choosing the action corresponding to the state of the world he perceives (e.g., a if the state of the world is A). In this case, both agents will share the same perception. However, under the assumption that each agent's perceptions are observable, there are other evolutionary equilibria where agents will have different (and inaccurate) perceptions, but act *as if* they have common perceptions. For example, suppose that both agents perceive A and C accurately, but agent 1 misperceives B as A, and agent 2 misperceives B as C. In this case, neither agent has accurate perceptions, and moreover, their perceptions differ. But consider the following behavior for both agents: play a if agent 2's perception is A, play b if agent 1's perception is A and agent 2's perception is C, and play c if agent 1's perception is C. This perception-play combination will give the same evolutionary fitness as the perception-play combination where each agent recognizes the underlying world correctly and responds to it. Therefore, in this simple example, it is possible for both players to have different (and incorrect) perceptions, but act *as if* they have common priors, and obtain the maximum payoffs.

Finally, it is useful to give an example where behavior is effective rational only with limited information; it does not correspond to equilibrium play of the underlying full-information game. This can happen because of *suppression of information* by all players. To illustrate this, consider once again a simple example. Suppose that players 1 and 2 play the following game:

Suppression of information (Game 3):

1\2	L	R
l	3,3	-1,1
r	w,-5	0,0

Here w is a random variable, which takes the value 0 and 5, each with probability 1/2. Both players observe w before making their decisions. It is clear that when $w = 5$, the unique Nash equilibrium outcome is (0,0), while when $w = 0$, there is also a Nash equilibrium with players receiving (3,3). Now imagine both players develop the misperception that $w = 0$ all the time. Then, the (true) payoffs of the perceived game become

1\2	L	R
l	3,3	-1,1
r	2.5,-5	0,0

and there is now an equilibrium with (l, L) all the time. So in this game developing a particular systematic misperception to suppress information will benefit both players and will not be eliminated by evolutionary forces. Because the agents are still playing according to the equilibrium of some coarsend version of the underlying game, the outcome is still effectively rational with this limited information, but does not correspond to the Nash equilibrium of the underlying full-information game.

The result regarding suppression of information relies on the assumption that there is a perfect mapping between players' true perceptions and what others observe about the players. When perceptions are not observable, there will still be evolutionary forces leading to effectively rational behavior, but we no longer obtain the suppression of information as an evolutionarily stable outcome.

2.2 Related Literature

This paper is related to a large body of research on evolutionary game theory. Whether evolution will lead to Nash equilibrium is a central question of this literature, and it is well-known that any rest point of a monotonic evolutionary process over types programmed to play different strategies corresponds to a Nash equilibrium outcome (see Weibull, 1997, or Samuelson, 1997, for surveys).

To the best of our knowledge, however, there has been little work on meta-games where evolution occurs over perceptions and the mapping of perceptions into actions.

The notable exceptions are Blume and Easley (1992) and Sandroni (2000) who analyze the benefits of developing accurate perceptions in non-game-theoretic situations,[9] and Dekel, Ely and Yilankaya (1998), who study the evolution of preferences under the (implicit) assumption of subjective rationality. Dekel et. al. (1998) show that, when agents' preferences are observable, any "stable" outcome will yield the same average payoff as that of the "efficient strategy." For example, in the Prisoners' Dilemma game, the unique stable outcome will be mutual cooperation.[10]

A number of contributions, including Banerjee and Weibull (1993), Dekel, Ely and Yilankaya (1998), Ely and Yilankaya (1999), and Ok and Vega-Rodondo (2000) show that in these types of models, when preferences are not observed, evolution takes us back to Nash equilibria. Our results are related to these findings, but enable us to analyze the co-evolution of perceptions and play. Furthermore, they are not driven by the assumption that preferences are unobserved. In fact, for the most part, we assume that perceptions are perfectly observed. Instead, our results follow because we relax the "subjective rationality" assumption embedded in many analyses of evolution of preferences. As a result, evolution favors effectively rational play, but does not wipe out misperceptions or systematic subjective irrationality. Agents can hold systematic misperceptions, and in interviews and experimental settings will display a variety of biases. This type of behavior would not be consistent with models of evolution of preferences, which leads to the same preferences for all agents.

We believe that an analysis of the evolution of perceptions and play is interesting, in part, because the whole discipline of psychology and the recent emerging field of behavioral economics are explicitly concerned with agents' perceptions and how these affect behavior. Moreover, such an analysis enables us to relax the subjective rational-

[9]For example, in Blume and Easley (1992) and Sandroni (2000) agents make forecasts and saving decisions, which affect market prices, but there is no further game-theoretic interaction. Blume and Easley (1992) compute agents' wealth in the long-run and drive a corresponding fitness function. They show that, when the preferences do not coincide with this fitness function, rational agents may lose their wealth and "irrational" agents may determine the prices in the long run, consistent with our argument that subjective rationality does not have an evolutionary basis when there are misperceptions. In the same vein, in an overlapping-generation model with risk averse agents, De Long et al. (1991) suggest that agents with inaccurate beliefs about the future prices may drive the agents with accurate beliefs out when the strategies imitated on the basis of ex-post wealth they generate rather than the ex-ante expected utility level.

[10]Similarly, Kockesen et al., 1999, analyze a class of games in which evolution rewards "negatively interdependent preferences".

ity assumption, which does not have good evolutionary foundations in the presence of misperceptions. Whether subjective irrationality is imposed or not has important implications regarding what types of behavior will be evolutionarily stable. Also, using this framework, we can investigate whether players have a common prior, which is of central importance to the game theory literature.

Our analysis is also related to recent independent work by Samuelson (2001), who illustrates that in the presence of deviations from Bayes' rule, an agent can increase his evolutionary fitness by distorting his utility function (e.g., by including the others' consumption levels to his own utility function). Nevertheless, our notion of "effective rationality despite misperceptions" is different from this idea. To see this, consider an agent who maximizes his fitness by distorting his utility function. According to our terminology, this agent has "accurate perceptions," since he will identify his preferred actions with those that maximize his fitness. In contrast, agents with inaccurate perceptions in our setup will report to prefer actions that do not maximize their fitness, but nevertheless end up taking actions that are good for their long-run fitness.

Finally, our results are also related to analyses of evolutionary dynamics with preplay communication (e.g., Warneyrd, 1993, Kim and Sobel, 1995, Banerjee and Weibull, 2000) because perceptions are similar to signals sent by players before the game is played. In these models, when observed signals do not restrict agents' characteristics or behavior, neutral stability leads to Nash equilibrium outcomes of the underlying game (see Banerjee and Weibull, 2000).[11] Nevertheless, perceptions are different from cheap talk because they also encapsulate what players know about the world, and whether they can distinguish between different worlds. In particular, when the underlying game consists of many different worlds or subgames—as we have here, a player's perceptions will restrict his possible plays by confining him to play the same (mixed) sub-strategy at any two worlds that are perceived to be the same. This is important in some of our results, in particular, for the result that jointly beneficial suppression of information may survive evolutionary pressure.

[11] When observed signals restrict agents' behavior, evolution may lead to non–Nash outcomes, as in Robson (1990), and to strictly dominated strategies as in Banerjee and Weibull (1993).

3 The Single-agent case

We start by analyzing the evolution of perceptions and play in a single-agent decision problem—i.e., in a setting without strategic interactions. This analysis will introduce some of the key concepts and illustrate the reasoning we use in the multi-agent case.

Throughout the paper we consider a large set of agent types subject to evolutionary selection, and assume that there is a process of replicator dynamics that relates the frequency of different agent types to a monotonic function of their fitness. We are interested in the stable outcomes of the replicator dynamics. Like many other evolutionary analyses, we limit ourselves to Neutrally Stable Strategies, which correspond to Lyapunov stable points of (strictly) monotonic replicator dynamics (see Bomze and Weibull, 1995, or Weibull, 1997).

Assume a world where we have a set C of consequences and a finite set S of strategies (actions) $s : \Omega \to C$ with uncertain consequences for some state-space Ω. The set of possible "worlds" is assumed to be finite and denoted by W with a generic member w. Intuitively, different worlds correspond to different types of decision problems that the individual is faced with, and the same action may have different payoff consequences in different worlds. We denote the probability distribution over W by Q, with $Q(w) > 0$ for each w. The expectation with respect to Q will be denoted by E_Q. For each $w \in W$, we consider a pair (u_w, q_w) of a utility function $u_w : C \to \mathbb{R}$ on the consequences, and a probability distribution q_w on Ω. The pair (u_w, q_w) represents the *base preferences* on actions, or the agent's evolutionary fitness. From a strict evolutionary point of view, q_w measures the frequency of the events and $u_w(c)$ is the evolutionary fitness following consequence c. We summarize the base preferences of the agent by

$$\hat{v}(s; w) \equiv E_{q_w}(u_w \circ s)$$

at each $s \in S$ and $w \in W$, where E_q denotes the expectation with respect to q. This notation will be used extensively in the next section.

