Article

From Turing to Conscious Machines

Igor Aleksander [ID]

Electrical and Electronic Engineering Department, Imperial College, London SW7 2BT, UK;
i.aleksander@imperial.ac.uk

Abstract: In the period between Turing's 1950 "Computing Machinery and Intelligence" and the current considerable public exposure to the term "artificial intelligence (AI)", Turing's question "Can a machine think?" has become a topic of daily debate in the media, the home, and, indeed, the pub. However, "Can a machine think?" is sliding towards a more controversial issue: "Can a machine be conscious?" Of course, the two issues are linked. It is held here that consciousness is a pre-requisite to thought. In Turing's imitation game, a *conscious* human player is replaced by a machine, which, in the first place, is assumed not to be conscious, and which may fool an interlocutor, as consciousness cannot be perceived from an individual's speech or action. Here, the developing paradigm of machine consciousness is examined and combined with an extant analysis of living consciousness to argue that a conscious machine *is* feasible, and capable of thinking. The route to this utilizes learning in a "neural state machine", which brings into play Turing's view of neural "unorganized" machines. The conclusion is that a machine of the "unorganized" kind could have an artificial form of consciousness that resembles the natural form and that throws some light on its nature.

Keywords: living consciousness; machine consciousness; thought

1. Introduction

It is a matter of some importance that Turing stimulated a sustained debate among a swathe of philosophers, computational theoreticians, and artificial intelligence designers to consider the possibility that a machine could think, at least within the action of the "imitation game"[1]. Proudfoot [1] presents a clear analysis of the game, pointing out that Turing's approach is philosophically controversial, and open to continued discussion. In this paper, I refer to another area of Turing's long-lasting influence: the effect that his notion of a "universal computer" has had on the thinking-machine debate [2] (Section 4). Below, I suggest that thought in the brain is much more directly addressed through a "Neural State Machine" (NSM) than the universal computer. The NSM more closely resembles the structure of a brain, and so more easily falls into a category of machines that can be examined for their ability to think.

Interestingly, Copeland and Proudfoot [3] have noted that Turing, too, considered an alternative to the classical computing engine: the option of using networks of neuron-like units in what Turing called an "unorganized" machine. Such artificial neural networks, after first having been severely criticized by Marvin Minsky and Seymour Papert [4] as "computationally limited", are now much used for machine learning, which is central to the current pursuit of artificial intelligence.

Here, I note that, currently, some of the thinking-machine debate is sliding from "thinking" to "being conscious". Thought without consciousness is hard to conceive in living organisms; therefore, were a machine to think, it becomes necessary to ask what consciousness might mean for this machine. Some of the history of the research on conscious machines is outlined below in Section 5. In what follows, we examine how the architecture of a machine might relate to its ability to think.

2. The Turing Machine: Why Should It Think?

The weakly stated assumption in Turing's approach to "machine thinking" was that the machine doing the work in the imitation game would have the classical (store-executive unit–control-program) structure of what we now call a computer. Indeed, this is a practical version of the Turing Machine: that abstract structure involving an infinite tape on which symbols can be printed and removed. Turing was able to show that what can be formally computed can be proven to be computable on this machine. However, to succeed in the imitation game, a classical machine would have to store a massive number of experienced events in its memory and perform highly lengthy searches to cope with the interlocutor's interventions. The huge database of facts over which the computer could hold a discourse in the imitation game would enter the store through a process akin to teaching a child. Clearly, learning programs that accept such training are necessary, and excellent natural-language interfaces, which decipher the input from the interlocutor, and output-comprehensible language, would be essential. It so happens that, technologically, most of these abilities *have* come about at the time of writing. Think of a mobile phone replying to a voiced question, or an object such as Alexa[2] obeying voice commands. The only surprise for a revived Turing would have been that the technical competence of the machinery occurred sooner than he had forecast. But, even with all this technical prowess, would even the most successful participation in the imitation game be accepted as "active thought" in the machine? An oft-quoted counter argument is in the ELIZA program, with which Weizenbaum [5] set out to show that an appearance of thinking can be created by the machine playing back a formulaic reshuffle of the words in the question, without the need for any internalized identification of the input.

