Seminar Paper No. 571

**THE SURVIVAL VALUE OF
ASSUMING OTHERS TO BE RATIONAL**

by

Johan Stennek

INSTITUTE FOR INTERNATIONAL ECONOMIC STUDIES

University of Stockholm

ISSN 0347-8769

Seminar Paper No. 571

**THE SURVIVAL VALUE OF
ASSUMING OTHERS TO BE RATIONAL**

by

Johan Stennek

**Seminar Papers are preliminary material circulated to
stimulate discussion and critical comment.**

March 1994

**Institute for International Economic Studies
S-106 91 Stockholm
Sweden**

The Survival Value of Assuming Others to be Rational*

by

Johan Stennek

March 21, 1994

Abstract

Are the cognitive, or information processing assumptions of non-cooperative game theory consistent with biological evolution? Could rationality and common knowledge of rationality have evolved as a result of natural selection? Previous studies, Banerjee and Weibull [3] and Stahl [29], give some support for the survival of "rational play," but not for information processing: "being right is as good as being smart". In contrast, this study provides an example of a multiple-situation environment where non-rational agents are eliminated. Further, evolution only selects agents with knowledge of rationality. These results contribute to the evolutionary foundation of *iterated elimination of strictly dominated strategies* and, in contrast to previous studies, to the cognitive interpretation of this solution concept.

*I thank Jörgen Weibull for helpful discussions. I am grateful to Ken Binmore, Jonas Björnerstedt, Henrik Horn, Harald Lang, Stefan Lundgren and Karl Wärneryd for their comments. I also thank the participants in the Workshop in Mathematical Economics and Game Theory, Institute of Economics, University of Copenhagen; and the participants in the seminar at IIES. Financial support from the Tore Browaldh, Jan Wallander and Tom Hedelius Research Foundation is gratefully acknowledged.

1 Introduction

Are the cognitive, or information-processing assumptions of non-cooperative game theory consistent with biological evolution? More precisely, could rationality and common knowledge of rationality have evolved as a result of natural selection? Previous studies, Stahl [29] and Banerjee and Weibull [3], give some support for the survival of "rational play," but not for information-processing: "being right is as good as being smart." In contrast, this study provides an example of a multiple-situation environment where non-rational agents are eliminated. Further, evolution only selects agents with knowledge of rationality. However, this process is "gradual" and it stops at a finite but "sufficiently high" level. These results contribute to the evolutionary foundation of *iterated elimination of strictly dominated strategies* and, in contrast to previous studies, to the cognitive interpretation of this solution concept.

By paying attention to the forces of natural selection one should be able to gain insight about human behavior, including human economic behavior. Natural selection does not, however, work on human behavior directly, but rather on the human psychological architecture. It is this innate psychological architecture, in turn, which produces behavior in a given situation. This is the view of evolutionary psychology (see for example Cosmides *et al.* [14]). One reason for adopting this view is that the forces of natural selection are slow. So, if we want to gain insights about human behavior in situations that occur today, but did not occur in "evolutionary" times, then we must rely on such an indirect mechanism.¹

The human psychology is studied using the cognitive, or information-processing approach (for an introduction see Anderson [1]). This means that the human mind is viewed as a set of information-processing rules governing, for example, perception, memories, rules of inference and learning mechanisms (see Cosmides

¹See Cosmides and Tooby [12] for a more detailed discussion of why the psychological level is needed in the discussion of evolution of behavior.

et al. [14, p.8]). I will use the term "cognitive design" to describe a specific set of information-processing rules. In the cognitive approach there is also a more specific idea that I will use. The human psychological architecture includes a so-called "theory of mind" module, used for theorizing about how other people reason (see Tooby and Cosmides [31]).

When I discuss if the cognitive assumptions of non-cooperative game theory are consistent with natural selection, I concentrate on the problems that arise from strategic interaction. I will use the replicator dynamic to model natural selection, and hence abstract from problems related to the biological machinery (see Hammerstein and Selten [18] and Selten [27]).

Non-cooperative game theory is built on two assumptions concerning human cognition. First, agents are assumed to be *rational*, in the sense of Savage [25]. According to this view, an agent maximizes his subjective expected utility. Second, agents are assumed to be *intentionalists*, in the sense of Dennett [16]. In forming beliefs about an opponent's choice, an agent bases his prediction on the assumption or theory of mind that his opponent is rational. A stronger version of intentionality is that agent *A* not only assumes that opponent *B* is rational, but also that *A* assumes that *B* assumes that his opponents are rational. By introducing more "belief-levels," in this way, a hierarchy of increasingly elaborated versions of intentionality is obtained².

²There are three differences between the idea of intentionality and the conventional notion of common knowledge of rationality. First, intentionality need not be *knowledge*. Intentionality is a probability-one belief, which may be true or false. Actually, Dennett describes intentionality as a short-cut method for predictions [16, p. 23], which works well but not perfectly [16, p. 28]. Second, only intentionality in the most elaborated version is *all-pervading*, *i.e.* prescribes the intentional strategy at all "belief-levels." Third, intentionality need not be *common*: Some agents in a population may be intentionalists while other agents form their beliefs in some other fashion, and still other agents do not form beliefs at all. Define knowledge as a probability-one belief that is true, then there exists common (and all-pervading) knowledge of rationality if and only if all agents are rational and intentionalists (in the most elaborated version). In the presence of rationality and common knowledge of rationality, iterated elimination of strictly dominated strategies is the relevant non-cooperative solution concept (see Bernheim [7]).

Related literature. In standard evolutionary game theory models³, one considers a *fixed* game in normal form. Further, *all* agents have a complete lack of information-processing powers. A "cognitive design" may be identified with the action that the agent is programmed to take⁴. In this framework it is only possible to discuss substantive rationality. The main result of interest here is that only iteratively strictly undominated actions survive, see Samuelson and Zhang [24, Thm. 1].

With this model as a starting point, one can introduce designs with more information-processing powers. Then, there are two types of questions, corresponding to Simons [28] distinction between *substantive* and *procedural* rationality. A substantive question is whether agents will *play* rationally, *i.e.* only choose iteratively strictly undominated (or rationalizable) actions in the long run. I mainly discuss the evolution of procedural rationality, which concerns how the surviving cognitive designs process information. There are many types of information an agent may process and condition his action on: (1) The game data, and in particular the payoffs. This is the basic information, considered in all studies. (2) The opponent's type (or the population distribution over types), in particular how the opponent processes information (see Banerjee and Weibull [3, sect. 2] and Stahl [29]). If information about the opponents type is not directly available, then an agent may consider indirect information: (3) The history of play, *e.g.* the actions taken in earlier matchings (see Banerjee and Weibull [3, sect. 3] and Dekel and Scotchmer [15]). (4) Pre-play communication, or cheap talk (see Banerjee and Weibull [5] and Robson [23]).

By and large, these studies give some support for substantive rationality, but not for procedural rationality. It is a foundation for the *as-if* approach: Without

³For a more detailed review of the discussions in evolutionary game theory about evolutionary selection of rationality, see Banerjee and Weibull [4].

⁴However, a possible interpretation of these studies is that they implicitly are built on the assumption that agents condition their behavior on the fact that they are in a given situation. A correct identification of the situation may be based on a considerable amount of information processing.

any processing of information, (some) agents take the same actions as they would if they were processing (payoff) information.

