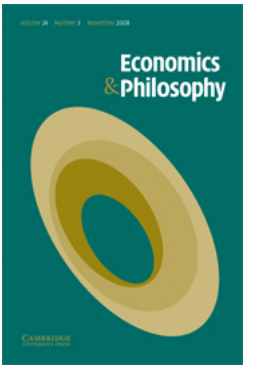

ON AN EVOLUTIONARY FOUNDATION OF NEUROECONOMICS

Burkhard C. Schipper

Economics and Philosophy / Volume 24 / Special Issue 03 / November 2008, pp 495 - 513
DOI: 10.1017/S0266267108002113, Published online: 05 November 2008

Economics and Philosophy, **24** (2008) 495–513
doi:10.1017/S0266267108002113

ON AN EVOLUTIONARY FOUNDATION OF NEUROECONOMICS

Burkhard C. Schipper*
University of California, Davis

Neuroeconomics focuses on brain imaging studies mapping neural responses to choice behaviour. Economic theory is concerned with choice behaviour but it is silent on neural activities. We present a game theoretic model in which players are endowed with an additional structure – a simple "nervous system" – and interact repeatedly in changing games. The nervous system constrains information processing functions and behavioural functions. By reinterpreting results from evolutionary game theory (Germano 2007), we suggest that nervous systems can develop to "function well" in exogenously changing strategic environments. We present an example indicating that an analogous conclusion fails if players can influence endogenously their environment.

1. INTRODUCTION

Neuroeconomics mainly focused on economic experiments using methods of brain imaging (for surveys see Glimcher and Rustichini 2004; Camerer, Lowenstein and Prelec 2005; McCabe 2008). Since neural activity is not explicitly modelled in economic theories, such theories may be of limited use for generating hypotheses that guide neuroeconomic experiments in an insightful way. To fill the gap, neuroeconomic theories are required that are more explicit on the biological constraints that the nervous system imposes on behaviour. In developing such theories, the formal tools of game theory may be a useful language for modelling complex phenomena of interaction within and between brains as it was similarly useful in the development of modern economic theory. The aim of this note is to outline how existing tools of evolutionary game theory and learning in games

* I thank Giacomo Bonanno, Fabrizio Germano and Amanda Kimball for providing useful comments on a short notice. All errors remain my responsibility.

may be reinterpreted to shed some light on the development of "brain" functions in a changing environment. No claim of originality is made: the main result has been developed elsewhere in the abstract context of evolution and learning in games by Germano (2007).

We consider a finite set of players who play repeatedly different strategic games selected randomly according to some exogenously given probability distribution on a finite set of games. The players are endowed with a "nervous system". This is a suggestive interpretation of a simple network-like structure with "neurons" as nodes and "synapses" as binary relation on neurons. The structure constrains the player's perception of the environment and her behavioural response – similar to incomplete information in games. The "richer" the nervous system, the better it can detect the variability of the environment and the more variability of behaviour it can generate. We ask the following question: Can such a nervous system be designed by evolution, development and learning to "function well" in the player's interaction with other players and the environment? Intuitively, a "well functioning" brain should be adapted to its environment in the sense of generating appropriate behavioural responses that enable the survival of the population of "brain-carriers". In this paper, we assume that "function well" means the brain's ability to play strategies that are not strictly dominated in the respective games and in the "average" game over the player's life-time. Reinterpreting a result by Germano (2007), we answer this question affirmatively. Yet, a simple example shows that if players can endogenously affect the change of the environment (like in non-trivial stochastic games), then this conclusion may not hold anymore.

At first glance, the evolutionary approach sketched in this note seems to be orthogonal to "mainstream" neuroeconomics today but we argue that it is relevant for the foundations of neuroeconomics. While economics studies optimal decision making, a typical neuroeconomic experiment will produce brain images of subjects when confronted with an economic decision task. These data are then interpreted with constructs that play a role in economic theories such as utility, expected utility, multiple selves etc. despite the fact that economic theory treats those as abstract constructs and optimizing behaviour "as if". So the implicit assumption in neuroeconomics is that the brain is the very machine that could produce in principle optimal or constrained optimal behaviour. More generally, the assumption behind *functional* magnetic resonance imagining (fMRI) is that different subsets of the brain are activated to fulfil different *functions* or goals. Glimcher (2003: Chapters 6 to 8) traced this assumption back to Marr, who according to Glimcher (2003: 142) suggested that "(i)n order to understand the relationship between behavior and brain, one had to begin by understanding the goals or functions of a behavior. Then one could begin to ask how the brain accomplishes a specific goal." Further

he writes (p. 167)[1] that "(t)he goal of the nervous system is to maximize the inclusive fitness of the organism". The question that we raise in this paper is whether or not evolution, development and learning can produce a nervous system that is capable of doing that. This answer seems to be not obvious to neuroscientists. According to Glimcher (2003: 166) a "major criticism that Marr's approach has faced is that it has been unclear whether evolution can be conceived of as a process that structures nervous systems to accomplish goals with enough efficiency to make the computational goal a useful starting point for neurobiological analysis."[2] This note may be seen as a very preliminary attempt to provide an answer to this criticism of the foundations of neuroeconomics with some tools of evolutionary game theory.

We are not the first to sketch some *neuroeconomic theory*. Others realized that hypotheses on how the brain constrains economic behaviour should be ideally grounded on models that integrate microeconomic theory with a theory of the brain. Recent papers by Benhabib and Bisin (2005), Bernheim and Rangel (2004), Brocas and Carrillo (2008a, 2008b), and Fudenberg and Levine (2006) build models with "multiple selves" motivated by the modularity of the brain but do not really attempt to represent physiological elements of the brain.[3] Hence, they are of limited use for generating

[1] Similarly, Glimcher (2003: 155) writes "(t)he other possibility, and the one implicitly advocated by Marr's approach, is to assume that the system was evolved to achieve a specifiable, and theoretically defined, mathematical goal so as to maximize the fitness of the organism."

[2] This criticism may be rooted in the first sentence of the following quote in Darwin (1859: 171–2.): "To suppose that the eye with all its inimitable contrivances for adjusting the focus to different distances, for admitting different amounts of light, and for the correction of spherical and chromatic aberration, could have been formed by natural selection, seems, I freely confess, absurd in the highest degree. When it was first said that the sun stood still and the world turned round, the common sense of mankind declared the doctrine false; but the old saying of Vox populi, vox Dei, as every philosopher knows, cannot be trusted in science. Reason tells me, that if numerous gradations from a simple and imperfect eye to one complex and perfect can be shown to exist, each grade being useful to its possessor, as is certainly the case; if further, the eye ever varies and the variations be inherited, as is likewise certainly the case; and if such variations should be useful to any animal under changing conditions of life, then the difficulty of believing that a perfect and complex eye could be formed by natural selection, though insuperable by our imagination, should not be considered as subversive of the theory. How a nerve comes to be sensitive to light, hardly concerns us more than how life itself originated; but I may remark that, as some of the lowest organisms in which nerves cannot be detected, are capable of perceiving light, it does not seem impossible that certain sensitive elements in their sarcode should become aggregated and developed into nerves, endowed with this special sensibility." (The first sentence only is quoted in Glimcher 2003: 152.)