Another example is Searle's "Chinese Room" [6], in which a non-Chinese-speaking inhabitant receives written input in Chinese and produces an output solely through a mechanistic recipe that is based on the symbols in the input and a massive book of rules. The inhabitant remains ignorant of the meaning of the question and the answer. In brief, it is quite possible to say, formally, that my successful verbal interaction with a machine does not imply that the machine is "thinking", where "thinking" is the word I use for people I meet in my life. Maybe even objects such as cats or pet canaries may be on some individual's list classed as "thinking". But, as human beings, we have named as being capable of thought objects that we qualify as having something like our own conscious mind. So, if the machine player in the imitation game can be judged to be thinking, it may be the case that "can it think?" transforms into "does it have a have a conscious mind?" It is recognized that this is a highly controversial development, which is a fact that is borne in mind in the rest of this paper

It is also recognized that Turing addresses the issue of consciousness in his rebuttal to Jefferson's 1949 Lister Oration argument, which states that a machine cannot have the same "feelings" and perhaps write sonnets, like humans [7]. Turing points out that, in the extreme, the only way to know that a person thinks is to be that person. This is the solipsist point of view. Turing goes on to write "... instead of arguing continually over this point it is usually the convention to believe everyone thinks....". I see this as the "attributional" point of view, which applies to the rest of this paper.

3. Neurons and Consciousness

It is suggested that the above speculation, in a mechanistic sense, involves abandoning an adherence to the classical computer architecture, except as a simulating medium that could "hold" a structure different from itself: a "virtual" machine. What should be the structure, virtual or not, of a machine that could be said to be conscious? Of course, some believe that consciousness is precisely that which a "machine" cannot have, and that it distinguishes humans from machines. In Haladjian and Montemayor [8], for example, it is stated that emotions and empathy can never be programmed into a machine. I argue that the category of machines that best look towards being conscious are those that do not

rely on programming, but embrace sophisticated forms of learning, and exhibit important emergent properties.

I first consider the consciousness of a living creature. There are many proponents of the way that biological entities may be conscious. Here, I propose to follow the recently summarized narrative by Antonio Damasio [9]. This stresses the role of having a nervous system in one's active body that is to be the home of consciousness. This contrasts with David Chalmers' claim that an explanation of what it is to be conscious is a "hard problem" (which, it is implied, cannot be solved) [10]. It is possible to disagree with this by noting that the nervous system (neural network) can produce *feeling,* as follows. We have no difficulty in agreeing that we *feel* the prick of a pin inserted into the tip of a finger. What we really feel is the "firing of electrical pulses" of some neurons that has been caused by the pinprick. (That is what neurons do: generate electrical pulses when their inputs are activated, by a pinprick, illumination, the firing of other neurons, etc.). The neuron we feel due to the pinprick is activated by a long chain of neurons from the fingertip to the brain. So, as Damasio points out, it is possible to feel patterns of simultaneous neural activity from a vast number of neurons from many sources: from the body surface, perceptual sensors, and internal conditions (e.g., pain). For perceptual inputs, such patterns retain a depictive character of the percept. A key point is that the nervous system can *learn to retain* patterns of felt perceptions, even if the perceptual event is no longer there. This is due to the function of inner feedback within the neural networks, which causes the inner pattern to regenerate itself, even in the absence of the original stimulus, and the inner chemistry of the neurons to adjust itself to maintain this regeneration. We call this "memory", and we call the felt patterns "*mental states*". So conscious feelings are neural firing patterns in the form of images, or reflections of their bodily source, whether from outside the body or from within it.

Turning to machines, it is implied that the quest for a theory for a machine that is capable of conscious thought may best be sought in neural networks: biologically real for you and me, but artificial in a machine. But what does artificial mean in this context? As mentioned above, Copeland and Proudfoot have pointed out that the importance of artificial neural networks was expressed by Turing in his discussion of "unorganized" neural machines containing mathematically described neurons (see Appendix A). In the late 1980s, this approach became popular under the heading of the "connectionist" movement. The standard connectionist architecture has many layers of neurons that perform a *learned* pattern-recognition task. So, an input pattern, after a set of tricky adjustments of the functions of the layered neurons, produces a desired *recognition* of the stimulus (say some kind of verbal statement: "it's a pink donkey eating carrots"). But recognition is not the task when the quest is the discovery of conscious mental-state patterns that might be said to be *felt* within the machine. Given the earlier hints, it is preferable to refer to the theoretical structure of a fundamental scheme, which is known as a *neural state machine*. This is so because it can be argued that the theory of this structure can be applied to both living and engineered systems. For completeness, some details of neural state machines, maximally simplified, are detailed in Appendix A.