Purpose and results. The present discussion is aimed at finding support for information-processing, rather than substantive rationality. To this end, I change two of the less attractive assumptions from the earlier literature. *First*, I assume that agents must confront many different situations, described by strategic form games. Previous discussions have assumed interaction in a fixed and symmetric game. Hence, one of the factors that presumably gives procedural rationality a high survival value, *viz.* the ability to adjust behavior to the situation, is absent. Hereby the distinction between cognition and behavior becomes interesting. *Second*, I will assume that an agent can observe neither the cognitive design of his opponent, nor the proportion of cognitive designs in the population.

The second assumption is a potential problem. If a player cannot see the cognitive design of his opponent, how can he predict the opponent's choice? My view is that an agent must use some *theory* of how his opponent reasons. Since I discuss the cognitive assumptions of non-cooperative game theory I focus the discussion on versions of intentionality. Like Stahl [29], I define a hierarchy of increasingly intelligent designs, here called IQ_d for $d = 0, 1, 2, \dots$ and ∞ which reason analogously to the iterative levels of strict dominance. IQ_0 is non-rational and chooses any action with the same probability; IQ_1 is rational and chooses any strictly undominated action with equal probability; IQ_2 is rational and intentionalist, he believes right or wrong that his opponent is rational and chooses any 2-iterations strictly undominated action with equal probability *etc.*

An important aspect of these designs is that intentionality (at "belief-level" d) is a dogma to IQ_d , and that an IQ_d -agent may have "problems of inference." If an intentionalist observes other agents' behavior, he may observe violations of his rationality assumption and should accordingly (if rational in the sense of Savage) modify his theory of mind. (A minimum is to accept a small probability of

mistakes.) The insistence on intentionality and the consequent inference problem lies in the very core of non-cooperative game theory. Binmore [9] defines Homo Ludens, or game theoretic man as a species that is rational, and with common knowledge of rationality within the species. Moreover, it is argued that Homo Ludens "believes in rationality despite any evidence." It is the survival of this creature which I discuss here⁵.

To keep the discussion simple I will not consider any other competing theory of mind. This is also the main weakness of the present study: intentionality does not get a tough match. How would intentional designs compete with a cognitive design that is capable of considering several rival theories of the mind, and that is using information to discriminate between them? However, also in my model, the survival of intentionality designs is not without complications, due to their dogmatic predictions.

The main conclusions concern an "environment" where all types of symmetric as well as asymmetric situations (2×2 strategic form games) occur with positive probability: In contrast with previous studies, non-rationality (IQ_0) is eliminated. In a complex environment "you must be smart to be right." Moreover, evolution selects for intentionality. However, due to the problem of false predictions, the evolution is "gradual"; if IQ_1 is not present in the population, then IQ_0 may out-compete IQ_d , $d \geq 2$. Further, the evolution stops discriminating between designs at a finite, $d = 2$, level. This level is, however, sufficiently high for guaranteeing rational play in 2×2 -games.

⁵Note also: (1) The motivation of the solution concepts sequential equilibrium (see Kreps and Wilson [19]) and perfect equilibrium (see Selten [26]). Deviations from rationality are thought of as highly improbable or slight mistakes, made by otherwise rational agents. No competing theories of human behavior are offered. (2) The problem to motivate backwards induction (see Aumann [2], and Bicchieri and Antonelli [8]).

2 The Model

Consider population with infinitely many agents, where each member is "neurally programmed" with a certain cognitive design d from a set of cognitive designs D^* . At each moment, individuals from the population are randomly matched into pairs for a situation of strategic interaction. Each interaction is described by a one-shot strategic-form game. The roles $r \in \{A, B\}$ in the game are allotted randomly. The subject of the study is the selection of cognitive design, and to differentiate (general-purpose) cognition from behavior (in specific situations) agents must face many different games. All games are finite, and to simplify I assume that both roles only have two available actions, $S_r = \{r_1, r_2\}$ for all r . Hence, the only formal difference between one game and another is the payoff function, $v : S_A \times S_B \rightarrow R^2$, or displayed in the game matrix:

$$\begin{array}{cc}
 & \begin{array}{cc} B_1 & B_2 \end{array} \\
 \begin{array}{c} A_1 \\ A_2 \end{array} & \begin{pmatrix} (a_1, b_1) & (0, 0) \\ (0, 0) & (a_2, b_2) \end{pmatrix}
 \end{array} \tag{1}$$

Note that I, without loss of generality for best reply behavior, have normalized all non-main-diagonal payoffs to zero. A matching game $(\{A, B\}, S_A \times S_B, v)$ is hence completely described by the payoff structure

$$g = (v_A(A_1, B_1), v_A(A_2, B_2), v_B(A_1, B_1), v_B(A_2, B_2)) = (a_1, a_2, b_1, b_2). \tag{2}$$

An *environment* is characterized by a function $e : R^4 \rightarrow R$ describing for all games g the probability density $e(g)$ with which they occur. The support of e is bounded by $0 \leq |v_r(\cdot, \cdot)| \leq u_{\max}$.

An agent cannot observe his opponent's cognitive design or even the population shares of different designs. Moreover, an agent does not take into account any aspect of history. An agent "knows" his and his opponents' payoffs in every situation, *i.e.* he "knows" g and whether he occupies role A or B .

However, although payoff information is *available* to all agents, this information may or may not be *used*, depending on the cognitive design. A design d may be identified with its implied strategy, where a strategy is the function $f^d : R^4 \times \{A, B\} \rightarrow \Delta^2 = \{(f_1^d, f_2^d) \in R_+^2 : f_1^d + f_2^d = 1\}$, that describes the frequency with which the design chooses actions r_1 and r_2 , as a function of payoffs only.

Payoff variations do not only describe variations in some quantitative aspect of the outcome in a certain situation. Rather, the variations in payoffs mainly reflect qualitative differences between different situations (*e.g.* market situations *vs.* family situations) in the environment. The actions of the underlying situation are given abstract labels A_1 and A_2 and so on. It is presumed that this labeling is arbitrary (the action *turn left* is sometimes called A_1 and sometimes A_2). As a consequence, a rule like "always choose A_1 " will not make sense in the present model even as a description of unintelligent play, since sometimes A_1 will mean *turn right*, sometimes *turn left*, sometimes *buy Volvo*, and sometimes *buy SAAB*. Rather, I will presume that no design condition its behavior on the labels of the actions - there is no "framing" effect. Formally, this condition is given by:

$$f_1^d(g, r) = f_2^d(g', r) \text{ if } \begin{aligned} &g = (a_1, a_2, b_1, b_2) \text{ and } g' = (-a_1, -a_2, -b_2, -b_1) \\ &\text{or} \\ &g = (a_1, a_2, b_1, b_2) \text{ and } g' = (-a_2, -a_1, -b_1, -b_2) \end{aligned} \quad (3)$$

The first condition guarantees that players do not condition their behavior on the labels of the role- A actions. To see this, it is sufficient to notice that the game in (1) and the second game matrices in (4) are equivalent representations in the normalized format of some underlying situation.

		B_1	B_2			B_1	B_2
$\xrightarrow{\text{re-label}}$	A_1	(0, 0)	(a_2, b_2)	$\xrightarrow{\text{normalize}}$	A_1	($-a_1, -b_2$)	(0, 0)
	A_2	(a_1, b_1)	(0, 0)		A_2	(0, 0)	($-a_2, -b_1$)

(4)

The second condition guarantees that players do not condition on the labels of the role- B actions. Note that design d uses the same strategy f^d in both role A and in role B . It is presumed that f^d satisfies some conditions to rule out framing effects resulting from the arbitrary labeling of roles. These will however not be used.

