
*Proposal full title:*

# INSIGHT: Darwinian Neurodynamics

*Proposal acronym:*

**INSIGHT**

*Type of funding scheme:*

Collaborative project

**FP7-ICT-2011-C FET Open**

## D2.2

## Bayesian insight in human search

*Name of the coordinating person:*
prof. Eörs Szathmáry
Coordinator email: **szathmary.eors@gmail.com**
Coordinator phone: **+49 89 45209 35-30**
Coordinator fax: **+49 89 45209 35-31**

# Revisions Table

| | | | |
|---|---|---|---|
| **Due delivered date** | 31 August 2016 | **Actual delivered date** | 31 August 2016 |
| **Lead beneficiary** | PARMENIDES | | |
| **Beneficiaries involved** | PARMENIDES | | |
| **Authors** | Michael Öllinger, Anna Fedor, Eörs Szathmáry | | |
| **Dissemination level** | PU | **Nature** | R |

| REV | Work performed | Reviewers | Beneficiary |
|---|---|---|---|
| **0** | Production of the document | Anna Fedor, Michael Öllinger, Eörs Szathmáry | PARMENIDES |

# Table of contents

## 1. Executive summary

We have analysed a classical cognitive task, the Wisconsin Card Sorting Test (WCST) in terms of Bayesian and evolutionary dynamics. Only the latter is able to generate novel hypotheses, in contrast to a simple Bayesian approach or the classical Dehaene-Changeux neuronal model. We suggest an evolutionary search architecture for this and similar tasks. The dynamical model will be simulated in the near future.

## 2. Introduction

The first thing to note is that only recently has the fundamental link between Darwinian dynamics and the Bayesian inference been realized. The clearest, and still pioneering, insight comes from Harper (2010) who calls attention to a remark by Ronald Fisher, pioneering statistician and one of the founders of population genetics:

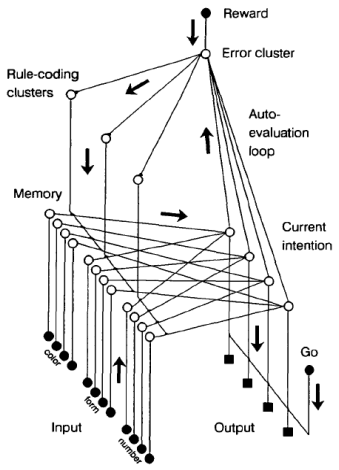
"*Inductive inference is the only process known to us by which essentially new knowledge comes into the world*" (Fisher, 1937).

Compare with the above with a quote from Dawkins:

"*The theory of evolution by cumulative natural selection is the only theory we know of that is in principle capable of explaining the existence of organized complexity*" (Dawkins, 1986)

As noted by Fernando et. al. (2012), this link is potentially tremendously important for neurobiology and cognitive science. This insight seems to percolate neurobiology: incidentally, this effect can be seen as a major contribution from the Insight project. We mention two examples: the recent opinion review on temporal relevance of brain anatomy (Friston & Buzsáki, 2016) and a paper in Frontiers in Systems Neuroscience (!) with the title "Universal Darwinism as a process of Bayesian inference" (Campbell, 2016).

**A classical cognitive task: The Wisconsin Card Sorting Test**




This task needs the integrity of the prefrontal cortex. A classical model of the task was provided several years ago by Dehaene and Changeux (1991). A few components of the model are absolutely important for the Insight project, namely: (i) the existence of the autoevaluation loop, meaning that we are dealing with offline thinking without action until the hypothesis considered satisfies a certain condition, (ii) the administration of reward, (iii) serial random search in hypothesis space by the application of WTA (winner-take-all) dynamics in the neuronal network. Note that even this simple model implements *elementary* Bayesian inference: the initial state of the neuronal network (the modelled prefrontal cortex) holds the candidate hypotheses, and reward gives the evidence that leads to the posterior distribution. An evolutionary version of this approach (not necessarily applicable to WCST) would require parallel search with copying. In order for the latter process to be effective WTA has to be replaced by WSA (Winner-share-all), which is implementable by reduced lateral inhibition (see Szilágyi *et al.* 2016 for the proposed neurobiological foundations).

