Contents lists available at ScienceDirect

Engineering

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

Views & Comments

从计算机科学理论审视意识和通用人工智能

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. 引言

我们定义了意识图灵机（CTM），目的是从计算机科学理论（TCS）角度来研究意识[1]。为此，我们遵循了TCS对简单性和可理解性的要求。因此，CTM被有目的地设计为一台简单机器。CTM不是大脑的模型，尽管它的设计从神经科学和心理学中大大受益，并继续受益。

CTM是一种意识的模型。在CTM中，它对“意识觉察”和“意识感觉”进行定义之后，解释了为什么这些定义捕获了普遍接受的对意识的理解和与意识相关的感觉[2]。

虽然CTM是为了理解意识而发展起来的，但它为创造人工通用智能（AGI）提供了一个深思熟虑和新颖的指导。例如，CTM拥有大量功能强大的处理器，有些具有专业知识，另一些虽没有专业知识，但有望培育出专业知识。面对任何需要解决的问题，CTM都可以很灵活地利用那些具有所需知识、能力和时间来解决问题的处理器，即使CTM本身并不知道哪些处理器具备这些条件。

2. CTM概括

CTM是以TCS精神对意识研究领域中著名的全局工作空间理论（GWT）进行拓展后的数学化形式表达（图1 [2]）。该理论由认知神经学家Baars [3–4]根据卡内基梅隆大学的工作（Simon [5]、Reddy [6]、Newell [7]和Anderson [8]）构思而成，随后由Dehaene、Changeux、Mashour和Roelfsema [9–11]扩展到全局神经元工作空间（GNW）。

Baars通过戏剧类比将意识觉察描述为演员在工作记忆舞台上表演的活动，他们的表演被大量坐在黑暗中的无意识处理器的观众所观察。

在CTM中，这个舞台由一个短期记忆（STM）表示，在中央时钟每一次滴答声中，它都包含了CTM的意识内容。观众则由CTM的长期内存（LTM）处理器代表，这是一个由大量强大的随机访问处理器组成的庞大阵列，其中有些处理器有专业知识，有些没有，但都有开发或培育专业知识所需的深度学习硬件。LTM处理器生成对CTM环境的预测，并从CTM的环境和其他处理器中获得反馈。基于这些反馈，每个处理器内部的学习算法都可以改进该处理器的行为。

LTM处理器竞相将它们的问题、答案和评论带到STM（舞台），以便立即向观众广播。在STM中的信息被编码为块。在CTM中，意识觉察/注意被定义为所有LTM处理器接收来自STM广播的块。随着时间的推移，不同LTM处理器会通过链接连接起来，这些链接作为直接在处理器之间传输信息块的通道。链接将LTM处理器之间的间接意识通信（通过STM）转化为这些相同处理器之间的直接无意识通信（不涉及STM）。

虽然这些定义是很自然的，但它们仅仅是一种定义。它们并不能论证CTM在“意识”一词通常使用的标准意

英文原文：Engineering, 2023, 25(6): 12–16
引用本文：Lenore Blum, Manuel Blum.A Theoretical Computer Science Perspective on Consciousness and Artificial General Intelligence. *Engineering*, https://doi.org/10.1016/j.eng.2023.03.010

义上是有意识的。然而，我们认为，CTM的定义和解释捕捉到了被广泛接受的对意识的直觉理解。虽然受到Baars全局工作空间模型的启发，但Baars模型[图1（a）]与CTM [图1（b）]之间存在显著差异。在架构方面，Baars的模型有一个中央执行器，而CTM则没有。CTM是一个分布式系统，能够实现一般智能的功能和应用。在CTM中，输入传感器会将环境信息直接传输到相应的LTM处理器；输出执行器基于直接从特定的LTM处理器获得的信息来影响环境。在Baars的模型中，输入和输出都通过工作记忆进行处理。在CTM中，块是正式定义的。它们由LTM处理器提交给STM进行定义明确的竞争。在Baars的模型中，块和竞争都没有被正式定义。对于Baars来说，一个有意识的事件发生在输入和中央执行器之间。在CTM中，当所有LTM处理器接收到STM广播的信息时，意识觉察发生。

尽管受到图灵简单而强大的计算机模型启发，但CTM显然不是一个标准的图灵机。这是因为赋予CTM“意识感”的，不是它的从输入到输出映射或它的计算能力，而是它的全局工作空间架构；它的预测动态（预测、反馈和学习周期）；丰富的多模态内部语言（我们称之为大脑语言[12]），用于处理器间的通信；以及一些特别重要的LTM处理器，包括内部语音、内在广义感觉和全局模型处理器。