Recall that an action r_i is *strictly dominated* for role r (in a generic⁶ 2×2 -game) if r_j gives strictly higher payoff to r independently of what action role p takes:

$$v_r(r_i, p_k) < v_r(r_j, p_k) \quad \forall p_k \in S_p. \quad (5)$$

Otherwise r_i is *strictly undominated*. Let $S_r^0(g) = S_r$ and define $S_r^n(g)$ recursively by

$$S_r^n(g) = \{r_i \in S_r^{n-1}(g) : \text{if } r_j \in S_r^{n-1}(g) \text{ then it is not the case that } v_r(r_i, p_k) < v_r(r_j, p_k) \quad \forall p_k \in S_p^{n-1}(g)\}. \quad (6)$$

Let also $S_r^\infty(g) = \bigcap_{n=0}^\infty S_r^n(g)$. If $r_i \in S_r^n$ I will say that r_i is *n-iterations strictly undominated*. In the case $n = \infty$ the action is *iteratively strictly undominated*⁷.

The Cognitive Designs. The cognitive designs $d = 1, 2, \dots$ and ∞ are defined in the following way. IQ_0 is intended to be the extreme opposite of rationality (in the sense of Savage). What is the *extreme* opposite of rationality? From a "substantive" point of view, the opposite is to choose the alternative with lowest expected utility. From an information-processing point of view, the opposite is to choose an alternative independent of its consequences. It is the

⁶To simplify, I will not consider situations where $a_1 = 0$ or $a_2 = 0$ or $b_1 = 0$ or $b_2 = 0$. These situations occur with probability zero, and will therefore not affect the evolution of the population.

⁷An action is a *best reply* to some action (profile) taken by his opponent(s), if there is no other action that gives him a higher payoff. An action is a *never best reply* if there is no action (profile) for the opponent(s) for which the action is a best reply. An action is *rationalizable* if it is not a "never best reply," nor a "never best reply" when other "never best replies" has been deleted. In two-role situations an action is strictly dominated if and only if it is not a best reply to any action taken by the opponent (see Weibull [33, Proposition 1]). Further, the set of iteratively strictly undominated actions is identical to the set of rationalizable actions (see Weibull [33, Corollary 1]).

latter view that is of interest here. Hence, IQ_0 does not process any relevant, *e.g.* payoff, information. The formal assumption is that the choice of action is distributed independently of payoff. This implies that IQ_0 always chooses all available actions with equal probability, *i.e.* one half⁸.

IQ_1 is rational in the sense of Savage. IQ_1 forms beliefs about the opponent's choice and maximizes expected utility with respect to these beliefs. IQ_1 processes information on his own, but not on the opponent's payoffs. The IQ_1 design always choose a best reply to some belief, *i.e.* a strictly undominated action. To simplify I assume that an agent with the IQ_1 design chooses among the (1-iterations) strictly undominated actions independently of payoffs, *i.e.* with equal probability.

IQ_2 is also rational in the sense of Savage. Moreover, IQ_2 is intentionalistic, *i.e.* beliefs are formed under the possibly false hypothesis that others also are rational (in the sense of Savage). IQ_2 uses data on both his own and the opponent's payoffs. To simplify I assume that an agent with the IQ_2 design chooses among the 2-iterations strictly undominated actions independent of payoffs, *i.e.* with equal probabilities⁹.

IQ_d for $d \geq 2$, is characterized by rationality and intentionality^d. To simplify I assume that all non-eliminated actions, d -iterations strictly undominated actions,

⁸To see that payoff independence implies equal probabilities, note that payoff independence is described by

$$f^0(g, r) = f^0(g', r) \quad \forall g, g' \in R^4. \quad (7)$$

Further, since $f_1^0(g, r) + f_2^0(g, r) = 1$, $\forall g$, the first part of condition (3) may be written $f_1^0((a_1, a_2, b_1, b_2), r) = 1 - f_1^0((-a_1, -a_2, -b_2, -b_1), r)$, or if $g = (a_1, a_2, b_1, b_2)$ and $g' = (-a_1, -a_2, -b_2, -b_1)$ then $f_1^0(g, r) + f_1^0(g', r) = 1$, and due to payoff independence $f_1^0(g, r) = \frac{1}{2}$, $\forall g \in R^4$. Hence, IQ^0 always choose both actions with equal probabilities,

$$f^0(g, r) = \left(\frac{1}{2}, \frac{1}{2}\right) \quad \forall g \in R^4. \quad (8)$$

⁹Note that IQ_2 assumes that his opponent is rational. This does not mean that he assumes his opponent to have design IQ_1 , which is also characterized by the auxiliary assumption that the choice among (1-iterations) strictly undominated actions is independent of payoffs. In the same way, although IQ_3 assumes that his opponent is intentionalist², it does not assume that the opponent has design IQ_2 , and so on.

are chosen independently of payoffs, *i.e.* with equal probabilities.

The designs are defined in terms of a "hierarchy" of decision theoretic assumptions¹⁰. They are ranked in the same order according to the amount of information they use (the number of variables) and according to the amount of reasoning (the number of information-processing steps, *i.e.* the number of iterations in eliminating strictly dominated strategies) they perform. If, however, we would measure, or rather partially order, the rationality of agents by the "size" of the set of information to which the agent plays a *best reply*, then IQ_2 is not more rational than IQ_1 or IQ_0 , since IQ_2 does not necessarily play a best reply, due to his dogmatic predictions.

To summarize, the strategy used by IQ_d , $d \in D^*$ in role $r \in \{A, B\}$ is described by

$$f^d(g, r) = \begin{cases} (1, 0) & \text{if } S_r^d(g) = \{r_1\} \\ (\frac{1}{2}, \frac{1}{2}) & \text{if } S_r^d(g) = \{r_1, r_2\} \\ (0, 1) & \text{if } S_r^d(g) = \{r_2\} \end{cases} \quad (9)$$

For the purpose of dominance and best-reply analysis the generic games may be classified into $2^4 = 16$ classes dependent on the sign of payoffs.

¹⁰The cognitive designs are not completely characterized by rationality (IQ_1), or rationality and intentionality (IQ_2), and so on. I have to add some *ad hoc* assumptions. To make the model as simple as possible, IQ_d is assumed to first eliminate all his d -iterations strictly dominated actions, then choose all remaining actions independent of payoff. An alternative is to place the *ad hoc* element in the beginning of the choice procedure. Let IQ'_1 's beliefs about his opponents' choice be independent of payoffs, and let IQ'_1 maximize against these beliefs. In the same way, let IQ'_2 's beliefs about his opponents' beliefs, about his choice be independent of payoffs, *et cetera*. The complexity of such analysis is higher due to the fact that the best-reply correspondence may cycle.

	a_1	a_2	b_1	b_2		
1	+	-	+	-	<i>Non - Strategic</i>	
2	+	-	+	+	<i>Simple Strategic</i>	
3	+	-	-	-	<i>Simple Strategic</i>	
4	+	+	+	-	<i>Simple Strategic</i>	
5	-	-	+	-	<i>Simple Strategic</i>	
6	+	-	-	+	<i>Non - Strategic</i>	
7	+	+	+	+	<i>Coordination</i>	
8	-	-	+	+	<i>Conflict</i>	(10)
9	+	+	-	-	<i>Conflict</i>	
10	-	-	-	-	<i>Coordination</i>	
11	-	+	+	-	<i>Non - Strategic</i>	
12	+	+	-	+	<i>Simple Strategic</i>	
13	-	-	-	+	<i>Simple Strategic</i>	
14	-	+	+	+	<i>Simple Strategic</i>	
15	-	+	-	-	<i>Simple Strategic</i>	
16	-	+	-	+	<i>Non - Strategic</i>	

Note that most classes of situations are redundant. For example all situations in classes 1, 6, 11 and 16 can be transformed into class 1, by re-labeling roles and actions. Similarly, re-labeling roles and actions shows that there is only need to consider four large classes of situations, *viz.* the classes NON-STRATEGIC (classes 1, 6, 11 and 16), SIMPLE STRATEGIC (classes 2, 3, 4, 5, 12, 13, 14 and 15), COORDINATION (classes 7 and 10), and CONFLICT (classes 8 and 9). Behavior can be summarized as follows.

