Four Preconditions for Solving MC4 Machine Consciousness

David Gamez[1]

[1] Department of Computer Science, Middlesex University, London, NW4 4BT, UK
d.gamez@mdx.ac.uk / www.davidgamez.eu

Abstract. A machine is MC4 conscious if it has phenomenal experiences that are comparable to human conscious experiences. From an ethical point of view it is important to know whether we have created MC4 consciousness in a machine. MC4 consciousness research can also contribute to the development of general theories of human consciousness. This paper discusses four problems that have to be solved before we will be able to address MC4 machine consciousness in a systematic way: We need more clarity about the measurement of consciousness, we need better ways of describing the physical world and consciousness, and we need to reach agreement about the final form that a theory of consciousness should take. When these problems have been addressed we will be able to develop scientific theories of consciousness that can make accurate believable predictions about MC4 consciousness in machines.

Keywords. consciousness, machine consciousness, artificial consciousness, science of consciousness, neural correlates

1 Introduction

It is often helpful to distinguish four types of machine consciousness [1]:

- **MC1**. *Machines with the same external behavior as conscious systems*. Humans behave in particular ways when they are conscious. For example, they are alert, they can respond to novel situations, they can inwardly execute sequences of problem-solving steps and they can learn. AI systems can exhibit some or all of these external behaviors.
- **MC2**. *Models of the correlates of consciousness*. Theories about the neural and functional correlates of consciousness in humans can be modeled in a computer.
- **MC3**. *Models of consciousness*. Phenomenal experiences have characteristic features, which can be modeled in computers and used to control robots.
- **MC4**. *Machines that are phenomenally conscious*. When humans are conscious they are immersed in a world of colors, smells, sounds, etc. These are not properties of the physical world – they are constituents of conscious experience. A machine that was immersed in a world of colors, smells and sounds would be MC4 conscious.

These different types of machine consciousness can be combined. For example, a robot that displayed conscious external behavior (MC1) could be controlled by a model of consciousness (MC3) and have phenomenal experiences (MC4).

This paper will address the question of MC4 consciousness: artificial systems that have phenomenal experiences that are similar to our spatially and temporally distributed experiences of color, smell, taste, etc. This type of machine consciousness is ethically significant because MC4 conscious machines could suffer. Research on MC4 machine consciousness could also be a key factor in the discovery of general theories of consciousness that are not limited to the neurons inside the human biological machine.

MC4 machine consciousness can be solved when we have developed scientific theories of consciousness that can make accurate detailed predictions about consciousness in any physical system. This will only become possible after four preconditions have been met. First, we need to reach agreement about how consciousness can be measured and abandon the idea that consciousness can be measured through machines' external behavior (Section 2). Next we need to move away from neural correlates of consciousness and find new ways of describing spatiotemporal physical patterns in the brain that could form the basis for generalizable theories of consciousness (Section 3). Third, we need less anthropomorphic ways of describing consciousness (Section 4). Finally we have to drop our desire for intuitively satisfying explanations of consciousness and search for mathematical relationships between formal descriptions of consciousness and formal descriptions of the physical world (Section 5). When these preconditions have been met we will be able to use human and animal experiments to discover mathematical theories of consciousness that can be generalized to make believable predictions about MC4 consciousness in artificial systems.

2 Measurement of Consciousness

We cannot directly measure other people's consciousness: we have to measure consciousness through external behavior (first-person reports). An inference from external behavior to consciousness only works when the external behavior is generated by a system that we believe is capable of consciousness – for example, when conscious external behavior is exhibited by normal humans and primates. The inference from external behavior to consciousness is less certain with brain-damaged patients, infants and cephalopods.

Some people believe that conscious external behavior can be used to infer the presence of consciousness in artificial systems. If this was the case, the problem of MC4 machine consciousness would be solved. Any machine that exhibited conscious external behavior would be judged to be phenomenally conscious and we could easily identify the correlates of consciousness in machines. The problem with this position is that it is extremely easy to write a computer program that mimics human first-person reports about consciousness. For example: "cout<<'I am conscious of a red apple.';" This behavior can be produced by a giant lookup table [2] and it is possible to interpret any sequence of physical states as the execution of this program [3, 4], including

a sequence of states in the brain of someone who is not conscious. When humans execute this program in their heads they are not conscious of the red apple [5]. These problems do not go away when the simple one line program is replaced by a super intelligent AI that passes the Turing Test, and they are also present when part of a person's brain is replaced with a functionally equivalent silicon chip [1].

