On identification of neural correlates of reflection in simple recurrent neural networks

S Bartsev¹,²,³, P Baturina² and G Markova²

¹ Institute of Biophysics SB RAS, Federal Research Center, Krasnoyarsk Scientific Center SB RAS, 50, Akademgorodok, Krasnoyarsk, 660036, Russian Federation
² Siberian Federal University, 79 Svobodny pr., Krasnoyarsk, 660041, Russian Federation
³ E-mail: bartsev@yandex.ru

Abstract. Reflection, that in a general sense means internal representation of the external world, refers to one of awareness levels observed in animals. In this paper we demonstrate the ability of a homogeneous recurrent neural network to solve a problem that requires a reflection. The delayed matching to sample test was chosen as a task which is impossible to pass without an internal representation of an external world. Experiments showed that simple recurrent neural networks can form these representations and store them as neuron firing patterns for several clock cycles. Although the trained network was able to distinguish these patterns easily, the identification of certain stimulus by neuron firing was not practically possible due to minor differences in the level of synchronous firing of a given neuron for different stimuli. Neural networks were shown to be applicable for modeling reflexive abilities, so these simple models may also be used for creation of general technique that ultimately can be applied to recognizing neural correlates of human consciousness.

1. Introduction
The problem of consciousness is one of the most difficult and intriguing problems of science. Attempts by philosophers and scientists to find out what the nature of consciousness is did not lead to significant and unambiguous results [1]. Moreover, there is a scientific position [2] that there is no need to understand what consciousness is, it is necessary to correctly describe it, as physics does, for example, in relation to electricity. We do not know what an electric charge is, but we are able to use knowledge about its properties. Physics has long used the formulas for the free fall of bodies without having an answer to the question of what gravity and mass are. In fact, the approach to identifying neural correlates of consciousness (NCC) proposed by F.Crick and C.Koch [3] and an attempt to reproduce the cognitive abilities of humans and animals using neural models completely conform to methodology of physics.

The identification of NCC in the course of experiments with people has obvious positive aspects, namely, the presence of self-report of the subjects about their feelings during measurements, and the ability to give tasks to the subjects and record the corresponding reaction. But, on the other hand, the great complexity of the human brain makes it difficult to identify the minimal neural ensembles and activity responsible for certain conscious phenomena, which is proclaimed the goal of research within the framework of the NCC concept [3].
At the same time, it becomes more and more obvious that there are different levels of consciousness. So in the review article [4], gradations of information processing in animals corresponding to different levels of awareness are introduced. The very first level of computation performed in the brain is “C0: unconscious processing”, which provides invariant pattern recognition, access to meaning, cognitive control, and reinforcement learning. These functions are modeled with varying degrees of success in deep learning neural networks.

The next level of computation is “C1: global availability of information”, which provides all-or-nothing choice and broadcasting of a relevant content, flexible routing of information, sequential execution of several tasks, since sequential computations or strategies require conscious perception. We are not aware of the reproduction of this type of computation in modern neural network models. Natural language processing performed by recurrent neural networks rather refers, in accordance with its definition, to C0.

The most advanced level of computing is "C2: Self-monitoring", which provide error detection, self-control, assessment of someone's knowledge, self-assessment. There is even no need to talk about reproducing this level of computation in artificial neural networks.

This allocation of levels of awareness leads to the assumption of the existence of consciousness not only in humans, but also in other organisms. So in the paper [5] it is directly stated that the detected activity of neurons in the brain of crows, which solve the problem of visual detection with a delayed response, is an empirical marker of avian consciousness.

But the ability to solve a task distributed over time is not the property of only higher vertebrates. It has been shown that honey bees are able to pass the delayed matching to sample (DMTS) test [6], which directly indicates the presence an internal representation of the external world in the bee's brain and their ability to manipulate it. Moreover, the number of neurons in the brain of a bee does not exceed 1 million - four orders of magnitude less than in humans! The discovery of such complex cognitive processes usually associated with consciousness in such different organisms provides a unique opportunity for comparative analysis of brain structures and the course of neural processes solving similar tasks. It can give additional possibility to find some similarity or even unity of the structures, and understand the principle of organizing these processes of external world mapping. This mapping can be called reflection in a general sense and its presence corresponds to the C1 level. Reflection in the narrow sense of the word (what psychologists mean by reflection) corresponds to level C2. Extending the list of studied animals by neural network models that reproduce phenomena of reflection can play an important role in identifying the minimum neural organization capable of reproducing the markers of consciousness.

