ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Systems with 'subjective feelings' – the perspective from weightless automata.

Igor Aleksander[1] and Helen Morton[2]

1 and 2- Department of Electrical and Electronic Engineering, Imperial College, London, SW7 2PE

Abstract. These are Christof Koch's [1] closing remarks at the 2001 Swartz Foundation workshop on Machine Consciousness, Cold Spring Harbour Laboratories:

"… we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artefacts designed or evolved by humans."

This account is aimed at identifying a formal expression of the 'subjective feelings in artefacts' that Koch saw as being central to the definition of a conscious machine. It is useful to elaborate 'artefacts' as the set of systems that have a physically realizable character and an analytic description. The weightless character of the description dispels the notion that cognition and consciousness lurk within the weight values of a system.

1 Introduction

A 'basic guess', first suggested in 1996 (Aleksander [2], p.10) and used since then, governs the progress of this tutorial paper: *"The acceptedly problematic mind-brain relationship may be found and analyzed in the operation of a specific class of neural, experience-building machines* of which **weightless neural automata** are a prime example and a prescription for construction.. This tutorial paper is a journey through a progressively refined logical description of the characteristics of such machines.

The set of machines central to this paper, $M(F)$, is characterized by having inner state structures that encompass subjective feelings (*M* for 'machine' *F* for 'feelings'). Then with M as the set of all machines that can have formal descriptions, that is, as in Koch [1], the set of all ‘artefacts designed or evolved by humans’, the aim of this account is to develop a logical description of the set $M(F) \subset M$.

This is based, first, on satisfying a logical requirement that the internal states of $M(F)$ be phenomenological (that is, be *about* the surrounding world – the set of *non*-phenomenological objects that can be said to be conscious, is here understood to be empty), second, to define a logical structure which leads to such states being the subjective inner states of the artefact, third, how such subjectivity becomes structured into an inner state structure that forms the unique ‘mind’ of the individual artefact and, finally, how ‘feelings’ can be identified in this state structure. The latter calls on the concept of ‘Cognitive Phenomenology’ which encompasses internal states that are phenomenological in a way that does not involve sensory or bodily experiences. As stated, the formal artefact used in the paper is the ‘weightless neural state machine’ (Aleksander [2], p. 97) in the belief that this gives a logical expression to the concepts of the paper.

2 Machine Phenomenology

M needs to be partitioned into those systems that have inner states, *M(I)*, (pendulums, state machines, brains … i.e. systems whose action is dependent on inner states that mediate perceptual input to achieve action) as against those that do not, *M(~I)*, (doorbells, perforated tape readers, translation machines … i.e. systems whose action depends on current input only). The ‘human machine’ must belong to $M(I)$ and some of its inner states are the ‘mental’ states that feature in definitions of human consciousness.

So, $M(F) \subset M(I)$ and to head towards a definition of $M(F)$, $M(I)$ needs refining, which comes from the fact that the inner state must be a subset of a phenomenological set $M(P)$, that is, a set of machines in which the inner states can be *about* events in the world and for which *there is something describable it is like to be in that state.* That is, $M(F) \subset M(P),\ M(P) \subset M(I)$.

Crucially, an ‘aboutness’ in $M(P)$-type machines can be characterized as follows.

A particular machine A, where $A \in M(P)$, is influenced by a world, which in a simplified way, but one that does not distort the flow of the argument, produces a sequence of perceptual inputs to *A*

$I^A = \{i_1^A, i_2^A, i_3^A, \dots\}$

To be phenomenological, there needs to be a sequence of internal states in *A*
$S^A = \{s_1^A, s_2^A, s_3^A, \dots\}$
where s_j^A *is about* the corresponding i_j^A. This implies a coding that uniquely represents i_j^A. Indeed, the coding can be the same for the two or so similar as not to lose the uniqueness. This relationship is made physically possible at least through the *learning* property found in a neural state machine (or neural automaton) as pursued below.

3 Achieving phenomenology in neural automata

Here one recalls that in conventional automata theory the finite state dynamics of a general system from $M(I)$with inner states $\{a_1, a_2$ …} is described by the dynamic equation
$a(t) = f[a(t-1), e(t)]$
where $x(t)$ refers to the value of a parameter at time *t* and $e(t)$ is an external influence on the system, at time *t*. To aid the discussions and without the loss of relevance, time is assumed to be discretized. To **learn,** an *automaton* in condition $[a(t-1), e(t)]$ **is given** $a(t)$to become an element of *f* in the sense that it 'stores' $a(t)$ as indexed by $[a(t-1), e(t)]$. That is, given the automaton in $[a(t-1), e(t)]$, the next state entered by the automaton is the stored state $a(t)$. This storing function is achieved in a class of neural networks dubbed *neural automata* that are trained in a so-called *iconic* way (Aleksander [2], p.151). The simplicity of this learning process with respect to weighted processes is noted.

