International Journal of Combinatorial Optimization Problems and Informatics, 17(1), Jan-April 2026, 90-100. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i1.1245

www.editada.org

The Computational Theory of Mind: Ethical and Philosophical Implications in the Age of Artificial Intelligence

Jorge A. Ruiz-Vanoye, Ocotlán Díaz-Parra, Marco A. Vera-Jiménez, Francisco R. Trejo-Macotela

[1] Universidad Politécnica de Pachuca. Carretera Pachuca - Cd. Sahagún km 20, Ex-Hacienda de Santa Bárbara, Zempoala Hidalgo, México. C.P. 43830. Tel: 771 211 1147 & 771 5477 510
* Corresponding author: ocotlan@diazparra.net

Abstract. The Computational Theory of Mind (CTM) posits that the human mind operates in ways analogous to a computer, processing information through symbolic representations and formal rules. With the advent of Artificial Intelligence (AI) and Large Language Models (LLMs), the scope and implications of CTM have broadened considerably. This paper examines the ethical and philosophical dilemmas associated with research on and applications of CTM, with particular attention to human enhancement, privacy, consent, moral responsibility, and free will. It also considers how CTM intersects with longstanding philosophical debates on consciousness and personal identity, while addressing challenges raised by alternative perspectives, such as embodied cognition. Rather than advancing a prescriptive stance, the paper argues for a balanced approach that seeks to leverage technological developments while safeguarding human values and identity in an increasingly AI-driven context.

Keywords: Computational Theory of Mind, Large Language Models.

Article Info
Received Jun 26, 2025
Accepted Oct 11, 2025

1 Introduction

Computational Theory of Mind (CTM) holds that the human mind is, in essence, a computer-like information-processing system. Within this framework, mental processes are typically understood as the manipulation of symbols in accordance with formal rules, in a manner analogous to how a computer processes data. Jerry Fodor (1975) advanced the view that the human mind operates in a computer-like fashion by relying on an internal representational system referred to as the language of thought, or Mentalese. This language is hypothetical and is generally assumed to possess a determinate syntax and semantics, which would allow thoughts to be formally structured, manipulated, and processed. On this basis, the core components of CTM are:

- Mental Representations. Symbolically encoded thoughts, beliefs, desires, perceptions etc.
- Algorithms and Manipulation Rules. Formal procedures and rules for processing mental representations.
- Memory System. Short-term and long-term memory for storing representations and rules.
- Input and Output Mechanisms. Sensory perception (sight, hearing, touch, etc.) and generation of actions or responses (movements, language).
- Executive Control. Coordination and regulation of mental activity; decision making on what algorithms to apply and how to manage representations.

The table 1 contains the comparative between CTP and the computers.

Table 1. CTM versus Computers.

Component CTM	Computer
Mental Representations	Data stored in bits, files, and documents in the file system.
Algorithms and Manipulation Rules	Software, programs, and algorithms that process data.
Memory System	RAM (random access memory) and hard disk/SSD storage.
Input and Output Mechanisms	Input devices (keyboard, mouse, sensors) and output devices (monitor, printer, speakers).
Executive Control	CPU (central processing unit) that executes and controls program operations.

Turing (1936) introduced the concept of abstract models of computation that manipulate symbols on an infinite tape in accordance with a set of formal rules. Building on this framework, the mind is often conceived as a system that processes symbols, in which thoughts function as internal symbolic representations and mental activity involves the manipulation of these symbols.

Chomsky (1957) argued that mental processes follow formal rules or algorithms comparable to computer programs. This principle derives from his work in linguistics, where computational ideas were applied to the analysis of human language through formal grammatical rules and syntactic structures. Within CTM, this position is commonly regarded as foundational, insofar as it maintains that the human mind operates in a manner analogous to a computer by manipulating symbols and processing information.

Fodor (1975) further developed the view that the mind represents information from the external world in an internal format that can be manipulated computationally. From this perspective, CTM maintains that human thought exhibits a language-like structure endowed with syntax and semantics, which would permit mental processes to operate in computational terms.

Putnam (1967) was a leading proponent of functionalism, the view that mental states are defined by their functional or causal roles within a cognitive system rather than by their physical composition. Within CTM, functionalism is widely treated as a central tenet, as it suggests that the mind could, in principle, be realised in any system capable of performing the relevant functions, rather than exclusively in the biological brain.

In this paper, we examine the foundational concepts of the Computational Theory of Mind (CTM) and consider its ethical and philosophical ramifications in the context of contemporary developments in artificial intelligence. Specifically, we analyse how CTM interprets the mind as a symbol-processing system and discuss the potential implications of this perspective for understanding human cognition, ethics, and the development of artificial intelligence.