In NON-STRATEGIC situations both roles have a strictly dominating action. In this terminology the Prisoners' Dilemma game is a NON-STRATEGIC situation! In NON-STRATEGIC situations an IQ_0 agent chooses both actions with equal

probability, whereas agents with designs IQ_1 and IQ_d , $d \geq 2$, choose the strictly dominating action with probability one (see expression (9)).

In SIMPLE STRATEGIC situations exactly one role has a strictly dominating action. In classes 2, 3, 14, and 15 the role A has a strictly dominating action, while in classes 4, 5, 12, and 13 the role B has a strictly dominating action. In the role with a strictly dominating action, an IQ_0 agent chooses both actions with equal probability, whereas agents with designs IQ_1 and IQ_d , $d \geq 2$, choose the strictly dominating action with probability one (see expression (9)). If the opponent has a strictly dominating action, an IQ_0 and an IQ_1 agent chooses both actions with equal probability, whereas an agent with IQ_d , $d \geq 2$, chooses whatever action is the best reply to the opponent's strictly dominating action (see expression (9)).

In COORDINATION situations no role has a strictly dominating action and there are three equilibria, two "pure" and one "mixed." An example is the game "battle of the sexes," but also the "dove-hawk" game. In COORDINATION situations all types choose both actions with equal probability (see expression (9)).

In CONFLICT situations no role has a strictly dominating action and there is only one (mixed) equilibrium. An example is the game "matching pennies." In CONFLICT situations all types choose both actions with equal probability (see expression (9)).

The Meta-Game. Evolutionary selection operates on cognitive designs, rather than directly on actions. To apply evolutionary game theory, I define a symmetric *meta-game* where the cognitive designs are the meta-strategies, and where meta-payoffs are the expected payoffs from an interaction between two designs.

Let $D = N = \{0, 1, 2, \dots\}$ and $D^* = D \cup \{+\infty\}$ be the set of designs. Let x_d , $d \in D^*$ denote the population shares of these designs. A population state at time t is described by the vector $(x_0(t), x_1(t), x_2(t), \dots) \in \Delta$ where

$\Delta = \left\{ x \in R_+^\infty : \sum_{d \in D} x_d \leq 1 \right\}$. Further, $x_{+\infty} = 1 - \sum_{d \in D} x_d$. Let $\dot{x}_d$ denote the time derivative of x_d . Let

$$u(d, b, g, A) = \sum_{i=1}^2 \sum_{j=1}^2 v_A(A_i, B_j) f_i^d(g, A) f_j^b(g, B) \quad (11)$$

denote the expected payoff to an agent with design $d \in D^*$, occupying role A , when he meets an opponent with design $b \in D^*$, occupying role B , in situation $g \in R^4$, and similarly for $u(d, b, g, B)$. Roles are allotted randomly with equal probabilities. Let

$$\bar{u}(d, b, g) = \frac{1}{2} \cdot u(d, b, g, A) + \frac{1}{2} \cdot u(d, b, g, B) \quad (12)$$

denote the expected payoff to an agent with design d when he meets an opponent with design b in situation g . Further, situations are drawn from distribution $e(g)$ with bounded support. Let

$$\bar{\bar{u}}(d, b | e) = \int \bar{u}(d, b, g) de(g) \quad (13)$$

denote the expected payoff to an agent with design d when he meets an opponent of design b in environment e . Further, every agent meets an opponent with design $b \in D^*$ with probability x_b . Let

$$\bar{\bar{\bar{u}}}(d | x, e) = \sum_{b \in D^*} \bar{\bar{u}}(d, b | e) \cdot x_b \quad (14)$$

denote the expected payoff to an agent of design d in environment e when the population state is x . Note that $0 \leq |u|, |\bar{u}|, |\bar{\bar{u}}|, |\bar{\bar{\bar{u}}}| \leq u_{\max}$.

Now, for a fixed environment e , define a two-person symmetric meta-game, denoted G_e , by letting each cognitive design constitute a (meta) strategy. Recall that $\bar{\bar{u}}(d, b | e)$ is the expected payoff to d in a meeting with b when roles are allotted randomly and the situation is drawn from density e . Hence,

$$G_e = (\{1, 2\}, D^* \times D^*, (\bar{\bar{u}}(d, b | e), \bar{\bar{u}}(b, d | e))) \quad (15)$$

Standard non-cooperative as well as evolutionary game theory may be used to analyze the meta-game. The evolutionary interpretation is of course that nature, rather than individuals, selects strategy, *i.e.* cognitive design.

Natural Selection. Natural selection is a process whereby nature "selects" one design $d \in D^*$ of an organism over other designs, in D^* , depending on how well it solves an adaptive problem, a problem that affects reproduction. I describe natural selection by the replicator dynamic, presupposing that the offspring inherits an exact copy of its single parent's design. Further, when I say that the human cognitive design has been selected by nature, I implicitly view this design as the *limit* result of the replicator dynamic. For a discussion of the abstractions made in this model see for example Hammerstein and Selten [18] and Selten [27].

An important question concerns my interpretation of payoffs as both utility and fitness. According to Symons [30, p.139], human beings as well as all other organisms have been designed by selection to strive for *specific* goals, and not the *general* goal of maximizing reproduction. However, preferences do relate positively to fitness: individuals who detected and *liked* sugar produced, on average, more progeny than did individuals who preferred the taste of green fruit. I will abstract from the problem of adapted preferences and their relation to fitness by assuming (1) that preferences already have been selected, and (2) that utility is equal to fitness. (It is sufficient to assume a positive relation.) Moreover, I will assume that preferences are common knowledge - which is a natural "corollary" assumption to the first two.

Hence, payoffs now represent reproductive fitness measured by the expected (approximately equal to average, in large populations) number of offspring. Hence, see Weibull [32, p.17],

$$\varphi_d(x, e) \equiv \frac{\dot{x}_d}{x_d} = \bar{\bar{u}}(d | x, e) - \sum_{b \in D^*} \bar{\bar{u}}(b | x, e) \cdot x_b \quad (16)$$

describes the evolution of population shares over time in meta-game G_e .

