Engineering 25 (2023) 12–16

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Views & Comments

A Theoretical Computer Science Perspective on Consciousness and Artificial General Intelligence

Lenore Blum [a,b], Manuel Blum [a,b]

[a] Department of Computer Science, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[b] Electrical Engineering and Computer Science Department, University of California, Berkeley, CA 94720, USA

1. Introduction

We have defined the Conscious Turing Machine (CTM) for the purpose of investigating a theoretical computer science (TCS) approach to consciousness [1]. For this, we have hewn to the TCS demand for simplicity and understandability. The CTM is consequently and intentionally a simple machine. It is not a model of the brain, though its design has greatly benefited—and continues to benefit—from neuroscience and psychology.

The CTM is a model of and for consciousness. Its definitions of "conscious awareness" and the "feeling of consciousness" in the CTM are followed by arguments explaining why the definitions capture commonly accepted understandings of consciousness and the associated feelings related to consciousness [2].

Although it is developed to understand consciousness, the CTM offers a thoughtful and novel guide to the creation of an artificial general intelligence (AGI). For example, the CTM has an enormous number of powerful processors, some with specialized expertise, others unspecialized but poised to develop an expertise. For whatever problem must be dealt with, the CTM has an excellent way to utilize those processors that have the required knowledge, ability, and time to work on the problem, even if it, the CTM, is not aware of which of the processors these may be.

2. CTM in a nutshell

The CTM is a mathematical formalization in the spirit of TCS of a modified version of the global workspace theory (GWT) of consciousness (Fig. 1 [2]) that was conceptualized by cognitive neuroscientist Baars [3,4] based on work done at Carnegie Mellon (by Simon [5], Reddy [6], Newell [7], and Anderson [8]), and was subsequently extended to the global neuronal workspace (GNW) by Dehaene, Changeux, Mashour, and Roelfsema [9–11].

Baars describes conscious awareness through a theater analogy as the activity of actors in a play performing on a stage of working memory, with their performance being observed by a large audience of unconscious processors sitting in the dark.

In the CTM, the stage is represented by a short-term memory (STM), which at each tick of a central clock contains the CTM's conscious content. Audience members are represented by CTM's long-term memory (LTM) processors, an enormous array of powerful random-access processors, some with specialized expertise and some without, but all with the deep learning hardware required to develop or expand the expertise. LTM processors generate predictions about CTM's environment and obtain feedback from CTM's environment and other processors. Based on that feedback, learning algorithms internal to each processor improve that processor's behavior.

LTM processors compete to bring their questions, answers, and comments to the STM (stage) for immediate broadcast to the audience. The information that passes through STM is coded as chunks. Conscious awareness/attention is defined in the CTM as the reception by all LTM processors of a chunk broadcast from STM. In time, various LTM processors become connected via links which serve as channels for carrying chunks directly between processors. Links turn indirect conscious communication (via STM) between LTM processors into direct unconscious communication (not involving STM) between those same processors.

While these definitions are natural, they are merely definitions. They are not arguments that the CTM is conscious in the standard sense that the word "conscious" is normally used. We argue, however, that the definitions and explanations from the CTM capture widely accepted intuitive understandings of consciousness.

Although inspired by Baars' global workspace model, there are significant differences between Baars' model (Fig. 1(a)) and the CTM (Fig. 1(b)). With respect to architecture, Baars' model has a central executive, while the CTM has none. The CTM is a distributed system that enables the emergence of functionality and applications to general intelligence. In the CTM, input sensors transmit environmental information directly to appropriate LTM processors; output actuators act on the environment based on information obtained directly from specific LTM processors. In the Baars' model, inputs and outputs are processed via working memory. In the CTM, chunks are formally defined. They are submitted by LTM processors to a well-defined competition for STM. In the Baars' model, neither chunks nor competition are formally defined. For Baars, a conscious event transpires between the input and the central executive. In the CTM, conscious awareness occurs

https://doi.org/10.1016/j.eng.2023.03.010
2095-8099/© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.

