MIT Press • Open Encyclopedia of Cognitive Science

Computational Theory of Mind

Matteo Colombo[1]

[1]Tilburg University

MIT Press

Published on: Jun 30, 2026

URL: https://oecs.mit.edu/pub/cmg9r89u

The *computational theory of mind* (CTM) claims that the mind is a computing system. Support for this claim comes from the success of artificial computing systems at performing certain learning, reasoning, and language processing tasks as well as the empirical and explanatory fruitfulness of computational modeling in psychology and neuroscience. The idea that reasoning is a kind of computation predates groundbreaking results in computability theory, the invention of computers, and the emergence of cognitive science in the 20th century. However, these advances played a decisive role in developing that seminal idea into a full-fledged CTM. In fact, there is no single CTM; there is a cluster of positions sharing the insight that the mind is somehow computational but disagreeing on such basic questions such as how computational models of mental capacities relate to physical descriptions of concrete systems or what kinds of computations best explain human mental capacities. Although CTM is currently the main paradigm in cognitive science, objections to CTM include the worry that it is explanatorily empty, the contention that CTM is a fundamentally misguided approach to understanding the mental capacities of organisms, and the alleged impossibility of explaining human common sense, creativity, understanding, and consciousness in computational terms.

History

The prehistory of CTM goes back to the 17th century, when Thomas Hobbes speculated that thinking can be analyzed into simple operations that a suitably organized physical mechanism could perform (Isaac, 2018). Two centuries later, Alan Turing's foundational work in computability theory extended this speculation into a formal account of an idealized computing system known as a *Turing machine* (Turing, 1936). Warren McCulloch and Walter Pitts (1943) leveraged Turing's mathematical notion of computation to hypothesize that networks of neurons in the brain implement computations constitutive of mental activities (on computing in neural networks, see also Turing, 1948/2004).

The first programmable, electronic, general-purpose digital computer was completed in 1945. After this invention, the hypothesis that the mind is a computing system—or, roughly, the software of the brain—propelled the emerging field of *cognitive science* in the 1950s (Boden, 2008) and began shaping philosophical debates about the mind–body problem (Putnam, 1960).

Since the 1970s, CTM has been variously formulated—most notably, in terms of rule-governed operations defined over discrete symbolic representations in a language of thought (Fodor, 1975) or the dynamics of networks of simple processing units, which can gradually learn and adapt from experience and compute nonsymbolic representations implemented as distributed patterns of activation, similarly to neural connections (Bechtel & Abrahamsen, 1991).

Today, CTM remains one of the most influential, productive, and contentious approaches to understanding mind, brain, and behavior.

Core concepts

CTM claims that the mind is a *computing* system. The most popular interpretation of this claim is in terms of *information processing*. However, merely appealing to information processing without explaining what *information* is and what kind of *processes* constitute computations is too generic and does not clarify the commitments and scope of CTM.

According to a second, *classical* version of CTM, there is "no computation without representation" (Fodor, 1975, p. 34). The key idea is that the computations constituting human mental capacities must be defined over a language-like system of symbolic mental representations, which, similarly to words in a natural language, can be (re)combined according to their logical structure to form complex thoughts [see Mental Representation; The Language of Thought Hypothesis]. One problem here is biological implausibility because neural computations do not seem to recruit language-like symbolic representations (Churchland & Sejnowski, 1990).

A third, *connectionist* version of CTM aims at more biological plausibility. Referring to artificial neural networks, which are idealized models of neural circuits (Rumelhart et al., 1986), this version claims that the mind is akin to a network of simple units (the analogs of neurons) linked together by excitatory and inhibiting connections (the analogs of synapses) changing in accordance with a learning algorithm [see Recurrent Neural Networks]. One traditional challenge here is that connectionism does not clearly offer an adequate account of such human mental capacities as logical reasoning, which seem to be better explained by computations defined over logically structured symbolic representations (Fodor & Pylyshyn, 1988). Proponents of connectionist CTM have suggested various ways in which connectionist networks can deal with these tasks (Smolensky, 1990), and there is now a large amount of literature probing symbolic and relational reasoning in modern neural networks (Geiger et al., 2023; Santoro et al., 2016).