2 Artificial Intelligence and Neuroscience

Artificial Intelligence (AI) is commonly defined as the capacity of a machine to mimic aspects of human intelligence, including learning, reasoning, problem solving, perception, and natural language understanding. John McCarthy (1956) described AI as the science and engineering of making intelligent machines, particularly intelligent computer programs. This definition is generally regarded as having laid the foundation for the field, establishing a focus on developing systems capable of performing tasks that would normally require human intelligence.

Neuroscience is the scientific study of the nervous system, encompassing its structure, function, development, genetics, biochemistry, physiology, pharmacology, informatics, and pathology. Santiago Ramón y Cajal (1894), widely recognised as the father of modern neuroscience, established the neuron doctrine, which postulated that neurons constitute the fundamental units of the brain and nervous system. His investigations into neural structure were instrumental in the emergence of neuroscience as a distinct scientific discipline.

The relationship between AI and neuroscience is often considered significant for several reasons. Neuroscience provides biological models that inspire the design of AI algorithms, such as artificial neural networks modelled on aspects of brain structure and function. Conversely, AI techniques can be applied to the analysis of large-scale neuroscientific datasets, potentially supporting the identification of patterns in brain activity and contributing to advances in the diagnosis and treatment of neurological disorders. Collaboration between these fields has also been associated with the development of brain–computer interfaces (BCIs), which enable direct communication between neural activity and external devices, thereby enhancing the quality of life for individuals with disabilities and advancing the study of mental and cognitive processes.

Some advances in artificial intelligence applied to neuroscience include:

- Neuroscientific Data Analysis. AI is widely used to analyse large volumes of neuroscientific data, including brain images and sequences of neural activity. Convolutional neural networks (CNNs) have been applied to functional magnetic resonance imaging (fMRI) in order to identify patterns of brain activity associated with different mental states and neurological conditions, such as Alzheimer's disease and epilepsy.
- Brain Modelling. AI contributes to the development of computational models of the brain that aim to replicate aspects of its structure and function for the study of neural behaviour. Projects such as the Human Brain Project employ AI-based simulations to support investigations into brain dynamics and the development of therapeutic approaches for mental disorders.
- AI-enhanced Brain–Computer Interfaces (BCIs). BCIs use AI techniques to improve the interpretation of neural signals and convert them into commands for external devices, including robotic prostheses.
- Drug Discovery. AI is increasingly applied to accelerate the drug discovery process by identifying candidate compounds and predicting their potential effects on brain function.
- Neuroimaging and Automated Diagnostics. AI is employed to enhance diagnostic accuracy in neuroimaging by automating the detection of abnormalities, tumours, lesions, and other pathological features.

Table 2 presents a summary of notable studies on the various applications of AI in neuroscience. Each row of the table shows the specific AI technology used, the main research contributions and the impact on the field.

Table 2. Related work on AI applied to neuroscience.

Research	AI	Contribution
Esteva et al. (2017)	Convolutional Neural Networks (CNNs)	Diagnosis of dermatological diseases at specialist level by means of image analysis.
Van Essen et al. (2013)	Functional Magnetic Resonance Imaging (fMRI) + IA	Detailed mapping of neural connections in the human brain.
Deisseroth (2011)	Optogenetics + AI modelling	Precise control of neuronal activity and progress in the study of specific neuronal circuits.
Lebedev & Nicolelis (2006)	Brain-Computer Interfaces (BCIs) + AI	Improvements in the quality of life of people with disabilities through robotic prosthesis control.
Markram (2012)	Computational Brain Modelling + AI	Detailed brain simulations to understand neural dynamics and develop new therapies.
Ekins & Puhl (2013)	Deep Learning in Pharmaceutical Research	Accelerating drug discovery by predicting chemical interactions.
Litjens et al. (2017)	Deep Learning in Medical Image Analysis	Automated and accurate diagnosis of pathologies in medical imaging, improving clinical efficiency.
Goh et al. (2017)	Deep Learning in Computational Chemistry	Identification of new chemical compounds for the treatment of neurological diseases.
Jang & LeBel (2023)	Semantic Reconstruction + AI Modelling	Accurate reconstruction of continuous speech from non-invasive brain recordings.
Xu et al. (2023)	BCIs + AI Algorithms	Improvements in the accuracy of BCIs and neural system rehabilitation techniques.