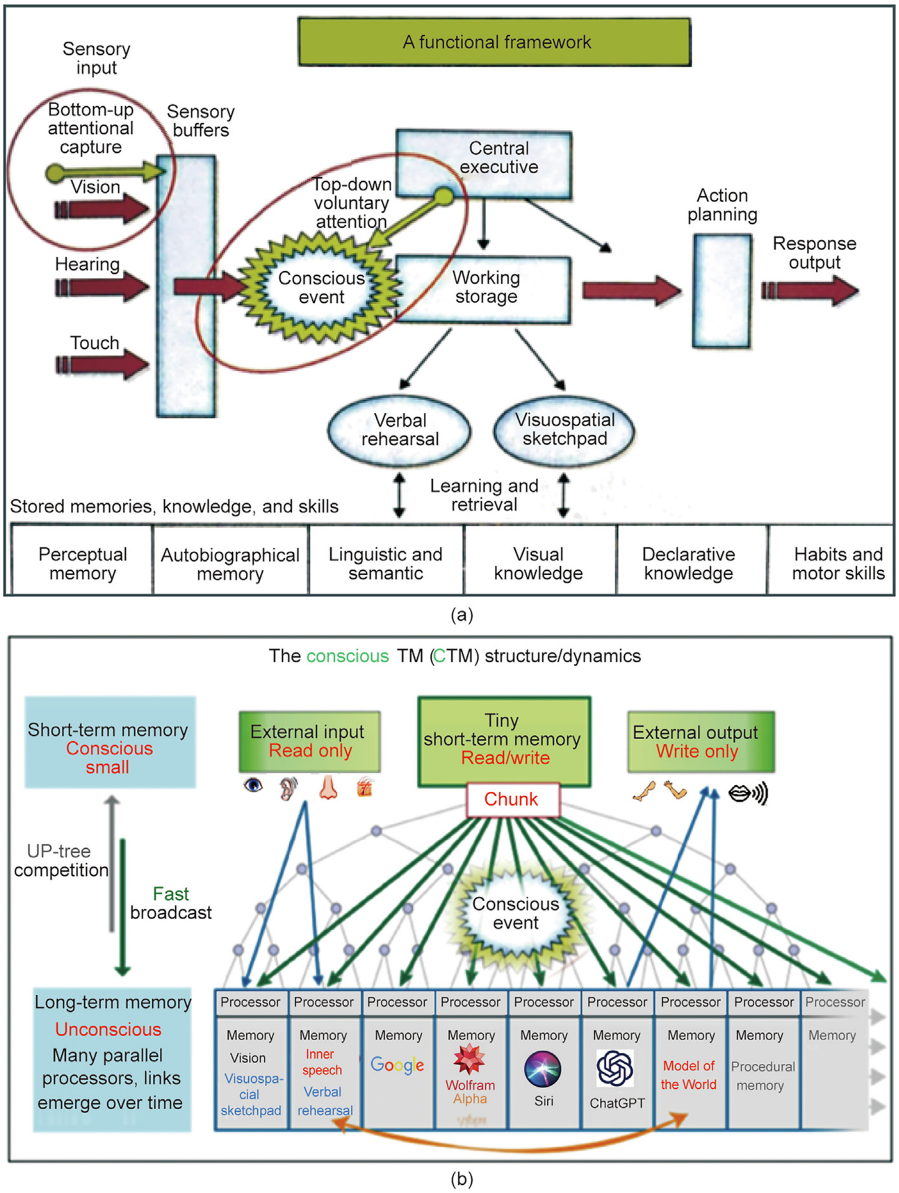

Fig. 1. Sketches of models: (a) Baars' GWT model and (b) the CTM [2]. TM: Turing machine.

at the moment when all LTM processors receive the information broadcast from STM.