Reverting to automaton *A,* say it is in some state s_{j-1}^A and receives the perceptual input i_k^A , then the dynamic equation may be rewritten to drop the superscript *A* as only one automaton is considered:

$$s_j = f[s_{j-1}, i_j]$$

To be *phenomenological* there needs to be a similarity relationship between s_j and i_j so that s_j can be said to *be about* i_j.
That is, using ≈ to mean 'is equal to or uniquely similar to', then

$$s_j \approx i_j$$

This achieves a phenomenological relationship between *S* and *I*. Finally, it is noted that *f* is a mapping $S \times I \xrightarrow{f} S'$, where $S'$is the set of 'next' states while S is the set of current states.

4 Achieving subjectivity in neural automata

So far, the automaton described is phenomenological to the extent that it has inner states that are about previously experienced external states. However, subjectivity (irrespectively of some differing definitions of what it means) includes the ability to make functional use of the created states 'owned' by the entity in what would colloquially be called 'thought'. This first requires that internal phenomenological states can exist without the presence of input: a 'perceptually unattended' situation. This input is given the symbol φ and is characterized by not creating a phenomenological state. So, say that the input i_a occurs more than once during the learning process then, starting in some state s_x when i_a occurs for the first time, we have

$$[s_x, i_a] \rightarrow s_a,$$

where $(\rightarrow)$ reads, 'causes a transition to'.
Then if φ is applied to the input, we have

$$[s_a, \varphi] \rightarrow s_a$$

The result of this entire action may be depicted by the commonly used state transition diagram (Figure 1.)

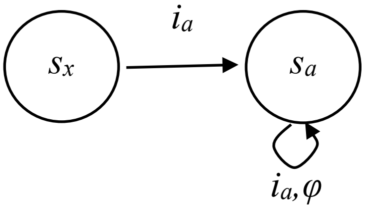

Figure 1: State diagram for the formation of subjective state s_a

So with input φ, s_a becomes a self-sustained state which is *about* the last-seen input i_a. A further step is that the same φ can occur in the creation of any single phenomenological state so that the automaton may be said to *own* the inner version of all externally experienced single events.

But generally, experience consists of sequences of external influences and the current formulation needs to be extended to internal representations of time dependent external experiences.

5 State structures and thought.

To make experiences incurred in time subjective, consider the input changing from $\boldsymbol{i_a}$ to $\boldsymbol{i_b}$. The relevant transition diagram then becomes (Figure 2)

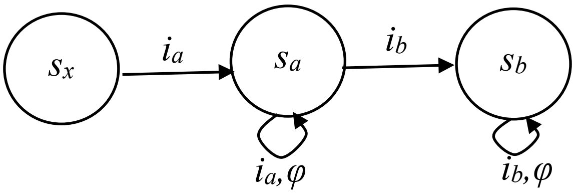

Figure 2: State diagram for the formation of the subjective experience of i_a followed by i_b

To take this further, it is recalled that these behaviours are subjective to the extent that they 'belong' to the automaton which physically performs the function

$$S \times I \xrightarrow{f} S'$$

where *f* is built up from experienced states and state transitions. It should be noted first that the automaton could be in some state s_p which on some occasions receives input i_q leading to s_q and other occasions i_r leading to s_r . Secondly it is asserted (but can be shown to be true in specific cases) that the neutrality of φ is such that it allows transitions to each of the learnt states in a probabilistic function. So, in the above examples, with φ as input, the automaton can change from state s_p to itself, s_q or s_r with probabilities determined by technological and learning exposure detail. The upshot of this is that the automaton develops a probabilistic structure of phenomenological states and transitions between them that are about past experience. This leads to the 'ownership' of *explorable* internal state structure, which, in the case of living entities, is called **thought**. One's life, and that of an artificial entity is based on a mix of inputs imposed by the world and φ, which allows thought to be driven by external perceptions, or previous internal states, that is, previous experience.

6 Attractors

Without going into detail about the statistical properties of neural networks, we note that for a particular input such as $\boldsymbol{i_a}$in figure

2, there is only one state that remains sustained in time, and that is s_a. It turns out that for some neural networks (including natural ones) starting in a state that is not about i_a, the states change getting more and more similar to $s_a$until s_a is reached. Then s_a is called an *attractor* under the input i_a. This issue returns in the consideration of volition below.