Metzger et al. (2023)	Generative Models (GANs and Diffusion Models)	Decoding of brain signals and generation of high-resolution brain images.
Yadollahpour et al. (2018)	Medical Decision Support Systems + AI	Prediction of chronic disease progression using decision support systems.
Schneider et al. (2023)	Latent Embeddings + Joint Analysis	Joint analysis of behaviour and neural activity using latent embeddings.
Zhu (2020)	Big Data + AI Modelling	Accelerating drug discovery through big data and AI-based modelling.
Wu Tsai Neurosciences Institute (2023)	Various AI Models	Advances in human health through the integration of neuroscience and AI.
Asgher et al. (2023)	BCIs + Human-Machine Interaction	Improvements in human-machine interaction and industrial applications through AI-powered BCIs.

LLMs are artificial intelligence systems that have been trained on extensive textual corpora to understand and generate human language in a coherent and contextually appropriate manner. These models rely on deep neural network architectures, such as transformers, to process and produce text. Prominent examples include OpenAI’s GPT-3 and GPT-4, as well as Google’s BERT and T5 models.

- Text generation: the ability to produce coherent and contextually relevant text from a given input.
- Natural language understanding: the ability to interpret and respond to questions, summarise texts, and perform machine translation.
- Conversational interaction: the ability to sustain natural and contextually appropriate interactions with human users.
- Sentiment analysis: the ability to identify and interpret emotions and opinions expressed in textual data.
- Personalisation and adaptation: the capacity to adjust style and tone in response to contextual and audience-related factors.

Given that the Computational Theory of Mind (CTM) holds that the human mind is, in essence, a computer-like information-processing system, mental processes are often characterised as the manipulation of symbols according to formal rules, in a manner analogous to computational operations. From this perspective, the potential contributions of LLMs to CTM can be described as follows:

- LLMs may be seen to simulate aspects of human thought processes by manipulating symbols and generating coherent linguistic output, which aligns with the view that cognition operates through formal rules and algorithms.
- LLMs rely on internal representations, such as word vectors and embeddings, to process and generate language, a feature that parallels the CTM claim that the mind internally represents information derived from the external world.
- The capacity of LLMs to generate adaptive and context-sensitive responses can be interpreted as reflecting the functionalist principle that mental states are defined by their roles within a cognitive system.

Large Language Models (LLMs) have increasingly been recognised as influential tools in research related to the Computational Theory of Mind (CTM). Table 3 summarises selected studies and current applications that illustrate the use of LLMs within CTM-related contexts.

Table 3. Relevant studies and current applications of LLMs in CTM.

Research	AI Technique	Contribution
Brown et al. (2020)	GPT-3 and GPT-4	Evaluation of GPT-3 and GPT-4 performance on complex theory of mind tasks, demonstrating similar abilities to those of a six-year-old child in false belief scenarios.

Kosinski (2024)	Evaluation of Language Models in Tasks	Use of a customised battery of false belief tasks to assess eleven LLMs, showing that models such as GPT-4 can make inferences about human mental states.
Sap et al. (2022)	Prompting and Assessment	Designing specific prompts to improve the comprehension and performance of LLMs in theory of mind tasks.
Chung et al. (2024)	Flan-PaLM and Instructional Models	Implementation of instructional models such as Flan-PaLM to improve the alignment of LLMs' responses with human preferences through reinforcement learning.
Michal Kosinski (2024)	Evaluation of LLMs in False Belief Scenarios	Discovery that LLMs can solve up to 75% of false belief tasks, suggesting a possible spontaneous emergence of theory of mind in these models.
Wang et al. (2023)	MoToMQA: Multi-Order Theory of Mind Q&A	Evaluation of LLMs in higher-order theory-of-mind tasks, showing that GPT-4 and Flan-PaLM achieve performance levels similar to those of adult humans.
Park et al. (2023)	Multi-agent Collaboration via LLMs	Use of LLMs to coordinate multi-agent collaboration, highlighting the importance of communication and belief state maintenance for effective cooperation.
Dasgupta et al. (2023)	Computational Psychiatry Perspective	Application of LLMs in computational psychiatry to model and predict human behaviour in clinical settings.

3 Ethical and Philosophical Implications

The extensive integration of technology into the human body raises concerns about the risk of dehumanisation, whereby aspects of human identity may be compromised. This concern stems from the possibility that technological enhancements could alter biological characteristics to such an extent that experiences and relationships commonly regarded as authentically human are significantly transformed. Increased dependence on technology may contribute to forms of emotional detachment, potentially affecting the quality of interpersonal relationships and lived emotional experience. There is also a risk of alienation, insofar as individuals may come to experience a sense of disconnection from their own humanity, perceiving themselves as partially artificial and distanced from what is understood as the human condition. The application of artificial intelligence (AI), particularly through Large Language Models (LLMs) and the Computational Theory of Mind (CTM), adds further complexity to these concerns. By simulating aspects of mental processes and generating highly personalised interactions, LLMs may encourage preferences for engagement with artificial systems over human relationships, thereby intensifying patterns of emotional disconnection and alienation. At the same time, CTM, which conceptualises the mind as an information-processing system, can influence how individuals interpret their own identity and consciousness.