Although inspired by Turing's simple and powerful model of a computer, the CTM is clearly not a standard Turing machine. That's because what gives the CTM its "feeling of consciousness" is not its input–output map or its computing power but rather its global workspace architecture; its predictive dynamics (cycles of prediction, feedback, and learning); its rich multi-modal inner language (which we call Brainish [12]) for inter-processor communication; and certain particularly important LTM processors including the Inner Speech, Inner Generalized Sensation, and Model of the World (MotW) processors.

As the introduction outlines, the CTM is not intended to be a model of the brain but to provide a simple model of consciousness. Even there, the model can hardly be expected to explain all phenomena of consciousness; it is too simple for that. The reasonableness of the model (and its TCS perspective) should be evaluated by its contribution to the discussion of consciousness and to the understanding of aspects of consciousness such as emotions of pain and pleasure and the potential of the CTM for use as an AGI.

The above paragraphs of Section 2 and Fig. 1(b) present an overview of the CTM model. We refer the reader to two papers for the formal definition of CTM as seven-tuple and chunk as six-tuple. The first [1] explores explanations for feelings of pain and pleasure in the CTM[†]. The second [2] explores additional phenomena generally associated with consciousness, such as free will and the disorder of blindsight[‡]. We provide plausible explanations derived from the formal model and draw confirmation from the model's high-level consistencies with the existing literature on psychology and neuroscience [4,13,14][††].

[†] For an update on pain and pleasure in the CTM, see Chapter 4 of the Blum's *The Hard Problem for Pain and Pleasure* (https://arxiv.org/abs/2011.09850).

[‡] For an update see https://arxiv.org/abs/2107.13704.

[††] A forthcoming monograph of Blum et al. entitled *Towards a Conscious AI: A Computer Architecture Inspired by Cognitive Neuroscience* describes in more detail how the CTM works. Its three appendices demonstrate how CTM can operate with no central executive, despite that all other global workspace models (such as Baars' functional model [4]) hypothesize a central executive. These same appendices show by example how CTM functions with just one chunk in STM instead of George Miller's 7 ± 2 [13] or Nelson Cowan's 3 or 4 [14]. No other models suggest that o ne chunk will suffice.

We note a historical synergy between theoretical computer science and neuroscience. Turing's simple computer model led neuroscientist Dr. Warren S. McCulloch and mathematician Walter Pitts to define their formal neuron, itself a simple model of a neuron [15]. Mathematics forced their model to have inhibition, not just excitation—because without inhibition, loop-free circuits of formal neurons can only compute monotonic functions—and these do not suffice to build a universal Turing machine. The McCulloch–Pitts neuron also gave rise to the mathematical formalization of neural nets and subsequent deep learning algorithms [16], further illustrating ongoing synergies.

3. CTM and AGI[†]

Although the CTM is defined to be a very simple model of consciousness, it being explicitly formally defined for generating definitions and understandings of consciousness, it also suggests a novel approach to AGI, giving a way to coordinate an enormous number of (special-purpose) artificial intelligence (AI) agents for the purpose of building the AGI. In particular, it suggests how to coordinate a huge number—10^7 or more[‡]—of processors, some specialized, most initially unspecialized but capable of being specialized, to solve a variety of unforeseen problems. In an AGI, specialized processors could be tasked to obtain information from a number of search engines such as ChatGPT or GPT-4, Wikipedia, Google Translate, Wolfram Alpha, the weather channel, newspapers, and HOL Light [17][††]. These are existing ready-made processors. Many more processors could and would be developed from scratch as needed by the CTM itself.

A principal contribution of the CTM is a way to coordinate processors that must solve a diverse collection of unforeseen problems. The CTM assigns tasks to its processors, even though it has no central executive and no single processor or collection of processors to keep track of which processors have the time and know-how to do the task. How it does that is an interesting koan.