[3] Brocas and Carrillo (2008b) write "The objective in this research is not to model the physiological elements involved in a brain process (neurons, synapses, neurotransmitters) but, instead, to capture the fundamental properties of those processes. The models are still 'as-if' representations of reality . . ."

hypotheses on neural data observable with modern technology (while being capable of generating hypotheses on economic behaviour).[4] Our approach here is different in that (besides taking an evolutionary approach) we seek to complement standard game theoretic models with a (crude) model representing physiological elements of the brain such as neurons and synapses. The hope is that an enhanced version could generate hypotheses that are eventually useful for empirical neuroeconomics.

2. BASIC BUILDING BLOCKS

2.1 Environment

Let Ω be a potentially large but finite space of states of nature. These states provide some description of the environment such as which game is to be played. The states of nature are drawn randomly and independently according to some probability distribution $\mu \in \Delta(\Omega)$, where $\Delta(\Omega)$ denotes the set of probability measures on Ω.

There is a finite game defined by a finite set of players $I = \{1, \ldots, m\}$, for each player i a finite set of actions A_i, and for each player i a fitness function $u_i : \times_{i \epsilon I} \Delta(A_i) \times \Omega \longrightarrow \mathbb{R}$, where $\Delta(A_i)$ denotes the set of probability distributions on A_i (i.e. mixed actions). Let us denote $a_i \in A_i$ an action of player i and $a_{-i} \in A_{-i} := \times_{j \in I \setminus \{i\}} A_j$ a profile of actions of player i's opponents. Similarly, let $\alpha_i \in \Delta(A_i)$ denote a mixed action of player i and $\alpha_{-i} \in \times_{j \in I \setminus \{i\}} \Delta(A_j)$ a profile of mixed actions of player i's opponents. We restrict the analysis to *symmetric games*, i.e. $A_i = A$ for all $i \in I$ and for all $\omega \in \Omega$, $u_i(\alpha_i, \alpha_{-i}, \omega) = u_{f(i)}(\alpha_{f(i)}, \alpha_{-f(i)}, \omega)$ for all bijections $f : I \to I$.

The framework is interpreted as follows: In each period, a state $\omega \in \Omega$ is drawn according to the probability distribution μ on Ω. The state ω determines a symmetric finite strategic game $G(\omega) := \langle I, A, (u_i(\omega))_{i \in I} \rangle$. That is, we assume that players may not just play one game in their life but at each period games are selected according to some exogenously fixed probability distribution μ.[5] We call (Ω, μ) the *environment*. We say a game $G(\omega)$ is *relevant* if $\mu(\{\omega\}) > 0$.

2.2 Nervous system

Each player $i \in I$ has a potentially large but finite set of *neurons*, $N_i = \{1, \ldots, n_i\}$. Let $\lhd_i$ be a binary relation defined on N_i for player i called "synapse". We interpret $j \lhd_i j'$ as "*for player i neuron j projects to neuron j'*",

[4] An exception is Chaplin and Dean (2008) who provide an axiomatic characterization of the dopamine reward prediction error hypothesis. (Dopamine is a neurotransmitter.)

[5] In section 6.3 we relax this assumption and allow players to influence the probabilities with which states are drawn.

$j, j' \in N_i$. Since such a synapse is directed, we let $\lhd_i$ be irreflexive (but it may not be transitive or complete). If $j \lhd_i j'$, then we call j the *presynaptic neuron and* j' the *postsynaptic neuron* for player i. Clearly, our model of the neuron abstracts from many interesting features (see Gazzaniga *et al.* 2002: Chapter 2).

There are special neurons called receptors used to obtain signals from the environment. Examples are the photoreceptor cells of the retina (see Gazzaniga *et al.* 2002: Chapter 5). Perhaps one way of featuring receptors in N_i would be to require that if $j \in N_i$ is a *receptor* then there is no $j' \in N_i$ such that $j' \lhd_i j$. That is, a receptor is a neuron which may project to other neurons but to which no other neuron projects to. Yet, this feature will not play a role in this note.

A *neuron sequence* for player i is $j^0, j^1, \ldots j^{\bar{l}}$ with $j^l \lhd_i j^{l+1}$ for $l \in \{0, 1, \ldots, \bar{l} - 1\}$. There can be loops. We call $\mathbf{N}_i = \langle N_i, \lhd_i \rangle$ the *anatomy* of player i's nervous system or bluntly player i's "brain". One may imagine it as a directed graph or network. The conception of the nervous system as a network has a long tradition in neuroscience that can be traced back at least to Exner (1894) and more recently to artificial neural networks. We will not use a neural networks approach here but just stick to primitive features of networks.

A *sensory correspondence* $s_i : \Omega \to 2^{N_i}$ for player i maps states of nature to neuronal responses thought of as neural "firing" or activation of a subset of neurons. We may want to impose conditions reflecting the constraints of the neural activity by the anatomy of the nervous system. To this extend, define for a brain $\mathbf{N}_i$, a particular set of subsets denoted $\mathcal{S}_i \subseteq 2^{N_i}$ by $N' \in \mathcal{S}_i$ if for all $j \in N'$ there exists a neuron sequence $j^0, \ldots, j \in N'$ with j^0 being a receptor. We explicitly let $\emptyset \in \mathcal{S}_i$. We may think of an element of $\mathcal{S}_i$ as a subset of neurons that is accessible by a receptor, i.e. a "module" (Glimcher, 2003: 150) or "neural circuit" accessible by a receptor. We let the sensory correspondence s_i be constrained by the anatomy of the brain by imposing the condition $s_i(\omega) \in \mathcal{S}_i$ for all $\omega \in \Omega$. If for $\omega \in \Omega$ the subset of neurons $s_i(\omega)$ is nonempty, then it must contain a receptor. Hence it can be activated by an environmental stimulus. If $s_i(\omega) = \emptyset$ for some $\omega \in \Omega$, then the stimulus ω does not activate any neurons.

To complete the model, we introduce a *behavioural function* $b_i : 2^{N_i} \to \Delta(A)$ for player i that maps neural activity to mixtures over actions. An example is the activation of motor structures inducing responses of what are called *effectors* such as arms, hands etc. (see Gazzaniga *et al.* 2002; Chapter 11). Note that since $\emptyset \in 2^{N_i}$, b_i defines a default behaviour if no neurons are activated. Note further that since b_i maps neural responses to *mixtures* of actions, we allow for randomness of behaviour. For instance, trichoplax adhaerens, a tiny marine animal, has no neurons (Schierwater 2005). Hence, its behaviour is not controlled by a brain. Still it displays variability in behaviour that we may view here as random.