Analogously to the base preferences of the agent, for any probability distribution q and preferences u, we define the payoff function

$$v(s) \equiv E_q(u \circ s)$$

This function gives the ("perceived") payoff of choosing action s given preferences u

and probability distribution q. We consider a finite set $V \subset \mathbb{R}^S$ of payoff functions $v : S \to \mathbb{R}$. We define a *perception* as a function

$$\pi : W \to V.$$

Intuitively, a perception $\pi(w)$ is an interpretation of a world w as a (possibly different) world w'. The set of actual worlds that can be realized is naturally contained in the set of possible worlds that the agent can perceive (i.e., $V \supset \{\hat{v}(\cdot, w) : w \in W\}$).

We call a perception π^* with $\pi^*(w) = \hat{v}(\cdot, w)$ at each w as *accurate*. Intuitively, an accurate perception enables the decision-maker to correctly recognize the world he is confronted with. Given any π and w, we say that there is a misperception at w if $\pi(w) \neq \hat{v}(\cdot, w)$. This corresponds to either to the perceived probability distribution differing from the underlying frequencies, i.e., $q \neq q_w$, or the perceived utility function differing from fitness function, i.e., $u \neq u_w$.[12]

Every time a world w is drawn, the agent perceives the world as $\pi(w)$ and takes an action $s \in S$, determining his fitness. Any function $\sigma : V \to S$ is called a *play*, and the set of all plays is denoted by $\Sigma = S^V$.

Definition 1 *A play σ is said to be* subjectively rational *if and only if*

$$\sigma(v) \in \arg\max_{s \in S} v(s).$$

A subjectively rational play maximizes the agent's perceived payoff. This is the standard definition of rationality in Game Theory. For many, this is a tautology, for the perceptions (or beliefs) are merely expressions of the choice, derived from the choice function using the principle of revealed preferences. In this paper, we take the perceptions and choice independent. We will take σ^* as a generic subjectively rational play.

Definition 2 *A pair (π, σ) of a perception and play is said to be* effectively rational (with full information) *if and only if $\sigma \circ \pi = \sigma^* \circ \pi^*$ for some subjectively rational σ^* and accurate π^*, i.e.,*

$$\sigma(\pi(w)) \in \arg\max_{s \in S} \hat{v}(s; w)$$

at each $w \in W$.

[12] Accurate perceptions may also correspond to a pair $(u, q) \neq (u_w, q_w)$ of inaccurate probability distribution and an inaccurate utility function, such that $E_{q_w}(u_w \circ s) = E_q(u \circ s)$. For an example, see Samuelson (2001).

That is, a pair (π, σ) of a perception and a play is effectively rational if we can never distinguish the resulting choice from that of an individual who perceives accurately and behaves rationally. Here, we use a strong notion of accurate perception; it implies that the agent distinguishes each state from the others (hence our qualification, with full information). In our next section, we will also define effective rationality with limited information, which only requires an agent to maximize his expected fitness, given the information available. Our next example illustrates that there will be many combinations of misperception and subjectively irrational play that yield a rational choice.

Example 1 Consider the case $W = \{w\}$, $S = \{a, b\}$ with $\hat{v}(a) = 1$ and $\hat{v}(b) = 0$. Suppose also that $V = \{v_1, v_2\}$ such that $v_1(a) = 1$ and $v_1(b) = 0$, and $v_2(a) = 0$ and $v_2(b) = 1$. Therefore, v_1 corresponds to accurate perception, while v_2 is a misperception. Then, there are four combinations of perceptions and play:

1. agent perceives accurately plays subjectively rationally, and chooses a;
2. he inaccurately perceives b better than a, and plays subjectively rationally, choosing b;
3. he accurately perceives a better than b, but plays subjectively irrationally, and chooses b;
4. he inaccurately perceives b better than a, but plays subjectively irrationally, and chooses a.

Combinations 1 and 4 are effectively rational with full information, and yield higher fitness than that of the other two perception-play combinations. With any monotonic replicator dynamics, effectively rational agents will survive, while the rest will die away. Some of these effectively rational players will hold misperceptions, they will make clearly irrational choices given their perceptions, but the overall choice will maximize their fitness.

More importantly, the concept of subjective rationality loses its appeal in the presence of misperceptions: the subjectively irrational agents in 4 do better than the subjectively rational agents in 2. Therefore, in a world where the agents may have misperceptions, subjective rationality does not necessarily have an evolutionary foundation.

Now, given a play σ, we can define effectively accurate perceptions, which are not necessarily accurate.

Definition 3 *Given any play σ, a perception π is said to be (effectively)* σ-accurate *if and only if $\sigma \circ \pi = \sigma \circ \pi^*$.*

With this terminology, (π, σ) is effectively rational if and only if π is σ-accurate.

Given any σ, we define the one-person meta-game

$$G_\sigma = (\Pi, g(\cdot; \sigma))$$

where

$$g(\pi; \sigma) = E_Q[\hat{v}(\sigma(\pi(w)); w)] \equiv E_Q[E_{q_w}[u_w(\sigma(\pi(w)))]]$$

at each $\pi \in \Pi$. Note that $g(\pi; \sigma)$ is the agent's expected payoff with respect to the base preferences, given that, at each w, the agent's perceptions will be $\pi(w)$ and he will play σ and choose $\sigma(\pi(w))$. In game G_σ, σ is fixed but the perceptions can vary. Even though G_σ is technically a game with a strategy space Π, it is not a usual game since the perceptions $\pi \in \Pi$ are not chosen; the agent happens to have them, perhaps because they inherited them or as a result of mutations. For this reason, we refer to it as a meta-game. When $g(\pi; \sigma) < g(\pi'; \sigma)$, given that he will choose according to σ, his evolutionary fitness will increase if his perceptions happen to change from π to π'.

We also consider the meta-game

$$G = (\Pi \times \Sigma, g)$$

where $g(\pi; \sigma)$ is defined as above. This game depicts a situation where both the perceptions and the play evolve together.

Definition 4 *Given any play σ, a perception $\pi \in \Pi$ is Neutrally Stable (NS) for G_σ iff $g(\pi; \sigma) \geq g(\pi'; \sigma)$ at each $\pi' \in \Pi$. Likewise, $(\pi, \sigma) \in \Pi \times \Sigma$ is NS for G iff $g(\pi; \sigma) \geq g(\pi'; \sigma')$ at each $(\pi', \sigma') \in \Pi \times \Sigma$.*

This definition simply states that a perception, or a combination of perception and play, will be Neutrally Stable if no other alternative gives higher fitness. Therefore, we require that there are no forces taking the system away from this configuration. This is a

simpler version of the definition of Neutral Stability we give below for meta-games with strategic interactions. Our interest in Neutrally Stable outcomes is motivated by the fact that the set of Lyapunov stable outcomes with respect to any monotonic replicator dynamics will be the set of Neutrally Stable outcomes (see, e.g., Weibull, 1997).

Proposition 1 *Given any subjectively rational σ^*, any perception $\pi \in \Pi$ is NS for G_{σ^*} iff it is σ^*-accurate.*

The proof of Proposition 1 is combined with that of Proposition 2. Proposition 1 states that under subjective rationality, the surviving agents will hold accurate perceptions (and will choose effectively rational behavior). This result implies that the example of misperceptions we found in Game 1 of Section 2 is due to strategic interactions and cannot happen in the single-agent case. This is intuitive, since as noted in Section 2, misperceptions are useful because they provide commitment, and in this single-agent meta-game, commitment has no value.

Next, we see that when the requirement of subjective rationality is relaxed, there will be some other effectively rational agents who, despite their misperceptions, do as well as these agents.

Proposition 2 *Any $(\pi, \sigma) \in \Pi \times \Sigma$ is Neutrally Stable for G iff it is effectively rational with full information.*

Proof. Take any effectively rational $(\hat{\pi}, \hat{\sigma})$ and any $(\pi, \sigma) \in \Pi \times \Sigma$. At each $w \in W$, by definition, we have $\hat{v}\left(\hat{\sigma}\left(\hat{\pi}\left(w\right)\right); w\right) = \max_{s \in S} \hat{v}\left(a; w\right) \geq \hat{v}\left(\sigma\left(\pi\left(w\right)\right); w\right)$. Hence, $g\left(\hat{\pi}, \hat{\sigma}\right) = E_Q\left[\hat{v}\left(\hat{\sigma}\left(\hat{\pi}\left(w\right)\right); w\right)\right] \geq E_Q\left[\hat{v}\left(\sigma\left(\pi\left(w\right)\right); w\right)\right] = g\left(\pi, \sigma\right)$, showing that $(\hat{\pi}, \hat{\sigma})$ is Neutrally Stable for G. Conversely, if (π, σ) is not effectively rational, then the former inequality will be strict at some w_0 with $Q\left(w_0\right) > 0$, rendering the latter inequality strict, and showing that (π, σ) is not Neutrally Stable for G.