These problems prevent us from using external behavior to measure consciousness in artificial systems, regardless of how humanlike that behavior is. If we cannot measure consciousness in artificial systems, we cannot do MC4 consciousness experiments on artificial systems. Instead, we will have to discover the relationship between consciousness and the physical world in humans (and similar systems) and then *generalize* these results to animals and artificial systems. To make this generalization possible we have to rethink the way in which we study consciousness in humans. We need to describe the physical states of human brains in a way that will enable us to apply our theories of consciousness to non-biological systems. We also have to describe consciousness without the anthropocentrism of natural language. These challenges are covered in the next two sections.

3 Description of the Physical World

The measurement problem forces us to develop our theories of consciousness through experiments on humans and similar animals. Although there have been promising results on the neural correlates of consciousness [6], it is difficult to generalize them to animals that have different brain architectures and neurons, and it is impossible to apply them to artificial systems that are controlled by synthetic or silicon neurons.

It might be thought that the generalization problem could be solved by linking consciousness to functions, computations or information patterns in the brain. If consciousness is linked to computations in the human brain, then a robot that executed the same computations would also be conscious. A key problem with functional, computational and informational theories of consciousness is that functions, computations and information are not objective properties of the physical world. Information appears when you apply a human-defined interface to the physical world [7] and different interfaces lead to different information sets. Functions are equivalent to computations, which are subjective uses that humans make of physical objects, not objective properties of physical objects [1]. The physical universe does not contain functions, computations or information. The subjectivity of functions, computations and information makes it impossible to prove that they are linked to consciousness in the human brain. It is also possible to interpret any sequence of physical states as a particular information pattern or computation, which leads to contradictory predictions about the consciousness that is associated with a sequence of physical states [1, 8, 9].

An objective scientific approach to consciousness has to look for connections between spatiotemporal physical patterns and consciousness. Much of the recent research on consciousness has focused on the relationship between *neural* activity patterns and conscious states. However, this will not lead to a theory that can be generalized to artificial systems. Neurons are defined in a specific biological context – the

brains of animals – we have no formal definition that would enable us to unambiguously identify neurons outside of this context. Suppose we genetically engineer a sequence of 100 hybrids between neurons and liver cells: the first cell is 100 % neuron; the middle cell is 50% neuron; the last cell is 100% liver. We have no systematic way of classifying intermediate cells in this sequence. Or suppose we synthesize an approximate neuron from basic biological components – we have no idea whether synthetic or silicon neurons could form correlates of consciousness. A generalizable theory of consciousness has to be based on precisely defined spatiotemporal structures that can be unambiguously identified in any physical system. We will have to define neurons more precisely if we want to generalize the results from the neural correlates of consciousness, or we could base our theories of consciousness on other properties of the physical world, such as electromagnetic waves.

4 Description of Consciousness

The majority of work on consciousness has been based on contrastive analyses of conscious and unconscious brains. This gives us valuable information about the neural activity patterns that are necessary for any conscious state to occur – neural activity patterns that are common to all conscious states. There has been less work on the detailed relationships between the contents of consciousness and physical states.

Conscious states are typically described in natural language. But generalizable theories of consciousness cannot be based on natural language because it is vague, highly compressed and context dependent. It is likely that artificial systems will have radically different experiences that will be impossible to describe in human language. Natural language descriptions of consciousness are also likely to be highly misleading because artificial systems might have different representations of the temporal and spatial properties of objects [10].

In some experiments conscious contents are specified using the stimuli that produced the conscious experiences. For example, in the work on brain reading with fMRI the decoded conscious experiences are presented as videos, which the subject compares with their own conscious experiences [11]. This approach works with humans because most humans have roughly the same conscious experience when they are exposed to the same stimulus. But the conscious experience that I have when I view a video is extremely unlikely to be the same as the consciousness of an artificial system that views the same video, so the video cannot be used to describe the consciousness of the artificial system.

To address this problem we need to find new less anthropocentric ways of describing consciousness that will enable us to generalize the experimental results on human consciousness to artificial systems.[1]

[1] One solution to this problem has been put forward by Balduzzi and Tononi, who suggested how states of consciousness could be described using high dimensional mathematical structures [12].

5 Theories of Consciousness

There is a lack of general agreement about the final form that theories of consciousness should take. A generalizable theory of consciousness should meet the following criteria:

1. Can generate testable predictions.
2. Is applicable to any physical system.
3. Compact (Occam's razor).
4. Based on objective properties of the physical world.