3. A DIGRESSION: NEUROECONOMICS VS. ECONOMICS

Functional neuroimaging may be viewed as mainly occupied with the description of s_i and b_i. That is, a subject i is exposed to some stimulus $\omega \in \Omega$, observations of brain activity $s_i(\omega)$ are made through MEG, EGG, PET or fMRI (for a discussion of those methods see Gazzaniga *et al.* 2002; Chapter 4) and a response in behaviour $b_i(s_i(\omega))$ is recorded. The implementation of such experiments is not as straightforward as it sounds here. To appreciate the difficulties involved, one needs to consider that the equipment requires large fixed costs. Moreover, the small sample sizes used in neuroeconomic experiments seem to suggest that the variable costs of experiments must be extremely high too. The experimental designs must meet additional challenges from potential confounding effects involved with brain scanners. Finally, typical neuroeconomic papers reveal that the data transformations and statistical analysis including their underlying assumptions are apparently difficult to report in a transparent manner.

Imaging studies of the brain yielded some empirical restrictions on s_i. For example let $\mathbf{N}_i = \langle N_i, \lhd_i \rangle$ be a brain. The condition $s_i(\omega) \neq N_i$ for all $\omega \in \Omega$ would capture a weak version of the *Principle of Functional Segregation*: No functions of the brain are performed by the brain as a whole. Similarly, the condition if $s_i(\omega) = E \neq \emptyset$ then $E = F' \cup F''$ with $F' \neq F''$ and nonempty F', $F'' \in \mathcal{S}_i$ would capture a weak version of the *Principle of Functional Integration*: No function is performed by a single "module" of the brain alone. For a discussion of those principles, see Cohen and Tong (2001).

Economics essentially follows a traditional behavioural paradigm and focuses in our game theoretic context on the optimality of *strategies* under complete or incomplete information. *Complete information* refers to the case where the player can perfectly observe the state of nature. In our framework, it would correspond to s_i being one-to-one or injective: for any ω, $\omega' \in \Omega$, $\omega \neq \omega'$ implies $s_i(\omega) \neq s_i(\omega')$. *Incomplete information* refers to the case where a player can not discriminate between some states of nature. That is, we do not rule out that for some ω, $\omega' \in \Omega$ with $\omega \neq \omega'$ we have $s_i(\omega) = s_i(\omega')$.

Under complete information, a strategy is simply a map $\sigma_i : \Omega \to \Delta(A)$. It assigns to each state of nature a mixture of actions. Under incomplete information, we need to restrict explicitly the strategies to private information. In our context it means that we need to constrain it by values of the sensory correspondence (analogous to the constraining strategies to types in games with incomplete information). That is, a strategy under incomplete information is a map $\sigma_i : \Omega \to \Delta(A)$ subject to for any ω, $\omega' \in \Omega$ with $s_i(\omega) = s_i(\omega')$ implies $\sigma_i(\omega) = \sigma_i(\omega')$.

The name "strategy" may be misleading here because it suggests that σ_i is the object of conscious choice by player i. Since we assume a

large number of states and at each period a random selection of states according to some probability distribution, such interpretation may not be appropriate in a descriptive sense. Rather, we may view a player as "programmed" to a heuristic or a rule (see Gigerenzer *et al.* 2000) that is then calibrated by an evolutionary learning process as outlined in section 6. While this "programming" perspective may not be the standard interpretation in economics, it is familiar to the economists from evolutionary game theory (see Weibull 1995).

Note that we allow for framing: Let $\omega, \omega' \in \Omega$ be such that $\omega \neq \omega'$ and $G(\omega) = G(\omega')$. That is, games at ω and ω' are formally identical but they may differ in their "colour" or "smell". Yet, we allow the values of the feasible strategy to differ between the states. For example, we allow that administering subjects oxytocin before playing trust games as in Kosfeld *et al.* (2005) or Zak *et al.* (2005) may alter the actions of the subjects as compared to a placebo.[6]

No matter whether we focus on complete or incomplete information, in our context we may view a strategy as a composition of the sensory correspondence and the behavioural function, $\sigma_i = b_i \circ s_i$. So an analogy between neuroeconomics and economics should become clear: When economics studies informational constraints on choice behaviour, neuroeconomics studies neurobiological constraints on choice behavior by adding the focus on how the nervous system constraints information processing. Which approach one should take depends largely on the type of question one wishes to ask. If one wants to study for instance the impact of brain lesions on behaviour (a question taken up in section 5), the standard economic approach does not suffice but a model on how the nervous system constrains information processing has to be added.

4. "WELL FUNCTIONING" BRAINS

Glimcher (2003: 167) writes "(t)he goal of the nervous system is to maximize the inclusive fitness of the organism". If a nervous system would play a strategy that is strictly dominated in the "average game of life", then clearly it would not maximize its fitness. Therefore we assume that in our context "functioning well" shall mean to play strategies that are not strictly dominated in the "average game of life". In experiments we usually judge a player's performance only in one isolated controlled game at a time but do not observe the player's performance in the "average game of life". Hence, we consider as a second criterion that "functioning well" refers to the ability of choosing in all relevant situations actions that are not strictly dominated.

[6] It is actually not clear whether oxytocin does not change the game (e.g. the fitness) as well since we are not specific here on what we mean by fitness.

More formally, an action $a_i \in A$ is strictly *dominated* in the game $G(\omega)$ if there is a mixed action $\alpha_i \in \Delta(A)$ such that[7] $u_i(a_i, a_{-i}, \omega) < u_i(\alpha_i, a_{-i}, \omega)$ for all $a_{-i} \in A_{-i}$.

For $\omega \in \Omega$, let D_ω be the set of actions that are not strictly dominated in the game $G(\omega)$. Define for $\Omega' \subseteq \Omega$, a set $D_{\Omega'} \subseteq A$ by (i) for all $\omega \in \Omega'$ there exists $a_i \in D_{\Omega'}$ with $a_i \in D_\omega$, and (ii) there is no $D \subsetneq D_{\Omega'}$ for which (i) holds. Condition (i) ensures that for each state $\omega \in \Omega'$ there exists an action a_i in $D_{\Omega'}$ that is not strictly dominated in $G(\omega)$. Condition (ii) requires that $D_{\Omega'}$ is "minimal" in the sense that there is no smaller set of actions satisfying condition (i). That is, $D_{\Omega'}$ is *a smallest set* of actions in A with the property that for each state $\omega \in \Omega'$ there is exactly one action in $D_{\Omega'}$ that is not strictly dominated in $G(\omega)$. $|D_{\Omega'} \cap D_\omega| = 1$ for all $\omega \in \Omega'$. Note that $D_{\Omega'}$ may not be unique.[8] For $\Omega' \subseteq \Omega$, let $\mathcal{D}_{\Omega'}$ denote the set of all sets of actions satisfying (i) and (ii). Note further that since Ω is finite, we must have that every $D \in \mathcal{D}_{\Omega'}$ is finite for every $\Omega' \subseteq \Omega$. In fact $|D| \leq |\Omega'|$ for all $D \in \mathcal{D}_{\Omega'}$ and all $\Omega' \subseteq \Omega$.

We define the *variability of the environment* (Ω, μ) by $\varepsilon(\Omega, \mu) := \min_{D \in \mathcal{D}_{supp\ \mu}} |D|$, where $supp\ \mu := \{\omega \in \Omega : \mu(\{\omega\}) > 0\}$ is the support of μ. Intuitively, $\varepsilon(\Omega, \mu)$ is the minimal number of actions required that enables the play of an action that is not strictly dominated in any relevant state. By definition, $\varepsilon(\Omega, \mu) \leq |\Omega|$. That is, the number of states of the environment provide an upper bound on the variability of the environment. Note that the definition of the variability of the environment depends on the choice of the solution concept (here actions that are not strictly dominated) and hence on the fitness "goal".