To prove Proposition 1, observe that (π, σ^*) is effectively rational iff π is σ^*-accurate. Hence, taking $\hat{\sigma} = \sigma = \sigma^*$ above, we obtain Proposition 1. ■

4 The Multi-person Case

In this section we analyze evolution when there are many players strategically interacting with each other.[13] In that case, when the perceptions are observable, under subjective rationality, a player may gain from his misperception, a fact that we established in Section 2 using Game 1 as an example. Nevertheless, we have also seen that in the presence of misperceptions, subjective rationality is no longer supported by evolution. Therefore, here we focus on a process in which both perceptions and play evolve together.

4.1 The environment

We consider a set $N = \{1, 2, \ldots, n\}$ of players and a set $S = \prod_{i \in N} S_i$ of strategy profiles. For each $w \in W$, we consider an underlying game $(N, S, \hat{v}(\cdot; w))$, where $\hat{v}(s, w) = (\hat{v}^1(s, w), \hat{v}^2(s, w), \ldots, \hat{v}^n(s, w)) \in \mathbb{R}^n$ is the players' base payoff vector at a strategy profile $s \in S$. Here $\hat{v}^i(s, w)$ measures the evolutionary fitness of a player $i \in N$ when the strategy profile s is played at w. The set of all feasible payoff functions is taken to be a finite set $V \subset (\mathbb{R}^n)^S$ that contains

$$E_Q[\hat{v}|W'] = \frac{1}{\sum_{w \in W'} Q(w)} \sum_{w \in W'} \hat{v}(\cdot, w) Q(w)$$

for each non-empty $W' \subset W$.

By a perception, we mean any function from $\pi : W \to V$. The set of all perceptions is, once again, denoted by $\Pi = V^W$. By a perception profile, we mean any vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n) \in \Pi^N$ of perceptions. We write $\pi_i^j(s, w)$ for the payoff of player j as perceived by player i when strategy profile s is played at world w. By a mixed perception we mean any probability distribution x on the set of perceptions. Here, $x(\pi)$ can be considered as the proportion of population who has perception π.

We focus on games where the realization of the payoff perceived by each player, $\pi_i(\cdot, w)$ for all i, is observable. That is, all agents see the payoff matrix as perceived by other agents. This is motivated by the discussion in the introduction where we pointed out that nonrational behavior is most likely to arise as an evolutionary stable outcome when perceptions/preferences are observed. With this assumption, a player's strategy will be a function that maps from the perception profiles into strategies: $\sigma_i : V^N \to$

[13] This corresponds to frequency-dependent evolution in the terminology of Maynard-Smith (1982).

$\Delta(S_i)$, where $\Delta(S_i)$ denotes the space of the probability distributions on S_i. Notice that any such strategy profile $\sigma : V^N \to \Delta(S)$ can also be thought as a *solution concept*, since it determines how the game will be played given the perceptions. It is also important that it is the realizations of the perceptions, $\pi_i(\cdot, w)$, not the perception function, π_i, itself that is observable.[14]

We consider a process in which at each date each player is randomly allocated to play role i in a randomly selected underlying game $(N, S, \hat{v}(\cdot; w))$. Each agent with perception π perceives the payoff function as $\pi(w)$. We then have a profile $(\pi_1(w), \cdots, \pi_n(w))$ of perceived payoff functions where $\pi_i(w)$ denotes the payoff function perceived by the player who plays role i. Observing his role i and the profile $(\pi_1(w), \cdots, \pi_n(w)) \in V^n$ of perceived payoff functions, a selected player plays some (possibly mixed) strategy $\sigma_i(\pi_1(w), \cdots, \pi_n(w)) \in \Delta(S_i)$ in the underlying game $(N, S, \hat{v}(\cdot; w))$. Given any $v \in V^n$ and any $s_i \in S_i$, $\sigma_i(s_i, v)$ denotes the probability of playing s_i according to $\sigma_i(v) \in \Delta(S_i)$. Since σ already depends on the player's own perceptions, a mixed strategy can be defined simply as a pair (x, σ) of a mixed perception and a mixed play. Finally, given any two strategies (x, σ) and (y, μ), the mixture $\lambda(x, \sigma) + (1 - \lambda)(y, \mu) = (\lambda x + (1 - \lambda) y, \lambda\sigma + (1 - \lambda)\mu)$ for any $\lambda \in [0, 1]$ is also a strategy.

To formalize the evolution of perception and play, we consider the meta-game G where a mixed strategy of a player is a pair (x, σ) of a perception $x \in \Delta(V^W)$ and a play $\sigma : V^N \to \Delta(S)$ determining which strategy $\sigma_i(v)$ he will play at each role i and at each profile $v = (\pi_1(w), \cdots, \pi_n(w))$ of perceived payoff functions. The expected payoff of a player with (x, σ) in a population with aggregate distribution (y, μ) of perception and play is

$$U((x, \sigma), (y, \mu)) = \frac{1}{n} \sum_{i \in N} \sum_{w \in W} \sum_{s \in S} \sum_{\boldsymbol{\pi} \in \Pi^n} \hat{v}^i(s, w) Q(w) \sigma_i(s_i, \boldsymbol{\pi}(w)) \prod_{j \neq i} \mu(s_j, \boldsymbol{\pi}(w)) x(\pi_i) \prod_{j \neq i} y(\pi_j),$$

where $s = (s_1, \ldots, s_n)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$. This equation states the following. The agent will be selected to play a role i (with probability $1/n$) in a randomly drawn underlying game $(N, S, \hat{v}(\cdot; w))$ (with probability $Q(w)$). He will have perception π_i

[14] If the perception function were observed, players would know not only how an agent perceives the current world, but also how he would have perceived each $w \in W$. For our qualitative results, it is not important whether the entire perception function or only the realization of this function is observable (so long as a player does not observe more about his perception than the others do—such "informational asymmetry" may allow agents to always distinguish between all $w \in W$, see Section 5).

with probability $x(\pi_i)$, while every other player j will have some perception π_j with probability $y(\pi_j)$. Observing the profile $\boldsymbol{\pi}(w) = (\pi_1(w), \ldots, \pi_n(w))$, he will play some $s_i \in S_i$ with probability $\sigma_i(s_i, \boldsymbol{\pi}(w))$, while every other player j will play some strategy s_j with probability $\mu(s_j, \boldsymbol{\pi}(w))$. The agent's true contingent payoff for such a play will be $\hat{v}^i(s, w)$.

Definition 5 *A pair* (x, σ) *is said to be* Neutrally Stable for G *iff, for each* (y, μ)*, there exists some* $\bar{\epsilon} > 0$ *such that*

$$U((x,\sigma), \epsilon(y,\mu) + (1-\epsilon)(x,\sigma)) \geq U((y,\mu), \epsilon(y,\mu) + (1-\epsilon)(x,\sigma))$$

for each $\epsilon \in (0, \bar{\epsilon})$*.*

As noted above, neutral stability corresponds to the Lyapunov stability in replicator dynamics. It is also clear that when (x, σ) is neutrally stable, $((x,\sigma),(x,\sigma))$ is a Nash equilibrium in the meta-game G, i.e., $U((x,\sigma),(x,\sigma)) \geq U((y,\mu),(x,\sigma))$ for each (y,μ) (see for example Weibull, 1997).

As in the single-agent case, we can define a meta-game to describe the evolution of perceptions for a fixed solution concept and the evolution of the solution concept (or play) given a fixed set of perceptions. The discussion of Game 2 in Section 2 illustrates that when evolution is only limited to perceptions, misperceptions can arise as (neutrally) stable outcomes of evolutionary processes. Here we discuss the evolution of both perceptions and play. To this end, we first analyze evolution of play under a common (and fixed) perception, and then show that any combination of uncommon perceptions actually embed a common perception. Using this insight, we then provide a general characterization of the simultaneous evolution of perceptions and play.

4.2 Evolution of play under a common perception

Consider the case where all agents are restricted to have a common perception of the payoff function, while how they play the game (given the perception) evolves according to the process we defined above. In this case, the common prior (perception) assumption holds by construction.

