A blueprint for conscious machines

Arlindo L. Oliveira

Although it has been the subject of human thought for many centuries, consciousness remains a mysterious and controversial topic. Every one of us is overly familiar with the phenomenon, but there is little agreement on what it is, what it entails, and how it is created. Certainly, no other phenomenon is simultaneously so familiar and so hard to explain. Researchers from fields as different as medicine, philosophy, psychology, neurobiology, and computer science have tried for centuries to frame, describe, and address the mind–body problem—how consciousness arises from purely physical processes—but success has been, to say the least, limited. Many believe that it is a phenomenon that will remain, forever, unknowable, outside the reach of human understanding. David Chalmers (1) famously argued that, while we may advance in our understanding of the different physical processes that are associated with consciousness (the easy problems), we will never be able to understand the process that leads to subjective experiences (the hard problem), since understanding this phenomenon represents a challenge of an entirely different nature. Other theories, such as mysterianism, defend that, while the phenomenon may be understandable in principle, it will never be accessible to the human mind (2), due to intrinsic limitations of the human brain. Despite these difficulties, many scientists and philosophers have advanced tentative explanations for the phenomenon (3–5), and proposed ways to assess it (6), measure it (7), create it in machines (8), and even use it to improve the performance of artificial intelligence systems (9). The paper by Blum and Blum (10), in PNAS, approaches the problem from an engineering perspective, by showing how a specific computational architecture, which they have proposed (11), can explain several phenomena that are closely related to consciousness.

Theories of Consciousness

Although there is widespread agreement that consciousness exists, little else is consensual about the phenomenon. Some explanations, such as panpsychism, associate different degrees of consciousness with all types of matter, but there is a reasonable agreement that consciousness is generated by human and, possibly, nonhuman minds. This implies that any successful theory of mind is dependent on an understanding of what consciousness entails. Consciousness is, in fact, a complex multidimensional phenomenon, closely related to sentience, self-awareness, feelings, and the ability to have subjective experiences, or qualia. According to a popular formulation by Thomas Nagel (12), an entity is conscious if there is “something that it is like” to be that entity, meaning that a being is conscious if there exists a subjective way of perceiving the world from that being’s point of view. For example, it is possible to imagine being a horse (i.e., seeing the world from a horse’s point of view), but it is not possible to imagine being a cup.

Dualist theories avoid the problem of explaining how consciousness arises from purely physical phenomena by postulating that the phenomenon falls outside the physical realm, resourcing to a nonphysical (and yet unknown) entity that is required to explain consciousness. Although dualism was championed by no one less than René Descartes (13) and is implicit in most religions, current views among scientists heavily favor physicalism, the theory that consciousness is purely a result of physical processes. Adepts of physicalism, however, disagree on the importance of the underlying physical structure. Functionalists (14), who draw inspiration from Turing’s seminal article (15), argue that function is the relevant criterion and that different realizations of the same mental functions should be equally conscious. Opponents of functionalism believe that specific substrates or structures may be required to elicit consciousness and that it is conceptually possible to create two systems with exactly the same behavior, one conscious and the other unconscious (16). Believers in functionalism, naturally, cannot conceive of the existence of zombies, conceptual creatures that would behave exactly like conscious humans but would be devoid of conscious experiences.

Despite the abundance of theories about consciousness—a recent survey identified 29 different actively researched theories (17)—two specific approaches have been gaining favor in recent decades and have even been pitted against each other as possible explanations for the phenomenon (18): IIT (integrated information theory) (7) and GWT (global workspace theory) (3, 19). IIT, which is derived from a set of postulates, is based on a measure of integrated information that quantifies how much more information is generated by a system when compared with the information generated independently by the system parts. The higher this measure, the higher the consciousness of a system. GWT is an architecture developed to describe how conscious and unconscious processes interact to create consciousness. GWT makes use of a “theater” metaphor, a spotlight of attention, which corresponds to the active conscious process. The bright spot shining on this stage corresponds to...
the contents of consciousness. Processes can move in and out of the spotlight, interacting with each other, and can be, at any moment, either conscious or unconscious. The GWT “theater” should not be confused with the Cartesian theater, since it does not assume an outside observer nor does it imply a dualistic view of the world. GWT does explain several features of consciousness but was not originally intended as a constructive theory, to be used as a recipe to build conscious systems.

An Architecture for Conscious Machines

Blum and Blum (11) have filled this niche, by proposing a specific architecture for conscious systems, the Conscious Turing Machine (CTM). The CTM is inspired by the GWT architecture and the ideas proposed by Alan Turing (20) in his seminal work where the Turing machine was proposed as a model for general computation. The view of the authors is physicalist, but not strictly functionalist. The architecture proposed implies that consciousness is a property of all properly organized computing systems, whether made of biological material or digital switches, but also implies that consciousness is a consequence of the specific architecture of the system and not only of its function.

In their PNAS article (10), Blum and Blum explore how the CTM model can help explain different phenomena that are associated with consciousness and hard to understand. These phenomena include blindsight (the ability to see things one is not conscious of), inattentional blindness (the inability to perceive objects in plain sight when attention is diverted), change blindness (the inability to perceive visible changes when attention is focused on specific processes), visual illusions, and even dreams. In their paper, Blum and Blum also use the CTM architecture to provide an explanation of the seemingly paradoxical concept of free will, a phenomenon that has been difficult to explain since at least the times of Lucretius: If our bodies and our brains follow the laws of physics, and their possible evolutions are either predetermined (in classical physics) or chosen randomly from a predetermined set (in quantum mechanics), why does it seem that we could have decided otherwise, every time we took a conscious decision in the past?

