Artificial Neuroconsciousness an Update

Igor Aleksander
Department of Electrical and Electronic Engineering, Imperial College
London, UK

Abstract

The concept of a theory of artificial neural consciousness based on neural machines was introduced at ICANN94 (Aleksander, 1994)[15]. Here the theory is developed by defining that which would have to be synthesized were consciousness to be found in an engineered artefact. This is given the name "artificial consciousness" to indicate that the theory is objective and while it applies to manufactured devices it also stimulates a discussion of the relevance of such a theory to the consciousness of living organisms. The theory consists of a fundamental postulate and a series of corollaries. In this paper the series of corollaries is extended and illustrated by means of characteristic state structures. Studies of artificial neuroconsciousness aim at two results: first to provide a single perspective on many mechanisms which perform cognitive tasks; and second, it provides an explanation of consciousness which stands alongside the many discussions found in the literature of the day[1-4].

1 Theory

The theoretical framework used in this work has one fundamental postulate from which follow 12 corollaries. This framework has been inspired by Kelly's[5] theory of "personal constructs"which explains the causes of personality differences in human beings.

The Fundamental Postulate: Consciousness and Neural Activity.

> *The personal sensations that lead to the consciousness of an organism are due to the firing patterns of some neurons, such neurons being part of a larger number which form the state variables of a neural state machine, the firing patterns having been learned through a transfer of activity between sensory input neurons and the state neurons.*

The words of this postulate are intended to have specific meanings which need to be stressed so that the corollaries which follow should make sense.

Personal sensation: Much of the controversy surrounding consciousness comes from the problem of infinite regress. Here it is implied that neural activity leads directly to personal sensation so dismissing the problem of infinite regress.

Firing patterns: Neurological terminology has been adopted to refer to the output activity of a group of neural elements. In an artificial system 'firing patterns' could refer to any measurement of the output quantity of the elements which constitute that system.

Neurons: This adoption of this neurological term is used to indicate that the theory is that of a cellular system where "neuron" is the name given to a basic cell.

Neural state machine: A state machine is the most general model of a finite computing process - it calls on the concept of an inner state which is a function of input sequences. Neural versions assume that neurons generate the variable values which, when taken together, form a state. (Corollary 1 formalises this notion and the generality of neural state machines has been argued elsewhere[6]).

Learned: Neurons are assumed to be plastic and it is this plasticity which allows them to learn meaningful, representational, firing patterns.

Iconic Transfer: This key property relates to the source of information which controls the learning of the neurons. It will be seen that distal, sensory information is postulated to impose output patterns on neurons so that these may be learned and recalled in the absence of input. It is this transfer that creates inner perception in the conscious organism.

Sensory Neurons: These are transducer neurons that transform energy from environmental input into the distal, sensory signals which control iconic transfer.

Corollary 1: The brain is a state machine.

The brain of a conscious organism is a state machine whose state variables are the outputs of neurons. This implies that a definition of consciousness be developed in terms of the elements of state machine theory.

Corollary 1 is a consequence of the intent in the fundamental postulate that the theory of artificial consciousness be based on state machine theory. State machines can model any system with inputs outputs, internal states and input-dependent links between such states. The states and their links form a state structure. Such machines can be probabilistic where links between states are defined as probabilities, they can have a finite or an infinite number of states. The fact that any conscious organism must have something called a brain with an attendant state structure is evidently true and not controversial. The key question is whether enough can be said about the nature of the state structure of organisms that are said to be conscious which distinguishes consciousness itself. This becomes the task for the corollaries which follow - to define the characteristics of state structure that are necessary for and specific to organisms that are said to be conscious.

Formalization of Corollary 1.
In any state machine, five items need to be defined:

i) The total ***input*** to the neural state machine is a vector **i** of input variables i_1, i_2 ...

$$\mathbf{i}=[i_1,i_2...]$$

The i_1, i_2 ..variables are the outputs of sensory neurons.
In living brains the number of such variables, being the number of neurons involved in the early layers of all sensory activity, is very large but finite. There is also some debate about whether it is important for these variables to be considered as binary (firing or not) or real (firing intensity per unit time). While it will be seen that this decision does not alter the course of the theory, it is assumed here that these variables are binary. This is done without loss of generality but with the gain that, using the methods of automata theory, it becomes possible to develop non-linear models.

Also, **I** is defined to be the set of all possible input vectors.

ii) The total ***output*** of the neural state machine is a vector **z** of output variables z_1, z_2 ...

$$\mathbf{z} = [\, z_1, z_2 \,...]$$

The z_1, z_2 .. variables are the outputs of 'actuator' neurons.

Again the variables z_1, z_2.... are considered to be binary, and, in living brains, would be seen as the output parts of the brain which are responsible for muscular action .

Also, **Z** is said to be the set of all possible output vectors.

iii) The ***inner state*** of the neural state machine is defined as a vector **q** of variables q_1, q_2 ..

$$\mathbf{q} = [\, q_1, q_2 \,...]$$

The q_1, q_2 .. variables are the outputs of 'inner' neurons.
Again, variables q_1, q_2 ... are binary, and, in brains, would be the states of neurons neither involved in input sensing nor output generation.

Also, **Q** is said to be the set of all possible input vectors.