The selection of levels of awareness made in [4] is in good agreement with Lefebvre's statement [7] that reflection, i.e. reflection of the external world (reflection in the general sense) and reflection of oneself in this world (reflection in the narrow sense) is a key property or function of consciousness. Hence it follows that namely conditions for the emergence, existence and external manifestation of reflection are the issues of fundamental importance in the study and reproduction in artificial systems the behavior that we associate with consciousness. It is possible that the search for neural correlates of reflection will facilitate understanding of the nature of consciousness. In a certain sense, reflection is easier to model, since it is possible to clearly define the conditions in which it should arise when solving certain behavioral tasks, some of which are given when describing the levels of consciousness - C1 and C2 [4].

It seems that modeling reflection on simple neural network models will allow to get an idea on the mechanisms of its implementation in natural neuron systems and to assess the possibilities of studying it through observation of neural activity.

The goal of the work is to train a homogeneous recurrent neural network solving a problem that requires an internal dynamic representation of the external world, and to determine whether it is possible to identify neural correlates of performing reflection.

The delayed matching to sample test (DMTS) was chosen as a task which is impossible to solve without an internal representation of an external world. This test issued in psychology for comparative studies of the cognitive abilities of children and animals, for the diagnosis of Alzheimer's disease and
depressive disorders. According to this test, the subject is presented with a certain stimulus (image), and then after some time a second stimulus is presented and the subject must note the similarity or difference of these stimuli. It is easy to see that this test corresponds to the C1 grade.

2. Materials and methods
To estimate the fundamental possibility of passing DMTS, we chose a simple stimulus and not very complex, uniformly arranged recurrent neural networks.

We used recurrent neural networks (RNN) with two inputs, two output neurons and different number of internal neurons. The RNN output – $y_o^{(t)}$ at a time step $t$ is calculated using the following equations:

$$y_h^{(t)} = f_h(W_h \cdot y_h^{(t-1)} + W_i \cdot x^{(t)})$$
$$y_o^{(t)} = f_o(W_o \cdot y_h^{(t)})$$

where $W_h$, $W_i$, $W_o$ are the hidden, input and output weight matrices, $x^{(t)}$ is the input vector at time step $t$, vectors $y_h^{(t)}$ and $y_h^{(t-1)}$ represent the hidden neuron activations at time step $t$ and $t-1$. Functions $f_h(\cdot)$ and $f_o(\cdot)$ are the non-linear activation functions. For simplicity, neuron biases are omitted in the equations.

Activation functions of neurons are:

$$f_h(x) = \frac{1}{2} \left( \frac{x}{a + |x|} + 1 \right), \quad f_o(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ b \cdot x, & \text{if } 0 < x < 1, \\ 1, & \text{if } x \geq 1 \end{cases}$$

Special output neurons with piecewise linear activation function are chosen to provide an accurate 0/1 output signal, and the activation function of hidden neurons is quite traditional.

Neural networks were trained to distinguish in a continuous stream of events matching pairs of stimuli spaced by arrival time. Specifically, one of three randomly selected short-term stimuli – input vectors: A – (01), B – (10) and C – (11) received at the input of the neural network at a random moment in time. Then there was a pause, the duration of which could randomly vary from 3 to 6 clock cycles, and then a second stimulus, also randomly selected, entered. After this, the neural network at the third step after the presentation of the second stimulus produced a single signal on the first output neuron, if these stimuli coincided and a single signal on the second output neuron, if the stimuli did not match. After this, the response cycle was repeated. A fragment of the input signal and the response of the neural network is shown in figure 1.

![Figure 1](image.png)

**Figure 1.** A fragment of continuously generated sequence of input signals and responses of the neural network. The plus chain at the top displays the quality of matching the objective function. The second line of symbols depicts the flow of stimulus signals entering the neural network input. Out_1 and out_2 lines represent signals taken from the output neurons. The circles with dashed lines mark examples of neural network responses to different (A) and coinciding (B) stimuli.
To modify the synapses the back-propagation algorithm was used with the error propagation depth to the past equal to 6 clock cycles. It should be noted right away that for RNN DMTS-test is rather difficult task. The maximum duration of a pause between stimuli, which was achieved with a neural network of 25 neurons, is 6 cycles, and for a neural network of 30 neurons, 8 cycles.