3 Symmetric Environments

Since much attention in evolutionary game theory has been focused on *symmetric* situations or environments, where $a_1 = b_1$ and $a_2 = b_2$ (for all g such that $e(g) > 0$) I will start by discussing these. In symmetric-situation environments all situations are either NON-STRATEGIC (classes 1 and 16) or COORDINATION (classes 7 and 10) situations. However, the first conclusion concerns all environments. I will say that two designs d and b are *behaviorally equivalent* in a meta-game G_e if they prescribe the same action in all "occurring" situations, *i.e.* if $f^d(g, r) = f^b(g, r)$ for all g with $e(g) > 0$, and all $r \in \{A, B\}$. Since I only consider environments with 2×2 - games the first observation and conclusion are obvious:

Observation 1 *IQ_d for $d \geq 2$ are behaviorally equivalent in all meta-games G_e .*

Hence, $\bar{\bar{u}}(d | x, e)$ are equal for all $d \geq 2$, and we may conclude:

Conclusion 1 *The population share of IQ_d in relation to the share of IQ_b , where $d, b \geq 2$, is constant over time.*

Proof. Define $g : \text{int}(\Delta) \rightarrow R$ by $g(x) = \log(x_d) - \log(x_b)$. Then $\dot{g}(x) = \dot{x}_d / x_d - \dot{x}_b / x_b = \bar{\bar{u}}(d | x, e) - \bar{\bar{u}}(b | x, e) = 0$. Hence $g(x(t))$ and x_d/x_b is constant. $\square$

Hence, evolution does not discriminate between high levels of intentionality, and common knowledge of rationality will not prevail even in the long run. In 2×2 - games, intentionality² (*if* selected for) is however sufficient to motivate the use of elimination of strictly undominated strategies, as a solution concept. Although trivial, this result is included here to be contrasted to Stahl's [29] model, where results are derived under the exogenous restriction that only finitely smart designs

are present in the population. Further, it is an open question whether the removal of this restriction would make evolution select for the "infinitely smart" design.

To simplify the subsequent presentation, let $y_0 = x_0$, $y_1 = x_1$ and $y_2 = \sum_{d \geq 2} x_d$, and consider the dynamic of y on Δ in R^3 . Since, IQ_1 and IQ_2 make identical choices in all $g \notin \text{SIMPLE STRATEGIC}$, we have

Observation 2 *IQ_1 and IQ_2 are behaviorally equivalent in a meta-game G_e with $\Pr\{\text{SIMPLE STRATEGIC}\} = 0$.*

Intuitively, only in SIMPLE STRATEGIC situations, when the opponent has a strictly dominating action, can an agent gain anything from a correct prediction of the opponent's action, or lose something from a false prediction. Moreover,

Observation 3 *IQ_0 is strictly dominated by IQ_1 in a meta-game G_e with $\Pr\{\text{NON-STRATEGIC} \cup \text{SIMPLE STRATEGIC}\} > 0$.*

Proof. For all $d \in D^*$ we have $\bar{u}(0, d | e) < \bar{u}(1, d | e)$, for all e such that $\Pr\{\text{NON-STRATEGIC} \cup \text{SIMPLE STRATEGIC}\} > 0$. First, note that $\bar{u}(0, d, g) < \bar{u}(1, d, g)$, for all $g \in \text{NON-STRATEGIC} \cup \text{SIMPLE STRATEGIC}$. If $g \in \text{NON-STRATEGIC}$ then $u(0, d, g, r) < u(1, d, g, r)$ for both roles since both roles have a strictly dominated action. If $g \in \text{SIMPLE STRATEGIC}$ then $u(0, d, g, r) < u(1, d, g, r)$ for the role which has a strictly dominated action (and $u(0, d, g, r) = u(1, d, g, r)$ for the other role, since then $f^0(g, r) = f^1(g, r)$). It is strictly better to choose the strictly dominated action with zero probability, independently of how the opponent plays (in a generic game). Second, note that $\bar{u}(0, d, g) = \bar{u}(1, d, g)$, for all $g \notin \text{NON-STRATEGIC} \cup \text{SIMPLE STRATEGIC}$, since then $f^0(g, r) = f^1(g, r)$. $\square$

Strictly dominated (meta) strategies die out in the replicator dynamic (see Samuelson and Zhang [24, Thm. 1]), and we can conclude:

Conclusion 2 *In a symmetric-situation environment, where NON-STRATEGIC situations occur with positive probability, and where the population share of IQ_d for some $d \geq 1$ is positive, the population share of the IQ_0 design converges to zero. Further, the population share of IQ_1 , in relation to the share of IQ_d , where $d \geq 2$ is constant over time.*

Proof. (The proof is adapted from Weibull [32, Prop. 4.5].) Since $\Pr\{\text{NON-STRATEGIC}\} > 0$ and $\Pr\{\text{SIMPLE STRATEGIC}\} = 0$, IQ_0 is strictly dominated by IQ_1 (observation 3), and hence by IQ_2 (observation 2), and hence by IQ_d , $d \geq 2$ (observation 1). Show that $y_0(t) \rightarrow 0$. Suppose $y_d > 0$ for $d \geq 1$. Let $\varepsilon = \min\{\bar{\bar{u}}(d | y, e) - \bar{\bar{u}}(0 | y, e) : y \in \Delta\}$. The minimum exists since the payoff function is continuous and Δ is compact. Note that $\varepsilon > 0$ since $\bar{\bar{u}}(d | y, e) > \bar{\bar{u}}(0 | y, e)$ for all $y \in \Delta$. Define $g : \text{int}(\Delta) \rightarrow \mathbb{R}$ by $g(y) = \log(y_d) - \log(y_0)$. Then $\dot{g}(y) = \frac{\dot{y}_d}{y_d} - \frac{\dot{y}_0}{y_0} = \bar{\bar{u}}(d | y, e) - \bar{\bar{u}}(0 | y, e)$. Note that $\dot{g}(y) \geq \varepsilon > 0$, $\forall y \in \Delta$. Hence $g(y(t))$ increases monotonically towards $+\infty$ from any $y(0) \in \Delta$, implying that $y_0(t) \rightarrow 0$. Show that $y_1/y_d = \text{constant}$ in the same way as in Conclusion (1). $\square$

Since IQ_d , $d \geq 1$ do not take any strictly dominated actions (or iteratively strictly dominated actions when the environment is symmetric) this conclusion is in line with the substantive rationality result of evolutionary game theory: iteratively strictly dominated actions are not played in the long run (see Samuelson and Zhang [24, Thm. 1]). It is however different from previous evolutionary results on procedural rationality (see Banerjee and Weibull [3] and Stahl [29]). In the present model primitive designs do not survive. The intuition is that, in a complex environment, you must be smart to be right in all situations.

In some sense this result is also contrary to Stahl's [29, p.15] "reasonable conjecture." Stahl argues that the fundamental characteristic that distinguishes any

smart_n for $n > 0$ agent from a smart_0 agent is not the ability to know the game data (payoffs), but the ability to think about how other agents will behave. Stahl conjectures that smartness has superior survival fitness over smart_0 agents in a multiple-situation environment. The present observation indicates that processing of (one's own) payoff information, but *not* strategic reasoning, has superior survival fitness. However, the complexity-reducing symmetry assumption is crucial.