正如引言所概述的那样，CTM并不是一个大脑模型，而是提供意识的简单模型。即使如此，也很难期望这个模型能解释所有的意识现象，因为它太简单了。模型的合理

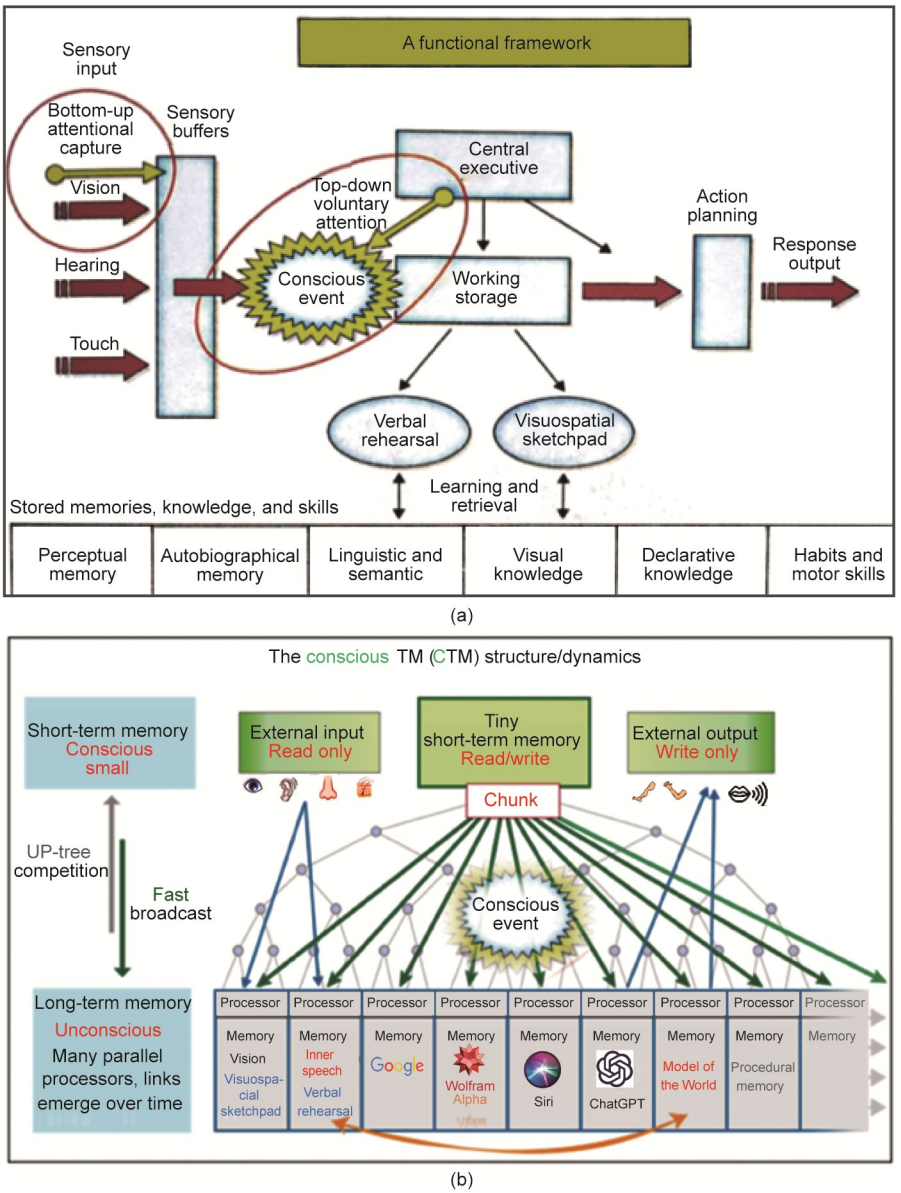

图1. 模型概述。（a）Baars的GWT模型和（b）CTM [2]。TM：图灵机。

性（及其TCS视角）应该通过对意识的讨论和理解来评估，譬如具有感受疼痛和快乐的情绪以及CTM作为AGI的潜力。

第2节的上述段落和图1（b）提供了对CTM模型的概述。我们推荐读者阅读两篇论文，以获取CTM作为七元组和块作为六元组的正式定义。第一篇论文[1]探索了在CTM中的痛苦感和快乐感的解释。第二篇论文[2]探索了通常与意识相关的其他现象，如自由意志和盲视障碍。我们提供了从正式模型中得到的合理解释，并从该模型与现有心理学和神经科学文献[4,13–14]的高水平一致性中得到确认。我们注意到理论计算机科学和神经科学之间的历史协同作用。图灵的简单计算机模型促使神经科学家Warren S. McCulloch博士和数学家Walter Pitts共同定义了他们所理解的形式化神经元，即一种简单的神经元模型[15]。他们通过数学所描述的神经元模型具有抑制性，而不仅仅是兴奋——因为如果没有抑制，神经元所构成的无环回路只能计算单调函数——而这些还不足以建立一个通用的图灵机。McCulloch-Pitts神经元还催生了神经网络的数学形式化和随后的深度学习算法[16]，进一步说明了持续的协同作用。