The prospect of enhancing physical and cognitive capacities through advanced technologies is a field that has received increasing attention. Commonly referred to as human enhancement, this domain encompasses a broad range of interventions, including advanced prosthetics, neurological implants, and cognitive augmentation mediated by AI and LLMs. Such technologies are often presented as offering opportunities not only to address biological limitations, but also to extend the range of human capabilities. Among the most widely discussed applications is the use of technology to mitigate physical disabilities. Mind-controlled prostheses, exoskeletons designed to support mobility in cases of paralysis, and artificial vision systems for individuals with visual impairments illustrate how technological interventions can transform everyday functioning. The integration of AI and LLMs into these systems may further enhance their efficiency and responsiveness, enabling more adaptive and personalised forms of interaction. These developments are frequently cited as demonstrating the potential of technology to expand human capacities beyond those afforded by biology alone.

The integration of CTM with advanced AI technologies and LLMs also raises concerns regarding mental manipulation and behavioural influence. Such technologies have the potential to shape thoughts and behaviours in ways that may be subtle or explicit, thereby creating risks of misuse or excessive control. The possibility of influencing cognition and behaviour through technological means generates substantial ethical challenges, particularly where individuals' decisions may be affected without explicit awareness or consent. This issue is especially salient in relation to neurotechnologies, including brain–computer interfaces (BCIs) and neural implants, which are designed to interact directly with neural activity and, in some cases, to modify it. The collection and use of neural and behavioural data raise further ethical and philosophical issues, particularly with respect to privacy and informed consent. High standards of transparency are required to ensure that individuals understand how their data are used and retain meaningful control over access and application. There is also a risk that such data could be misused or inadequately protected, potentially compromising mental privacy and creating vulnerabilities that affect autonomy and personal well-being.

Debates concerning determinism and free will in the context of CTM and AI raise fundamental questions about human agency. If the human mind is conceptualised as a computational system governed by algorithmic rules, it becomes necessary to ask whether space remains for free will. Deterministic interpretations suggest that actions and decisions are fully determined by prior conditions, implying that choices could be understood as the outcome of preconfigured computational processes rather than autonomous deliberation. This raises an ethical dilemma: if decisions can be modelled and predicted by algorithms, to what extent can individuals be considered free agents? The deployment of advanced AI systems capable of predicting or influencing behaviour may intensify these concerns by further challenging perceptions of autonomy and moral responsibility. Addressing these questions from ethical and philosophical perspectives is therefore widely regarded as essential to ensuring that the integration of CTM and AI respects human dignity. CTM also raises questions regarding the attribution of moral responsibility for human actions. By framing the mind as an information-processing system governed by formal rules, CTM can be interpreted as suggesting that decisions and behaviours are, to some degree, predetermined. This perspective challenges traditional accounts of moral responsibility, particularly in legal and ethical contexts that rely on assumptions of individual agency. The increasing use of AI systems that predict or influence behaviour further complicates this issue, as it may contribute to a diminished sense of personal autonomy. If decisions can be systematically modelled and anticipated by algorithms, the basis for attributing responsibility becomes less clear. Developing ethical and legal frameworks capable of addressing these challenges is therefore often seen as necessary to preserve human agency while acknowledging the role of computational processes in decision-making.

Finally, CTM provides a framework for approaching consciousness as an outcome of computational processes. By conceptualising the mind as an information-processing system, CTM proposes that consciousness emerges from complex interactions among computational operations. This approach offers one way of engaging with longstanding philosophical debates concerning the nature of consciousness. From a CTM perspective, conscious states are understood as internal symbolic representations governed by formal rules, analogous to computational processes in artificial systems. This view allows for the exploration of subjective experience in computational terms, contributing to scientific and philosophical discussions of consciousness. In doing so, CTM not only engages with traditional debates on the relationship between mind and brain, but also opens avenues for interdisciplinary research involving philosophy, neuroscience, and artificial intelligence.