Consider a meta-game G_π where all agents share a fixed pure perception, π, and where the strategy of an agent is some $\sigma^\pi : V \to \Delta(S)$ determining which strategy $\sigma_i^\pi(v)$

he will play at each role i when the commonly perceived payoff function is v. Define $\bar{\boldsymbol{\pi}} = (\pi, \pi, \ldots, \pi)$ as the perception profile that assigns the common perception π to all players. Let also $\sigma : V^N \to \Delta(S)$ be such that $\sigma^\pi(., \pi(w)) = \sigma(., \bar{\boldsymbol{\pi}}(w))$, and similarly, let $\mu : V^N \to \Delta(S)$ and $\mu^\pi : V \to \Delta(S)$ be such that $\mu^\pi(., \pi(w)) = \mu(., \bar{\boldsymbol{\pi}}(w))$. That is, σ and μ induce the same behavior as σ^π and μ^π when all players have common perception. Then, we can define expected (true) fitness in the meta-game U^π as:

$$U^\pi(\sigma^\pi, \mu^\pi) = U((\bar{\boldsymbol{\pi}}, \sigma), (\bar{\boldsymbol{\pi}}, \mu)) = \frac{1}{n} \sum_{i \in N} \sum_{w \in W} \sum_{s \in S} \hat{v}^i(s, w) Q(w) \sigma_i(s_i, \bar{\boldsymbol{\pi}}(w)) \prod_{j \neq i} \mu_j(s_j, \bar{\boldsymbol{\pi}}(w)),$$

where $\bar{\boldsymbol{\pi}} = (\pi, \pi, \ldots, \pi)$. Definition 5 immediately implies that a play σ is *Neutrally Stable for* G_π iff, for each μ, there exists some $\bar{\epsilon} > 0$ such that $U((\bar{\boldsymbol{\pi}}, \sigma), \epsilon(\bar{\boldsymbol{\pi}}, \mu) + (1-\epsilon)(\bar{\boldsymbol{\pi}}, \sigma)) \geq U((\bar{\boldsymbol{\pi}}, \mu), \epsilon(\bar{\boldsymbol{\pi}}, \mu) + (1-\epsilon)(\bar{\boldsymbol{\pi}}, \sigma))$, or alternatively such that $U^\pi(\sigma, \epsilon\mu + (1-\epsilon)\sigma) \geq U^\pi(\mu, \epsilon\mu + (1-\epsilon)\sigma)$ for each $\epsilon \in (0, \bar{\epsilon})$.

The information incorporated in a common perception π is represented by the information partition

$$I^\pi = \left\{ \pi^{-1}(\pi(w)) \,|\, w \in W \right\}.$$

The cell that contains a given w is denoted by $I^\pi(w) = \pi^{-1}(\pi(w))$. No $w' \in I^\pi(w)$ can be distinguished from w using the knowledge of π. Knowing the function π and the realization $\pi(w)$, a Bayesian can infer that the true world is in $I^\pi(w) = \pi^{-1}(\pi(w))$, but when $I^\pi(w)$ is not a singleton, he cannot distinguish between different worlds contained in $I^\pi(w)$. Therefore, the expected true fitness given the realization $\pi(w)$ and the function π is

$$E_Q[\hat{v}|I^\pi(w)] \equiv \frac{1}{\sum_{w' \in I^\pi(w)} Q(w')} \sum_{w' \in I^\pi(w)} \hat{v}(\cdot, w') Q(w').$$

In the case where π is an invertible function, $\pi^{-1}(\pi(w)) = w$, and $E_Q[\hat{v}|I^\pi(w)] = \hat{v}(\cdot; w)$. If Bayesian players knew only the perception function π and the perceived payoffs $\pi(w)$, using this coarse information they could compute expected payoffs as $E_Q[\hat{v}|I^\pi(w)]$.

Definition 6 *Given any* $(\pi, \hat{\sigma})$*, play* $\hat{\sigma}$ *is said to be* effectively rational with respect to I^π *iff*

$$\hat{\sigma} \in \arg\max_{\sigma^i} \sum_{s \in S} E_Q\left[\hat{v}^i(s) \,|I^\pi(w)\right] \sigma^i(s_i) \mu^{-i}(s_{-i})$$

for some belief μ_{-i} about the other players' play s_{-i} for each $i \in N$ and $w \in W$. A play $\hat{\sigma}$ is said to be effectively rational with limited information *iff it is effectively rational with respect to some I^{π}. A play $\hat{\sigma}$ is said to be* effectively rational with full information *iff it is effectively rational with respect to I^{π^*} where $\pi^*(w) = w$ at each w.*

Because $\hat{\sigma}$ maximizes the agent's utility given some belief, μ_{-i}, regarding other players' behavior, it also rules out the play of donated strategies given the information set I^{π}. In this sense, Definition 6 generalizes Definition 2 from previous section to the case with strategic interactions. When the payoff to each agent is independent of the others' actions, the definition of effectively rational behavior with full information coincides with that in Definition 2.

Next, given the limited information set $I^{\pi}(w)$, one can construct a new game $(N, S, E_Q[\hat{v}|I^{\pi}(w)])$ with a set of players N, strategy spaces S, and payoff functions $E_Q[\hat{v}|I^{\pi}(w)]$, from the underlying game $(N, S, \hat{v})$. We refer to this as a *coarsened* game, since the payoff functions are the averages of the true payoff functions over an arbitrary information set I^{π}, which may not distinguish all underlying worlds from each other. We now establish that there is a close link between the Nash equilibria of this derived game and the neutrally stable outcomes of G_{π}.

Proposition 3 *Given any $\pi : W \to V$ and any $\hat{\sigma} : V \to \Delta(S)$, if $\hat{\sigma}$ is neutrally stable for G_{π}, then $\hat{\sigma}(\pi(w))$ is a Nash equilibrium of game $(N, S, E_Q[\hat{v}|I^{\pi}(w)])$ for each $w \in W$. Therefore, $\hat{\sigma} \circ \pi$ is effectively rational with limited information.*

Proof. We will prove the contraposition of the first part. Let $\hat{\sigma} : V \to \Delta(S)$ be such that $\hat{\sigma}(\pi(w))$ is not a Nash equilibrium of $(N, S, E_Q[\hat{v}|I^{\pi}(w)])$ for some $w \in W$. Then, there exists some $k \in N$ and $\sigma_k(\pi(w)) \in \Delta(S_k)$ such that

$$INC = \sum_{s \in S} E_Q\left[\hat{v}^k(s, w)\,|I^{\pi}(w)\right] \sigma_k(s_k, \pi(w)) \prod_{j \neq k} \hat{\sigma}_j(s_j, \pi(w)) - \sum_{s \in S} E_Q\left[\hat{v}^k(s, w)\,|I^{\pi}(w)\right] \hat{\sigma}_k(s_k, \pi(w)) \prod_{j \neq k} \hat{\sigma}_j(s_j, \pi(w)) > 0.$$

Now, consider $\tilde{\sigma} : V \to \Delta(S)$ with

$$\tilde{\sigma}_i(v) = \begin{cases} \sigma_k(\pi(w)) & \text{if } i = k \text{ and } v = \pi(w), \\ \hat{\sigma}_i(v) & \text{otherwise.} \end{cases}$$

Using the definitions, we compute that

$$U^{\pi}(\tilde{\sigma},\hat{\sigma}) = U^{\pi}(\hat{\sigma},\hat{\sigma}) + INC \cdot \frac{1}{n} \sum_{w' \in I^{\pi}(w)} Q(w') > U^{\pi}(\hat{\sigma},\hat{\sigma}),$$

showing that $(\hat{\sigma},\hat{\sigma})$ is not a Nash equilibrium of the meta game G_{π}. But $(\hat{\sigma},\hat{\sigma})$ is a Nash equilibrium of G_{π} whenever $\hat{\sigma}$ is neutrally stable for G_{π}. Therefore, $\hat{\sigma}$ is not neutrally stable for G_{π}. That $\hat{\sigma} \circ \pi$ is effectively rational with limited information then immediately follows from the fact that no dominated strategy is played in a Nash equilibrium. ■

This proposition implies that only Nash equilibrium strategies in the "coarsened" game $(N, S, E_Q[\hat{v}|I^{\pi}(w)])$ are candidate neutrally stable outcomes of the common perception meta-game, G_{π}. The intuition for why Nash equilibria of the underlying game can be neutrally stable is straightforward: suppose all players share the same accurate perception. Then, standard arguments from evolutionary game theory imply that Nash equilibria of the underlying game will be stable outcomes of the meta-game G_{π}, and therefore, evolution leads to behavior that is effectively rational with limited information. The same reasoning immediately generalizes to the case where π is not accurate: even if players have misperceptions, what matters is their fitness, determined by payoffs in the underlying game, $\hat{v}$, and the information incorporated in these perceptions. Finally, to see why Nash equilibria of the "coarsened" game $(N, S, E_Q[\hat{v}|I^{\pi}(w)])$ can be neutrally stable, recall the example in the introduction where suppression of information was mutually beneficial. When all players have the same perception not enabling them to distinguish between different worlds, Nash equilibria of this coarsened game can also be neutrally stable. We will provide a more detailed analysis of this in Example 2 below.

Motivated by the idea that agents can adjust their behavior much faster than their preferences, Dekel, Ely and Yilankaya (1998) assume that agents will always play a Nash equilibrium of the game where the payoffs are determined by their preferences, even though these preferences may not reflect their true fitness. Our proposition, in contrast, can be interpreted as showing that when players adjust their behavior faster than their preferences (in the sense that their behavior can change while their preferences are fixed), the play will approach a Nash equilibrium of the underlying game, where the payoffs are players' true fitness, not their perceived payoffs.