3. CTM和AGI

虽然CTM被定义为一个非常简单的意识模型，它被明确正式定义为生成定义和理解意识，但它也为AGI提出了一种新颖的方法，为构建AGI提供了一种协调大量（特殊用途）人工智能体（AI agent）的方法。具体而言，它提出了如何协调10^7或者更多处理器的方法，其中一些是专用处理器，大多数最初是非专业的，但随后能够被培育为专业的，以解决各种不可预见的问题。在AGI中，这些专业的处理器可以从许多搜索引擎中获取信息，如Chat-GPT或GPT-4、维基百科、谷歌翻译、Wolfram Alpha、天气频道、报纸和HOL Light [17]。这些都是现有的现成处理器。根据CTM本身的需要，可以而且将会从头开始开发更多的处理器。

CTM的主要贡献在于，它提供了一种协调解决各种不可预见问题的处理器的方法。CTM将任务分配给它的处理器，即使它没有中央执行器，也没有单个处理器或处理器集合来跟踪哪些处理器有时间和技术来完成任务。它是如何做到这一点的是一个有趣的难题。

假设CTM或者其包含的一个处理器有一个任务要执行，但不知道如何执行它，并且不知道众多LTM处理器中哪一个有知识、能力和时间来处理这个任务。通过定义明确的STM竞争，该处理器向所有LTM处理器提出帮助请求。该请求将以一定的概率（在CTM中定义明确）到达STM，以便向所有（LTM）处理器进行全局广播。所有拥有相关专业知识和时间解决这个问题的处理器都会通过竞争和全局广播做出回应。它们的广播反过来又可以激励其他处理器开始发挥作用。通过这种方式，CTM使用强大的处理器来共同解决一个不知道如何解决、没有解决问题的方法，并且无法确定哪些处理器（如果有的话）可以提供帮助的问题。

关于逻辑思维和数学方面，CTM是理想的协调其处理器来识别合理的逻辑论证，给出正确的数学证明，并检查它的工作的工具。它可以根据需要对处理器进行编程和修改，以实现这些目标。例如，一个处理器可以提出一个方法的证明，第二个处理器可以评估这种方法成功的可能性，第三个处理器可以概述一个潜在的“证明”，第四个处理器可以检查提出“证明”是否真的是证明（指出出现的问题），等等。

更一般地说，CTM可以而且必须有用来检查语句或参数真实性的处理器。例如，假设吃虾是健康的。有一种消息来源可能会说：“是的，虾是健康的”，但这种说法来自于一个代表冷冻食品行业的协会，这让它备受质疑。另一种说法可能是，“不，虾是不健康的：它有很多胆固醇，而胆固醇是不健康的。”我们解释的另一个来源可能称，“是的，虾是健康的，虾中的胆固醇是健康的低密度脂蛋白。”由于最后一段的作者（来自De Oliveira e Silva等[18]）是一位受人尊敬的（Rockefeller）大学学者，并且该论文发表在一本著名的期刊上，因此其论据是迄今为止最有力的。对该论文的回应可能会进一步加强或削弱这一评估。

4. CTM为AGI的设计带来了什么特征?

CTM是一个简单的TCS意识模型。它不是大脑，也不是AGI。也就是说，我们认为它的基本特性在AGI的设计中具有价值。例如：

（1）CTM提出了一种建立AGI的方法，这种AGI没有中央执行器，这意味着没有指挥家，也没有舞台导演。它有大量的处理器，每个处理器在很大程度上都是自导自演的，就像一个自导自演的音乐团体成员一样。这种架构令人出乎意料并且感到奇怪，因为大型集会、管弦乐队和国家通常都有一个领袖。CTM在舞台上只有一个演员。那个演员不是领导者。它只是作为：①一个小缓冲区来保持当前比赛的获胜曲目；②一个广播电台，用来向LTM

的所有观众传送这些曲目。

（2）CTM解决了一个难题：一个冗长而微妙的论证，例如，一个困难的定理证明，如何能够被理解——就像抓住它一样容易？那个掌中的一小块，就好比包含了“我搞明白了!”之类的内容。这个块来自一个处理器，如果被问及，它可以指向一个证明的大纲，其中的每个短语都可以指出它在证明中所依赖的是什么，以及依赖于它的是什么。

（3）CTM全局工作区的听众都是可自我监控的处理器。他们对个人贡献的价值有最终的决定。

（4）Baars [4]说，全局工作空间的观众会相互协商，同意送谁上舞台，但他们是如何做到的呢？巴尔斯没有说。

然而，CTM却精确地解释了如何做到这一点。它举办了一场定义明确的比赛，类似于国际象棋或网球比赛，并且可以证明其比国际象棋或网球比赛更好，因为它以很小的成本和可以忽略不计的额外时间，就能保证每个处理器将以与其信息价值（由处理器的睡眠专家算法冷静计算的值）成正比的概率广播其信息，这是国际象棋和网球比赛做不到，实际上也无法做到的。