The explanations provided in the PNAS paper (10) provide evidence that the GWT architecture and the CTM could be useful as explanatory models for the phenomenon of consciousness. At the least, they can explain some phenomena that remain strange and hard to understand. At the most, the CTM can, in time, become a blueprint for the construction of artificial consciousness, in the same way that Turing machines paved the way for modern computers.

However, one must proceed with care and not jump to the conclusion that these proposals are the final theories for consciousness and that only minor details need to be ironed out, since many objections remain. One objection that is particularly hard to address has been repeatedly leveled against these, and other, theories of consciousness. If the functionalist view is correct, and consciousness is only dependent on the function and not on structure or substrate, then the particular architecture of the CTM cannot be the only one that can support consciousness. As the authors are well aware, as renowned computer science experts, their CTM can be simulated by any universal Turing machine (UTM) or, in practice, by any modern computer. The ability of Turing machines to simulate each other is a well-known feature of these computational models, having been demonstrated in the 1937 work of Alan Turing (20), which has inspired the present PNAS paper (10).

So, if one could, in principle, simulate a CTM on a standard UTM, what is so special about this architecture? This question is also related to the Chinese room argument, devised by John Searle (21), where he imagines a system consisting of a man sitting in a room, who can use a large set of rules to answer fluently, in Chinese, questions posed in that language, of which he knows nothing. Searle argues that the system does not understand Chinese, even though it can fluently answer the questions it receives in that language.

In their PNAS article, Blum and Blum explore how the CTM model can help explain different phenomena that are associated with consciousness and hard to understand.

The answer to these questions remains open and vividly debated, but I venture to suggest that, in both cases, it comes down to resource usage, a theme that has been extensively explored in computer science literature. The assumption, made by Searle (21), that it is possible to write down a set of rules that would enable a non-Chinese-speaking person to answer questions in Chinese is, simply, not realistic. Such a set of rules could never be written down, nor would it fit the space available in any reasonable circumstances. In the same vein, a simulation of a CTM by a UTM, or by a program running in a standard von Neumann architecture, would simply be too slow to respond to changes in the environment and behave in a conscious way, in a time scale that makes sense. The CTM, with its particular architecture, can respond to internal and external stimuli rapidly, reacting to these requests in an expedited manner, which is hard to accomplish with a purely serial architecture. Such an explanation is consistent with the assumption that consciousness is a phenomenon that developed over evolutionary time to provide creatures with the ability to react, rapidly, to changes in the internal and external environments.

The CTM theory, which is itself based on the GWT, is unlikely to be the final explanation for the phenomenon of consciousness or even a workable blueprint for conscious machines. However, it may provide us with some of the tools required to advance our understanding of this mysterious but critically important phenomenon.

ACKNOWLEDGMENTS. A.L.O. would like to thank Mário Figueiredo, Tiago Domingos, Pedro Lima, Lenore Blum, and Manuel Blum for many discussions on these topics.
1. D. J. Chalmers, Facing up to the problem of consciousness. *J. Conscious. Stud.* 2, 208–219 (1995).
2. C. McGinn, *The Problem of Consciousness* (Blackwell, 1991).
3. B. J. Baars, *A Cognitive Theory of Consciousness* (Cambridge University Press, 1988).
4. D. C. Dennett, *Consciousness Explained* (Little, Brown, 1991).
5. A. R. Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Houghton Mifflin Harcourt, 1999).
6. C. Koch, M. Massimini, M. Boly, G. Tononi, Neural correlates of consciousness: Progress and problems. *Nat. Rev. Neurosci.* 17, 307–321 (2016).
7. G. Tononi, Consciousness as integrated information: A provisional manifesto. *Biol. Bull.* 215, 216–242 (2008).
8. S. Dehaene, H. Lau, S. Koeuder, What is consciousness, and could machines have it? *Science* 358, 486–492 (2017).
9. Y. Bengio, The consciousness prior. *arXiv* [Preprint] (2017). https://arxiv.org/abs/1709.08568 (Accessed 1 May 2022).
10. L. Blum, M. Blum, A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1072/pnas.2115934119 (2022).
11. M. Blum, L. Blum, A theoretical computer science perspective on consciousness. *J. Artificial Intell. Consciousness* 8, 1–42 (2021).
12. R. Descartes, *Meditationes de Prima Philosophia* (in Latin) (Michaelem Soly, 1641).
13. H. Putnam, “Minds and machines” in *Dimensions of Mind: A Symposium*, S. Hook, Ed. (New York University Press, 1960), pp. 138–164.
14. A. Turing, Computing machinery and intelligence. *Mind* 59, 433–460 (1950).
15. D. J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, 1996).
16. D. Satin et al., Theoretical models of consciousness: A scoping review. *Biol. Sci.* 11, 535 (2021).
17. S. Reardon, Rival theories face off over brain’s source of consciousness. *Science* 366, 293 (2019).
18. S. Dehaene, L. Naccache, Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79, 1–37 (2001).
19. A. Turing, On computable numbers, with an application to the Entscheidungs problem. *Proc. Lond. Math. Soc.* 2, 230–265 (1936).
20. J. Searle, Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424 (1980).