Steels, "Is Artif. Csness Possible?"

CONSCIOUSNESS:

Distinction and Reflection

Edited by
Giuseppe Trautteur

BIBLIOPOLIS

ISBN 88-7088-341-8

Printed in Italy by Arte Tipografica s.a.s., Napoli

Table of Contents

Is artificial consciousness possible?

Luc Steels

1. Introduction

The construction of artificial systems which exhibit behavior usually classified as intelligent, has resulted in the last decades in a large number of fundamental insights into the nature of cognition, communication, perception, and action. It has also resulted in technologies based on these insights such as expert systems, vision systems, natural language processing applications, and so on (Shapiro 1992). These technologies cannot compete with human intelligence but have shown that there is some value to the insights obtained so far. There is however one phenomenon that Artificial Intelligence researchers have stayed away from, which is (self)-consciousness.

Some critics of artificial intelligence, notably Penrose (Penrose 1989), have argued that artificial consciousness is not possible, and – because consciousness is essential for *true* intelligence – this impossibility presents a fundamental limitation on artificial intelligence itself. This paper argues the opposite, namely that artificial consciousness is possible – even necessary, when we want to reach complex autonomous agents. To see why, we must take up the same question as posed by Penrose (Penrose 1989, p. 405): "What selective advantage does a consciousness confer on those who actually possess it?" The flow of the argument will be as follows. First the phenomenon itself is better circumscribed (section 2) Then I discuss the reasons why it has been argued, by Penrose amongst others, that artificial systems can never be conscious (section 3). Next I present four theses (section 4) and develop briefly a scenario (section 5). Some conclusions and implications end the paper.

2. The phenomenon of consciousness

The subjective experience of consciousness cannot be denied, although there cannot be said to be a consensus about what

constitutes consciousness. Let us start from characterizations given by Penrose, particularly because we will examine later his arguments why artificial consciousness is not possible. Penrose puts consciousness in the context of common sense, judgement of truth, understanding and artistic appraisal: "Consciousness is needed in order to handle situations where we have to form new judgements, and where the rules have not been laid down beforehand" (Penrose 1989, p. 411). A typical example is mathematical discovery, more specifically the sudden experience of having hit a proof or the definition of a concept. Such an experience is for example described by Poincaré: "...At the moment when I put my foot on the step, the idea came to me, without anything in my former thoughts seeming to have paved the way for it, that the transformation I had used to define the Fuchsian functions were identical with those of non-Euclidean geometry. I did not veryify the idea; I should not have had time, as, upon taking my seat in the omnibus, I went on with a conversation already commenced, but I felt a perfect certainty. On my return to Caën, for convenience sake, I verified the result at my leisure" (Poincaré quoted by Penrose 1989, p. 419).

It seems then that prior to the experience of consciousness there are a number of alternatives. The experience itself is related with the coming to the foreground of one alternative and the associated feeling of certainty that this is the right alternative. The alternatives could be concerned with an aesthetic judgement, a belief that something is the case, an opinion about somebody, an understanding of how the pieces of a puzzle fit together. A theory of consciousness must therefore explain how one alternative happens to come to the foreground. It must also explain why there is a real experience associated with it.

3. Arguments against artificial consciousness

Penrose develops two arguments why artificial consciousness (and hence artificial intelligence) is impossible. The two arguments are somewhat contradictory, because one assumes a Platonistic philosophy, and the other one seeks a physicalist explanation.

Penrose admits that he is a Platonist, in other words he subscribes to the view that there is a world populated by

mathematical concepts and necessary truths and that this world has existence similar to the existence of physical objects: "To speak of Plato's world at all, one is assigning some kind of reality to it which is in some way comparable to the reality of the physical world" (Penrose 1989, p. 430). Platonism implies dualism. The mind is part of the mental Platonic world. The brain is part of the physical world. This generates a mind-body problem: "How is it that a material object (a brain) can actually evoke consciousness?" and "How is it that a consciousness, by the action of its will, actually influences the (apparently physically determined) motion of material objects?" (Penrose 1989, p. 405) Penrose sees the role of consciousness as a way to resolve the mind-body problem: "I imagine that whenever the mind perceives a mathematical idea, it makes contact with Plato's world of mathematical concepts. ...When one *sees* a mathematical truth, one's consciousness breaks through into this world of ideas, and makes direct contact with it.... When mathematicians communicate, this is made possible by each one having a direct route to truth, the consciousness of each being in a position to perceive mathematical truths directly, through this process of *seeing*" (Penrose 1989, p. 428). It is unclear however what this *making contact* is like, or how it is that only mathematicians have this *direct route to truth*.