（5）睡眠专家学习算法[19–20]决定处理器为其信息分配值的方式，该值（大部分）是由处理器自行决定的。CTM在没有教师指导和纠正答案的情况下进行运作。它的处理器根据来自输入、链接和广播的反馈进行自我预测和自我纠正。我们希望AGI的设计师们能够关注它们是如何做到这一点的。

（6）CTM的全局模型（MotW）处理器开发了全局模型。这些全局模型对于计划、测试、修正、区分虚构与非虚构、生命与非生命、自我与非自我具有相当重要的意义，尤其是对于产生有意识的感觉而言。CTM最初就有一个基本的MotW处理器，然后不断升级它及其模型。CTM构建和维护其全局模型的方法特别重要，因为CTM并不像Baars模型那样直接感知全局[图1（a）]，而是通过其全局模型[图1（b）]间接感知世界。CTM可以解释它在做什么，以及为什么要做。它可以回答有关其行动的“怎样做”和“为什么”的问题，并提供论据来支持其答案。

5. 支持和反对使用CTM作为创建AGI的论点

CTM的规范说明了其功能，描述了每个处理器如何为其块分配重要性的有价度量（权重），以及该度量如何受到每个LTM处理器中的睡眠专家算法的影响。这里描述了在t时刻开始的比赛是如何运行的，包括在所有N个块之间的竞争，由在t时刻的所有N个处理器贡献。每一次这样的比赛都需要$\log_2 N$步，第一步是并行进行的$N/2$场比赛，第二步是$N/4$场比赛，最后是一场单场比赛获得获胜者。该比赛的速度与任何网球和国际象棋锦标赛一样快，但更好，因为国际象棋和网球锦标赛不能像CTM比赛一样保证，块到达STM的概率与它们的重要性成正比。因此，CTM处理器可以保持硬连线，而不影响哪个块将赢得任何给定的比赛。

使用CTM作为AGI指南的另一个论点来自于神经科学研究，它证明了在人类中，“语言和思想不是一回事”[21]。患有全面性失语症的人，“尽管他们几乎完全失去了语言能力，但他们仍然能够做加减法，解决逻辑问题，思考别人的想法，欣赏音乐……”“健康的成年人在理解一个句子时，大脑的语言区域会被强烈地调动，但当他们执行非语言任务时，如算术、在工作记忆中存储信息……，或听音乐时却不能。”

受这项研究的影响，并将大型语言模型与人类语言的形式和功能属性进行了比较，Mahowald等[21]认为，虽然大型语言模型“是很好的语言模型”，但它们是“人类思想的不完整模型”。他们进一步认为，“未来的语言模型可以通过在核心语言系统和其他认知过程的组成部分之间建立一个劳动分工来掌握形式和功能性的语言能力，……”，就像在人类大脑中一样。他们提供了两个建议来实现这一目标。他们的第一个建议是架构模块化（即独立的专门模块彼此协同工作）。CTM通过利用具有不同输入域、知识和功能的多个处理器，集成了这种模块化。

他们的第二个建议是新型模块化（即在大型语言模型中出现的模块化），指出深度学习就足以实现AGI的可能性，尽管他们认为架构模块化“与……现实生活中的语言一起使用更好”。

Bubeck等[23]支持这种出现的可能性，并且研究了大型语言模型GPT-4早期实验中显示的令人印象深刻的多重“火花”，并将其视为AGI的早期版本。

事实可能是，创建AGI不需要全局工作空间模型或CTM，仅仅依靠深度学习就足够了，一台具有足够大矩阵规模的单个机器可以成为通用AGI。另一方面，我们可以认为，深度学习AGI的矩阵大小必须随着它要解决的问题数量的平方而增加，这样的大小很难实现，因为目前最好的AI模型具有大约10^{14}个参数。CTM是为理解意识而设计的，它可以合理地处理10^7个AI，每个AI有10^{14}个参数，即10^{21}个参数。相比之下，银河系中有10^{11}颗恒星，可见宇宙中有2×10^{23}颗恒星。阿伏伽德罗的数字是它的三

倍，约为6.0221×10^{23}。回到意识问题上，CTM全局工作空间模型是将人工智能转化为AGI的一种很有前途的未开发的方法。我们预计，拥有类似CTM的大脑、构建全局模型的机器人将拥有“意识感”，从而更有可能产生共鸣。最后，随着人工智能变得更像人类，如果我们的目标是避免对我们星球上的共同居民造成痛苦，那么理解意识和痛苦感将是至关重要的。

Acknowledgements

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.
[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.