We next extend Proposition 3 to the case of a mixed common perception (rather than a common pure perception). A mixed common perception is a probability distribution

$x \in \Delta(\Pi)$ as in G. Unlike in G, however, we require the agents to share the same perception π at each realization π. Thus, by construction, players always share a common perception. Now construct the meta-game G_x with payoff function

$$\begin{aligned} U^x(\sigma,\mu) &= \sum_{\pi\in\Pi} U^\pi(\sigma,\mu)\, x(\pi) \\ &= \frac{1}{n}\sum_{\pi\in\Pi}\sum_{i\in N}\sum_{w\in W}\sum_{s\in S} \hat{v}^i(s,w)\, Q(w)\, x(\pi)\, \sigma_i\left(s_i, \bar{\boldsymbol{\pi}}(w)\right) \prod_{j\neq i} \mu_j\left(s_j, \bar{\boldsymbol{\pi}}(w)\right), \end{aligned}$$

where $\bar{\boldsymbol{\pi}} = (\pi,\ldots,\pi)$. Naturally, a play σ is *Neutrally Stable for* G_x iff for each μ, there exists some $\bar{\epsilon} > 0$ such that $U^x(\sigma, \epsilon\mu + (1-\epsilon)\sigma) \geq U^x(\mu, \epsilon\mu + (1-\epsilon)\sigma)$ for each $\epsilon \in (0,\bar{\epsilon})$.

Take any commonly perceived payoff function $v = \hat{\pi}(w)$ for some $w \in W$ and $\hat{\pi}$ with $x(\hat{\pi}) > 0$; the profile of perceived payoffs is $(v, v, \ldots, v)$. Denote the set of perception functions that can lead to perceived payoff v by $\Pi_v = \{\pi | \pi(w) = v \text{ for some } w \in W\}$. Observing v and knowing the distribution x on perceptions, a Bayesian can update his beliefs about the perception functions and compute the expected payoff functions given this coarse information. Given any $\pi \in \Pi_v$, the posterior probability that the common perception function is π is

$$\Pr(\pi|x,v) = \frac{x(\pi)}{\Pr(v|x)} \equiv \frac{x(\pi)}{\sum_{\pi'\in\Pi_v} x(\pi')}.$$

When $\pi \notin \Pi_v$, we have $\Pr(\pi|x,v) = 0$. Now, given any common perception π, and the commonly perceived payoff $v = \hat{\pi}(w)$, the expected true payoff (as perceived by a Bayesian) will be $E_Q\left[\hat{v}|\pi^{-1}(v)\right] = E_Q\left[\hat{v}|I^{\hat{\pi}}(w)\right]$. Then, the expected payoff function given $v = \hat{\pi}(w)$ and x will be

$$E[\hat{v}|x,v] = \sum_{\pi\in V^W} \Pr(\pi|x,v)\, E_Q\left[\hat{v}|\pi^{-1}(v)\right] = \sum_{\pi\in V^W} \Pr(\pi|x,\hat{\pi}(w))\, E_Q\left[\hat{v}|I^{\hat{\pi}}(w)\right].$$

Effective rationality with respect to x is defined as in Definition 6, using $E[\hat{v}|x,v]$ as the payoff function.

For any perceived payoffs are $v = \hat{\pi}(w)$, we can now construct another derived game $(N, S, E[\hat{v}|x,v])$. Once again, we are interested in this game because there is a close link between the Nash equilibria of this derived game and the neutrally stable outcomes resulting from evolution over play.

Proposition 4 *Given any mixed common perception x, and any $\hat{\pi}$ with $x(\hat{\pi}) > 0$, if $\hat{\sigma} : V \to \Delta(S)$ is neutrally stable for G_x, then $\hat{\sigma}(\hat{\pi}(w))$ is a Nash equilibrium of game $(N, S, E[\hat{v}|x, \hat{\pi}(w)])$ for each $w \in W$. Therefore, $\hat{\sigma} \circ \hat{\pi}$ is effectively rational with limited information.*

This proposition extends Proposition 3 to the case of mixed common perceptions. It is useful because when we consider evolution over perceptions and play, we will encounter mixed perceptions. We omit the proof of this proposition which is similar to that of Proposition 3.

4.3 Co-evolution of perceptions and play

Now we consider the meta game G, where the perceptions and play evolve together, and show that at any neutrally stable outcome the agents play a Nash equilibrium of a coarsened version of the underlying game. Towards this goal, we first derive the common mixed perception $\tilde{x}$ *embedded* in (possibly non-common) mixed perceptions x. Loosely speaking, the common perception $\tilde{x}$ embedded in x will contain exactly the same information regarding the underlying state as the initial uncommon (mixed) perception profile x. We will then show that, for any neutrally stable (x, σ) in the meta-game G, the play σ is neutrally stable for the meta-game $G_{\tilde{x}}$, where the agents' perceptions are fixed at the mixed common perception $\tilde{x}$ embedded in x. Proposition 4 proves that for each neutrally stable (x, σ) for G, and for each realization $\boldsymbol{\pi}(w)$ of agents' perception profile, $\sigma(\boldsymbol{\pi}(w))$ is a Nash equilibrium of the coarsened game $(N, S, E[\hat{v}|\tilde{x}, \tilde{\pi}(w)])$, where agents have a common perception (prior), and estimate their payoffs via Bayes' rule using all available information. This will establish our claim in the introduction that despite different perceptions, agents will play *as if* they share a common perception, and will choose some effectively rational behavior with limited information.

We take a set $\tilde{V} \supset V$ with cardinality $|\tilde{V}| = |V^n|$. The set $\tilde{V}$ will be the image of the common perceptions embedded in any perception profile $\boldsymbol{\pi}$.[15] We also fix a one-to-one and onto function

$$\phi : V^n \to \tilde{V}.$$

[15]Note that this theory is useful when we approximate the contiunuum with grid V, in which case $|V^n|$ and $|V|$ will have the same limit.

The function ϕ will provide the isomorphism between the perception profiles and the common perceptions embedded in them.

Common perceptions embedded in perception profiles Given any perception profile $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \in \Pi^N$, we consider a (generalized) perception $\tilde{\pi} : W \to \tilde{V}$ with

$$\tilde{\pi} = \phi \circ \boldsymbol{\pi}.$$

The set of all such (generalized) perceptions is denoted by $\tilde{\Pi} = \tilde{V}^W$. For each such perception $\tilde{\pi} \in \tilde{\Pi}$, we also have the perception profile $\boldsymbol{\pi} = \phi^{-1} \circ \tilde{\pi}$. We will say that the common perception $\tilde{\pi}$ is *embedded* in the perception profile $\boldsymbol{\pi}$: it contains exactly the same information as the uncommon perception profile $\boldsymbol{\pi}$. That is, at each $w \in W$, we have

$$I^{\tilde{\pi}}(w) = \tilde{\pi}^{-1}(\tilde{\pi}(w)) = \boldsymbol{\pi}^{-1}\left(\phi^{-1}(\phi(\boldsymbol{\pi}(w)))\right) = \boldsymbol{\pi}^{-1}(\boldsymbol{\pi}(w)) = I^{\boldsymbol{\pi}}(w). \tag{1}$$

Hence, the information partitions generated by $\tilde{\pi}$ and $\boldsymbol{\pi}$ are the same. Here, $I^{\boldsymbol{\pi}}(w) = \bigcap_{i \in N} I^{\pi_i}(w) = \{w' | \pi_i(w') = \pi_i(w) \, \forall i \in N\}$ is the set of all worlds, w', that cannot be distinguished from world w by looking at the realization of $\boldsymbol{\pi}$.

The mixed common perceptions embedded in mixed (uncommon) perceptions are defined similarly. For each mixed perception x, we define a mixed common perception $\tilde{x}$ on $\tilde{\Pi}$ by setting

$$\tilde{x}(\tilde{\pi}) = \Pr\left(\phi^{-1}(\tilde{\pi}) | x\right) \equiv \prod_{i \in N} x(\pi_i) \tag{2}$$

at each $\tilde{\pi} \in \tilde{\Pi}$, where $\phi^{-1} \circ \tilde{\pi} = (\pi_1, \ldots, \pi_n) \equiv \boldsymbol{\pi}$. The mixed common perception $\tilde{x}$ is embedded in the mixed (uncommon) perception x in the sense that $\tilde{x}$ assigns the same probability to the common perception $\tilde{\pi} = \phi \circ \boldsymbol{\pi}$ embedded in each perception profile $\boldsymbol{\pi}$ as the x assigns to the probability profile $\boldsymbol{\pi}$ itself.

As was the case with $\tilde{\pi}$ and $\boldsymbol{\pi}$, $\tilde{x}$ and x generate the same information partitions. Given any $\tilde{\pi}' = \phi \circ \boldsymbol{\pi}'$, and any $w \in W$, the probability that we have a common perception $\tilde{\pi}'$ given $\tilde{x}$ and $\tilde{\pi}(w)$ is the same as the probability that we have perception profile $\boldsymbol{\pi}'$ given x and $\boldsymbol{\pi}(w)$, i.e., $\Pr(\tilde{\pi}' | \tilde{x}, \tilde{\pi}(w)) = \Pr(\boldsymbol{\pi}' | x, \boldsymbol{\pi}(w))$. Hence, by (1), we have

$$\begin{aligned} E[\hat{v} | \tilde{x}, \tilde{\pi}(w)] &= \sum_{\tilde{\pi}' \in \tilde{\Pi}} \Pr(\tilde{\pi}' | \tilde{x}, \tilde{\pi}(w)) \, E_Q\left[\hat{v} | I^{\tilde{\pi}}(w)\right] \\ &= \sum_{\boldsymbol{\pi}' \in \Pi^N} \Pr(\boldsymbol{\pi}' | x, \boldsymbol{\pi}(w)) \, E_Q[\hat{v} | I^{\boldsymbol{\pi}}(w)] = E[\hat{v} | x, \boldsymbol{\pi}(w)]. \end{aligned}$$

That is, the expected payoff function given the perceived payoffs will be the same under x and $\tilde{x}$.