There are other challenges to Computational Theory from Alternative Philosophical Perspectives:

- Embodied cognition (Lakoff & Johnson, 1999; Clark, 1997). The embodied cognition perspective argues that the mind cannot be fully understood without reference to the body and its interactions with the environment. In contrast to Computational Theory of Mind (CTM), which conceptualises the mind as an abstract information-processing system, embodied cognition maintains that mental processes are intrinsically connected to bodily experience and environmental context. From this viewpoint, cognition does not occur in isolation but is deeply shaped by sensory perception, physical action, and continuous engagement with the external world. This position implies that theories of mind that rely exclusively on computational explanations may be incomplete, insofar as they fail to integrate the bodily dimensions of human cognition.
- Phenomenology (Husserl, 1931; Merleau-Ponty, 1962). Phenomenology places emphasis on direct subjective experience and raises questions about whether computational accounts can adequately capture the full scope of consciousness. From this perspective, conscious experience cannot be reduced to computational operations alone, as it encompasses subjective and phenomenological qualities that are intrinsically personal. Phenomenologists argue that CTM, by focusing primarily on symbolic manipulation and data processing, overlooks essential aspects of lived experience, including emotion, temporality, and intentionality. As a result, this approach casts doubt on the capacity of CTM to address dimensions of the human mind that resist quantification or algorithmic replication.
- Information theories (Shannon, 1948; Bateson, 1972). Information-theoretic approaches examine whether the information processed by the human mind can be equated with the data handled by computational systems. While both minds and machines process information, proponents of this view suggest that the nature and meaning of such information differ in significant ways. Critics contend that CTM oversimplifies mental activity by treating it as purely computational, without sufficient attention to how information is interpreted and contextualised in human cognition. Moreover, these theories emphasise that mental information is shaped by factors such as memory, emotion, and perception, which do not have direct counterparts within standard computational architectures.

4 Network or Grid Theory of Mind Consciousness

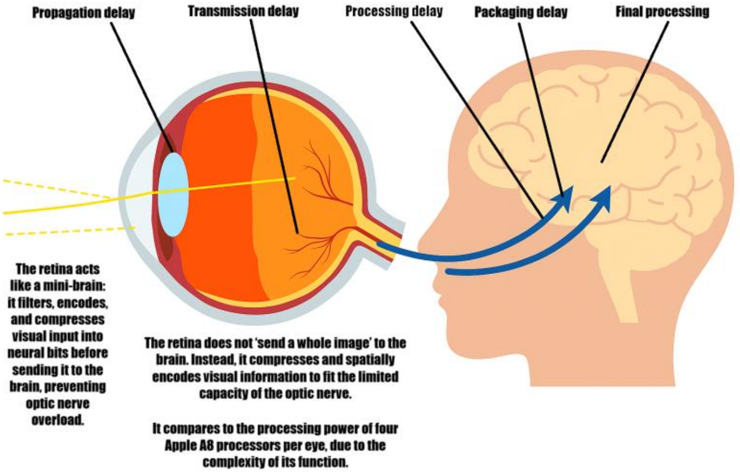
A conscious visual system is unlikely to be optimised for immediate reactions, as its processing is relatively slow, with integration windows of up to 400 ms (Fleming and Michel, 2025). Instead, its evolutionary role may be to support offline cognition, such as model-based planning. This capacity may have emerged during the transition from water to land, when sensory horizons expanded. The authors associate consciousness with the function of reality monitoring, that is, distinguishing between real external signals and internal representations.

To illustrate this idea, the brain can be compared to a computer network, while visual awareness can be compared to an application that requires incoming data to function (a video call) (Ruiz-Vanoye, Sossa-Azuela, Díaz-Parra & Trejo-Macotela, n.d.). As in a networked system, before data reaches an application, it passes through multiple stages. Each stage introduces its own latency. When an end system needs to transmit a packet over a shared medium (such as Wi-Fi or a cable modem), it may intentionally delay transmission as part of a protocol that reduces collisions or interference with other devices using the same medium. The following provides a step-by-step comparison:

Table 4. Network or Grid Theory of Mind Consciousness

In a network	In your brain (conscious vision)	What does it mean?	Functional consequence
Propagation delay: the time (≈ 1 ms) it takes for the signal to travel along the wire.	The time it takes for the visual signal (the light from the object) to get from the eye to the brain. The retina acts like the NIC of the computer: it translates physical signals (photons) into neural bits.	The signal journey. Consciousness does not occur in the eye. That information must still travel through the optic nerve.	Photons arrive at retina and action potentials travel ~20 cm of optic nerve. Physical light pathway + nerve conduction. (retina → optic nerve → thalamus).
Transmission delay: the time (10–20 ms) it takes to send all the bits.	The time it takes for the visual signal (light converted to neural impulses) to travel from the retina through the optic nerve to the brain. Like signal propagation along a cable.	Convert the signal into data. It means that your brain needs time to transfer the visual message from the eye to the processing areas. Even before understanding what you see, the signal must be fully delivered—just like all bits must arrive before data can be used.	The eye converts light into electrical impulses, and these impulses are sent as data to the brain. It's like when a computer starts sending packets over the network. You still don't know what you're seeing.
Processing delay: the time (50-150 ms) it takes a device to examine the packet and decide where to send it.	The time it takes for the brain to process the image step by step (shape, colour, movement).	Decoding the data. The cortex extracts edges, shapes, colour, depth. This costs 50-150 ms. If this delay increases (e.g. cognitive load), the 'application latency' goes up.	The brain starts to work: it analyses the colours, shapes, movements. It's as if your computer receives a file and is decompressing it to understand it.