Suppose (a processor of) CTM has a task to perform but nary a clue how to perform it, and no idea which, if any, of the many LTM processors has the knowledge, ability, and time to deal with the task. Through a well-defined competition for STM, that processor submits a request for help to all LTM processors. The request will with some probability **[**well-defined in the CTM**]** reach STM for global broadcast to all (LTM) processors. All processors that have relevant expertise and time to work on the problem respond, again through the competition and global broadcast. Their broadcasts in turn can motivate other processors to come into play. In this way, the CTM engages powerful processors to collectively solve a problem that CTM had no idea how to solve, no approach to solving the problem, and no sense which processors, if any, could be helpful.

With regard to logical thinking and mathematics, the CTM is ideal for orchestrating its processors to recognize a sound logical argument, write a correct mathematical proof, and check its work. It can program its processors and modify them as needed to reach such goals. For example, one processor can suggest an approach to a proof, a second can evaluate the likelihood that the approach will work, a third can outline a potential "proof," a fourth can check if a proposed "proof" is really a proof (pointing out what problems arise) if not, and so on.

More generally, the CTM can and must have processors for checking the truth of statements or arguments. Consider for example the assertion that shrimp is healthy to eat. One source may state "YES, shrimp is healthy," but that statement comes from an association that represents the frozen foods industry, making it suspect. Another may state, "NO, shrimp is unhealthy: it has lots of cholesterol and cholesterol is unhealthy." Yet another source may state, and we paraphrase, "YES, shrimp is healthy, and the cholesterol in shrimp is the healthy LDL kind." As that last quote (from De Oliveira e Silva et al. [18]) is authored by a scholar from a respected (Rockefeller) University and published in a reputable refereed journal, its case is the strongest thus far. Responses to the paper may further strengthen or weaken the assessment.

4. What features does the CTM bring to the design of an AGI?

The CTM is a simple TCS model of consciousness. It is not a brain, and it is also not an AGI. That said, we suggest that its basic features have value in the design of an AGI. For example:

(1) The CTM suggests an approach to building an AGI that has no central executive, meaning no conductor, and no stage director. It has an enormous number of processors, each of which is largely self-directing, rather like the members of a self-conducting musical ensemble. This architecture is unexpected and strange because large assemblies, orchestras, and political states, generally have a leader. The CTM has only one actor on stage. That actor is not the leader. It serves merely as ① a small buffer to hold the winning chunk of the current competition and ② a broadcasting station to beam that chunk to the entire LTM audience.

(2) The CTM solves a conundrum: How is it possible for a long and subtle argument, for example, the proof of a difficult theorem, to be understood—grasped as it were in the palm of one's hand? That handful is the final chunk that contains something like "Eureka! I got it." That chunk is from a processor that, if asked, can point to the outline of a proof, each phrase of which can point to what in the proof it depends on and what depends on it.

(3) The audience members of the CTM global workspace are self-monitoring processors. They have the final word on the value of their personal contribution[‡‡].

(4) Baars [4] says that the audience members of the global workspace consult among themselves to agree on who to send to the stage, but how do they do that? Baars does not say.

The CTM, however, explains precisely how to do it. It hosts a well-defined competition that is similar to and provably better than a chess or tennis tournament in that, at minor cost and negligible extra time, it guarantees that each processor will have its information broadcast with probability proportional to the value of its information (a value computed dispassionately by the processor's sleeping experts algorithm), something that chess and tennis tournaments do not, indeed cannot do.

(5) Sleeping experts learning algorithms [19,20] determine the manner in which a processor assigns a value to its information, a value that is (mostly) self-determined by the processor. The CTM makes do with no teacher at the head of the class and no answer sheet to make corrections. Its processors self-predict and self-correct based on feedback from inputs, links, and broadcasts. We expect AGI designers to be interested in how they manage that.

[†] AGI is the ability of an intelligent agent to understand or learn any intellectual task that can be learned by human beings or other animals.

[‡] 10^7 is the estimated number of cortical columns in the brain.