Although there are other models of computation and other versions of CTM (Colombo & Piccinini, 2023), recent developments combine insights from classical and connectionist versions of CTM with evidence from neuroscience to better understand brain computation (Doerig et al., 2023).

Questions, controversies, and new developments

Some basic questions for proponents of CTM are how to determine whether a physical system computes and, if it does, what functions it computes and how. In addressing these questions, one challenge that has recently received sustained attention is how to avoid trivializing CTM by implying that any physical system, including walls and buckets of water, can accurately be described as implementing computations (Sprevak, 2018).

There are several new accounts of computation in physical systems that try to meet this *triviality challenge*. For example, according to *semantic* accounts, the computational states of a physical system are individuated by what they mean (by their representational content; Shagrir, 2022). This entails that only physical systems that manipulate representations can compute. So, if there are no neural representations, then the brain cannot compute; similarly, if there are no mental representations, then CTM cannot be true. Instead, according to *mechanistic* accounts, physical computing systems are multilevel mechanisms whose components are functionally organized to manipulate physical properties of the system in accordance with a step-by-step, algorithmic procedure (Piccinini, 2020). Thus, if the brain computes, then one of its biological functions is to manipulate certain properties of neural action potentials, such as their rate or timing, according to algorithms. Because the rate and timing of a sequence of action potentials can be implemented in multiple kinds of mechanisms besides the brain, the mechanistic account allows for the same kind of computation to be implemented in multiple kinds of physical mechanisms—for example, silicon chips and neurons—if the mechanisms have the right kind of functional organization.

Another set of issues of ongoing interest concerns the scope of CTM. Does it apply to all, some, or no mental capacities? For example, some claim that CTM cannot explain creativity because computing systems can only do what the algorithms they implement instruct them to do (Lovelace, 2015). Others contend that computing systems cannot have conscious experiences because consciousness is not reducible to the functional properties of any physical system (Chalmers, 1995) [see Consciousness and AI]. Other critics argue that computing systems cannot have capacities for understanding, common sense, and skillful behavior (Dreyfus, 1992). Finally, several authors reject CTM altogether because it would be a misguided approach to understanding the dynamic, embodied, and metabolically grounded nature of the mind of organisms (Varela et al., 2017).

Broader connections

Clarifying the relationship between evolutionary, psychological, and neurophysiological levels of explanation is challenging (Colombo & Knauff, 2020). CTM can help address this challenge by offering unified analyses of computing systems bringing together mutually constraining hypotheses about the kind of problem a mental capacity is adapted to solve, the algorithmic procedures and representational structures involved in the capacity, and the causal mechanisms implementing those procedures and structures (Marr & Poggio, 1976).

CTM also bears on whether the same kind of mental state can be realized by different kinds of physical systems, including biological and nonbiological systems [see Multiple Realizability]. CTM entails this possibility, but critics have cast doubt on the empirical plausibility of the multiple realizability of mental kinds, which would, in turn, cast doubt on CTM (Cao, 2022).

The relationship between metabolism and computation is another topic of broad interest. The brain is far more energy efficient than any artificial computing system engineered so far, which raises the issues of whether biological processes that seem merely metabolic might play some computational role (Chirimuuta, 2024, chapter 4) and how computational models of human mental capacities should consider metabolic factors (Haueis & Colaço, 2025).

One final exciting connection is between CTM and illness. If neural computation explains mental capacities, then errors in neural computing (or *miscomputations*) might explain mental dysfunctions (Colombo, 2021). If neural prostheses and brain computer interfaces implementing neural computations can successfully restore such capacities as motor control, perception, and speech to various patients (Lebedev & Nicolelis, 2017), then computing technologies may well turn humans into cyborgs (Clark, 2004).