## 3. Core report

Bayesian update does not account for the generation of new candidate hypotheses; it only accounts for the selection of already existing variant hypotheses. This is where Darwinian Neurodynamics could come into play, because it provides a mechanism just for that. Hypotheses are evolutionary units that are selected just like in Bayesian models, but they are also capable of multiplying with heredity and variation, thus implementing full evolutionary search. As Friston and Buzsáki (2016) remarked: *"The Bayesian brain falls short in explaining how the brain creates new knowledge"* (p. 9). We suggest that neuronal evolutionary dynamics might serve as a remedy.

### 3.1. Bayesian update and replicator dynamics (Harper, Campbell)

Harper (2010) describes how Bayesian update is equivalent to selection. Bayesian update is based on the following equation:

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{P(E)} = \frac{P(E \mid H_i)P(H_i)}{\sum_{i=1}^{n} P(E \mid H_i)P(H_i)} \; for i = 1,2,\ldots,n, where$$

- $H_i$ stands for hypothesis i
- $P(H_i)$ is the prior probability of $H_i$ and all hypotheses constitute the state space (the sum of the prior probability of all hypotheses equals to 1)
- E is evidence
- P(E) is the marginal probability of E; this can be calculated as in the third part of the equation
- $P(H_i|E)$ is the posterior probability of $H_i$ given E

The posterior probability calculated by the above equation serves as the prior probability for the next step of update based on new evidence.

The replicator equation is as follows:

$$x_i' = \frac{f_i(x)x_i}{\bar{f}(x)} = \frac{f_i(x)x_i}{\sum_{i=1}^{n} f_i(x)x_i}, \; for i = 1,2,\ldots,n, where$$

- $x_i$ is the frequency of type i and the types completely describe the population (the sum of all frequencies equals to 1)
- $f_i(x)$ is the fitness of type i dependent on the population distribution
- f(x) is the average fitness of the population
- $x_i'$ is the frequency of type i in the next generation

It is easy to see how the two equations are similar. Both progress is steps, in case of Bayesian models, these are called updates, whereas in case of selection, the steps are called generations. Bayesian update calculates the probability of the hypotheses whereas selection calculates the relative frequencies, but if types are probabilistically chosen in proportion to their relative frequencies, then these two are the same. The probability of the evidence given $H_i$ is equivalent to the fitness of $H_i$ (or $x_i$).

## 3.2. An example: The Wisconsin Card Sorting Test

We would like to use a standard psychological test, the Wisconsin Card Sorting Test (WCST) to show how Bayesian update and the replicator equation leads to the same inference in a given problem. The Wisconsin Card sorting Test is a classic test in psychology, which mainly tests functions of the prefrontal cortex. The task is to sort a deck of cards into one of four possible positions (position A, B, C, D from left to right), occupied by four key cards (Figure 1). The cards have different number of different figures on them in different colours. The cards only differ in these three dimensions, namely, colour of figures, number of figures and shape of figures. Each dimension has four possible features, see Table 1, i.e., there are $4^3 = 64$ possible cards.

The participant can move the top card from the deck on top of one of the key cards. After the participant moves a card, he gets feedback whether the move was correct or not. A move is considered correct if it complies with the current rule. There are three possible rules: colour rule, shape (or form) rule and number rule. The game goes through all three rules in random order. The rule changes when either the participant moves a certain number of cards correctly or when he exceeds the number of allowed moves.

All three rules are similarity-based rules taking into account one of the three dimensions, e.g., when the current rule is the shape rule, any card with a star on it should go to the key card with the stars on it, when the current rule is the number rule, every card with three figures on it should go to the key card with three figures on it. It seems that for most healthy participants, these rules are easy to find.