This Platonist and dualist position of Penrose naturally leads to a rejection of the idea that computers can be used as the basis for artificial intelligence. A computer (just like a brain) is a physical device. If there is an unbridgeable gap between the physical and the mental, then a computer can of course never exhibit intelligence. But one can also assume a physicalist viewpoint, as indeed the majority of AI researchers is doing. Mental phenomena are then seen as the consequence of physical phenomena and no independent existence of a mental world is required. This resolves at once the mind-body problem. (Physical) signals flow in response to external phenomena to the more central processing structures from where signals flow to the actuators. We can interpret these flows and the processing of them in terms of information processing, but essentially all there is physically speaking are electromagnetic states and continuous changes to them. This is similar to the way that we can interpret chemical activities

in a cell in information processing terms (e.g. as the decoding of the DNA) but that is a (useful) interpretation, no more, no less.

In a physicalist view, mathematical or other concepts have another status. They are constructed by minds and transmitted culturally. Proofs or algorithms are also constructed by minds, and do not need to exist prior to their discovery or design. An examination of the history of mathematics as conducted by Lakatos (Lakatos 1979) illustrates clearly that proofs and mathematical concepts have their own evolutionary history, which contradicts the sudden almost mystical insight attributed by Penrose to mathematicians.

Computational processes are physical (electromagnetic) processes taking place in the electronic circuits of the computer. The only thing which makes these processes computational is that they happen to be executed in a physical device with a certain oraganization which we call a computer. Analogous processes could take place in quite different physical systems with the same organisation (for an example in optical media, see (Feitelson 1988)) and we would still call them computational. The optimism of AI that computers can be programmed to exhibit intelligent behavior (assuming that they are part of a more global system which also includes a body, sensors and actuators) is based on the hypothesis that there is nothing peculiar about the physical phenomena causing mental phenomena except their organisation, and that computers are sufficiently complex to be able to generate the required physical phenomena. This is similar to the viewpoint shared by all biologists that it is not the physical components and chemistry which distinguishes life from non-life but the way these components are organised.

The counterarguments of Penrose against the possibility of artificial intelligence all assume a Platonic viewpoint, and are therefore invalid, if one assumes a physicalist viewpoint. For example, Penrose states: "algorithms inhabit Plato's world, and hence that world, according to the strong AI-view, is where minds are to be found" (Penrose 1989, p. 429). But AI researchers do not put algorithms in Plato's world (in fact it would be hard to find a programmer who believes that algorithms exist prior to his or her design) and hence Penrose's

conclusion would not be accepted. Here is another example: "They (i.e. AI researchers) might try to take the line that algorithms can exist as marks on a piece of paper ... But such arrangements of material do not in themselves actually constitute an algorithm. In order to become algorithms, they need to have an interpretation, i.e. it must be possible to decode the arrangements; and that will depend upon the *language* in which the algorithms are written. A pre-existing mind seems to be required in order to *understand* the language" (Penrose 1989, p. 429). Again this sounds odd if one understands, as one should, that algorithms are descriptions of processes. These descriptions are physically represented in the machine and can therefore be used by another physical process (the machine interpreter) to cause other physical processes such as changes in registers or a jump of control flow. The DNA is in this sense also an algorithm because it *defines* the processes in the cell that give rise to proteins. Biochemistry has shown that no pre-existing mind is needed to decode the genetic messages in the DNA, although the genetic messages clearly assume a particular language.