Let $S_i(\Omega, \mathbf{N}_i)$ be the set of all sensory correspondences from Ω to $\mathcal{S}_i$. Similarly, let $B_i(\mathbf{N}_i, A)$ be the set of all behavioural functions from $2^{\mathcal{N}_i}$ to $\Delta(A)$. A strategy $\sigma_i : \Omega \to \Delta(A)$ is *feasible* for the brain $\mathbf{N}_i$ if $\sigma_i = b_i \circ s_i$ with $s_i \in S_i(\Omega, \mathbf{N}_i)$ and $b_i \in B_i(\mathbf{N}_i, A)$. As mentioned in section 3, we don't view here a strategy as an object of conscious choice by the brain but rather as a heuristic or rule to which a player is "programmed".

We define the *size of the brain* by $\beta(\mathbf{N}_i) := |\mathcal{S}_i|$. Note that the size is not necessarily increasing in the number of neurons but such increase requires also appropriate synapses and the connectivity to receptors and effectors. The larger the size of the brain, the more variability in behaviour it may generate and the better it can gather information about the environment. The following example is used to motivate the above definition:

[7] We abuse notation when writing u_i both as a function of pure actions and mixed actions.

[8] An example is easily constructed: Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Moreover, let $D_{\omega_1} = \{a_1, a_2\}$, $D_{\omega_2} = \{a_2, a_3\}$ and $D_{\omega_3} = \{a_4\}$. Then both $\{a_1, a_3, a_4\}$ and $\{a_2, a_4\}$ satisfy the definition for $D_{\{\omega_1, \omega_2, \omega_3\}}$.

Example 1. Consider the brain $N_i = \{1\}$. $\mathcal{S}_i = \{\emptyset, \{1\}\}$. Hence $\beta(\mathbf{N}_i) = 2$. The environment is given by $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\mu(\{\omega\}) > 0$ for all $\omega \in \Omega$ and a single person game as follows:

State	s_i	Activation	b_i	Action	Fitness
ω_1	$\rightarrow$	$\emptyset$	$\rightarrow$	a_1	$u_i(a_1, \omega_1) > u_i(a_2, \omega_1) > u_i(a_3, \omega_1)$
ω_2	$\nearrow$				$u_i(a_2, \omega_2) > u_i(a_3, \omega_2) > u_i(a_1, \omega_2)$
ω_3	$\rightarrow$	$\{1\}$	$\rightarrow$	a_3	$u_i(a_3, \omega_3) > u_i(a_1, \omega_3) > u_i(a_2, \omega_3)$

The table shows also one possible assignment of the sensory correspondence s_i and behavioural function b_i. There are no possible assignments of s_i and b_i that would allow the individual to choose her most preferred action in each state. The reason is simply that the size of the brain is not large enough given the variability of the environment, $\varepsilon(\Omega, \mu) = 3$. One more neuron would be sufficient to solve the problem.

This observation can be generalized to characterize brains that "function well" in all relevant situations.

Remark 1. *The size of the brain* $\mathbf{N}_i$ *is strictly lower than the variability of the environment* (Ω, μ) *if and only if for any feasible strategy* σ_i *of the brain* $\mathbf{N}_i$ *there exists a relevant game* $G(\omega)$ *for which* σ_i *prescribes a strictly dominated action*.

Proof. "$\Rightarrow$": Suppose to the contrary that there exists a strategy σ_i feasible for $\mathbf{N}_i$ such that $\sigma_i(\omega)$ is not strictly dominated for all $\omega \in \Omega$ with $\mu(\{\omega\}) > 0$. Then $|\text{range}\,\sigma_i| \geq \varepsilon(\Omega, \mu)$. Since σ_i is feasible, $\sigma_i \in b_i \circ s_i$ with $s_i \in S_i(\Omega, \mathbf{N}_i)$ and $b_i \in B_i(\mathbf{N}_i, A)$. Thus $|\text{range}\,\sigma_i| = |\text{range}\,b_i| \leq \beta(\mathbf{N}_i)$, a contradiction to $\beta(\mathbf{N}_i) < \varepsilon(\Omega, \mu)$.

"$\Leftarrow$": Suppose to the contrary that $\beta(\mathbf{N}_i) \geq \varepsilon(\Omega, \mu)$. Then construct a strategy σ_i such that for each $\omega \in \Omega$ with $\mu(\{\omega\}) > 0$, $\sigma_i(\omega)$ is not strictly dominated in $G(\omega)$. Such a strategy is feasible for $\mathbf{N}_i$, a contradiction.

We denote by $\Sigma(\mathbf{N}_i)$ *a finite set of strategies feasible for the brain* $\mathbf{N}_i$. Moreover, in light of Remark 1 we assume that *if the size of the brain* $\mathbf{N}_i$ *is at least as large as the variability of the environment* (Ω, μ) *then* $\Sigma(\mathbf{N}_i)$ *contains a strategy prescribing for each relevant game* $G(\omega)$ *an action that is not strictly dominated*. Finally, we assume that if $N_i = N_j$ then $\Sigma(N_i) = \Sigma(N_j)$.

A brain may be well adapted to its environment in the sense of not playing a strictly dominated action in any relevant situation. Yet, such a strategy may be strictly dominated by another strategy in the overall "average game of life". Let $U_i(\sigma) := \sum_{\omega \in \Omega} \mu(\{\omega\})\ u_i(\sigma(\omega), \omega)$ denote the expected fitness of player i from playing strategy σ_i when opponents play σ_{-i} (i.e. expected over the entire life for a fixed strategy profile). This is

the payoff function in the "*average game of life*" denoted by Γ defined for a given environment (Ω, μ), the set of players I, a given profile of brains $(\mathbf{N}_i)_{i\in I}$ and for each player i/brain $\mathbf{N}_i$ a set of feasible strategies $\Sigma(\mathbf{N}_i)$.

A feasible strategy $\sigma_i \in \Sigma(\mathbf{N}_i)$ is strictly dominated by a mixture of feasible strategies $\rho_i \in \Delta(\Sigma(\mathbf{N}_i))$ in Γ if[9] $U_i(\sigma_i, \sigma_{-i}) < U_i(\rho_i, \sigma_{-i})$ for all $\sigma_{-i} \in \times_{j\in I\setminus\{i\}}\Sigma(\mathbf{N}_j)$. Note that according to this definition, a strategy of player i may become strictly dominated if player i's size of the brain increases or the sizes of her opponents' brains increase. That is, a player with a previously "well functioning" brain may find it impossible to adapt herself well after opponents evolve more sophisticated brains.