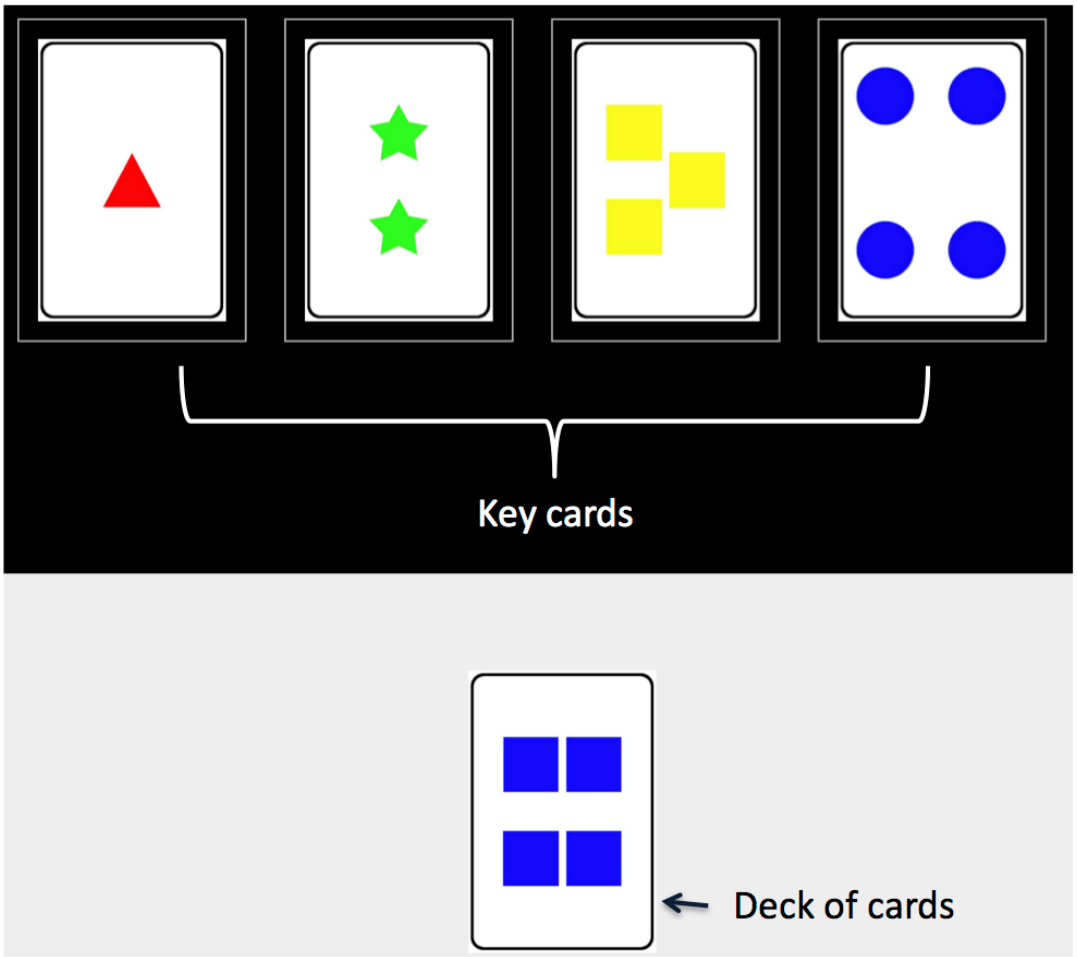



Figure 1. The layout of the Wisconsin Card Sorting Test.

Table 1. Dimensions and features in the WCST.

| Number | Colour | Shape | Index |
|---|---|---|---|
| One | Red | Triangle | 1 |
| Two | Green | Star | 2 |
| Three | Yellow | Square | 3 |
| Four | Blue | Circle | 4 |

Let's start from the assumption, that the three basic rules are readily available for the problem solver and that he does not consider any other rule. The current rule is the shape rule and we will show how the problem solver finds this rule in two steps based on Bayesian inference and based on discreet replicator dynamics. Both models have the same assumptions and will go through the same steps. If the probability of two or more rules are the same, the model chooses from them probabilistically, but for the sake of demonstration, it will choose the same rules in the two models. The first card will be a card with one green rectangle on it, and the second card will be a card with two red circles on it.

Table 2. Comparison of Bayesian update and discrete replicator dynamics for a hypothetical case of the WCST.