Further reading

- Colombo, M., & Piccinini, G. (2023). *The computational theory of mind*. Cambridge University Press.
- Haugeland, J., Craver, C. F., & Klein, C. (Eds.). (2023). *Mind design III: Philosophy, psychology, and artificial intelligence*. MIT press.
- Miłkowski, M. (2013). The computational theory of mind. In J. Fieser & B. Dowden (Eds.), *Internet encyclopedia of philosophy*. https://iep.utm.edu/computational-theory-of-mind/

References

- Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parallel processing in networks*. Basil Blackwell.

  ↩
- Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press.

  ↩
- Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese*, *200*(6), 506. https://doi.org/10.1007/s11229-022-03524-1

  ↩
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200-219.

  ↩
- Chirimuuta, M. (2024). *The brain abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press.

  ↩
- Churchland, P. S., & Sejnowski, T. J. (1990). Neural representation and neural computation. *Philosophical Perspectives*, *4*, 343-382. https://doi.org/10.2307/2214198

  ↩

- Clark, A. (2004). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

  ↩
- Colombo, M. (2021). (Mis)computation in computational psychiatry. In F. Calzavarini & M. Viola (Eds.), *Neural mechanisms. New challenges in the philosophy of neuroscience* (pp. 427-448). Springer.

  ↩
- Colombo, M., & Knauff, M. (2020). Editors' review and introduction: Levels of explanation in cognitive science: From cultures to molecules. *Topics in Cognitive Science*, *12*(4), 1224-1420. https://doi.org/10.1111/tops.12503

  ↩
- Colombo, M., & Piccinini, G. (2023). *The computational theory of mind*. Cambridge University Press.

  ↩
- Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431-450. https://doi.org/10.1038/s41583-023-00705-w

  ↩
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.

  ↩
- Fodor, J. A. (1975). *The language of thought*. Thomas Y. Crowell.

  ↩
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

  ↩
- Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2023). Relational reasoning and generalization using nonsymbolic neural networks. *Psychological Review*, *130*(2), 308-333. https://doi.org/10.1037/rev0000371

  ↩
- Haueis, P., & Colaço, D. J. (2025). Metabolic considerations for cognitive modeling. *Behavioral and Brain Sciences*. Advance online publication. https://doi.org/10.1017/S0140525X25103956

  ↩
- Isaac, A. M. C. (2018). Computational thought from Descartes to Lovelace. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 9-22). Routledge.

  ↩

- Lebedev, M. A., & Nicolelis, M. A. L. (2017). Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation. *Physiological Reviews*, *97*(2), 767-837. https://doi.org/10.1152/physrev.00027.2016

  ↩
- Lovelace, A. A. (2015). 1842 notes to the translation of the sketch of the analytical engine. *Ada User Journal*, *36*(3), 152-180.

  ↩
- Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry* (Report No. AIM-357). MIT Artificial Intelligence Library. http://hdl.handle.net/1721.1/5782

  ↩
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115-133. https://doi.org/10.1007/BF02478259

  ↩
- Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.

  ↩
- Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind* (pp. 57-80). New York University Press.

  ↩
- Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. MIT Press.

  ↩
- Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. *ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning*, *48*, 1842-1850. https://dl.acm.org/doi/10.5555/3045390.3045585

  ↩
- Shagrir, O. (2022). *The nature of physical computation*. Oxford University Press.

  ↩
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1-2), 159-216. https://doi.org/10.1016/0004-3702(90)90007-M

  ↩

- Sprevak, M. (2018). Triviality arguments about computational implementation. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 175-191). Routledge.

  ↩
- Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*(1), 230–265. https://doi.org/10.1112/plms/s2-42.1.230

  ↩
- Turing, A. M. (2004). Intelligent machinery. In B. J. Copeland (Ed.), *The essential Turing: The ideas that gave birth to the computer age* (pp. 410–432). Oxford University Press. (Original work published 1948)

  ↩
- Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind, revised edition: Cognitive science and human experience*. MIT Press.

  ↩