Remark 2. *Suppose that for every player $i \in I$ the size of her brain $\mathbf{N}_i$ is weakly larger than the variability of the environment. If $\sigma_i \in \Sigma(\mathbf{N}_i)$ is not strictly dominated in the average game of life Γ by some other strategy feasible for $\mathbf{N}_i$, then $\sigma_i(\omega)$ is not strictly dominated in $G(\omega)$ for all $\omega \in \Omega$ with $\mu(\{\omega\}) > 0$.*

Proof. Suppose by contradiction that $\sigma_i \in \Sigma(\mathbf{N}_i)$ is not strictly dominated in Γ but that there exist a state $\omega \in \Omega$ such that $\sigma_i(\omega)$ is strictly dominated in $G(\omega)$. Construct a new strategy σ_i^* that agrees with σ_i on all games $G(\omega')$ with $\mu(\omega') > 0$ where $\sigma_i(\omega')$ is not strictly dominated in $G(\omega')$. In any other games $G(\omega'')$ with $\mu(\omega'') > 0$ where $\sigma_i(\omega'')$ is strictly dominated in $G(\omega'')$ let $\sigma_i^*(\omega'')$ strictly dominate $\sigma_i(\omega'')$. Since $\beta(\mathbf{N}_i) > \varepsilon(\Omega, \mu)$, such strategy is feasible for $\mathbf{N}_i$ and by assumption such strategy is contained in $\Sigma(\mathbf{N}_i)$. Note that σ_i^* strictly dominates σ in Γ, a contradiction.

The converse is not true. A counter example can be constructed similarly to Germano (2007: Example 2).

5. BRAIN LESIONS

The motivation for this section is twofold: first, in neuroscience, *lesion* studies are common. A lesion is a damage of brain tissue possibly separating projections between neurons or destroying neurons altogether. The effect of such lesions is then studied in patients. While some brain functions are lost due to lesions, patients are often quite well calibrated to the environment. For instance, the patient N.R. who suffered from the Balint's syndrome caused by a right pariental lesion due to a stroke can not see two objects shown to him at the same time but only sees one object at a time while speech and comprehension are normal (see Gazzaniga *et al.* 2002: 245, 292). The second purpose of this section is to define a "set of brains" that we will use in the next section on the evolution of brains.

Given a brain $\mathbf{N}_i = \langle N_i, \lhd_i\rangle$, define a brain $\mathbf{N}_i' = \langle N_i, \lhd_i'\rangle$ by $N_i' \subseteq N_i$ and for $j, j' \in N_i'$, $j \lhd_i' j'$ implies $j \lhd_i j'$ (but not necessarily vice versa).

[9] Again, we abuse notation when writing U_i both as a function of strategies and mixtures of strategies.

We can view $\mathbf{N}_i'$ as a brain obtained from $\mathbf{N}_i$ by a lesion. By definition, $\beta(\mathbf{N}_i') \leq \beta(\mathbf{N}_i)$. That is, the size of the brain without the lesion is weakly higher than the size of the brain with the lesion. Naturally, we assume $\Sigma(\mathbf{N}_i') \subseteq \Sigma(\mathbf{N}_i)$. A brain with a lesion has a lower number of feasible strategies available than the brain without the lesion.

Let $\mathcal{N}_i$ denote the (partially ordered) set of all brains that can be obtained from $\mathbf{N}_i$ by lesions. We call $\mathcal{N}_i$ *the set of brains derived from* $\mathbf{N}_i$. In the next section, we do not necessarily interpret a brain $\mathbf{N}_i' \in \mathcal{N}$ as a brain obtained from $\mathbf{N}$ by a lesion. Rather, $\mathcal{N}_i$ is just a set of brains with weakly lower size than the size of brain $\mathbf{N}_i$.[10]

Do lesions always matter? The following example illustrates that this depends on the kind of lesion of the player's brain and the environment.

Example 2. Consider a brain given by $N_i = \{1, 2, 3\}$ with $1 \lhd_i 2$, $1 \lhd_i 3$ and $2 \lhd_i 3$. Thus $\mathcal{S}_i = \{\emptyset, \{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\}$ and $\beta(\mathbf{N}_i) = 5$. Consider now a lesion that severs the synapse between 2 and 3. Note that S_i and $\beta(\mathbf{N}_i)$ remains unchanged. Hence such a lesion won't affect information processing and behaviour no matter how rich the environment is. Consider now a lesion that severs the synapse between 1 and 2. The size of the brain is now reduced to 3 even though no neuron was removed. Despite this "brain damage", the player still can "function well" in below environment $\Omega = \{\omega_1, \omega_2, \omega_3\}$ with $\mu(\{\omega\}) > 0$ for all $\omega \in \Omega$ because it is feasible for her to be programmed to s_i and b_i given in the table:

State	s_i	Activation	b_i	Action	Fitness
ω_1	$\longrightarrow$	$\emptyset$	$\longrightarrow$	a_1	$u_i(a_1, \omega_1) > u_i(a_2, \omega_1) > u_i(a_3, \omega_1)$
ω_2	$\longrightarrow$	$\{1\}$	$\longrightarrow$	a_2	$u_i(a_2, \omega_2) > u_i(a_3, \omega_2) > u_i(a_1, \omega_2)$
ω_3	$\longrightarrow$	$\{1, 3\}$	$\longrightarrow$	a_3	$u_i(a_3, \omega_3) > u_i(a_1, \omega_3) > u_i(a_2, \omega_3)$

For such an environment, the lesion won't affect her ability to play actions that are not strictly dominated in each state. This holds true even if the lesion would remove either neuron 2 or 3 altogether.

Lesions may have an *externality* on other players (as caretakers of patients sometimes note).

Example 3. Let the environment consist of two states, $\Omega = \{\omega_1, \omega_2\}$, with $\mu(\{\omega\}) = \frac{1}{2}$ for all $\omega \in \Omega$. The game at each state is given by the following payoff matrices:

$$\begin{array}{cc} \omega_1 & \omega_2 \\ \begin{pmatrix} 10, 10 & 1, 9 \\ 9, 1 & 0, 0 \end{pmatrix} & \begin{pmatrix} 0, 0 & 9, 1 \\ 1, 9 & 10, 10 \end{pmatrix} \end{array}$$

[10] Note that there may be several different brains in $\mathcal{N}_i$ with an identical size.

Let both players' brains be given by $N_1 = N_2 = \{1\}$. Such a brain enables each player to play a feasible strategy given by

$$\sigma_1(\omega) = \begin{cases} \text{up} & \text{if } \omega = \omega_1 \\ \text{down} & \text{if } \omega = \omega_2 \end{cases} \qquad \sigma_2(\omega) = \begin{cases} \text{left} & \text{if } \omega = \omega_1 \\ \text{right} & \text{if } \omega = \omega_2 \end{cases}$$

that selects the strict dominant action and the Pareto efficient outcome in each game. That is, for player 1 we let $s_1(\omega_1) = \emptyset$, $b_1(\emptyset) = \text{up}$, $s_1(\omega_2) = \{1\}$, $b_1(\{1\}) = \text{down}$ and analogously for player 2. Suppose now that player 1 suffers a brain lesion such that her brain with the lesion is $N_1' = \emptyset$. The above strategy is not feasible any more for player 1 with the brain damage. Only constant strategies are feasible that prescribe either up or down or a constant mixture thereof at both states. Since player 2 sticks to his dominant strategy, in expectations any constant strategy yields a fitness of $9\frac{1}{2}$ to player 1. Yet, player 2 incurs a much bigger fitness loss since she receives in expectations only $5\frac{1}{2}$. While player 1 suffered the brain damage, the healthy player 2 incurs most of the costs.[11]

6. DEVELOPMENT AND EVOLUTION OF BRAIN FUNCTIONS

We are not born with a fully developed brain. For instance, in newborns the optic nerves are not developed completely but reach typical adult patterns only at the age of about 2 years. But even the nervous systems of adults maintain some neural plasticity as indicated by learning of new skills or the development of phantom sensation of amputees (see Gazzaniga *et al.* 2002: Chapter 15). More generally, if the nervous system regulates the interaction of the organism with other organisms and the complex changing environment, there should be an evolutionary selection of nervous systems. First, we will try to analyse the question whether "successful" brain functions s_i and b_i can develop among interacting brains in a changing environment. Second, we focus on the evolution of brains.