| **Bayesian update** | **Discrete replicator dynamics** |
|---|---|
| The prior probability of all hypotheses is the same:<br>$P(H_C) = P(H_N) = P(H_S) = 0.33$<br>Let's assume that the model chooses the shape rule first. | The frequency of all types are the same:<br>$x_C = x_N = x_S = 0.33$<br>Let's assume that the model chooses the shape rule first. |
| Move 1: The problem solver moves the first card to key card B according to the shape rule. It's incorrect.<br>one green rectangle -> two green stars -> INCORRECT | |
| The model calculates the posterior probabilities of all three hypotheses:<br>$P(H_C) = 0*0.33 / (0*0.33 + 1*0.33 + 1*0.33) = 0$<br>$P(H_N) = 1*0.33 / (0*0.33 + 1*0.33 + 1*0.33) = 0.5$<br>$P(H_S) = 1*0.33 / (0*0.33 + 1*0.33 + 1*0.33) = 0.5$<br>Let's assume that the model chooses the number rule from the two rules with the same probability. | The model calculates the frequencies of all three types for the next generation:<br>$x_C' = 0*0.33 / (0*0.33 + 1*0.33 + 1*0.33) = 0$<br>$x_N' = 1*0.33 / (0*0.33 + 1*0.33 + 1*0.33) = 0.5$<br>$x_S' = 1*0.33 / (0*0.33 + 1*0.33 + 1*0.33) = 0.5$<br>Let's assume that the model chooses the number rule from the two rules with the same frequency. |
| Move 2: The problem solver moves the second card to key card B according to the number rule. It's incorrect.<br>two red circles -> two green stars -> INCORRECT | |
| The model calculates the posterior probabilities of the remaining two hypotheses. Posterior probabilities calculated in the previous step become prior probabilities for this step.<br>$P(H_N) = 0*0.5 / (0*0.5 + 1*0.5) = 0$<br>$P(H_S) = 1*0.5 / (0*0.5 + 1*0.5) = 1$ | The model calculates the frequencies of the remaining two types. Frequencies calculated in the previous generation become frequencies for this step.<br>$x_N'' = 0*0.5 / (0*0.5 + 1*0.5) = 0$<br>$x_N'' = 1*0.5 / (0*0.5 + 1*0.5) = 1$ |
| The model found the shape rule. | |

Note, that if the whole fitness landscape is known in advance, the selectionist model can choose the correct hypothesis in just one step, i.e., after the first move.

## 3.3. Evolutionary approach

The problem with the above described approaches is their main assumption: (1) that the three basic rules are readily available, and (2) that the problem solver does not consider any other rule. The problem becomes more apparent, if we modify the task and introduce a new rule. None of the above approaches would be able to find the new rule, since there are no processes that would generate new rules. This is where our idea of applying true evolution has an advantage: it is not only capable of selecting from the already existing rules, but is also capable of generating new rule variants.

For evolution to take place, evolutionary units need to replicate with heredity and variation and hereditary traits need to influence the fitness of evolutionary units. Evolutionary units in the case of the WCST would be different rule variants and their fitness could be calculated based on already available evidence, i.e., feedback from previous moves. Rule variants could replicate by copying with errors, which would take care of new variations just like mutations. To explore this possibility, we added a new rule to the WCST. We describe our new experiment in the following section.

## 3.4. The modified WCST

We hypothesize that the three standard rules are easy to find for two reasons: (1) the dimensions are salient, visual cues and (2) the similarity based-rule is readily available without learning. A more difficult rule would be where the matching along a dimension is arbitrary (e.g., all red cards go to the yellow key card, all green cards go to the red key card, etc.), or where the dimension is hidden (e.g., the sequential order of cards), see Table 3. It is also possible to compose rules along more than one dimensions, but we will not deal with these complex rules here.

Table 3. Factors that influence the difficulty of a rule.

| | **Dimension is salient** | **Dimension is hidden** |
|---|---|---|
| Feature matching is similarity-based | Classic WCST with the three basic rules | Modified WCST with an index rule of 1234 sequence |
| Feature matching is arbitrary | Modified WCST | Modified WCST with an arbitrary index rule: insight task? |

We were interested in the search behaviour of participants, when none of the standard rules apply. In our modified version of the WCST task, after passing the three standard rules, participants had to learn a fourth rule, the index rule. This rule is based on a hidden or less salient dimension: the index of cards in a sequence of four cards. In one version of the task, a small number counts from 1 to 4 below the cards, so the index is visible, but probably less salient than the shape, colour and number of figures on the cards; in another version, the index is not shown. Moreover, the matching between the index and the target position is arbitrary, i.e., it has to be learnt.