Although Penrose is a Platonist, he somewhat surprisingly seeks nevertheless a physical analogue of consciousness, and finds in this a second reason why artificial intelligence is not possible (today). Penrose seeks the physical analogue in quantum physics. He compares the jump from the quantum level to the classical level, which occurs when the linear superposition of alternatives is resolved by an observer, as similar to the resolution of the different alternatives which exist prior to the experience of consciousness: "I am speculating that the action of conscious thinking is very much tied up with the resolving out of alternatives that were previously in linear superposition. This is all concerned with the unknown physics...which, I am claiming, depends upon a yet-to-be discovered theory of quantum gravity" (Penrose 1989, p. 438). For Penrose, it therefore follows that only a quantum computer can ever be conscious and hence intelligent. In this sense, Penrose is part of a tradition of physicists who give a role to consciousness in physical theory: see (Kafatos and Nadeau 1990). But, as Edelman (Edelman 1992, pp. 212-218) has clearly demonstrated, the quantum level is the wrong level for studying brain processes, and the

analogue does not explain anything about what the processes of consciousness might be (simply because quantum gravity theory is itself not yet worked out).

The Platonic and dualist position of Penrose as well as the quantum theory hypothesis do not hold up under scrutiny. The physicalist hypothesis is a much simpler explanation of how an agent may interact with the world in an intelligent way but then mathematicians must be willing to give up their special status of having a direct route to truth. The quantum theory hypothesis is stated too vaguely to be a testable theory and it seems to be at the wrong level.

4. Four theses

Although the major thrust of Penrose seems ill-founded, the insight that consciousness is related to the resolution of alternatives is valuable and will be retained. But rather than assuming a mystical contact to the Platonic world or a quantum physical process to explain how this resolution takes place, we can take the current AI theories (without regard for detail) as a basis and add a mechanism which has been well studied in biology and chemistry. This is the mechanism of self-organisation (Babloyantz 1986).

Concretely, let us start from the following four theses:

1. Mental activity consists in the construction of models and the use of these models in decision-making and action.

For example, diagnosis of a motorcycle implies the construction of a model of the symptoms, the internal structure and functioning of the motorcycle, the causal relationships between symptoms and malfunctions, the possible remedies to a particular malfunction, and so on. This insight underlies almost all the current work in AI. Although there are some arguments against the over-use of complex models in physically embodied agents see e.g. (Brooks 1991b), nobody denies that agents need models of the world.

2. Models can (and need to) be represented in physical structures, and mental activity then corresponds to the physical activity of manipulating these physical structures.

A model can be represented in physical structures. This means that for every concept used in the model, we need to postulate a (physical) symbol and for every model operation,

we need to introduce a physical analog operating over the physical representations of concepts. This is now a standard technique in AI and used on large scale applications involving literally hundreds of thousands of symbols and symbol processing operations. It has been expressed by Newell and Simon as the "Physical Symbol Systems Hypothesis" (Newell 1990).

3. Consciousness is the (physical) mechanism that helps to manage the parallel processes that build up and use representations of models. In particular, consciousness is necessary to help reach coherence and force decisions when there are time constraints.

When looking at concrete problem domains, it quickly becomes clear that there is a large number of possible operations that can be performed. This generates a combinatorial explosion of possible directions in which models can be developed, particularly because there are usually multiple viewpoints on a problem. An intelligent system therefore has to choose which avenues it will explore. This introduces the so called control problem, which is one of the key problems in AI.

The control problem has been approached in many ways. In the earliest systems (such as the General Problem Solver), and in logical reasoning systems, there is a unique and general-purpose control structure. This strongly limits the kind of problems that can be tackled. Later on, more and more control knowledge has been represented explicitly, so that a system can reason at the control level. This is for example possible in the SOAR architecture (Newell 1990). The problem with this kind of explicit control of reasoning is that more and more complexity is introduced and that successful control relies strongly on the designer.

4. Consciousness is a self-organising process, in the sense of non-linear complex dynamical systems theory.

An alternative is that control is achieved using a dynamical process which operates on quantities such as the strength of connections between steps in the deduction, the strength of the evidence available for a theory, etc. No homunculus needs to be postulated to regulate this process. Instead consciousness can be self-organising in the sense of (Nicolis and Prigogine 1977). Self-organising processes have been observed in

chemical reactions (as in the Belouzov-Zhabotinski reaction) in physical media (for example lasers) and in higher order systems (such as cells or societies of ants). Each time we find the same principle. Self-organisation only occurs in open systems, i.e. systems that have a continuous interaction with an external environment. There are a set of processes that break down a structure and another set of processes that build up a structure. Each of these processes is local, but the structure itself is global. There is also an autocatalytic step so that the process of building up the structure is enforced (see the examples in (Babloyantz 1986)). Typically there are different possible alternatives present in a system (e.g. spiral waves may go in one direction or the other). But due to the autocatalytic step, one alternative will eventually dominate in competition with the others. Self-organisation is therefore an alternative to the quantum process for resolving one alternative out of many. Because it operates at a macroscopic level (instead of the quantum level) it is a much more plausible candidate.