6.1. Development

Starting from an initial distribution of feasible strategies $\bar{\rho} = (\rho_i, \rho_{-i}) \in \times_{j \in I} \Delta(\Sigma(\mathbf{N}_j))$ among brains, we assume that brains *develop* feasible strategies according to a discrete-time stochastic aggregate log-monotone dynamics defined by

$$(1) \quad \rho_i^{t+1}(\sigma_i) = \frac{\rho_i^t(\sigma_i) e^{\lambda_i(\bar{\rho}^t)(u_i(\sigma_i(\omega^t), \rho_{-i}^t, \omega^t) - u_i(\bar{\rho}^t, \omega^t))}}{\sum_{\sigma_i' \in \Sigma(\mathbf{N}_i)} \rho_i^t(\sigma_i') e^{\lambda_i(\bar{\rho}^t)(u_i(\sigma_i'(\omega^t), \rho_{-i}^t, \omega^t) - u_i(\bar{\rho}^t, \omega^t))}}$$

[11] Similarly, one can find examples in which the value of a brain damage is strictly positive because the brain damage works like a commitment device.

where $\lambda_i : \times_{j \in I} \Delta(\Sigma(\mathbf{N}_j)) \longrightarrow \mathbb{R}_+$ is a positive continuous function bounded away from zero. This dynamics is just one learning dynamics reflecting the "law of effect": The probability of playing a certain strategy increases in the relative performance of the strategy in randomly drawn games among brains. Note that the propensity to use a certain strategy is updated with respect to randomly drawn games (instead of the average game of life). This dynamics has been studied by Cabrales and Sobel (1992) in a standard evolutionary game setting and for our stochastic environments by Germano (2007).

Proposition 1. *Fix an environment* (Ω, μ) *and a profile of brains* $(\mathbf{N}_j)_{j \in I}$. *Let* $\sigma_i \in \Sigma(\mathbf{N}_i)$ *be a feasible strategy of the brain* $\mathbf{N}_i$ *for some player* i, *which is strictly dominated in the average game* Γ. *If for every player* $j \in I$ *there is initially positive probability that* $\mathbf{N}_j$ *uses any feasible strategy in* $\Sigma(\mathbf{N}_j)$, *then the brain* $\mathbf{N}_i$ *develops to use* σ_i *with zero probability almost surely*.

Suppose further that for every player $i \in I$ *the size of the brain* $\mathbf{N}_i$ *is weakly larger than the variability of the environment* (Ω, μ). *If for some player* $i \in I$, $\sigma_i \in \Sigma(\mathbf{N}_i)$ *is a feasible strategy for* $\mathbf{N}_i$ *that prescribes a strictly dominated action in some relevant game and for every player* $j \in I$ *there is initially positive probability that* $\mathbf{N}_j$ *uses any feasible strategy in* $\Sigma(\mathbf{N}_j)$, *then the brain* $\mathbf{N}_i$ *develops to use* σ_i *with zero probability almost surely*.

Proof. The first conclusion is a reinterpretation of Germano (2007: Proposition 1). The second conclusion follows from the first conclusion using Remark 2.

Since the statement is for a fixed profile of brains, the interpretation is restricted to learning and development of brains. In light of Proposition 1 it would be interesting to study the correlation between behavioural changes and the development of nervous systems in children. For instance, Harbaugh, Krause and Berry (2001) examine to which extent consumption choices by 7- and 11-year-old children and college undergraduates satisfy the axioms of revealed preference. They find that choices by even the 7-year-olds are considerably more likely to obey revealed preference axioms than would be true if they were choosing randomly. Eleven-year-olds do better still, while college students do no better than 11-year-old children. They argue that this evidence suggests that the ability to choose rationally is not innate, but that it does develop quickly.

6.2. Evolution

Now we turn our attention to the evolution of brains. Consider a sufficiently large population of players. Each player is endowed with a brain $\mathbf{N} \in \mathcal{N}$, where $\mathcal{N}$ is a set of brains derived from some brain $\bar{\mathbf{N}}$ as discussed in section 5. We assume that the size of $\bar{\mathbf{N}}$ is weakly larger then

the variability of a fixed environment (Ω, μ), $\beta(\bar{\mathbf{N}}) \geq \varepsilon(\Omega, \mu)$. We denote by $\eta \in \Delta(\bar{\mathcal{N}})$ the distribution of brains within the population. For example, $\eta(\mathbf{N})$ denotes the fraction of the population endowed with the brain $\mathbf{N} \in \bar{\mathcal{N}}$.

At each period t players are randomly and anonymously matched to play the game at ω^t. If a player's brain is $\mathbf{N} \in \bar{\mathcal{N}}$, then he is programmed to some feasible strategy $\sigma \in \Sigma(\mathbf{N})$. Let $\rho_{\mathbf{N}} \in \Delta(\Sigma(\mathbf{N}))$ be the distribution of strategies in the population of players endowed with brain $\mathbf{N}$. For example, $\rho_{\mathbf{N}}(\sigma)$ is the fraction of players programmed to $\sigma \in \Delta(\Sigma(\mathbf{N}))$ among all players in the population with brain $\mathbf{N}$. (If $\eta(\mathbf{N}) = 0$, then $\rho_{\mathbf{N}}$ can be arbitrary.) We define $\rho \in \Delta(\Sigma(\bar{\mathbf{N}}))$ by $\rho(\sigma) = \sum_{\mathbf{N} \in \bar{\mathcal{N}}} \eta(\mathbf{N}) \rho_{\mathbf{N}}(\sigma)$. This is the fraction of the entire population programmed to σ.

We assume that the evolutionary selection of strategies within the entire population follows equation (1), i.e.

$$\rho^{t+1}(\sigma) = \frac{\rho^t(\sigma) e^{\lambda(\bar{\rho}^t)(u_i(\sigma(\omega^t), \overset{m-1}{\rho^t, \ldots, \rho^t}, \omega^t) - u_i(\bar{\rho}^t, \omega^t))}}{\sum_{\sigma' \in \Sigma(\mathbf{N})} \rho^t(\sigma') e^{\lambda(\bar{\rho}^t)(u_i(\sigma'(\omega^t), \overset{m-1}{\rho^t, \ldots, \rho^t}, \omega^t) - u_i(\bar{\rho}^t, \omega^t))}}.$$

This equation may be viewed as a discrete-time version of the replicator dynamics used in standard evolutionary game theory (see Cabrales and Sobel 1992). By Remark 1, the evolutionary selection of strategies has implications on the *evolution of brains*:[12] If $\rho(\sigma) = 0$ for all feasible strategies $\sigma \in \Sigma(\mathbf{N})$ of the brain $\mathbf{N}$, then $\eta(\mathbf{N}) = 0$.