4 Asymmetric Environments

The rest of the discussion concerns environments where asymmetric situations occur with positive probability. The first thing to note is that IQ_d , $d \geq 2$, does not strictly dominate IQ_1 in all environments (meta-games). Moreover, IQ_d , $d \geq 2$, does not even strictly dominate IQ_0 in all environments (meta-games), although IQ_d , $d \geq 2$, always chooses a strictly undominated action, while IQ_0 just picks an action at random and may choose a strictly dominated action. The reason is that the IQ_d , $d \geq 2$, designs may choose an action based on wrong predictions about the choice of an IQ_0 agent. In some SIMPLE STRATEGIC situations, for example for agent A in

$$\begin{array}{cc}
 & \begin{array}{cc} B_1 & B_2 \end{array} \\
 \begin{array}{c} A_1 \\ A_2 \end{array} & \begin{pmatrix} (\varepsilon, \varepsilon) & (0, 0) \\ (0, 0) & (M, -\varepsilon) \end{pmatrix}
 \end{array} \tag{17}$$

where $M > \varepsilon > 0$ this false prediction may be crucial. If B is IQ_0 , then A will earn a higher payoff if he is IQ_0 than if he is IQ_2 , $u(0, 0, g, A) = \frac{1}{4}M + \frac{1}{4}\varepsilon > \frac{1}{2}\varepsilon = u(2, 0, g, A)$. Moreover, there exist environments such that IQ_0 out-competes IQ_d , for all $d \geq 2$, even if the initial aggregate population share of IQ_d , $d \geq 2$, is very high, provided the population share of IQ_1 is zero.

Conclusion 3 *There exist environments (meta-games) such that if $y_1(0) = 0$, then $y_0(t) \rightarrow 1$ for all $y_0(0) \geq \lambda$, where λ may be close to zero.*

Proof. Consider a meta game with $\Pr\{(\varepsilon, M, \varepsilon, -\varepsilon)\} = 1$. To make things interesting let $M > 3 \cdot \varepsilon$. Take $\lambda > \frac{4 \cdot \varepsilon}{M + \varepsilon} \in (0, 1)$. Let $\Delta_\lambda = \{y \in \Delta : y_1 = 0 \text{ and } y_0 \in [\lambda, 1]\}$ and $d \geq 2$. Let $\varepsilon_\lambda = \min\{\bar{\bar{u}}(0 | y, e) - \bar{\bar{u}}(d | y, e) : y \in \Delta_\lambda\}$. The minimum exists since the payoff function is continuous and Δ_λ is compact. Note that $\varepsilon_\lambda > 0$ since $\bar{\bar{u}}(0 | y, e) > \bar{\bar{u}}(d | y, e)$ for all y such that $y_1 = 0$ and $y_0 \geq \lambda$. Define $g : \text{int}(\Delta_\lambda) \rightarrow \mathbb{R}$ by $g(y) = \log(y_0) - \log(y_d)$. Then $\dot{g}(y) = \frac{\dot{y}_0}{y_0} - \frac{\dot{y}_d}{y_d} = \bar{\bar{u}}(0 | y, e) - \bar{\bar{u}}(d | y, e)$. Note that $\dot{g}(y) > \varepsilon_\lambda > 0$, $\forall y \in \Delta_\lambda$. Note that Δ_λ is positively invariant since $\dot{g}(y) > 0 \Rightarrow \dot{y} > 0$. Moreover, $g(y(t))$ increases monotonically towards $+\infty$ from any $y(0) \in \Delta$ such that $y_1(0) = 0$ and $y_0(0) \geq \lambda > \frac{4 \cdot \varepsilon}{M + \varepsilon}$, implying that $y_d(t) \rightarrow 0$. $\square$

One may even suspect that the population share of the IQ_d , $d \geq 2$, designs can converge to zero in the original meta-game. However, IQ_1 strictly dominates IQ_0 , if NON-STRATEGIC or SIMPLE STRATEGIC situations occur with positive probability (observation 3). Hence, just as in symmetric-situation environments, IQ_1 will gradually out-compete IQ_0 . Moreover, IQ_d , $d \geq 2$, will gradually start to make correct predictions more often than false (in SIMPLE STRATEGIC situations), thereby earning a higher payoff than IQ_1 . As a consequence, IQ_d , $d \geq 2$, will start to out-compete IQ_1 . Formally,

Observation 4 *IQ_d , $d \geq 2$, strictly dominates IQ_1 in the reduced meta-game G'_e obtained from the original meta-game G_e by eliminating IQ_0 , in any meta-game G_e with $\Pr\{\text{SIMPLE STRATEGIC}\} > 0$.*

Proof. For $d \geq 1$ we have $\bar{\bar{u}}(1, d | e) < \bar{\bar{u}}(2, d | e)$, for all e such that $\Pr\{\text{SIMPLE STRATEGIC}\} > 0$. (Also, recall that IQ_d , $d \geq 2$, are behaviorally

equivalent.) First, note that $\bar{u}(1, d, g) < \bar{u}(2, d, g)$, for all $g \in \text{SIMPLE STRATEGIC}$. In the role with a strictly dominated strategy we have $u(1, d, g, r) = u(2, d, g, r)$ since then $f^1(g, r) = f^2(g, r)$. In the role without a strictly dominated strategy we have $u(1, d, g, r) < u(2, d, g, r)$. Assume for example that B_2 is 1-iterations strictly dominated and that A_2 is 2-iterations strictly dominated, then $u(1, d, g, A) = \frac{1}{2}v_A(A_1, B_1) + \frac{1}{2}v_A(A_2, B_1)$ and $u(2, d, g, A) = v_A(A_1, B_1)$ which is higher. Second, note that $\bar{u}(1, d, g) = \bar{u}(2, d, g)$, for all $g \notin \text{SIMPLE STRATEGIC}$, since then $f^1(g, r) = f^2(g, r)$. $\square$

Iteratively strictly dominated (meta) strategies die out in the replicator dynamic (see Samuelson and Zhang [24, Thm. 1]), and we can conclude:

Conclusion 4 *In an environment (meta-game) where SIMPLE STRATEGIC situations occur with positive probability, and where the initial population shares of the IQ_1 and IQ_d , for some $d \geq 2$, designs are positive, the population share of the IQ_0 and IQ_1 designs converges to zero over time in the replicator dynamic.*

Proof. (The proof is adapted from Weibull [32, Corollary 4.5.1].) Since $\Pr\{\text{SIMPLE STRATEGIC}\} > 0$, IQ_0 is strictly dominated (observation 3), and IQ_1 is iteratively strictly dominated (observation 4). Show that $y_0(t) \rightarrow 0$. Let $\varepsilon = \min\{\bar{\bar{u}}(1 | y, e) - \bar{\bar{u}}(0 | y, e) : y \in \Delta\}$. The minimum exists since the payoff function is continuous and Δ is compact. Note that $\varepsilon > 0$ since $\bar{\bar{u}}(1 | y, e) > \bar{\bar{u}}(0 | y, e)$ for all $y \in \Delta$. Define $g : \text{int}(\Delta) \rightarrow \mathbb{R}$ by $g(y) = \log(y_1) - \log(y_0)$. Then $\dot{g}(y) = \frac{\dot{y}_1}{y_1} - \frac{\dot{y}_0}{y_0} = \bar{\bar{u}}(1 | y, e) - \bar{\bar{u}}(0 | y, e)$. Note that $\dot{g}(y) \geq \varepsilon > 0$, $\forall y \in \Delta$. Hence $g(y(t))$ increases monotonically towards $+\infty$ from any $y(0) \in \Delta$, implying that $y_0(t) \rightarrow 0$. Show that $y_1(t) \rightarrow 0$. The state $y(t)$ approaches the face of the simplex where $y_0 = 0$. Hence, we can think of $y(T)$, for T sufficiently large, as an initial state arbitrarily close to this face, and, by continuity of $\bar{\bar{u}}$, apply the same proof as above, to the dynamics near this face. $\square$

5 The Punch line

In asymmetric environments: Non-rationality is eliminated, $y_0 \rightarrow 0$ (Conclusion 4). Moreover, evolution selects for intentionality, $y_1 \rightarrow 0$ (Conclusion 4). However, the evolution towards higher intentionality is "gradual" in the sense that IQ_1 plays an *essential* role to assist IQ_d , $d > 1$ (Conclusion 3). Further, the evolution towards higher intentionality stops discriminating between designs at a finite, $d = 2$, level (Conclusion 1). This level is however sufficient for guaranteeing rational play in 2×2 - games.