7 Action

The stated purpose of having subjective mental states is to enable the organism to act in some appropriate way in its world. This is closely connected to the concept of volition, as will be seen. Automata *action* is a concern in automata theory as, in general, an automaton, in addition to performing the next-state function $S \times I \xrightarrow{f} S'$ also performs an output function $S \xrightarrow{g} Z$, where Z is a set of output actions which in the most primitive entities, causes locomotion in its world. In more sophisticated entities language falls within the definition of Z. As with *f*, an automaton can learn to build up g as experience progresses. Here is an example. Say that the automaton can take four actions: movement in four cardinal directions, that is $Z = \{n, s, e, w\}$. The automaton can then either be driven in its (2D) world or it can explore it at random. In either case an element of $Z = \{n, s, e, w\}$ is associated with the state of the automaton and this determines the next input and associated state. Therefore, the state trajectory is now about a real world trajectory. The same principle applies to any other form of action, including language, in the sense that action, movement, utterances or, indeed, inaction, become associated with the state structure of the automaton leading, through the exploration of state trajectories to the ability to fix the brakes on a car, play the piano or plan an escape from jail.

8 Volition and attractors

Referring to the paragraphs on attractors, the input or an internal state could represent something that is wanted. The resulting trajectory to an attractor in a system that performs actions, internally represents the necessary actions for achieving the desired event. In the case of the automaton in the last section, this trajectory indicates the steps necessary to find that which is wanted. This is substantial topic and previous literature on this functioning may be found (Aleksander [2] pp

181-189, Aleksander [3] pp. 130 - 139). In fact, this activity is part of a set of five requirements for the presence of consciousness in an automaton (Aleksander [3] pp 23-39). (Detail of this is not necessary for the current discussion)

9 Feelings and Cognitive Phenomenology

It is the contention of a group of philosophers, Tim Bayne (Bayne and Montague [4]) Galen Strawson (Bayne and Montague [4] pp 285- 235) and Michelle Montague (Montague [5]) that classical phenomenology is too closely allied to perceptual and sensory events and therefore avoids the discussion of mental states related to meaning, understanding and abstract thought. Such states, it is argued, are *felt* alongside the sensory/perceptual. For example, were someone to utter a word in their own language, there is something it is like to understand such words hence there is a cognitive character to this phenomenology. Advocates of cognitive phenomenology argue that this feeling is common to all utterances that are understood. Similarly, an utterance that is not understood is accompanied by a feeling that is common to all non-understood utterances Within our work with automata it has been suggested that feelings of understanding or not, the presence or absence of meaning in perceptual input, language understanding and abstract thought are parts of the *shape* of state trajectories which affect the 'what it's like' to be in these trajectories (Aleksander [6]). For example, a heard word that is understood, will have a state trajectory (in the weightless state machine) that ends stably in an attractor. If not understood, the trajectory will be a random walk. In a machine, these differences in state behaviour warrant different descriptions which can be expressed in the action of the machine. This mirrors the way that perceptions and feelings warrant different actions in ourselves. Indeed, the effect of the two felt events on action can be very similar in machine and human. For example, an understood utterance (attractor) can lead to action whereas a non-understood one (random walk) may not.

10 Summary and conclusion: subjective feelings in machines.

In this tutorial paper , through the advocacy of weightless neural automata it has been shown that machines with progressive characteristics that lead to 'artefacts with subjective feelings' may be formally defined, within the challenge issued by Koch. The account relates to the design of machine phenomenology in neural automata, links of automata state structures to subjectivity, the importance of attractors in what can be termed 'thought', and how action is found in automata, leading to a consideration of volition as an influence on state trajectories. The paper includes a presence of 'feelings' through a consideration of cognitive phenomenology modelling. In sum, using the 'basic guess' that a neural automaton's physical neural structure relates to mental structure in living organisms has led to a description of how subjective feelings may be incorporated in a machine, that is, the engineering of a machine that is conscious of being a machine.

References

[1] C. Koch, Closing Remarks, Proceedings of the Swartz Foundation workshop on Machine Consciousness, Cold Spring Harbour Laboratories 2001.

[2] I. Aleksander, *Impossible Minds: My Neurons My Consciousness, Revised Edition,* London, Imperial College Press, 2015.

[3] I. Aleksander, *The World In My Mind, My Mind In The World,* Exeter, Imprint Academic, 2005

[4] T. Bayne and M. Montague, *Cognitive Phenomenology,* Oxford, Oxford University Press, 2011.

[5] M. Montague, Perception and Cognitive Phenomenology, Philosophical Studies (8):2045-2062, 2017.

[6] I. Aleksander, Cognitive Phenomenology: A Challenge For Neuromodelling, Proc AISB 17, 2017