To find this new rule, participants have to extend their search space beyond the three basic dimensions and the similarity-based matching rule. This is a feature shared with insight tasks: In insight tasks, most participants start searching for the solution in a restricted search space, which they have to extend (representational change) to find the solution. They unintentionally restrict the search space because some

features of the task mislead them, or because their previous experience with similar tasks tells them that the solution is in the restricted search space. Insight tasks can be pragmatically defined by this need to extend the search space. A phenomenological approach would set the criteria of a subjective feeling of Aha! which usually accompanies the representational change or finding the solution. Of course, both criteria depend on the individual: it is possible to start the search in the extended search space from the beginning, and it is possible to solve a task without experiencing the Aha!

We constructed the new rule to comply with the pragmatic definition of insight tasks, and were interested whether it elicits the Aha! feeling from participants.

## 3.5. Experimental design

The current design is a 2x2 design (Table 4), with two manipulations: using a special card and changing the order of rules. The order of rules will be either color rule -> number rule -> shape rule -> index rule, or index rule -> color rule -> number rule -> shape rule. We hypothesize that when the index rule is the last one, going through all three standard rules first would reinforce the importance of the salient basic cues of color, number and shape, possibly evoking a mental set, which would make finding the index rule harder.

The special card is a card in Figure 2. It differs from all other cards in the colour, number and shape dimensions, because it has five white half-moons on it. This card is used when the current rule is the index rule. Our hypothesis is that since this card does not match any of the key cards along the basic dimensions, when using this card, participants will not be misled by the salient basic cues, so they would find the index rule easier.

Table 4. Experimental design.

| | **Standard cards for all four rules** | **Special card for the index rule** |
|---|---|---|
| Order of rules: color rule, number rule, shape rule, index rule | 1: Index last with standard cards | 3: Index last with special card |
| Order of rules: index rule, color rule, number rule, shape rule | 2: Index first with standard cards | 4: Index first with special card |

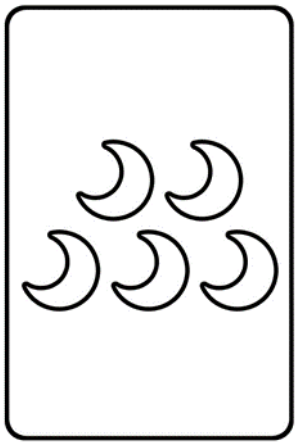


Figure 2. The special card.

Using unambiguous cards, each move complies with exactly one of the four rules. Of course it is possible that participants consider other than these four simple rules, but we won't be able to detect it. Extension of the search space is possible not only by considering the index dimension and arbitrary matchings along a dimension, but also by considering a rule which takes into account more than one dimension. More than 4 dimensions are possible to imagine within the task too, like matching between the current card and cards before the last card: "if the card before the last card had a star on it, I put the current card on position B"

It is also possible to consider further dimensions outside the task, like if I hear a siren wailing outside on the street I put the card on the left, if I don't hear a siren, I put the card on the right. We think, that it is unlikely that participants would consider rules outside of the computerized game.

## 3.6. Analyses

The modified WCST experiment is in the piloting phase. To improve reproducibility and transparency, we are going to preregister the experiment on the Open Science Framework, which means that we have to write all analysis scripts in advance.

The experiment serves a dual purpose. From the cognitive psychological point of view, we can test whether the modified WCST is an insight task at least to some of the participants. It tests the pragmatic hypothesis that the need to extend the search space makes a task an insight task. We will also test specific hypothesis for the task, namely:

1. Going through the standard rules first makes the task of finding the index rule harder
   a. Condition 1 will be more difficult than condition 2
   b. Condition 3 will be more difficult than condition 4

2. Using the standard cards for the index rule makes the task of finding the index rule harder
   a. Condition 1 will be more difficult than condition 3
   b. Condition 2 will be more difficult than condition 4

From the point of view of Darwinian Neurodynamics this task is also interesting, since it is a task where the fitness landscape only reveals itself through many steps, or iterations of the game. We will compare a Bayesian and an evolutionary model for this task and see which one is closer to human behaviour.

## 3.7. Connectionist model of WCST

We will combine the attractor network based Darwinian Cognitive Architecture with the model of Changeux to provide a model that is capable of solving the modified WCST.

The Changeux model (Figure 3) is one of the connectionist models that successfully replicated some features of human behaviour in the WCST. The three standard rules are hard-wired in the model: it has input in only three dimensions, and also within these dimensions feature matchings are one-to-one.