Note that IQ_d , $d \geq 2$, out-competes its less sophisticated rivals, although IQ_d , $d \geq 2$, itself are relatively unsophisticated: IQ_d , $d \geq 2$, does neither know its opponent's type, nor the relative frequencies of different types in the population. Moreover, an agent with an IQ_d , $d \geq 2$, design does not even learn anything about the cognition of others, from his individual experience of other agents play. An intentionalist believes in rationality, despite any evidence.

In symmetric environments: Non-rationality is eliminated, $y_0 \rightarrow 0$ (Conclusion 2). Although evolution stops discriminating between designs at level $d = 1$ (Conclusion 2), this level is sufficiently high for guaranteeing rational play.

So, the punch line is that evolution selects rationality and enough knowledge, but not necessarily common knowledge, of rationality to guarantee rational play. This result contributes to the evolutionary foundation of *iteratively strictly undominated strategies*, and the "usual" cognitive interpretation of this solution concept.

6 Concluding Remarks and Possible Extensions

Cosmides and Tooby [13] argue that the human mind consists of a collection of information-processing mechanisms that are *specialized* for solving various *families* of problems. For example, they report experimental evidence indicating that

subjects more often reason according to the rules of formal logic when confronted with situations of social exchange than in other situations. My assumption about general-purpose cognitive mechanisms used in all games may conflict with this view. If cognitive mechanisms are specialized for families of situations, then rules like "always choose A_1 " makes sense as a description of unintelligent but possibly well-adapted play. The action A_1 could, for example, be "cooperate" in the "social exchange" family of situations. That rule is not procedurally rational since the agent does not process any outcome (or payoff) information¹¹. The rule may however be well adapted, and substantially rational, if all social exchange situations are such that cooperation is profitable. I think it would be worthwhile also to consider a model with specialized information-processing mechanisms¹². Indeed, the level of generality is an interesting question *per se*: What levels of generality in information-processing mechanisms are favored by natural selection? To be more specific, can specialized mechanisms decrease the amount of information-processing (and hence time consumption) that is needed for making a good choice?

It would be possible to generalize the idea of (simple) programmed behavior, like "always choose A_1 ," to a more complex function that prescribes an action based on outcome (or payoff) information. If behaviors can be as complex as the environment then, as Stahl [29, Section 5] argues, it is hard to escape the conclusion that "having the right behavior is as good as being smart." However, this should not be taken as an objection to the conclusions of the present study. To analyze Stahl's view we should distinguish two levels of discussion: The first level

¹¹Identifying the situation as one of social exchange may involve processing information. It also includes abstract thinking in deciding which action corresponds to the idea of cooperation. Note that the classification of situations to families (and classification of actions) should not depend on processing information about outcomes (or payoffs). If such processing should occur the classification would be redundant (actions could be conditioned on payoffs directly).

¹²Note, however, then the discussion would not concern a 'general purpose' game theory, but one applicable only to the aforementioned family of situations. Moreover, only those of today's situations that are members of some family, which also have members in hunter-gatherer times, can be analyzed with some evolutionary-psychology motivated game theory.

concerns what language we use to describe information-processing, and the second level concerns the actual procedure of information-processing. First, we may describe the same information-processing either in *mental* terms such as "maximizing subjective expected utility" and "theory of mind" or in terms of *behavioral hypotheticals* (if-contingency-then-action statements). Using mental terms simply gives an efficient, high-level language. As compared with a neurophysiological description of information-processing, behavioral hypotheticals also form a high-level language. Second, the actual information-processing programme may either be a "smart" procedure for choice, exemplified by the iterated elimination of strictly dominated strategies, or a "more mechanical" checking of the if-clause in some list of if-contingency-then-action statements. In the present model, from which for example, the time consumption of decision making is abstracted, it is simply not possible to distinguish "smart procedures for choice" and "programmed behavior": If the "smart procedures for choice" and the "programmed behavior" take the same information as input and prescribe the same action under all contingencies, then they will be represented by the same strategy in the meta-game¹³.

Note also that the distinction between higher and lower degrees of information-processing does *not* coincide with the distinction between controlled (conscious) and automatic processes (see Anderson [1]). The capacity of attention for separate tasks is limited and must be divided among competing processes. Making a skill more automatic is a "rational" way to economize on limited cognitive recourses.

It may seem paradoxical that it is the occurrence of relatively simple situations (NON-STRATEGIC and SIMPLE STRATEGIC) that drive the evolution to select rationality and "knowledge" of rationality. These cognitive devices are, however,

¹³For a further discussion see for example Block [10] or Fodor [17]. A more fundamental question is whether rationality, or intelligence, contains aspects that are not algorithmic, see for example Penrose [21].

too simple to have any "cutting power" in analyzing more complex situations (COORDINATION and CONFLICT). Consequently, the occurrence of more complex situations does not increase the relative fitness of rationality and "knowledge" of rationality. The problem is one of coordination. Intentionality results in some, but far from complete, coordination. If this coordination failure results in a loss of efficiency, there should be forces for further coordinating mechanisms to emerge. In the present study efficiency losses occur in COORDINATION situations. In at least some such situations, when decision making cannot effectively be decentralized to agents, a coordinating *culture*, which is something greater than and distinct from individual cognition, may do the job.

Another approach would be to have more coordination built into the cognitive designs. For example, one could introduce the *Theorist* design, which analyzes situations according to the Nash equilibrium concept. Another possibility would be to introduce the *Empiricist* design, which form beliefs in an extrapolative fashion. Consider a meta-game with only the *Empiricist* and *Theorist* designs competing against each other. My guess is that both players "choosing" to be *Theorists* is a strict Nash equilibrium of the meta-game. However, I do also believe that both players "choosing" to be *Empiricists* could be a strict Nash equilibrium, in some environments, or meta-games. Hence, both outcomes could be asymptotically stable in the replicator dynamics (see [32, propositions 3.1 and 4.2]).

I believe that it would be possible to generalize my results. Consider an environment of finite 2-role games. There are k_r^g strategies for role r , in game g . Define $\bar{k}$ as the smallest integer such that $k_r^g \leq \bar{k}$, $\forall g, r$. Let $\bar{\bar{k}} = 2 \cdot (\bar{k} - 1)$. The environment is supposed to be *rich* in the sense that all "types of situations" occur with positive probability. I want to show that the evolution of rationality (information-processing) is characterized by the following four properties: (1) non-rationality is eliminated, and (2) evolution selects for intentionality, however (3) the evolution towards higher intentionality is gradual, and further (4) the

evolution towards higher intentionality stops at a finite, but "sufficiently high" level. This should be proved by showing that only IQ_d , $d \geq \bar{k}$ survives iterated elimination of strictly dominated strategies, and that all IQ_d , $d \geq \bar{k}$ are behaviorally equivalent. However, to do so would require a slight change in the definition of the designs. When design d makes its b 'th iteration and eliminates some action r_i , the probability mass "owned" by r_i should be transferred only to those actions that strictly dominate r_i , and not to all uneliminated actions. To see why, consider the game matrix with the payoffs of role A displayed.

$$\begin{array}{cc}
 & \begin{array}{cc} B_1 & B_2 \end{array} \\
 \begin{array}{c} A_1 \\ A_2 \\ A_3 \end{array} & \begin{array}{cc} 1 & 1 \\ 0 & 0 \\ 2 & -2 \end{array}
 \end{array} \tag{18}$$

Let p be the probability of B_1 . Then IQ_0 would earn $\frac{4}{3}p - \frac{1}{3}$ and IQ_1 would earn $2p - \frac{1}{2}$ and IQ_0 would have a higher payoff if $p \leq \frac{1}{4}$. Hence, IQ_0 is not strictly dominated by IQ_1 in all possible environments. However, if IQ_1 transfers all probability from A_2 to A_1 then he would earn $\frac{4}{3}p$.