The next lemma clarifies the relationship between the meta-game G, where the perceptions and play evolve together, and the meta-game $G_{\tilde{x}}$, where the play evolves under the mixed common perception $\tilde{x}$.

Lemma 1 *Let $\tilde{\sigma} : \tilde{V} \to \Delta(S)$, $\tilde{\mu} : \tilde{V} \to \Delta(S)$, $\sigma : V^N \to \Delta(S)$, and $\mu : V^N \to \Delta(S)$ be such that $\sigma = \tilde{\sigma} \circ \phi$ and $\mu = \tilde{\mu} \circ \phi$. Then,*

$$U^{\tilde{x}}(\tilde{\sigma}, \tilde{\mu}) = U((x, \tilde{\sigma} \circ \phi), (x, \tilde{\mu} \circ \phi)) = U((x, \sigma), (x, \mu)).$$

Proof. We have

$$\begin{aligned} U^{\tilde{x}}(\tilde{\sigma}, \tilde{\mu}) &= \frac{1}{n} \sum_{\tilde{\pi} \in \tilde{\Pi}} \sum_{i \in N} \sum_{w \in W} \sum_{s \in S} \hat{v}^i(s, w) Q(w) \tilde{x}(\tilde{\pi}) \tilde{\sigma}_i(s_i, \tilde{\pi}(w)) \prod_{j \neq i} \tilde{\mu}_j(s_j, \tilde{\pi}(w)) \\ &= \frac{1}{n} \sum_{\boldsymbol{\pi} \in \Pi^N} \sum_{i \in N} \sum_{w \in W} \sum_{s \in S} \hat{v}^i(s, w) Q(w) \left[\prod_{i \in N} x(\pi_i) \right] \tilde{\sigma}_i(s_i, \phi(\boldsymbol{\pi}(w))) \prod_{j \neq i} \tilde{\mu}_j(s_j, \phi(\boldsymbol{\pi}(w))) \\ &= U((x, \tilde{\sigma} \circ \phi), (x, \tilde{\mu} \circ \phi)) = U((x, \sigma), (x, \mu)). \end{aligned}$$

■

This lemma states that the outcome of the evolution of play in any situation with widely differing beliefs and perceptions can be represented as the outcome of evolution of play in a situation with common perceptions. Perhaps, the fact that there is a common perception that captures the same information as a set of uncommon perceptions is not surprising. But this lemma will enable us to next prove the stronger result that strategies will evolve, and the long run, precisely *as if* all agents are using this common *embedded* perception.

Evolution of perceptions We can now state and prove the main result of the paper, which links the neutrally stable outcomes of G to neutrally stable outcomes of the common perception game $G_{\tilde{x}}$, and to Nash equilibria of a coarsened version of the underlying game $(N, S, E[\hat{v}|x, \boldsymbol{\pi}(w)])$. We write $\Pr(\boldsymbol{\pi}|x)$ for the probability of a perception profile $\boldsymbol{\pi}$ under a mixed perception x, i.e., $\Pr(\boldsymbol{\pi}|x) = \prod_{i \in N} x(\pi_i)$.

Proposition 5 *For every neutrally stable (x, σ) for G, the play $\tilde{\sigma} = \sigma \circ \phi^{-1}$ is neutrally stable for $G_{\tilde{x}}$, where $\tilde{x}$ is the mixed common perception defined by (2). Therefore, given any $\boldsymbol{\pi}$ with $\Pr(\boldsymbol{\pi}|x) > 0$, and any $w \in W$, $\sigma(\boldsymbol{\pi}(w))$ is a Nash equilibrium of game $(N, S, E[\hat{v}|\tilde{x}, \tilde{\pi}(w)])$, where $\tilde{\pi} = \phi \circ \boldsymbol{\pi}$. Therefore, $\sigma \circ \boldsymbol{\pi}$ is effectively rational with limited information.*

Proof. Take any neutrally stable (x, σ) for G. Take any $\tilde{\mu} : \tilde{V} \to \Delta(S)$, and let $\tilde{\sigma} = \sigma \circ \phi^{-1}$ and $\mu = \tilde{\mu} \circ \phi$. Since (x, σ) is neutrally stable for G, there exists some $\bar{\epsilon} > 0$ such that for each $\epsilon \in (0, \bar{\epsilon})$,

$$U((x, \sigma), (x, \epsilon\mu + (1-\epsilon)\sigma)) \geq U((x, \mu), (x, \epsilon\mu + (1-\epsilon)\sigma)).$$

But, by Lemma 1, we have

$$\begin{aligned} U^{\tilde{x}}(\tilde{\sigma}, \epsilon\tilde{\mu} + (1-\epsilon)\tilde{\sigma}) &= U((x, \tilde{\sigma} \circ \phi), (x, \epsilon\tilde{\mu} \circ \phi + (1-\epsilon)\tilde{\sigma} \circ \phi)) \\ &= U((x, \sigma), (x, \epsilon\mu + (1-\epsilon)\sigma)) \end{aligned}$$

and $U^{\tilde{x}}(\tilde{\mu}, \epsilon\tilde{\mu} + (1-\epsilon)\tilde{\sigma}) = U((x, \mu), (x, \epsilon\mu + (1-\epsilon)\sigma))$. Therefore, for each $\epsilon \in (0, \bar{\epsilon})$,

$$U^{\tilde{x}}(\tilde{\sigma}, \epsilon\tilde{\mu} + (1-\epsilon)\tilde{\sigma}) \geq U^{\tilde{x}}(\tilde{\mu}, \epsilon\tilde{\mu} + (1-\epsilon)\tilde{\sigma}),$$

showing that $\tilde{\sigma}$ is neutrally stable for $G_{\tilde{x}}$.

The second statement follows from Proposition 4: since $\tilde{\sigma} = \sigma \circ \phi^{-1}$ is neutrally stable for $G_{\tilde{x}}$, by Proposition 4 (for $G_{\tilde{x}}$), $\tilde{\sigma}(\tilde{\pi}(w)) = \sigma(\boldsymbol{\pi}(w))$ is a Nash equilibrium of $(N, S, E[\hat{v}|\tilde{x}, \tilde{\pi}(w)])$ at each $w \in W$ whenever $\tilde{x}(\tilde{\pi}) = \Pr(\boldsymbol{\pi}|x) > 0$. Effective rationality of $\sigma \circ \boldsymbol{\pi}$ again immediately follows from the Nash equilibrium result. ■

This proposition states that evolution of play will ensure the emergence of behavior corresponding to all agents having common perceptions and playing the Nash equilibrium behavior of some underlying game. That is, evolution will lead agents to play *as if* they have *a common prior* and are effectively rational (in fact, they play an equilibrium). This is despite the fact that each agent may hold different misperceptions and act "irrationally" given their perceptions. Their behavior will effectively undo these perception differences, leading to effectively rational behavior and to a situation with effectively common perceptions.

Notice that the proposition maps the neutrally stable outcomes of evolution to the Nash equilibria of some coarsened game, yielding effectively rational behavior with limited information. This may not, however, correspond to the equilibrium of the full-information game, because some suppression of information may take place (hence our emphasis on limited information). This is because all agents may develop misperceptions that suppress information that would reduce the payoff to all parties, destroying some "insurance" opportunities—as in Game 3 of Section 2. We will now present such a neutrally stable outcome of this game in more detail.

Example 2: Suppression of information Consider the same example as in Game 3 in the introduction, with true payoff function $\hat{v}$ for this game is

$1\backslash 2$	L	R
l	3,3	-1,1
r	w,-5	0,0

where x can take values 0 and 5 with equal probabilities. We take $W = \{w_1 = 0, w_2 = 5\}$ with $Q(w_1) = Q(w_2) = 1/2$, where w takes 0 at w_1 and 5 at w_2. We take $V = \{\hat{v}(\cdot; w_1), \hat{v}(\cdot; w_2), \bar{v}\}$ as the set of possible payoff functions, where $\bar{v} = (\hat{v}(\cdot; w_1) + \hat{v}(\cdot; w_2))/2$, which corresponds, in this case, to $w = (0+5)/2 = 2.5$. We consider the pure perception-play pair $(\bar{\pi}, \bar{\sigma})$ where $\bar{\pi}(w) = \bar{v}$ at each $w \in W$ and

$$\bar{\sigma}^i(v_1, v_2) = \begin{cases} (1,0) & \text{if } v_1 = v_2 = \bar{v}, \\ (0,1) & \text{otherwise} \end{cases}$$

for each $i \in N$. Here, the first entry is the probability that $\bar{\sigma}^i$ assign to the strategy l or L when he plays rows or columns, respectively. According to this strategy, a player plays l (or L) if each player perceives the game as the average game, and plays r (or R) otherwise. When everyone has perception-play pair $(\bar{\pi}, \bar{\sigma})$, each player gets 3.