We note also that our goal is to prove the feasibility of the internal representation of external world in small RNNs with simple structure. Determination of an effective structure of networks or effective options for the learning algorithms is beyond the scope of current study.

After achieving good training quality, a comparative analysis of neuron excitation patterns for various stimuli was performed for several neural networks. Obviously, with these input signals, neuronal firing patterns are determined by the structure of the RNN. However, there is one complication - we are dealing with fully-connected RNNs, where all neurons are equal and not arranged. For definiteness, we assign to neurons numbers when we form matrices of weight coefficients. It means we can have completely identical trained neural networks that perform this function equally well, but there will be no portrait similarity of the weighting matrices. However, we can apply special procedure for modifying the synapse matrix that preserves the functioning of the RNN [8-10] at the transposition of neurons. Then we can choose the appropriate criterion to rank the neurons and bring the different RNNs performing the same function to a similar form.

In this work the criterion of total contribution of each neuron to the functioning of a trained RNN was used for ranking. It was determined as follows: 1) the average signal of each neuron was calculated in the course of sufficiently long sequence of signals; 2) the sum of absolute values of outgoing (axon) synapses of each neuron was found; 3) the product of these two quantities for each neuron was found. The neurons of RNN were renumbered according to the rating of their contribution to the RNN operating. This procedure led to a certain ordering of neuronal firing patterns.

3. Results

An important result is that a homogeneous RNN can be trained to pass DMST and in a fairly wide interval of delays (from 3 to 6 clock cycles), i.e. a RNN does not “know” when the second stimulus will arrive exactly; it can appear with equal probability in any of the clock cycles, starting with the 3rd and ending with the 6th. In RNN an image of albeit an extremely primitive external world is formed and retained for a rather long time, which is then interacts with external stimulus and used to make a decision.

It was unexpected that the very way of presenting and preserving this image turned out to be quite complex and dynamic. For example, a diagram of the levels of neurons firing is shown with the duration of the pause between the appearance of the first and second C-type (11) stimuli in the case of 3 and 6 cycles (figure 2). It can be seen that the NN produced a clear response indicating the similarity of both stimuli.

Surprisingly, a clear correct answer is formed with a completely different general picture of firing on the third and sixth clock cycles after the arrival of the first stimulus. It is easy to see that the patterns of neuronal firing are also constantly changing at intermediate delay times. To assess the similarity of neuron patterns while waiting for the second stimulus with a 6-clock pause, we calculated the MSE(mean squared error (discrepancy)) between the level of excitation of each neuron at successive times for a given stimulus:

\[
MSE_i = \frac{1}{M(K - 1)N} \sum_{j=1}^{M} \sum_{k=1}^{K-1} \sum_{l=1}^{N} (\alpha^k_j - \alpha^{k+1}_j)^2
\]

where \(i\) is the number of hidden neuron, \(k\) is the number of the beat during the waiting period for the second stimulus, \(j\) is the conditional number of the combination (A, B, C), \(\alpha^k_j\) is the level of excitation of the \(i\)-th neuron at the \(k\)-th time step for the \(j\)-th type of stimulus, \(M\) is the number of stimuli (in our case, 3), \(K\) is the number of variable pause cycles (4), \(N\) is the number of hidden neurons.
Figure 2. Dynamics of neuron firing during the DMTS test with a delay of the second stimulus: A) 3 and B) - 6 clock cycles. Leaders indicate: 1) the moment of arrival of the second stimulus with a delay of 3; 2) the beginning of the formation of the neural network solution after the arrival of the second stimulus; 3) the beginning of the relaxation of the NN after the generation of the response.

And also the MSE discrepancy between the level of excitation of each neuron at synchronous moments of time, but with different stimuli (the designations are the same):

\[ MSE_2 = \frac{1}{MKN} \sum_{j \neq j} \sum_{k=1}^{M} \sum_{i=1}^{N} (x_i^k - x_{i+1}^k)^2 \]

The results of MSE calculations are shown in the table 1.