Corollary 1. *Given the environment* (Ω, μ), *consider the set of brains* $\bar{\mathcal{N}}$ *derived from a brain* $\bar{\mathbf{N}}$ *whose size is weakly larger than the variability of the environment. If initially there is a completely mixed distribution of brains* $\eta \in \Delta(\bar{\mathcal{N}})$ *in the population of players and for each brain any feasible strategy has initially strict positive probability in the population, then evolution lets the fraction of players with a brain of strictly smaller size than the variability of environment go to zero almost surely.*

Proof. For all brains $\mathbf{N}$ with $\beta(\mathbf{N}) < \varepsilon(\Omega, \mu)$ it follows from Remark 1 that any feasible strategy $\sigma \in \Sigma(\mathbf{N})$ must prescribe a strictly dominated action $\sigma(\omega)$ for some game $G(\omega)$ with $\mu(\{\omega\}) > 0$. Then the result follows from Proposition 1. That is, the result is just a reinterpretation of Germano (2007: Proposition 1).

Empirically, there is quite some variation of the number of neurons (a proxy for our measure of brain size) in organisms. For instance, trichoplax adhaerens, a tiny marine animal, has no neurons at all (Schierwater, 2005) whereas human beings are estimated to have about 95 billion neurons and about 100 trillion synapses. While humans do not

[12] So the evolution of brains is "indirect" in the spirit of the indirect evolution of utility functions in an approach pioneered by Güth and Yaari (1992).

have the largest brain both in terms of relative or absolute volume or weight or the total number of neurons, they have the highest number of cortical neurons (for a survey see Roth and Dicke 2005). The cerebral cortex is often associated with "thinking", "perceiving", "producing" and "understanding" language but it is also involved in more basic functions such as vision, hearing, touch, movement and smell (Gazzaniga *et al.* 2002: 70). It is the most recent structure in the history of brain evolution. The following table provides a comparison of numbers of cortical neurons in some mammals (see Haug, 1987; Roth and Dicke 2005):

Animal taxa	**Number of cortical neurons**
Man	11 500 000 000
African elephant	11 000 000 000
False killer whale	10 500 000 000
Chimpanzee	6 500 000 000
Bottlenose dolphin	5 800 000 000
Gorilla	4 300 000 000
Horse	1 200 000 000
White-fronted capuchin	610 000 000
Rhesus monkey	480 000 000
Squirrel monkey	480 000 000
Cat	300 000 000
Dog	160 000 000
Opossum	27 000 000
Hedgehog	24 000 000
Rat	15 000 000
Mouse	4 000 000

In light of Corollary 1, it would be an interesting empirical exercise to investigate beside the brain sizes of organisms also a measure of the variability of their environment, and check for a correlation. Note however that this does not provide a test for the result because it could well be that organisms 1 and 2 are such that the brain size of 1 is smaller than the brain size of 2 and the variability of 1's environment is higher than the variability of 2's environment but both organisms' brain sizes are larger than their respective environment's variability.

6.3. Endogenous changes of environments

Today there are signs that human behaviour changes the environment more and more. For instance, the industrial revolution may cause global warming. Even in more primitive societies, actions today impact the

environment tomorrow; for example hunters need to move on once animals are hunted, nomads need to move depending on the grazing activity of their livestock, wars destroy potentials of future production etc. Do the conclusions of the previous section remain intact in such a more realistic setting?

More formally, in contrast to section 2.1 suppose now that at each period players can influence interactively through their actions the probability with which the next state is drawn. In particular, we assume that $\mu(\omega^{t+1}|\omega^t, a)$ is the *transition probability* that the state is ω^{t+1} in period $t+1$ given that the state in t is ω^t and the players' profile of actions at t is $a = (a_1, \ldots, a_m)$. Essentially, these transition probabilities together with games $\{G(\omega)\}_{\omega\in\Omega}$ render the environment into a *stochastic game*. Analogous to the theory of stochastic games, we call player i's strategy σ_i *Markov* if at any period of time it just depends on the current state.

Example 4 (**apokalupsis eschaton**). There are two players. Their environment consists of two states $\Omega = \{\omega_1, \omega_2\}$. In any of those states, either player can take either of two actions. The payoffs in each state are given by the payoff matrices. The transition probabilities associated with each state and each profile of actions are given below the payoff matrices. (Each component of the matrix corresponds to the state and action profiles above, assigning the probability of transiting to ω_1 and ω_2 respectively. We let $\epsilon > 0$.)

$$\begin{array}{cc} \omega_1 & \\ \begin{pmatrix} 3,3 & 0,4 \\ 4,0 & 2,2 \end{pmatrix} & \\ \begin{pmatrix} (1,0) & (1-\epsilon,\epsilon) \\ (1-\epsilon,\epsilon) & (1-\epsilon,\epsilon) \end{pmatrix} & \end{array} \qquad \begin{array}{cc} \omega_2 & \\ \begin{pmatrix} -10,-10 & -10,-10 \\ -10,-10 & -10,-10 \end{pmatrix} & \\ \begin{pmatrix} (0,1) & (0,1) \\ (0,1) & (0,1) \end{pmatrix} & \end{array}$$

$G(\omega_1)$ is a standard Prisoner's dilemma with *down* and *right* being strictly dominant. In $G(\omega_2)$ any action is not strictly dominated.

We assume that the initial state is ω_1. No matter whether players have a brain or not, the dynamics in equation (1) should lead players to play the strictly dominant action in $G(\omega_1)$ starting from a completely mixed action profile. Such play leads at some point to the absorbing game $G(\omega_2)$ with very low fitness to both players. Yet, playing the strictly dominated action in $G(\omega_1)$ is part of the strategy that is strictly dominant in the average game. So there is no way in which brains as modelled in this note can develop to "function well" in the "average game of life" with the dynamics in equation (1) because "functioning well" would mean to avoid the "bad life" in game $G(\omega_2)$ altogether. Note that we could slightly perturb the payoffs and transition probabilities and the same conclusion would obtain. Thus such class of games is not negligible.

What to make of it? On one hand, we can dismiss adaptive play given by equation (1) as extremely mechanistic and backward looking and our model of the "brain" as a meaningless caricature. What would a model of a brain need to look like that is able to generate foresight required to "function well" in problems like Example 4? What enables the imagination of consequences without having to experience similar consequences beforehand? On the other hand, stories like the one of Adam and Eve show that we may (even religiously) believe that humans are created in such a way that they fail to envision consequences of their actions (despite being told about them beforehand).