The rule-coding clusters have a misleading name, because rules are actually coded in the memory -> intention connections, where every memory unit has a connection with only one intention unit. Each rule-coding cluster gates these connections in one dimension. We think that a different interpretation is equally

consistent with human behaviour. We would rename rule-coding clusters to attentional clusters, and they would gate input to the memory clusters.

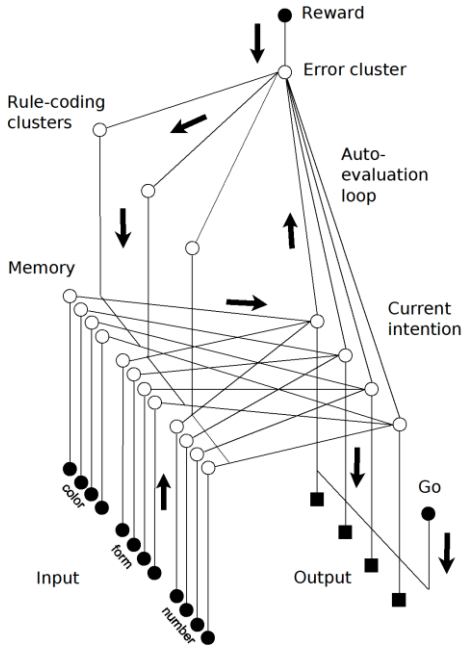



Figure 3. The Changeux model.

## 3.8. The modified Changeaux model

We are interested what happens, if none of the standard rules are sufficient. Participants need to extend their search space to find the new rule. They can do this in three basic ways (and the combinations of these):

1. They can consider an arbitrary rule instead of the similarity based rule within a dimension. This means that they need more than the hard-wired connections, i.e., the memory and the intention units would potentially form a fully connected network (all-to-all connections). Through trial-and-error certain connection would strengthen (Hebbian learning when the feedback is positive), and a new rule could emerge, e.g., red goes to yellow, green goes to red, etc.
2. They can consider more than one dimension at the same time. This means, that more than one attentional cluster would be inactive and allow activation to flow from memory to intention units. If they still consider only the hard-wired similarity-based connections this cannot result in a sufficient rule that has a matching to all cards. E.g., if the color and the form units are active, red triangles would be matched to the red triangle, but there would be no unambiguous matching for red stars. Because of this, this extension would probably occur together with the arbitrary rule extension.
3. They can consider more than the original three dimensions. In our experiment this dimension would be the index of the card in their serial order. This dimension is hidden/less obvious than the other dimensions.

To accomodate these possibilities in the model, we made the following modifications (Figure 4):

1. All-to-all connections between the memory and intention layers.
2. The possibility of more than one active dimension.
3. A fourth dimension, namely, index.

The red components represent the neural replicator systems and its connections. The neural replicator systems consists of a population of attractor networks that perform a Darwinian search as it is described in our previous paper. There are two connections between the neural replicator system and the Changeux model.

First, the best selected pattern from the neural replicator system is fed into a layer of 64 neurons which perform synaptic gating on the memory->intention connections. With these connections any rule can be coded that is based on either one or more of the four dimensions .

Second, the reward feeds into the neural replicator system. The current rule should modify the system based on the reward. It is obvious that some part of the current rule is correct if there was a positive feedback, and incorrect if there was a negative feedback. But the rules should code all matchings between cards, not just the current card. Moreover, the system does not know which dimension(s) are important.