			Here you almost know what you are looking at, but you don't consciously realise it yet.
Packaging delay: waiting time (80–120 ms) for a packet to be filled before sending it (as in a video call).	The brain needs to accumulate enough information before something “jumps” to consciousness.	Wait for enough information. Consciousness is not immediate, it needs to gather a minimum amount of information first.	The brain needs to gather a bit more information to be sure. It's like when an app waits until it has enough content to start displaying it (like when YouTube loads). Only when there is enough ‘data’ together does awareness kick in.
Final processing (App): when the application finally receives the data and displays it.	The moment you become aware of what you saw (e.g. ‘an apple!’).	Conscious access / conscious perception. You are only now aware of what happened, but some time has passed since it happened.	Finally, the brain “shows it on the screen”. Now you are conscious. But all that took time: maybe 300 ms after it started.



The retina acts like a mini-brain: it filters, encodes, and compresses visual input into neural bits before sending it to the brain, preventing optic nerve overload.

The retina does not 'send a whole image' to the brain. Instead, it compresses and spatially encodes visual information to fit the limited capacity of the optic nerve.

It compares to the processing power of four Apple A8 processors per eye, due to the complexity of its function.

Fig. 1. Visual consciousness works like a time-delayed network.

Visual consciousness can be described as operating in a time-delayed manner, analogous to a networked system. Because conscious perception depends on physical processes that unfold over time, seeing something consciously is not instantaneous, even if this delay is not subjectively apparent. In a similar way to how network performance is assessed through latency, the temporal dynamics of consciousness constrain the processes in which it can participate. Conscious awareness is generally understood to lag slightly behind ongoing events, comparable to the delay experienced during a video call. When events unfold very rapidly—for example, when an object is suddenly thrown and an evasive movement occurs—the response may be initiated without conscious awareness. Only when awareness emerges sufficiently quickly (within approximately 100 milliseconds) can it meaningfully influence ongoing action.

Two related concepts are central to this account. First, visual latency (τ_v) refers to the interval between the moment a visual stimulus reaches the retina and the point at which the resulting information is processed and utilised by motor or cognitive systems. This latency reflects both sensory transmission delays and early stages of cortical processing. Second, expectation weight (β) denotes the extent to which the brain relies on internal predictions, or generative models, when interpreting sensory input. This parameter plays a key role in Bayesian accounts of perception, including predictive inference frameworks. These concepts are particularly relevant in clinical contexts. In individuals with psychosis, a marked increase in both visual latency (τ_v) and expectation weight (β) has been observed, suggesting an increased reliance on internal models relative to incoming sensory evidence (Adams et al., 2013; Hong et al., 2005). By contrast, individuals with autism spectrum disorder (ASD) tend to exhibit reduced reliance on expectancy and shorter visual delays, which aligns with a stronger orientation towards immediate sensory input and diminished dependence on predictive mechanisms (Pellicano & Burr, 2012).

5 Conclusions

Computational Theory of Mind (CTM) has been widely regarded as a robust theoretical framework for examining human mental processes from a computational perspective. Through its integration with advanced technologies, such as artificial intelligence (AI) and large language models (LLMs), CTM is often understood to offer not only structured accounts of cognitive functioning, but also new possibilities for addressing neurological conditions and supporting specific cognitive capacities. For these reasons, continued theoretical and technological development of CTM is frequently considered important for several purposes:

- CTM enables the modelling and simulation of complex mental processes, which may contribute to improved understanding of neurological disorders such as Alzheimer's disease, Parkinson's disease, and other cognitive impairments. The application of AI and LLMs in this context has the potential to support pattern identification and the development of more precise intervention strategies.
- The integration of brain–computer interfaces (BCIs) with AI technologies may facilitate the development of novel therapeutic approaches for neurological conditions. BCIs and neural implants can draw on computational accounts of cognition to enhance assistive technologies and improve quality of life for individuals with disabilities.
- CTM provides a theoretical basis for technological cognitive enhancement, including the development of cognitive prostheses that could extend certain mental capacities and help mitigate physical or cognitive limitations.