5. Consciousness as a self-organising process

I will now try to make all this a little bit more concrete, although much more detail is obviously needed:

(1.) We must first of all assume a large number of parallel structure formation processes at work. These processes construct and change representations, possibly making use of other representations. For example, there can be a set of processes capable to make an analogy with existing structures. Or there could be a set of processes performing something similar to logical reasoning. Other processes could establish more direct relationships between sensing and acting. They are directly linked to flows of sensory inputs, and directly linked to action parameters for actions. All these processes require energy. Maintaining a structure, more precisely the links between nodes in the structure, requires energy as well. The more energy can be drawn to a particular activity, the stronger its links will be and the faster and more reliable a process can take place. So we need to extend AI systems such that energy becomes an explicit element.

(2.) Attention is like an additional energy supplier. If at-

tention is focused on a particular activity, i.e. if we become *aware* of that activity, more energy is channeled to it and the structures and processes involved become potentially stronger. Attention is coupled to the real world through sensory modules (unusual phenomena detected by sensors *draw* the attention towards the structures related to these phenomena). Attention is also coupled to verbalisation/visualisation. We cannot verbalise something unless we are paying attention to it. When verbalisation/visualisation happens, we become aware. Attention needs to shift from one area to another to remain responsive to the pressures of the real world.

(3.) What we call consciousness, is a completely distributed self-organising phenomenon in which patches of structure (and the activities associated with them) solidify into coherent and stronger structures after having been loosely coupled in competition with other potential structures. The process is auto-catalytic in the sense that structures/activities that are reaching strength feed on themselves, so that they become stronger. Moreover strong structures/processes draw attention, i.e. they cause more energy to be drawn into them. This enforces the autocatalytic effect, because as the structures/processes draw attention they become even stronger because they receive more energy.

The subjective experience of the resolution of different alternatives, as reported by Poincaré, corresponds to the sudden *clicking together* of a structure which therefore grabs attention. Because this coming together is a physical process that involves energy, it can be experienced (and should be physically observable).

(4.) There is also a potential explanation for self-consciousness. Meta-level architectures and computational as well as knowledge-level reflection have already been shown feasible in artificial systems (Weyhrauch 1980; Maes and Nardi 1988; Weyhrauch's chapter, this book). Self-consciousness can then be seen to be similar to consciousness, except that the activities and structures concern the models that the agent makes of itself, i.e. it is a self-organising phenomenon operating at the agent's meta-level.

6. Conclusions

At this point, the theory proposed here is pure speculation (as all other theories of consciousness, for that matter). However there are some testable implications. First of all, we can expect to discover the physical basis of consciousness. It must be situated at another level than that of current neural network theories which focus on the neuron or neuronal group level. At the right macroscopic level, the self-organisation process should be measurable in terms of continuously evolving attractor states corresponding to the reaching of coherence of particular substructures and activities. Perhaps the neuronal oscillations that have been observed even at large distances are a first piece of the evidence. Second, there are some potential explanations of psychological phenomena. For example, the Buddhist tradition involves meditation practices to focus on a better control of awareness/attention and hence give the agent more control over consciousness (Varela *et al.* 1992). The possibility to better control consciousness is compatible with the theory proposed here, because it can be done by exercising the awareness process, and thus supplying energy to a certain group of activities. Third, we can try to build artificial agents that use a similar mechanism to help manage the many parallel distributed processes necessary to maintain the agent. Self-organisation phenomena have already been shown to be possible in artificial systems so that this is not an obstacle (Langton 1989). A more important obstacle is that a sufficiently complex agent must be built which has real-time interactions with the world through sensors and actuators (Steels 1993). A simple agent does not need consciousness.