The only way a mutant may have a higher payoff is that when he is the row player he chooses r at w_2, and the column player continues to play L, giving him a payoff of 5. Strategy $\bar{\sigma}^i(v_1, v_2)$ dictates that the column player will play L only if its opponent, the mutant in this case, perceives the payoff function at w_2 as $\bar{v}$. In that case, our mutant must play r at $(\bar{v}, \bar{v})$. At w_1, if the mutant's perception differs from $\bar{v}$, the column player will play R, in which case the highest our mutant gets is 0 (when he plays r). Hence, if he distinguishes w_1 and w_2, the highest our mutant can get is $(0.5)(5)+(0.5)(0) = 2.5$, lower than his payoff of 3 under $(\bar{\pi}, \bar{\sigma})$.

If the mutant does not distinguish w_1 and w_2, he must perceive the payoff function as $\bar{v}$ at w_1, too. In that case, the profile of perceived payoff functions will always be $(\bar{v}, \bar{v})$, hence the incumbents will always play L, and the mutant will always play r and again obtain $(0.5)(5)+(0.5)(0) = 2.5$, lower than 3. In either case, a mutant gets strictly lower payoff, unless his perception function is $\bar{\pi}$ and he plays l at $(\bar{v}, \bar{v})$. Now we need to check whether these mutants, who do as well as incumbents against incumbents, cannot do any better when they meet each other. But, when any two such mutants meet, each will get exactly 3, not higher than the incumbents' payoff. Therefore, $(\bar{\pi}, \bar{\sigma})$ is neutrally stable.

This example shows that some Nash equilibrium of a coarsened game may be neutrally stable (even when it does not correspond to a Nash equilibrium of the full-information game). Not all Nash equilibria of all coarsened games can be neutrally stable outcomes, however. For example, Proposition 2 shows that in one-person games, each neutrally stable outcome leads to effectively rational behavior with full information, ruling out all non-trivial coarsenings. We next investigate this issue in games with strategic interactions.

4.4 A further restriction

In deriving Proposition 5, we considered only mutations of play, and did not consider mutations of perceptions. We will now consider mutations of both perceptions and play, and derive a stronger restriction on the set of neutrally stable outcomes.

Let us write

$$U^w(x, \sigma) = \frac{1}{n} \sum_{i \in N} \sum_{s \in S} \sum_{\boldsymbol{\pi} \in \Pi^n} \hat{v}^i(s, w) \prod_{j \in N} \sigma_j(s_j, \boldsymbol{\pi}(w)) \prod_{j \in N} x(\pi_j)$$

for the expected value of having perception x and playing σ, while all the other agents do the same, given that the underlying world is w. We also define

$$\underline{U}^w = \frac{1}{n} \sum_{i \in N} \min_{\sigma_{-i}} \max_{\sigma_i} \sum_{s \in S} \hat{v}^i(s, w^*) \prod_{j \in N} \sigma_j(s_j)$$

as the minimum expected utility an agent gets if he recognizes that the underlying world is w and gives his best response. Finally, write

$$V_x = \{v = \pi(w) \,|\, x(\pi) > 0, w \in W\}$$

for the set of payoff functions perceived with positive probability under x.

Proposition 6 *Let $(\hat{x}, \hat{\sigma})$ be neutrally stable for G. If $V_x \neq V$, then*

$$U^w(\hat{x}, \hat{\sigma}) \geq \underline{U}^w$$

at each $w \in W$.

Proof. Assume $V_x \neq V$ and take any $\bar{v} \in V \backslash V_{\hat{x}}$. To derive a contradiction consider a neutrally stable $(\hat{x}, \hat{\sigma})$ such that $\underline{U}^{w^*} > U^{w^*}(\hat{x}, \hat{\sigma})$ for some $w^* \in W$. Consider the following mutant $(\bar{x}, \bar{\sigma})$:

1. he recognizes w^*: $\bar{x}(\bar{\pi}) = \hat{x}(\pi)$ for each π and $\bar{\pi}$ with $\hat{x}(\pi) > 0$, $\bar{\pi}(w^*) = \bar{v}$, and $\bar{\pi}(w) = \hat{\pi}(w)$ whenever $w \neq w^*$; and

2. he gives his best response at w^*: for each $i \in N$, when $v_i = \bar{v}$, $\bar{\sigma}_i(v_1, \ldots, v_n) \in \arg\max_{\sigma_i} \sum_{s \in S} \hat{v}^i(s, w) \sigma_i(s_i) \prod_{j \neq i} \sigma_j(s_j, v_1, \ldots, v_n)$, and $\bar{\sigma}_i(v_1, \ldots, v_n) = \hat{\sigma}_i(v_1, \ldots, v_n)$ otherwise.

When $w \neq w^*$, the mutant $(\bar{x}, \bar{\sigma})$ does as well as the incumbents. We will now show that he will do strictly better than the incumbents at w^*. At w^*, the mutant gets

$$\begin{aligned}
& \frac{1}{n} \sum_{i \in N} \sum_{v_{-i} \in V^{n-1}} \Pr(v_{-i} | \hat{x}, w^*) \max_{\sigma_i} \sum_{s \in S} \hat{v}^i(s, w^*) \sigma_i(s_i) \prod_{j \neq i} \hat{\sigma}_j(s_j, (\bar{v}, v_{-i})) \\
\geq & \frac{1}{n} \sum_{i \in N} \sum_{v_{-i} \in V^{n-1}} \Pr(v_{-i} | \hat{x}, w^*) \min_{\sigma_{-i}} \max_{\sigma_i} \sum_{s \in S} \hat{v}^i(s, w^*) \prod_{j \in N} \sigma_j(s_j) \\
= & \frac{1}{n} \sum_{i \in N} \min_{\sigma_{-i}} \max_{\sigma_i} \sum_{s \in S} \hat{v}^i(s, w^*) \prod_{j \in N} \sigma_j(s_j) \equiv \underline{U}^{w^*} > U^{w^*}(\hat{x}, \hat{\sigma}),
\end{aligned}$$

where $(\bar{v}, v_{-i})$ is the vector of perceived payoff functions where i and any other j perceive the payoff function as $\bar{v}$ and v_j, respectively, and $\Pr(v_{-i} | \hat{x}, w^*)$ is the probability of $(\bar{v}, v_{-i})$ given that the incumbents have mixed perceptions $\hat{x}$ and the underlying world is w^*. (Note that $\sum_{v_{-i} \in V^{n-1}} \Pr(v_{-i} | \hat{x}, w^*) = 1$; and at w^*, each incumbent gets $U^{w^*}(\hat{x}, \hat{\sigma})$.) This establishes that his expected payoff will be strictly better than the incumbents', contradicting our hypothesis that $(\hat{x}, \hat{\sigma})$ is neutrally stable for G. ■

This proposition shows that in any Neutrally stable outcome an individual cannot increase his payoff by introducing more information. This restricts the set of coarsened games whose Nash equilibria can correspond to Neutrally stable behavior.

5 Evolution When Perceptions Are Not Observable

We have so far limited the analysis to situations in which perceptions are observable. The previous literature has also investigated situations in which agents' preferences may be only imperfectly observable (e.g., Dekel, Ely and Yilankaya, 1998, Ely and Yilankaya, 1999, and Ok and Vega-Rodondo, 2000). We now extend our analysis to this case.