In conclusion, the continued theoretical refinement and technological development of Computational Theory of Mind can be seen as playing an important role in advancing the understanding and treatment of neurological conditions, as well as in supporting specific aspects of human cognition. By leveraging the capabilities of AI and LLMs, while engaging with ethical and philosophical considerations, it may be possible to move towards forms of technological integration that contribute to human well-being.

Possible Future Work in the Integration of AI and the Computational Theory of Mind (CTM)

- Research on how to integrate CTM with advanced AI systems to improve the understanding of human mental processes and their replication in machines, robots or metaverse elements.
- Creating detailed simulations of human cognitive processes using LLMs and other AI algorithms to study the mind from a computational perspective.
- Enhance the development of BCIs using AI to improve communication between the human brain and machines, enabling the control of external devices by thought.
- Implementation of AI assistants that provide real-time cognitive support, improving memory, learning and decision-making.
- Use of LLMs or similar to provide advanced cognitive support, such as idea generation, thought organisation and complex problem solving.
- Development of communication systems based on LLMs or similar to enable more natural and fluid interactions between humans and machines.

- Establishment of ethical and legal frameworks to guide the development and implementation of human-machine integration technologies.
- Establishment of committees and procedures to assess the ethical impacts of new technologies prior to their implementation.
- Design of devices that enhance human cognitive capabilities, such as memory and information processing.
- Integration of advanced technologies to overcome physical and mental limitations, improving quality of life and human capabilities.
- Development of protocols and technologies to ensure privacy and security of brain data collected and processed by AI systems.
- Creation of consent systems that allow users to adjust their data permissions in real time according to their preferences and needs.
- Educational programmes that inform users about the risks and benefits of integration with advanced technologies in their body.
- Research on how to assign moral responsibility in situations where human decisions are influenced by AI algorithms.
- Development of new legal and ethical frameworks that clearly define responsibility in the use of advanced technologies.
- Study of how AI technologies affect the autonomy and decision-making capacity of individuals.
- Analysis of the compatibility between free will and the influence of computational systems on human behaviour.
- Research into the creation of models that can replicate or simulate human consciousness.
- Use of AI to explore and map different states of consciousness and their relationship with computational processes.
- Study of how the integration of advanced technologies affects the perception of personal identity and continuity of self.
- Development of technologies that help people maintain a coherent and authentic identity in a highly technological environment.
- Projects exploring how technologies can support embodied cognition, ensuring that physical and sensory experiences remain integrated into cognitive processing.
- Development of sensorimotor interfaces that enhance the interaction of the body with the environment, using advanced technologies to improve perceptual and motor skills.
- Creation of ethical guidelines and recommendations for the development and use of neurocognitive technologies.
- Study of the ethical implications of the integration of AI and neural technologies, with an emphasis on protecting human dignity and human rights.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). *The computational anatomy of psychosis: A predictive coding account*. *Frontiers in Psychiatry, 4*, 47.

Asgher, U., Qamar, A., & Rizwan, M. (2023). Advances in artificial intelligence (AI) in brain computer interface (BCI) and Industry 4.0 for human machine interaction (HMI). *Frontiers in Human Neuroscience, 17*, 583256. https://doi.org/10.3389/fnhum.2023.583256

Bateson, G. (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University of Chicago Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. https://arxiv.org/abs/2005.14165

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Raffel, C. (2024). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2207.10551*. https://arxiv.org/abs/2207.10551

Clark, A. (1997). Being There: Putting Brain, Body, and World Together Again. MIT Press.

Dasgupta, I., Wang, J. X., & Botvinick, M. (2023). Computational psychiatry and the theory of mind: Applications of large language models. *Nature Neuroscience*, 26, 1278-1286. https://doi.org/10.1038/s41593-023-01154-8

Deisseroth, K. (2011). Optogenetics. *Nature Methods*, 8(1), 26-29. https://doi.org/10.1038/nmeth.f.324

Ekins, S., & Puhl, A. C. (2013). The next era: deep learning in pharmaceutical research. *Pharmaceutical Research*, 30(8), 2121-2130. https://doi.org/10.1007/s11095-013-1040-8

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118. https://doi.org/10.1038/nature21056

Fleming, S. M., & Michel, M. (2025). *Sensory Horizons and the Functions of Conscious Vision*. *Behavioral and Brain Sciences*, 1–53. https://doi.org/10.1017/S0140525X25000068

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal of Computational Chemistry*, 38(16), 1291-1307. https://doi.org/10.1002/jcc.24764

Hong, L. E., Summerfelt, A., McMahon, R., Adami, H., Francis, G., Elliott, A., & Thaker, G. K. (2005). *Response to unexpected target changes during sustained visual tracking in schizophrenic patients*. *Experimental Brain Research, 165*(2), 125–134.