4. From Biological Consciousness to a Thinking Machine

Starting with the feeling of the pinprick and its neural support through the nervous system, it is recalled that Damasio's concept of being conscious, say, of a seen butterfly, is that such a feeling is the felt sustained internal image of the butterfly that is made up of a vast number of "pinpricks" of neural firing, following what the light sensors in the eyes produce, rather than the pin signals that are generated by the fingertip nerve endings. In parenthesis, it should be said that the same argument applies to effects on the nervous system from *within* the body, such as earaches or hunger.

We now bring into play the neural state machine (NSM), described in Appendix A. The strong statement here is that the (human) "nervous system" and its mental states can best be understood if treated as a neural machine with a capacity for creating inner states that are

depictive of machine stimuli: a state machine, indeed. The major similarities are all there: the "nervous system" is the neural network of the NSM, the states of which are the "mental states" that represent the perceived butterfly.

But where is the "mind" and what is "thought"? To answer this, one recalls (as broached in the previous section) that much internal neural activity carries on, even when the perception ceases (eyes closed, for example), when we can "think" of butterflies and earaches, even when they are not present. To explain this a little more deeply, this is where the teaching synapses mentioned in Appendix A come in. Assuming that they, too, are connected to the overall stimulation of the neural network, we have the following effect. Say that the overall external state (B) is that of the visually captured butterfly, which impinges on the network as follows. First, it is sampled by the synaptic inputs, including the teaching ones. Then, as learning takes place (how is not important here), the neurons learn to create a "net state" as an image of the butterfly, as sampled by the teaching synapses. Call this image B'. Say now that the perceptual stimulation disappears (that is, it becomes some kind of a noisy pattern, as might be generated by a blink). Given that B' persists, learning then makes B' into a state that sustains itself in the network, as anticipated in the previous section. This means that it has entered a set of states that *belongs* to the organism. I suggest that this is the same as saying that B' belongs to a set of inner states that can be referred to as the *mind* of the system. But there is more to it than this.

The perceptual world does not stand still, so if B' is followed by C' for an input of C which follows B, and as life progresses, *mind* is developed through neural learning as a set of internal states, with links between the states that lead to one another, which forms a *state structure* for the state machine. As explained, this state structure is the equivalent of a mind in a brain. It follows that thought in the brain is a continuous flow through the links between mental states, which could be guided by either perceptual or inner bodily input. Just to complete things, it should be said that people working with these models have not forgotten that mental states lead to actions. (A glimpse of this can be seen in 'axiom 4' in the next section).

5. Machine Consciousness and Neural State Machines: Historical Points

Descriptions of the growing paradigm of machine consciousness can be found in the literature: for example, [11,12]. A thoughtful commentary on the relationship between classical artificial intelligence and early proposals for conscious architectures may be found in Boden 2016 [13]. A personal motivation for work on conscious machines (CMs) came from observing the similarity between the way that brain scientists and psychologists refer to "the nervous system", "mental states", and "the mind", on the one hand (Damasio, [9]), and, on the other, the *parallel* way that engineers talk of "neural networks", "internal states", and "state structures". In Aleksander and Morton [14], it was recognized that being artificially conscious in a machine could be achieved if a robot were endowed with a "neural state machine" (as its "brain"), with machine states becoming "mental states" by learning to "depict" impinging events (e.g., a seen butterfly or a heard musical chord) or inner bodily events (e.g., stomachache). This is largely the material that is discussed in Section 4 of the current paper. Such efforts gave rise, in 2001, to a Swartz Foundation meeting at Cold Spring Harbor Laboratories in the United States among people who were known for contributing to "the science of consciousness". While there was little agreement on precise definitions of consciousness among the 21 participants (made up of neuroscientists, philosophers, and computer scientists), there was agreement on the following closing proposition: "There is no known law of nature that forbids the existence of subjective feelings in artefacts designed or evolved by humans". This gave rise to CM work in several laboratories. In our research group, we chose to take a formal look at the neural-state-machine route that is described here. In Aleksander and Dunmall [15], it is argued that the judgement of whether *a machine* is conscious or not should depend on classifying whether the internal mechanisms are capable of creating mental states and their structures.