A more interesting generalization would be to consider games with three roles, and define a new type of cognitive design based on the independence axiom (of Bernheim [7]). Such a model would make it possible to discriminate between iterative dominance and rationalizability as non-cooperative solution concept.

An alternative to the idea of a "fixed world" is to assume that whatever action is chosen, the worst alternative will occur, and hence to choose the "maxi-min" action. Another alternative is to assume that whatever action is chosen, the most favorable alternative will occur, and hence to choose the "maxi-max" action. Would IQ_2 completely out-compete *Maxi-Min* and *Maxi-Max*?

References

- [1] ANDERSON, J. *Cognitive Psychology and its Implications*, Third edition, W. H. Freeman and Company, New York, 1990.
- [2] AUMANN, R. "Backwards Induction and Common Knowledge of Rationality," paper presented at the Nobel Symposium on Game Theory, Björkborn, Sweden, June 18-20, 1993.
- [3] BANERJEE, A. AND WEIBULL, J. "Evolutionary Selection and Rational Behavior," forthcoming in KIRMAN, A. AND SALMON, M. (eds.), *Rationality and Learning in Economics*, Blackwell.
- [4] BANERJEE, A. AND WEIBULL, J. "Evolution and Rationality: Some Recent Game-Theoretic Results," Research Papers in Economics, 1993:2 WE, Department of Economics, Stockholm University.
- [5] BANERJEE, A. AND WEIBULL, J. "Evolutionary selection with discriminating players," Research Papers in Economics, 1993:6 WE, Department of Economics, Stockholm University.
- [6] BARKOW, J.; COSMIDES, L. AND TOOBY, J. (eds.) *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, New York, Oxford, Oxford University Press, 1992.
- [7] BERNHEIM, D. "Axiomatic Characterization of Rational Choice in Strategic Environments," *Scandinavian Journal of Economics*, 88 (3), 473-488, 1986.
- [8] BICCHIERI, C. AND ANTONELLI, G. A. "Game-Theoretic Axioms for Local Rationality and Bounded Knowledge," paper presented at the Nobel Symposium on Game Theory, Björkborn, Sweden, June 18-20, 1993.
- [9] BINMORE, K. *Models of Man*, lecture held at the Department of Economics, Stockholm University, June 1993.

- [10] BLOCK, N. "What is Functionalism?" Block, N. (ed.) *Readings in Philosophy of Psychology*, Harvard University Press, Cambridge, Massachusetts, 1980.
- [11] CARLSSON, H. "How Common is Common Knowledge? Boundedly Sophisticated Players Play the Finitely Repeated Prisoners' Dilemma," in CARLSSON, H. AND LARSSON, B. (eds.) *Problems of the Mixed Economy: Cooperation, Efficiency and Stability*, Elsevier Science Publishers P.V. (North-Holland), 1990.
- [12] COSMIDES, L. AND TOOBY, J. "From Evolution to Behavior: Evolutionary Psychology as the Missing Link," in: DUPRÉ, J. (Ed.) *The Latest on the Best: Essays in Evolution and Optimality*. The MIT Press, Cambridge, Massachusetts, 1987.
- [13] COSMIDES, L. AND TOOBY, J. "Cognitive Adaptations for Social Exchange," in: BARKOW, J.; COSMIDES, L. AND TOOBY, J. (Eds.) *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, New York, Oxford, Oxford University Press, 1992.
- [14] COSMIDES, L.; TOOBY, J. AND BARKOW, J. "Introduction: Evolutionary Psychology and Conceptual Integration," in: BARKOW, J.; COSMIDES, L. AND TOOBY, J. (Eds.) *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, New York, Oxford, Oxford University Press, 1992.
- [15] DEKEL, E. AND SCOTCHMER, S. "On the Evolution of Optimizing Behavior," *Journal of Economic Theory* 57, 392-406 (1992).
- [16] DENNETT, D. *The Intentional Stance* MIT Press, Cambridge, Massachusetts, 1987.
- [17] FODOR, J. "The Mind-Body Problem," *Scientific American* Vol. 244 No. 1, January 1981.

- [18] HAMMERSTEIN, P. AND SELTEN, R. *Game Theory and Evolutionary Biology*, unpub. ms. May, 1993.
- [19] KREPS, D. AND WILSON, R. "Sequential Equilibria," *Econometrica*, Vol. 50, No. 4 (July, 1982).
- [20] MASTERTON, R.B.; HODOS, W. AND JERISON, H. (eds.) *Evolution, Brain and Behavior: Persistent Problems*. Lawrence Erlbaum Associates, Publishers. Hillsdale, New Jersey, 1976.
- [21] PENROSE, R. *The Emperor's New Mind*, Vintage, London, 1990.
- [22] PINKER, S. AND BLOOM, P., Natural Language and Natural Selection," in: BARKOW, J.; COSMIDES, L. AND TOOBY, J. (Eds.) *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, New York, Oxford, Oxford University Press, 1992.
- [23] ROBSON, A. J. "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology* 144, 389-396 (1990).
- [24] SAMUELSON, L. AND ZHANG, J. "Evolutionary Stability in Asymmetric Games," *Journal of Economic Theory* 57, 363-391 (1992).
- [25] SAVAGE, L. *The Foundation of Statistics*. New York: John Wiley and Sons, 1954.
- [26] SELTEN, R. "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *Int. Journal of Game Theory*, Vol. 4, Issue 1, 25-55 (1975).
- [27] SELTEN, R. "Evolution, Learning and Economic Behavior," *Games and Economic Behavior* 3, 3-24 (1991).

- [28] SIMON, H. "From Substantive to Procedural Rationality," in *Methods and Appraisal in Economics*. Ed. S. J. Latis, pp. 129-148, Cambridge University Press, 1976.
- [29] STAHL, D. "Evolution of Smart_n Players," *Games and Economic Behavior* Vol. 5, No. 4, October 1993, pp 604-617.
- [30] SYMONS, D. "On the Use and Misuse of Darwinism in the Study of Human Behavior," in: BARKOW, J.; COSMIDES, L. AND TOOBY, J. (Eds.) *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, New York, Oxford, Oxford University Press, 1992.
- [31] TOOBY, J. AND COSMIDES, L. "The Psychological Foundations of Culture," in: BARKOW, J.; COSMIDES, L. AND TOOBY, J. (Eds.) *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, New York, Oxford, Oxford University Press, 1992.
- [32] WEIBULL, J. "An Introduction to Evolutionary Game Theory," unpub. ms. August, 1992.
- [33] WEIBULL, J. "Chapter 2, Non-Cooperative Game Theory," unpub. ms. Department of Economics, Stockholm University, July 30, 1993.