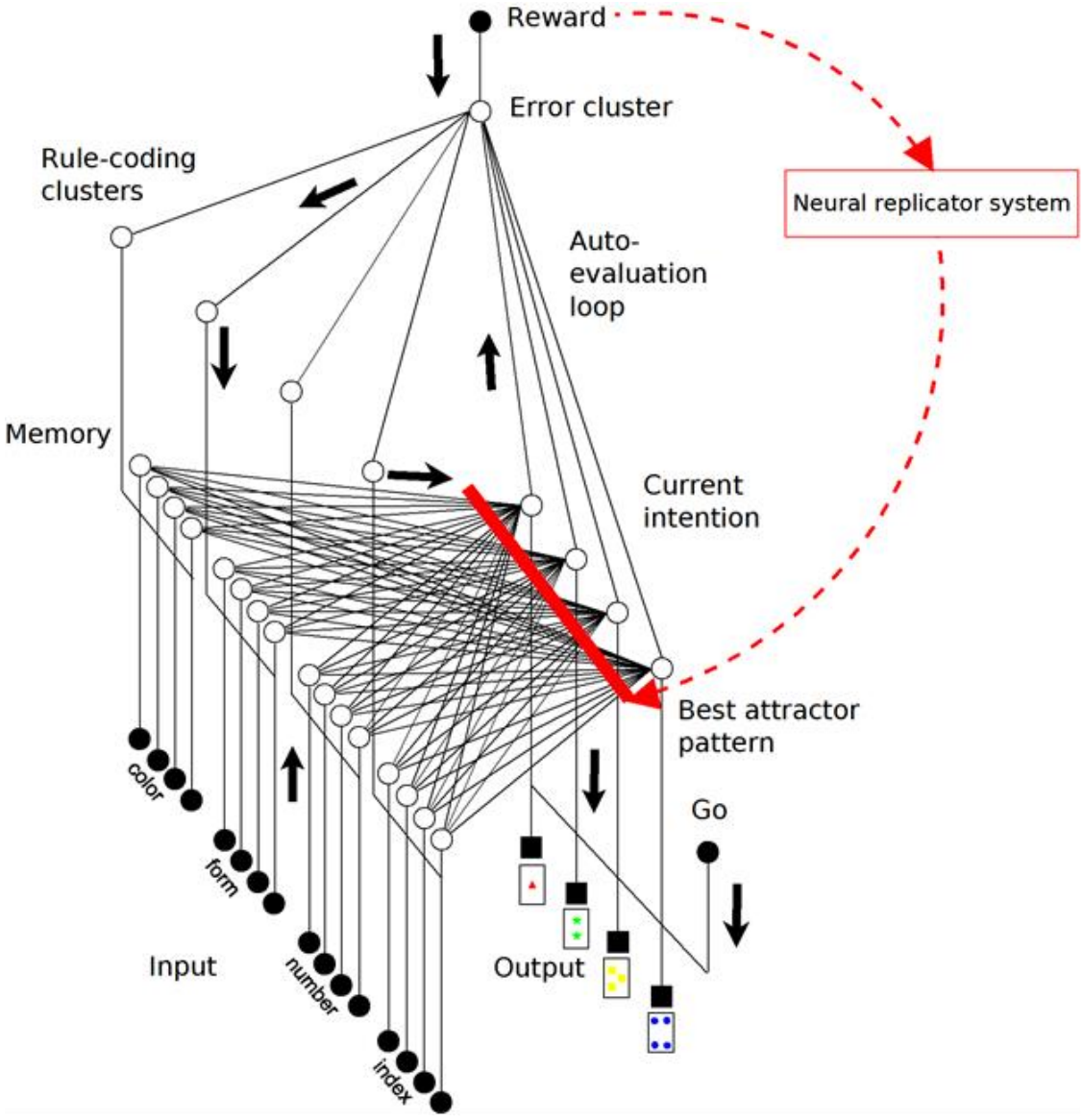



Figure 4. Combination of the Changeux model and Darwinian Cognitive Architecture.

Simulation of this model will happen in the near future.

## References

Campbell, J.O. (2016) Universal Darwinism as a process of Bayesian inference. *Front. Syst. Neurosci.* 10:49.

Dawkins R. (1986) *The Blind Watchmaker.* New York: W. W. Norton & Company.

Dehaene S & Changeux J.-P. (1991) The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cereb. Cortex* 1, 62-79.

Fernando,C., Szathmáry,E., andHusbands,P. (2012) Selectionist andevolutionary approaches to brainfunction: a critical appraisal. *Front. Comput. Neurosci.* 6:24.

Fisher,R.A. (1937). *The Design of Experiments,* 9th Edn. NewYork,NY: Macmillian.

Friston, K., Buzsáki, G (2016) The Functional Anatomy of Time: What and When in the Brain. *Trends Cogn. Sci.* 20, 500-511.

Harper, M. (2009). The Replicator Equation as an Inference Dynamic. *arXiv:0911.1763* [math.DS].

Szilágyi A, Zachar I, Fedor A, de Vladar HP, Szathmáry E (2016) Breeding novel solutions in the brain: A model of Darwinian neurodynamics. *F1000Research* accepted.