In the exogenously changing environments studied in the previous section, larger brains have an evolutionary advantage in more complex environments. When the change of environment depends endogenously on the players' actions, a larger brain can also generate more variability in behaviour and hence make the environment more variable as well. Therefore, it is not clear any more, whether larger brains maintain an evolutionary advantage over smaller brains in endogenously changing environments. It is possible to build more sophisticated examples where only the presence of a large brain in a population of "no-brainers" triggers the transition to "bad" absorbing sets of games. It is also possible to construct examples, where large brains are needed to enter relatively small absorbing sets of games and then once entered evolutionary drift reduces the brain size in the population over time because the evolutionary selection pressure is not present anymore in the small absorbing set of games.

7. SOME FURTHER DISCUSSION

What is really the relevance of such an evolutionary model? It gives a preliminary answer to the "major criticism that Marr's approach has faced". Namely, that "it has been unclear whether evolution can be conceived of as a process that structures nervous systems to accomplish goals with enough efficiency to make the computational goal a useful starting point for neurobiological analysis," (Glimcher 2003: 166). It does shed some light on the dependence of the appropriate brain size on the variability of the environment but such a relationship is far from surprising and the model falls short of generating a hypothesis that is really testable. It does also question the ability of evolution and development to adapt appropriately to an environment that can be changed by the players themselves. But given the crude model of the nervous system and the evolutionary process, how seriously should it be taken?

One important aspect from an economic point of view – which is not considered here at all – is that large brains in humans constitute large investments. This large investment does not only come in form of bodily

capital (the extreme rapid growth requires prenatally about 60% of the metabolism, see Roth and Dicke 2005: 254) but large brains also demand education and hence further investment by society into human capital. Moreover, the "maintenance" of such large brains consumes about 20% of the total metabolism while it constitutes only 2% of the body weight (Roth and Dicke 2005: 254). A more comprehensive theory of the development and evolution of the brain needs to take into account the trade off between the higher costs of a larger brain and the more sophisticated behaviour it may generate. Robson and Kaplan (2003) present such a "brain-capital" theory.

This note uses results by Germano (2007) and he may have anticipated such use when he wrote (p. 324) "(I)t seems that some of the main challenges lie in the characterizing 'good' rules that ideally apply to a wide range of games and environments, and linking them to actual cognitive (or genetic) behaviour". He also presents additional results such as on the elimination of strategies that are not rationalizable, Nash equilibria in the average game of life as limit points under convergence etc. It would be interesting to consider such strategically more sophisticated solution concepts because Dunbar and Shultz (2007) suggest that the strategic demand from living in complex societies selected for sophisticated brains whereas our focus on actions/strategies that are not strictly dominated covers mainly the demands upon the brain made by the ecological variability.

Our model has nothing to say about internal mental conflicts modelled in recent papers on neuroeconomic theory by Benhabib and Bisin (2005), Bernheim and Rangel (2004), Brocas and Carrillo (2008a, 2008b) and Fudenberg and Levine (2006). Our hope is that a more sophisticated evolutionary approach could shed some light on the evolution of multiple selves. A first attempt is presented by Livnat and Pippenger (2006) who show what computational constraints give optimally rise to "multiple selves". However, they do not model the evolutionary selection of players with multiple selves in the spirit of evolutionary game theory.

REFERENCES

Benhabib, J. and A. Bisin. 2005. Modeling internal commitment mechanisms and self-control: a neuroeconomics approach to consumption-saving decisions. *Games and Economic Behavior* 52: 460–92.

Bernheim, B. D. and A. Rangel. 2004. Addiction and cue-triggered decision processes. *American Economic Review* 94: 1558–90.

Brocas, I. and J. D. Carillo. 2008a. The brain as a hierarchical organization. *American Economic Review,* forthcoming.

Brocas, I. and J. D. Carillo. 2008b. Theories of the mind. *American Economic Review Papers and Proceedings* 98: 175–80.

Cabrales, A. and J. Sobel. 1992. On the limit points of discrete selection dynamics. *Journal of Economic Theory* 57: 407–19.

Camerer, C., G. Loewenstein and D. Prelec. 2005. Neuroeconomics: how neuroscience can inform economics. *Journal of Economic Literature* 43: 9–64.

Chaplin, A. and M. Dean. 2008. Dopamine, reward prediction error, and economics. *Quarterly Journal of Economics*, forthcoming.

Cohen, J. D. and F. Tong. 2001. The face of controversy. *Science* 293: 2405–7.

Darwin, C. 1859. *The origin of species by means of natural selection or the preservation of favoured races in the struggle for life*. New York: Penguin, 1958.

Dunbar, R. I. M. and S. Shultz. 2007. Evolution in the social brain. *Science* 317: 1344–7.

Exner, S. 1894. *Entwurf zu einer physiologischen Erklärung der psychologischen Erscheinungen*. Leipzig, Wien: F. Deuticke.

Fudenberg, D. and D. Levine. 2006. A dual self model of impulse control. *American Economic Review* 96: 1449–76.

Gazzaniga, M. S., R. B. Ivry and G. R. Mangun. 2002. *Cognitive neuroscience. The biology of the mind*, 2nd edn. New York: Norton.

Germano, F. 2007. Stochastic evolution of rules for playing finite normal form games. *Theory and Decision* 62: 311–33.

Gigerenzer, G., P. M. Todd and the ABC Research Group. 2000. *Simple heuristics that make us smart*, New York: Academic Press.

Glimcher, P. W. 2003. *Decisions, uncertainty, and the brain*. Cambridge, MA: MIT Press.

Glimcher, P. W. and A. Rustichini. 2004. Neuroeconomics: the consilience of brain and decision. *Science* 306: 447–52.

Güth, W. and M. Yaari. 1992. Explaining reciprocal behavior in a simple strategic game: an evolutionary approach. In *Explaining process and change – approaches to evolutionary economics*, ed. U. Witt 23–34. Ann Arbor: The University of Michigan Press.

Harbaugh, W. T., K. Krause, and T. Berry. 2001. On the development of rational choice behavior. *American Economic Review* 91: 1539–45.

Haug, H. 1987. Brain sizes, surfaces, and neuronal sizes of the cortex cerebri: a stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *American Journal of Anatomy* 180: 126–42.

Kosfeld, M., M. Heinrichs, P. Zak, U. Fischbacher and E. Fehr. 2005. Oxytocin increases trust in humans. *Nature* 435: 673–6.

Livnat, A. and N. Pippenger. 2006. An optimal brain can be composed of conflicting agents. *Proceedings of the National Academy of Sciences*, USA 103: 3108–202.

McCabe, K. A. 2008. Neuroeconomics and the economic sciences. *Economics and Philosophy*. 24.

Robson, A. and H. S. Kaplan. 2003. The evolution of human life-expectancy and intelligence in hunter-gatherer societies. *American Economic Review* 93: 150–69.

Roth, G. and U. Dicke. 2005. Evolution of the brain and intelligence. *Trends in Cognitive Sciences* 9: 250–7.

Schierwater, B. 2005. My favourite animal, trichoplax adhaerens. *Bioessays* 27: 1294–302.

Weibull, J. 1995. *Evolutionary game theory*. Cambridge, MA: MIT Press.

Zak, P., R. Kurzban and W. T. Matzner. 2005. Oxytocin is associated with human trustworthiness. *Hormones and Behavior* 48: 522–7.