Husserl, E. (1931). *Ideas: General Introduction to Pure Phenomenology*. George Allen & Unwin.

Jang, J., & LeBel, A. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5), 858-866. https://doi.org/10.1038/s41593-023-01154-8

Kosinski, M. (2024). Evaluation of language models in tasks of theory of mind. *Nature Communications*. (Este artículo aún no está disponible en acceso abierto, por lo que se sugiere buscarlo en la revista *Nature Communications*).

Kosinski, M. (2024). Evaluation of large language models in false belief scenarios. *Nature Machine Intelligence*. (Este artículo aún no está disponible en acceso abierto, por lo que se sugiere buscarlo en la revista *Nature Machine Intelligence*).

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books.

Lebedev, M. A., & Nicolelis, M. A. L. (2006). Brain–machine interfaces: past, present and future. *Trends in Neurosciences*, 29(9), 536-546. https://doi.org/10.1016/j.tins.2006.07.004

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., ... & van der Laak, J. A. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005

Markram, H. (2012). The human brain project. *Scientific American*, 306(6), 50-55. https://doi.org/10.1038/scientificamerican0612-50

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence".

Merleau-Ponty, M. (1962). *Phenomenology of Perception*. Routledge & Kegan Paul.

Metzger, C., Galibert, S., Lu, J., Chakravarty, M. M., & Lajoie, G. (2023). Generative AI for brain imaging and brain signal decoding. *arXiv*. https://arxiv.org/abs/2403.07721

Park, J., Kim, S., Lee, K., & Choi, J. (2023). Multi-agent collaboration via large language models. *Journal of Artificial Intelligence Research*. https://arxiv.org/abs/2302.01442

Pellicano, E., & Burr, D. (2012). *When the world becomes 'too real': A Bayesian explanation of autistic perception*. *Trends in Cognitive Sciences, 16*(10), 504–510.

Putnam, H. (1967). Psychological Predicates. En W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37-48). University of Pittsburgh Press.

Ramón y Cajal, S. (1894). *The Croonian Lecture: La Fine Structure des Centres Nerveux*. Proceedings of the Royal Society of London.

Ruiz-Vanoye, J. A., Sossa-Azuela, J. H., Díaz-Parra, O., & Trejo-Macotela, F. R. (n.d.). *Quality of Consciousness Service (QoCS): Toward a Measurable Standard for Conscious Processing* [Unpublished manuscript].

Sap, M., LeBras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., & Choi, Y. (2022). Social Bias Frames: Reasoning about Social and Power Implications of Language. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://www.aclweb.org/anthology/2022.emnlp-main.681/

Schneider, S., Amvrosiadis, T., Buhmann, J., & Radtke, F. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature, 619*(7970), 321-329. https://doi.org/10.1038/s41586-023-04450-3

Shannon, C. E. (1948). *A Mathematical Theory of Communication*. Bell System Technical Journal, 27, 379-423, 623-656.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1), 230-265. https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf].

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *Neuroimage*, 80, 62-79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Wang, A., Kordi, Y., Mishra, S., Liu, P., Smith, N., & Khashabi, D. (2023). Multi-order theory of mind question answering: Machine theory of mind in a multi-agent environment. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://arxiv.org/abs/2301.10911

Wu Tsai Neurosciences Institute. (2023). Research projects link neuroscience and AI to advance human health. Stanford University. https://neuroscience.stanford.edu/research/projects

Xu, F., Ming, D., Jung, T. P., Xu, P., & Xu, M. (2023). The application of artificial intelligence in brain-computer interface and neural system rehabilitation. *Frontiers in Neuroscience*, 17, Article 1290961. https://doi.org/10.3389/fnins.2023.1290961

Yadollahpour, A., Nourozi, J., Mirbagheri, S. A., Simancas-Acevedo, E., & Trejo-Macotela, F. R. (2018). Designing and implementing an ANFIS based medical decision support system to predict chronic kidney disease progression. *Frontiers in Physiology, 9*, 1753. https://doi.org/10.3389/fphys.2018.01753

Zhu, H. (2020). Big data and artificial intelligence modeling for drug discovery. *Annual Review of Pharmacology and Toxicology, 60*, 573-589. https://doi.org/10.1146/annurev-pharmtox-010919-023301