Suppose that in addition to the observable perceived payoff function $\pi(\cdot, w)$, each agent also has a private perception function π^p. As a result, each agent will observe his own $\pi^p(w)$, and can condition his behavior on this perception, but he will not observe the private perception of other players. We can incorporate this into our analysis by simply redefining the play function to have a larger domain, $\sigma : V^{N+1} \to \Delta(S)$. Similarly, we define a perception profile for player in role j, $\boldsymbol{\pi}_j$, as $\boldsymbol{\pi}_j = (\pi^p, \pi_1, ..., \pi_n)$.[16] This implies that each agent can condition his play on all observed perceptions and his own private perception. We can now state

Proposition 7 *Given any neutrally stable $(\hat{x}, \hat{\sigma})$ for G and any $\hat{\pi}$ with $x(\hat{\pi}) > 0$, $\hat{\sigma}(\hat{\pi}(w))$ is a Nash equilibrium of the underlying game $(N, S, \hat{v}(\cdot, w))$ for each $w \in W$. Therefore, $\sigma \circ \boldsymbol{\pi}$ is effectively rational with full information.*

Proof. An immediate generalization of Proposition 5 implies that all neutrally stable outcomes correspond to Nash equilibria of some coarsened game. We only have to rule out Nash equilibria of these coarsened games that are not Nash equilibria of the underlying game $(N, S, \hat{v}(\cdot, w))$. Suppose that $(\hat{x}, \hat{\sigma})$ is neutrally stable, and $\hat{\sigma}(\boldsymbol{\pi}(w^*))$ is a Nash equilibrium of a game $(N, S, E_Q[\hat{v}|x, \boldsymbol{\pi}(w^*)])$, but not a Nash equilibrium of $(N, S, \hat{v}(\cdot, w^*))$ for some $w^* \in W$. This implies that there exist a role j, and a strategy s'_j such that

$$\hat{v}^j\left(s'_j, \sigma_{-j}(\boldsymbol{\pi}(w)), w\right) > E_Q\left[\hat{v}^j|x, \boldsymbol{\pi}(w)\right]$$

Now consider a mutant with $\pi^p = \pi^*$, where π^* is the accurate perception. This deviation is not observed by other players, so their strategies are still given by $\hat{\sigma}_{-j}(\boldsymbol{\pi}(w))$. This then enables the mutant to choose s'_j when he plays the role j at w^*, increasing his fitness. Hence, $(\hat{x}, \hat{\sigma})$ is not neutrally stable. Since $\hat{\sigma}(\hat{\pi}(w))$ is a Nash equilibrium of

[16] It will turn out to be redundant that $(\pi_1, ..., \pi_n)$ are observable, given that π^p is private information. Nevertheless, this more general form highlights that adding a flexible piece of private information to our setup so far removes many of the neutrally stable outcomes of Proposition 5.

the full information game, it rules out the play of all dominated strategies, hence $\sigma \circ \boldsymbol{\pi}$ is effectively rational with full information. ■

Therefore, once we allow perceptions not to be observed, we obtain a stronger proposition than Proposition 6: now only Nash equilibria of the underlying game are candidate neutrally stable outcomes. This is intuitive, since starting from any outcome that is a Nash equilibrium of some coarsened game, but not a Nash equilibrium of the underlying game, a player can take a deviation to expand his information set and increase his payoff.

6 Concluding Remarks

There are many examples of potentially irrational behavior that may have evolutionary benefits by acting as a commitment device. For example, the tendency to seek revenge may be a useful commitment to punish those who break their promises. In this paper we argue that a key link in this reasoning, that of *subjective rationality*, does not have strong evolutionary foundations, undermining much of the appeal of this argument. Subjective rationality requires that agents always choose actions that maximize their perceived payoffs. Although this appears almost tautological, we show that when there are misperceptions, subjectively irrational agents can have greater evolutionary success than subjectively rational agents.

Once we allow mutations to lead to subjective irrationality as well as misperceptions, we find that selection will lead to effectively rational behavior, albeit with limited information: perceptions or preferences act as commitment devices only when they are informative about future behavior, and relaxing subjective rationality breaks this link. Thus somewhat paradoxically, subjective irrationality creates stronger evolutionary forces towards "rational behavior".

Interestingly, there is no immediate link between rational behavior and accurate perceptions. Evolution will select agents who will act "rationally", but these agents will have very different perceptions, and sometimes will make systematic mistakes (yet their mistakes will cancel each other). This reasoning suggests that systematic misperceptions in experimental settings or in situations with little relevance to long-run fitness do not necessarily translate into widespread "irrational" behavior.

We also show that even though, in an "evolutionary equilibrium," agents will have

very different perceptions, they will possess effectively common perceptions of the game and the payoffs, in the sense that they will play *as if* they have *a common prior*. This result is of interest since there is disagreement over whether the common prior assumption in game theory has good theoretical foundations. Our analysis suggests that this assumption may have good evolutionary foundations.

It is important to emphasize that our analysis does not imply that agents will behave "rationally" in all situations. First, according to our results, when perceptions are observable, behavior will be effectively rational only with limited information, and may deviate from the Nash equilibrium of the full-information game. This is because suppression of information may be mutually beneficial for all agents. Second, our analysis has been limited to the case where there are no cognitive restrictions on behavior, and therefore should be interpreted as implying that there will not be systematic biases given the cognitive resources available to the agents. This does not rule out boundedly rational behavior because of cognitive limitations. Finally, as with all evolutionary analyses, our results apply in the "long run", and adjustment, especially after important changes in environments, may take a long time, during which behavior that is not neutrally or evolutionarily stable can be observed.

References

[1] Alchian, A. (1950) : "Uncertainty, evolution, and economic theory," *Journal of Political Economy*, 58, 211-221.

[2] Aumann, R. (1987): "Correlated equilibrium as an expression of Bayesian rationality," *Econometrica*, 55, 1-18.

[3] Aumann, R. (1998): "Common priors: A reply to Gul", *Econometrica*, 66-4, 929-938.

[4] Aumann, R. and A. Brandenburger (1995): "Epistemic conditions for Nash equilibrium," *Econometrica*, 63, 1161-80.

[5] Banerjee, A. and J. Weibull (1993) : "Evolutionary selection with discriminating players, " Economics Department Working Paper 1616, Harvard University.

[6] Banerjee, A. and J. Weibull (2000) : "Neutrally stable outcomes in cheap-talk coordination games," *Games and Economic Behavior*, 32-1, 1-24.

[7] Becker, G. (1962): "Irrational Behavior and Economic Theory" *Journal of Political Economy*, 70,1-13.

[8] Bendor, J., D. Mookherjee, and D. Ray (2001): "Reinforcement learning in repeated interaction games," *Advances in Theoretical Economics*, Vo1. 1, No. 1, Article 3.

[9] Blume, E. and D. Easley (1992): "Evolution and Market Behavior," *Journal of Economic Theory*, 58, 9-40.

[10] Bomze I. and J. Weibull (1995): " Does Neutral Stability Imply Lyapunov Stability?" *Games and Economic Behavior*, Vol. 11, No. 2 , 173-92.

[11] Dekel, E., J. Ely, and O. Yilankaya (1998) : "Evolution of preferences and Nash Equilibrium," mimeo, Northwestern University.

[12] Dekel, E. and F. Gul (1997): "Rationality and knowledge in game theory," in Kreps and Walls (editors): *Advances in economics and econometrics: theory and applications, Seventh World Congress*, Vol. 1.

[13] De Long, J., A. Shileifer, L. Summers, and R. Waldman (1990) "Noise Trader Risk in Financial Markets" *Journal of Political Economy* Vol. 98, No. 4, 703-38.

[14] Ely, J. and O. Yilankaya (1999) : "Evolution of preferences," *Journal of Economic Theory.*

[15] Friedman, M. (1953): "The Methodology of Positive Economics" in *Essays in Positive Economics*, Chicago: University of Chicago Press.

[16] Harrison, M. and D. Kreps (1978): "Speculative Investor Behavior in a Stock Market with Heterogenous Expectations" *Quarterly Journal of Economics*, 92, 323-36.

[17] Gul, F. (1998): "A comment on Aumann's Bayesian view", *Econometrica*, 66-4, 923-927.

[18] Kim Y.-G. and J. Sobel (1995): "An evolutionary approach to pre-play communication," *Econometrica*, 63, 1181-1193.

[19] Kockesen, L., E. Ok, and R. Sethi (2000) : "The strategic advantage of negatively interdependent preferences," *Journal of Economic Theory*, 92, 274-299.

[20] Maynard-Smith, John (1982): *Evolution and the Theory of Games*, Cambridge University Press, Cambridge.

[21] Morris, S. (1996): "Speculative Investor Behavior and Learning" *Quarterly Journal of Economics*, Vol. 111, No. 4, 1111-33.

[22] Ok, E. and F. Vega-Redondo (2000) : "On the evolution of individualistic preferences: An incomplete information scenario," *Journal of Economic Theory*, forthcoming.

[23] Robson (1990) : "Efficiency in evolutionary games: Darwin, Nash and the secret handshake," *Journal of Theoretical Biology*, 144, 379-396.

[24] Samuelson Larry (1997): *Evolutionary games and equilibrium selection* Cambridge, Mass. : MIT Press.

[25] Samuelson Larry (2001): "Information-based relative consumption effects," mimeo.

[26] Sandroni, A. (2000): "Do markets favor agents able to make accurate predictions?," *Econometrica*, 68-6, 1303-1341.

[27] Stenneck, J. (2000): "The Survival Value of Assuming Others to be Rational," *International Journal of Game Theory*, Vol. 29, No. 2, 147-163.

[28] Van Den Steen (2001): "Organizational beliefs and managerial vision," mimeo, Stanford University.

[29] Warneyrd K. (1993): " Cheap Talk, Coordination, and Evolutionary Stability" *Games and Economic Behavior* v5, n4 (October): 532-46.

[30] Weibull J. (1997): *Evolutionary Game Theory*, MIT Press, Cambridge.

[31] Yildiz, M. (2000) : "Sequential bargaining without a common prior on the recognition process," mimeo, MIT.