[††] HOL Light is a formal mathematical programming system for generating proofs that are logically and mathematically correct, and/or for checking any "proofs" given it. It was used, for example, by Tom Hales to prove the correctness of his solution to the Kepler conjecture.

[‡‡] The idea reminds us of the visiting Admiral at Massachusetts Institute of Technology who told McCulloch's neurophysiology group in 1959 that in the navy, it is not the flagship that commands the fleet: It is the ship with the information. The CTM's processors are the ships of an enormous fleet. The workings of the CTM give precise meaning to the Admiral's words.

(6) CTM's MotW processor develops world models. These world models have considerable importance for planning, testing, making corrections, distinguishing fiction from nonfiction, living from nonliving, self from not-self, and most importantly for contributing to feelings of consciousness. The CTM has at birth a rudimentary MotW processor, then continuously upgrades it and its models. The CTM's approach to constructing and maintaining its world models is particularly essential, since the CTM does not directly perceive the world like the Baars model (Fig. 1(a)) but indirectly through its world models (Fig. 1(b)).

(7) The CTM can explain what it is doing and why. It can answer questions regarding the "how" and "why" of its actions and provide arguments to support its answers.

5. Arguments for and against using CTM as a guide for creating an AGI

The specification of CTM[†] elucidates its functioning. Descriptions are provided for how each processor assigns a valenced measure of importance (a weight) to its chunks and how that measure is affected by the sleeping-experts algorithm in each LTM processor. There is a description of how the tournament that starts at time t is run, involving a competition among all N chunks[‡], contributed by all N processors at time t. Each such tournament takes ($\log_2 N$) steps, the first step being $N/2$ matches performed in parallel, the second being $N/4$, and the final being a single match to crown the winner. The tournament is as fast as any tournament for tennis and chess but better as chess and tennis tournaments do not guarantee, as does the CTM tournament, that chunks get to STM with probability proportional to their importance. On that account, CTM processors can remain hard-wired and in place without affecting which chunk will win any given tournament.

Another argument for using CTM as a guide for AGI comes from neuroscience research demonstrating that in humans, "language and thought are not the same thing" [21]. Individuals with global aphasia, "despite their near-total loss of language are nonetheless able to add and subtract, solve logic problems, think about another's thoughts, appreciate music, ..." Healthy adults "strongly engage the brain's language areas when they understand a sentence, but not when they perform non-linguistic tasks such as arithmetic, storing information in working memory,...., or listening to music."

Influenced by this research and comparing large language models to formal and functional properties of human language, Mahowald et al. [22] argue that while large language models "are good models of language," [they are] "incomplete models of human thought." They further argue that "future language models can master both formal and functional linguistic competence by establishing a division of labor between the core language system and components for other cognitive processes, ..." as in the human brain. They provide two suggestions to accomplish this. Their first suggestion is architectural modularity (whereby separate specialized modules work together with each other). The CTM incorporates such modularity by utilizing multiple processors with different input domains, knowledge, and functionalities.

Their second suggestion, emergent modularity (modularity that emerges within a large language model), points to the possibility that deep learning alone will suffice for AGI, though they argue that architectural modularity is "much better aligned with... real-life language use."

The possibility of emergence is supported by Bubeck et al. [23] who examined the impressive and multiple "sparks" of general intelligence demonstrated by early experiments with the large language model GPT-4 and viewed it as an early version of an AGI.

It may turn out that no global workspace model or CTM is needed to create an AGI, that deep learning independently suffices, and that a single machine with a sufficiently large matrix size can be a universal AGI. One can argue, on the other hand, that the matrix size of a deep learning AGI must increase with the square of the number of problems it is to solve, and such a size would be difficult to achieve since the best current AIs currently use about 10^{14} parameters. The CTM, which is designed for understanding consciousness, can reasonably handle 10^7 AIs with 10^{14} parameters per AI, for 10^{21} parameters. For comparison, there are 10^{11} stars in the Milky way galaxy and 2×10^{23} stars in the visible universe. Avogadro's number is three times as large as that at approximately 6.0221×10^{23}.