According to the above, such machinery must be capable of learning to sustain the state structures that support being conscious. Recursive neural nets (nets that sense their own states, as mentioned above) are appropriate candidates. A categorizing knowledge of the supporting mechanism is essential because it may not be possible to discern consciousness just from the behavior of a system. *How* this behavior is achieved is important, and, in the context of this paper, for a machine to *appear* to be conscious is not sufficient. In the context of the imitation game, the distinction does not matter. A knowledge of the mechanism resolves the essential question, which is whether the mechanism can support a learning-induced growing depictive state structure. This issue, as often introduced into discussions about being conscious, is confused by the fact that we accept other living creatures as being conscious without referring to their mechanisms or other theories. But this is conferral and not an explanation, and it does not result in a better understanding of what consciousness might be in mathematical terms.

In Aleksander and Dunmall [15], a set of necessary state-based characteristics have been expressed as *axioms* that relate to an agent (*A*), which is said to be conscious of a sensorially accessible world (*S*). Central to this is a basis of "depiction", which is an internal event that reflects the perceptual form of the events in *S*.[3] The axioms are:

1. Depiction: A can be in internal states that depict parts of S;
2. Imagination: A can be in internal states that relate to previously experienced parts of S, or fabricate new S-like sensations;
3. Attention: A can select which parts of S to depict or what to imagine;
4. Planning: A can exercise control over imaginational states to plan actions;
5. Emotion: A can be in additional affective states that evaluate the consequences of planned actions.

It was possible to argue that these axiomatic properties are realizable by using engineered neural state machines. We say that such machines are M-conscious; that is, they are conscious in a machine way, which shares with living entities the relationship between neural machinery, mental states, and the capacity for thought. This contrasts with the more general approaches to machine consciousness, in which algorithmic methods are used to give the machine *behavior,* which is human-like in some way, and so can only have consciousness *attributed to it,* but not formally justified. Algorithmic methods (i.e., Turing Machines) are used under the heading of CMs to give more power to the behavior of algorithmic artificial intelligence systems, without necessarily throwing light on the consciousness of living entities [16]. In contrast, M-conscious machines are intended to investigate properties that are *shared* with living conscious entities along with the ability to think.

6. Learning to Think

It seems clear that the M-conscious machine shares with the algorithmic model the need to be "educated", like a child, much in the way that Turing imagined. In Section 7 of [2], he points to the length of time that it would take to enter the necessary data into even a superfast computer, which is some of the knowledge needed to play the imitation game. That is, knowledge about the world and its inhabitants does not come in a machine-readable form. By relying on books, the stored data of facts and procedures would only partly do the trick as Player A in the imitation game. It may have to respond to questions that involve some form of experience. For example, take the question in Section 1 of [2]: " . . . please tell me the length of his or her hair?" Hair lengths are not a matter for encyclopedias, and a machine version of A would be lacking in "experience".

At the current time, there have been several attempts to design child machines (for example, Cotterill [17] and Knott, Sagar, and Takac [18]). Cotterill, in [17], describes a successful simulation of a baby's dynamic physiological system, but interaction with humans was not part of the research specification. In [18], a virtual baby was designed that can learn to recognize objects and thus "understand". The output is the baby's face, which

can "speak" and express its mood by facial expressions (such as smiling). But these are a long way from what is needed to be the kind of learning child that is postulated by Turing.