Functional, computational and informational theories of consciousness meet criteria 1-3. However, as explained in Section 3, they fail to meet Criteria 4 because they are not based on objective properties of the physical world. Theories about the neural correlates of consciousness meet Criteria 4, but they are often not compact or generalizable, and they have a weak ability to generate testable predictions.

Many people are looking for an intuitively satisfying explanation of the relationship between consciousness and the physical world. They want to make an imaginative transition between a physical state (for example a mental image of a physical brain) and a conscious state (for example, the color red). We are never going to get this type of theory because we can only imagine conscious experiences, not the invisible physical world as it is in itself without any conscious properties. In principle we might be able to discover a theory that will enable us to make an imaginative transition from a conscious experience of a brain to another conscious experience. But this would only become intuitively convincing once we had learnt which brain patterns are linked to conscious experiences [1]. The limitations of brain scanning technology and human memory are likely to rule this out for the foreseeable future.

The most plausible type of theory that meets all of the criteria is a mathematical relationship between a formal description of the physical world (see Section 3) and a formal description of consciousness (see Section 4). Compact mathematical theories are the gold standard in many sciences, they can generate testable predictions and they can be applied to any physical system. The most impressive mathematical theory of consciousness that has been developed so far is Tononi's information integration theory [13]. Although it has serious limitations, it does show how we might be able to develop workable mathematical theories of consciousness in the future.

To develop fine grained mathematical theories of consciousness we need high resolution data from the brain. However, the vast amount of data that could potentially be recorded from a brain could not be comprehended by a single human brain. This suggests that humans are not likely to be capable of discovering mathematical theories of consciousness. We could address this problem by using artificial intelligence to search for mathematical relationships between consciousness and the physical world.

6 Conclusion

This paper has discussed four problems that have to be addressed before we can develop a scientific solution to MC4 machine consciousness. We need to reach agreement about how consciousness can be measured and abandon the idea that consciousness can be inferred from a machine's external behavior. Scientific work on consciousness needs to move away from neural correlates and start to look for more accurately defined spatiotemporal structures in the human brain that will lead to generalizable theories of consciousness. We need to develop ways of describing consciousness that avoid the anthropocentrism and context dependence of natural language. We have to drop our desire for intuitively satisfying explanations of consciousness and search for mathematical relationships between formal descriptions of consciousness and formal descriptions of the physical world. If these challenges can be addressed, we will have made significant progress towards tackling MC4 machine consciousness in a scientifically plausible way.

7 References

1. Gamez, D.: Human and Machine Consciousness. Open Book Publishers, Cambridge (2018)
2. Block, N.: Troubles with Functionalism. In: Eckert, M. (ed.) Theories of Mind: An Introductory Reader, pp. 97-102. Rowman & Littlefield, Maryland (2006)
3. Putnam, H.: Representation and Reality. MIT Press, Cambridge, Massachusetts; London (1988)
4. Bishop, J.M.: A Cognitive Computation Fallacy? Cognition, Computations and Panpsychism. Cognitive Computation 1, 221-233 (2009)
5. Searle, J.R.: Minds, Brains, and Programs. Behavioral and Brain Sciences 3, 417-457 (1980)
6. Koch, C., Massimini, M., Boly, M., Tononi, G.: Neural correlates of consciousness: progress and problems. Nat Rev Neurosci 17, 307-321 (2016)
7. Floridi, L.: The Method of Levels of Abstraction. Minds and Machines 18, 303-329 (2008)
8. Gamez, D.: Are Information or Data Patterns Correlated with Consciousness? Topoi 35, 225-239 (2016)
9. Gamez, D.: Can we Prove that there are Computational Correlates of Consciousness in the Brain? Journal of Cognitive Science 15, 149-186 (2014)
10. Chrisley, R.: Taking Embodiment Seriously: Nonconceptual Content and Robotics. In: Ford, K.M., Glymour, C., Hayes, P.J. (eds.) Android Epistemology. AAAI Press/ The MIT Press, Cambridge and London (1995)
11. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L.: Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology 21, 1641-1646 (2011)
12. Balduzzi, D., Tononi, G.: Qualia: the geometry of integrated information. PLoS Computational Biology 5, e1000462 (2009)
13. Oizumi, M., Albantakis, L., Tononi, G.: From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS Computational Biology 10, e1003588 (2014)