Returning to consciousness, the CTM global workspace model is a promising untapped approach for transforming AI into AGI. We expect that robots with CTM-like brains that construct models of the world will have "feelings of consciousness," and hence be more likely to experience empathy. Finally, as AIs become more human-like, understanding consciousness and feelings of pain will be critical if we aim to avoid inflicting suffering on our planet's co-inhabitants.

Acknowledgments

This work of Lenore Blum and Manuel Blum was supported in part by Carnegie Mellon University (CMU), in part by a sabbatical year from CMU at the Simon's Institute for the Theory of Computing, and in part by a generous gift from UniDT. We are grateful to Jean-Louis Villecroze for his ongoing work to simulate CTM, Paul Liang for his insight into multimodal Brainish, our students at CMU and Peking University who constantly challenge us, and our friends and colleagues Raj Reddy and Michael Xuan for their suggestions, personal support, and extraordinary encouragement.

References

[1] Blum M, Blum L. A theoretical computer science perspective on consciousness. J Artif Intell Res Conscious 2021;8(1):1–42.
[2] Blum M, Blum L. A theory of consciousness from a theoretical computer science perspective: insights from the Conscious Turing Machine. Proc Natl Acad Sci USA 2022;119(21):e2115934119.
[3] Baars BJ. A cognitive theory of consciousness. Cambridge: Cambridge University Press; 1988.
[4] Baars BJ. In the theater of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. J Conscious Stud 1997;4(4):292–309.
[5] Simon HA. The sciences of the artificial. Cambridge: MIT Press; 1969.
[6] Reddy DR. Speech recognition by machine: a review. Proc IEEE 1976;64(4):501–31.
[7] Newell A. Unified theories of cognition. Cambridge: Harvard University Press; 1990.
[8] Anderson JR. ACT: a simple theory of complex cognition. Am Psychol 1996;51(4):355–65.
[9] Dehaene S, Changeux JP. Experimental and theoretical approaches to conscious processing. Neuron 2011;70(2):200–27.
[10] Dehaene S. Consciousness and the brain: deciphering how the brain codes our thoughts. New York City: Viking Press; 2014.
[11] Mashour GA, Roelfsema P, Changeux JP, Dehaene S. Conscious processing and the global neuronal workspace hypothesis. Neuron 2020;105(5):776–98.
[12] Liang PP. Brainish: formalizing a multimodal language for intelligence and consciousness. 2023. arXiv:2205.00001.
[13] Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956;63(2):81–97.
[14] Cowan N. George Miller's magical number of immediate memory in retrospect: observations on the faltering progression of science. Psychol Rev 2015;122(3):536–41.
[15] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33.

[†] In Refs. [1,2] and in the upcoming monograph of Blum et al. (title: *Towards a Conscious AI: A Computer Architecture Inspired by Cognitive Neuroscience*).

[‡] We assume that $N = 2^k$ for some positive integer k.

[16] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.

[17] Hales TC. A proof of the Kepler conjecture. Ann Math 2015;162(3):1065–185.

[18] De Oliveira e Silva ER, Seidman CE, Tian JJ, Hudgins LC, Sacks FM, Breslow JL. Effects of shrimp consumption on plasma lipoproteins. Am J Clin Nutr 1996;64(5):712–7.

[19] Blum A. Empirical support for winnow and weighted-majority algorithms: results on a calendar scheduling domain. Mach Learn 1997;26:5–23.

[20] 5wBlum A, Hopcroft J, Kannan R. Foundations of data science. Cambridge: CambridgeUniversityPress; 2020.

[21] Fedorenko E, Varley R. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. Ann N Y Acad Sci 2016;1369(1):132–53.

[22] Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, Fedorenko E. Dissociating language and thought in large language models: a cognitive perspective. 2023. arXiv:2301.06627.

[23] Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. 2023. arXiv:2303.12712.