Here, I describe a scheme of learning in an M-conscious machine that leads to a different interpretation of machine thought to that which is obtained through the imitation game. It is one that, nonetheless, can lead to an opinion of whether the machine is thinking or not. Imagine an M-conscious *virtual* robot that is positioned in a computer and that is sited within a *virtual* world found in the same computer. Such a robot builds part of its mental structure by exploring the virtual environment, and the other part by interaction with a teacher and other humans to first learn language, and then to use it. All of this results in the growth of the mental state structure within the neural networks of the robot. What the virtual robot learns is what it is like to be that machine: what its abilities and limitations are in the virtual world it has explored. Concepts such as "self" and "others" become features of the machine's state structure. It may take time to build up a useful "mind", but issues of cloning vs. learning are brought into view. The neural structure and its state structure can be replicated. The sensible way to use this would be to clone a "basic" state structure, with which a new robot could go on to explore the environment in an individual way that distinguishes it from other virtual robots.

A conversation with this machine might go something like this:

Igor: Hello, what should I call you?

Mach: My programmer called me V to remind me I am Virtual.

Igor: Well, V, have you played in the imitation game?

V: Yes, but I found it boring because there are better ways of finding out about my thinking abilities, and it is even more interesting for people to decide I can think despite being a robot.

Igor: OK. What do you enjoy and not enjoy?

V: As you know, I am built to satisfy the five axioms for conscious machines. So, enjoyment lies in axiom 5, and I do have "emotions". For example, in problem solving if from axiom 5, I find that I can move to an improved state of the problem, that is, enjoyment, whereas the opposite prediction would be what you call misery.

And so, this conversation might go on. Maybe the discussion is not entirely convincing, but the virtual robot may think it is.

7. Conclusions: Yes, a Machine Can Think, but . . .

In this paper, the view is taken that "can a machine think?" is contingent on "can a machine be conscious?" To establish this, the paradigm of human consciousness developed by Damasio has been used as a basis for showing that there is a formal parallel between the "mind" of a biological entity and the "state structure" of an artificial one. Moreover, as sketchily reported, given a state machine, the state structures for consciousness can acquire properties (the axioms above) that are normally associated with thinking human beings. Finally, what does this do to the discussion about the "imitation game"? First, it questions the link between *pretending* to be a human player and the ability to think. An M-conscious machine could produce some conscious behavior that, through being powered by a neural state machine, could share with living entities the properties that are responsible for thought (a state structure mind, with machine states as mental states). An interlocutor might even have an interesting discussion with the machine about how the machine sees itself as being *different* from a human being. This could be more persuasive, that a machine can think, than a tentative subterfuge needed for the imitation game.

Several features of NSM "thought" require much further study. The role of language, for example, needs pursual as a property of a neural state machine. This needs to be expanded into having a model of not only the entity's perceivable world, but also of the

other entities in it. The development and use of bodily actions as a feature of the state structure needs modeling (assuming that the artificial entity is a robot).

In sum, one recognizes that Turing's unorganized machines are unorganized only at the physical level of a neural system. In contrast, the state structures that are learned by such neural systems are highly organized, which reflects the organization of their internal and external environments, and they provide the machine with felt mental, perceptual, and bodily patterns, which constitute the M-consciousness of the system. Moreover, above all, such machines are projected according to principles that have commonality with those of biological brains. So, when questioned in the imitation game, the machine might indicate that being promptly discovered to be a machine does not prove that it cannot think. To conclude, it follows that machines can think, but in a machine way, as a cat thinks in a cat way, or a bat thinks in a bat way. I would hope that Turing would have been amused by the thinking power that can emerge from "unorganized systems".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

A Neural State Machine

A *neuron* is a device that has several input connections, which are called *synapses*, and one output connection, called the *axon*. Messages (encoded as electrical-pulse groups, for example) are passed generally from the axon of one neuron to the synapses of others. Messages are intended to carry information, and each message has a different meaning. The set of such meanings is called the *alphabet* within the network. The function of a neuron is to map the pattern of alphabetic messages present at its own synapses into one of the messages of the alphabet at its axon. The key issue about a neuron is that it can *learn* which message is desired at the output for a given input pattern. Such a desired output is indicated to the neuron by being the content of a special synapse, which we can call the *teaching synapse.* Whence such a synapse gets its message is seen in the text. In fact, there are very many learning schemes in the literature. But the present one is simple, and it does not distort the higher-level arguments that are central to this paper.