| Difference between levels of neuron firing at different time moments and the same stimulus (\(MSE_1\)) | Clock cycles after the first stimulus | N1 | N2 | N3 | Sum |
|---|---|---|---|---|---|
| 3 | 0.111 | 0.082 | 0.032 | 0.225 |
| 6 | 0.158 | 0.112 | 0.105 | 0.374 |

| Difference between synchronous levels of neuron firing at different first stimuli (\(MSE_2\)) | Clock cycles after the first stimulus | N1 | N2 | N3 | Sum |
|---|---|---|---|---|---|
| 3 | 0.139 | 0.015 | 0.066 | 0.220 |
| 6 | 0.131 | 0.013 | 0.050 | 0.195 |

As can be seen from the table, the differences between the level of neuron firing at different times for a given stimulus are, on average, greater than the differences in the level of synchronous firing of a given neuron for different stimuli. From this it follows that there is practically no possibility, based on
the appearance of neuronal firing, to understand what stimulus was received by the neural network (artificial or natural), unless a rigid binding to the given points in time is provided and the neural network does not change during the experiment, as it was for the considered RNN.

Student’s test for comparing means shows no significant difference between MSEs. We do not conduct rigorous statistical estimates on the reliability of differences. The main fact that we consider is that neuronal firings are comparable. Moreover, the similarity between patterns associated with different stimuli is shown. On the other hand, we demonstrate the problem of neural correlates detecting even concerning such extremely simple RNNs.

Furthermore, in the context of revealing neural correlates of reflection it is important to note that three considered neural networks have formed their own presentation of stimuli (figure 3). This may be a hint that the characteristic patterns of conscious phenomena may differ significantly from person to person, and the experience gained on one person cannot be directly transferred to another.

![Figure 3](image)

**Figure 3.** Fragments of records of neuron firing for three different neural networks. For clarity, we selected a mode with a fixed pause duration between two stimuli – 6 clock cycles.

### 4. Conclusion

Experiments show that simple recurrent neural networks can form the internal representations of external stimuli and apply them for passing the delayed match to sample test. It is possible for trained networks to store these representations, which are presented in the neuron firing patterns, for several clock cycles (from 3 to 6 in our case).

Although the trained network is able to distinguish these patterns easily, the identification of certain stimulus by neuron firing is not practically possible due to little differences in the level of synchronous firing of a given neuron for different stimuli. This let us to conclude that, in general case, there is no simple correlation between the function performed by neural network and its neuron firing portrait (that in the way of DMTS passing can be called neural correlate of reflection).

The situation with detecting neural correlates of consciousness associated with specific function in the brain of humans and higher animals is more complicated because of incomparably more complex organization of their nervous system. Moreover, experimental results show that characteristic patterns of functioning are individual even regarding the same function.
We believe that the development of the special method of firing pattern recognizingshould solve this problem. As we demonstrate, neural networks are suitable for modeling reflexive abilities, so these simple models may also be used for creation of general technique that ultimately helps to recognize neural correlates of human consciousness.

Acknowledgments
The reported study was funded by FASIE, project number 138GRCTC10-D5/61873.

References
[1] Revonsuo A 2010 Consciousness: the Science of Subjectivity (Hove: Psychology Press)
[2] Khrennikov A Yu 2004 Modeling Thinking Processes in p-adic Coordinate Systems (Moscow: Fizmatlit)
[3] Crick F and Koch C 2003 A framework for consciousness Nature Neuroscience 6(2) 119-26
[4] Dehaene S, Lau H and Kouider S 2017 What is consciousness, and could machines have it? Science 358(6362) 486-92
[5] Nieder A, Wagener L and Rinnert P 2020 A neural correlate of sensory consciousness in a corvid bird Science 369(6511) 1626-9
[6] Loukola O J, Perry C J, Coscos L and Chittka L 2017 Bumblebees show cognitive flexibility by improving on an observed complex behavior Science 355(6327) 833-6
[7] Lefebvre V A 2003 Reflexion (Moscow: Cogito-Centre)
[8] Bartsev S I and Bartseva O D 2002 Symmetric Structures and Equifinality of Evolution Outcomes in Simple Neural Network Models DokladyBiochem. and Bioph. 386 235-8
[9] Bartsev S I and Bartseva O D 2006 Functional Invariant Approach to the Biological System Uniqueness: a Simple Neuronet Model DokladyBioch. and Bioph.406 15-8
[10] Bartsev S I and Bartseva O D 2010 Heuristic Neural Network Models in Biophysics: Application to the Problem of Structural and Functional Conformity (Krasnoyarsk: Sib. Fed. Univ.)