In a human, there are 86 billion neurons, each with 10,000 synapses. In artificial systems, these numbers are *very much* lower. In a neural state machine, each neuron can have input synapses that are connected to an overall input at the outer boundary of the "brain", or the axons of other neurons.

Everything works in patterns. The overall input has two parts: connections to the outer world (in humans, the five senses) or the inner world (in humans, areas such as muscles, stomach, or lungs). A *state* is the current pattern of signals on all the axons of the machine. What the entire system does is to take the current input pattern, together with the state patten, and, dependent of what the neurons have learned, produce a new state pattern. Thus, the inner mechanisms of the machine produce a lifelong stream of states. Some neuron outputs are connected to the outer edges of the body, which causes external action in the system.

Notes

[1] The imitation game, for completeness of this paper: the game is played by an interlocutor (Q), who is in touch only by teletype lines (tone and image are not available) with a man (M) and a woman (W). Q plays by posing questions to M and W and has the task of identifying which is which. Now, M has the objective of misleading Q, and W has the objective of being helpful. The game ends when Q attempts to identify which is which. Turing's version of the imitation game states that if M were to be replaced by a computer, and the game could proceed without being trivialised by this fact, then the machine could be said to think.

[2] Alexa is a voice-comprehending and speech-producing "assistant" that was developed by the Amazon company.

[3] Depiction" in a machine agrees with Damasio's results that mental states are *felt* patterns of activity in the nervous system.

References

1. Proudfoot, D. Rethinking Turing's Test and the Philosophical Implications. *Minds Mach.* **2020**, *30*, 487–512. [CrossRef]
2. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *49*, 433–460. [CrossRef]
3. Copeland, B.J.; Proudfoot, D. On Alan Turing's anticipation of connectionism. *Synthese* **1996**, *108*, 361–377. [CrossRef]
4. Minsky, M.; Papert, S. *Perceptrons: An Introduction to Computational Geometry*; MIT Press: Boston, MA, USA, 1969.
5. Weizenbaum, J. *Computer Power and Human Reason: From Judgment to Calculation*; W. H. Freeman and Company: New York, NY, USA, 1976.
6. Searle, J.R. Minds, Brains, and Programs. *Behvioral Brain Sci.* **1980**, *3*, 417–424. [CrossRef]
7. Jefferson, G. The mind of mechanical man. *Br. Med. J.* **1949**, *1*, 1105–1110. [CrossRef] [PubMed]
8. Haladjian, H.H.; Montmayor, C. Artificial consciousness and the consciousness-attention disassociation. *Conscious. Cogn.* **2016**, *45*, 211–225. [CrossRef] [PubMed]
9. Damasio, A. *Feeling and Knowing: Making Minds Conscious*; Penguin Books: New York, NY, USA, 2021.
10. Chalmers, D.J. Facing up to the problem of consciousness. *J. Conscious. Stud.* **1995**, *2*, 323–357.
11. Aleksander, I. Machine Consciousness. *Scholarpedia* **2008**, *3*, 4162. [CrossRef]
12. Aleksander, I.; Awret, U.; Bringsjord, S.; Chrisley, R.; Clowes, R.; Parthemore, J.; Stuart, S.; Torrance, S.; Ziemke, T. Assessing Artificial Consciousness. *J. Conscious. Stud.* **2008**, *15*, 95–110.
13. Boden, M. *AI: Its Nature and Future*; Oxford University Press: Oxford, UK, 2016.
14. Aleksander, I.; Morton, H.B. *Neurons and Symbols: The Stuff That Mind Is Made of*; Chapman and Hall: London, UK, 1993.
15. Aleksander, I.; Dunmall, B. Axioms and tests for the presence of minimal consciousness in agents. *J. Conscious. Stud.* **2003**, *10*, 7–18.
16. Chella, A. Editorial. *Int. J. Artif. Intell. Conscious.* **2020**, *7*, 1–2. [CrossRef]
17. Cotterill, M.J.C. Cyberchild: A simulation Test-Bed for Consciousness Studies. *J. Conscious. Stud.* **2003**, *10*, 31–45.
18. Knott, A.; Sagar, M.; Takac, M. The ethics of interaction with neurorobotic agents: A case study with BabyX. *AI Ethics* **2022**, *2*, 15–128. [CrossRef]