Westcliff International Journal of Applied Research
WIJAR

**Westcliff International Journal of Applied Research**

*Vol. 10, No. 1. Spring 2026*

https://doi.org/10.47670/wuwijar20261HEQ

# Building Socially Aware Machines:

# The New Frontier of AI Theory of Mind

*Hajar El Qasemy*
Westcliff University

## Abstract

The research focus was motivated by a curiosity about what lies beyond current artificial intelligence (AI) capabilities and an interest in exploring the next frontier of AI evolution. This research is conducted as a literature review, with the purpose of bringing to light ongoing advances in theory of mind (ToM) AI and outlining future AI directions, specifically for readers who are seeking insights into what's on the horizon in the field of AI. Concretely, the literature review explores the emergence of ToM in AI, tracing its evolution from traditional AI systems towards ToM models capable of comprehending and predicting human mental states. Through a discussion of the current landscape, challenges, and future directions, this literature review clarifies how close AI is to achieving fully-fledged ToM and what's beyond the ultimate realization of ToM. This research reviews recent peer-reviewed sources: empirical and theoretical/conceptual journal articles, review papers, book chapters, conference proceedings, preprints, textbooks, and extended abstracts that were published within the last five years. The findings indicated that ToM in AI is still under development, with its current applications being either in research or experimental phases. Excellent progress was observed in areas such as emotion recognition, predictive modeling, conversational AI, multi-agent systems, simulation, and cognitive modeling. However, constructing mental models with ToM capabilities remains challenging, particularly in leveraging meta-learning to accurately represent existing intelligent entities, whether artificial or human. The conclusion showed that our world still does not fully grasp complex human thoughts, and creating AI systems that can adapt to our ever-changing minds and infer internal mental states is an ambitious milestone beyond which lies self-awareness, a drastic shift in technology that the world may not be ready for.

*Keywords:* Theory of mind, artificial intelligence, mental models, meta-learning, self-aware AI

## Introduction

Artificial intelligence (AI), specifically computer vision, discriminative models, and neural networks (NNs), attracted the public's attention through autonomous vehicle navigation, speech recognition, and complex gaming, respectively (Zhang & Gosline, 2023). With recent advances in generative AI, specifically ChatGPT OpenAI, the public was impressed and more people became familiar with the term "AI," which has been around since 1950 but was less used in daily speech among the public (Zhang & Gosline, 2023; Haenlein & Kaplan, 2019). AI is evolving at a high pace and both scientists and technology researchers are continuously pushing the limits of AI by attempting to mimic human intelligence and feed

it into machines (Bhuiyan, 2024). Yet, areas such as self-awareness and theory of mind have not been achieved so far (Nebreda et al., 2024). As of 2025, self-aware AI is hypothetical, whereas the theory of mind is under development (El Qasemy, 2025). Therefore, this review paper will focus on ToM, with the purpose of bringing to light ongoing advances in ToM AI and outlining future AI directions, particularly for readers who are seeking insights into what's on the horizon in the field of AI.

This research is structured as a narrative literature review synthesizing recent peer-reviewed sources, including empirical and theoretical/conceptual journal articles, review papers, book chapters, conference proceedings, preprints, textbooks, and extended abstracts that were published within the last five years. The literature review is organized thematically and was evaluated according to the criteria of accuracy, objectivity, and currency. The inclusion criteria restricted the literature review to studies that shared reliable evidence, maintained a neutral perspective, and reflected the most recent developments in ToM AI.

**Definition of Terms**

1. Artificial Intelligence: Also referred to as AI, it is the simulation of human intelligence in machines with the purpose of creating AI systems capable of thinking, reasoning, learning, and solving problems with or without explicit human instructions (Sheikh et al., 2023).

2. Computer Vision: A field of artificial intelligence that allows machines to detect and interpret visual information (Szeliski, 2022). Through computer vision, AI systems can successfully perform tasks such as image classification, facial recognition, and scene reconstruction (El Qasemy, 2025).

3. Discriminative Models: A type of machine learning model. It differentiates between the classes of data. For instance, in binary classification, discriminative models define whether an input belongs to class A or B, e.g., logistic regression and support vector machines (Gordon & Hernandez-Lobato, 2020).

4. Neural Networks: A set of computing systems formed by layers of interconnected nodes (Montesinos López et al., 2022). Neural networks are inspired by the human brain, specifically the neurons, and are the foundation of deep learning. Neural networks are efficient in natural language processing and speech recognition (Samek et al., 2021).

5. Generative AI: An AI system that generates content. Generative AI models learn patterns from the data they receive and utilize the knowledge to generate realistic outputs, including images, texts, audio-visual media, and codes (Oluwagbenro, 2024).

6. Meta-learning: The process by which a system learns how to adapt to new tasks or environments more efficiently, using knowledge gained from learning many related tasks (Luo et al., 2022).

7. Theory of Mind: The ability to understand that others have their own thoughts, beliefs, emotions, intentions, perspectives, etc. In AI, the theory of mind implies the creation of machines that can both model and predict human mental states, a level of social-emotional intelligence that's still being developed (Mao et al., 2024).

## Discussion

Theory of mind in artificial intelligence is a fascinating multidisciplinary concept that caught the attention of both researchers and scientists. With its multidisciplinary nature tapping into psychology, neuroscience, machine learning, and computer science, ToM is challenging researchers and scientists to convert recognition of human feelings and expressions into understandable computer models for AI agents to rely on. This type of technology is referred to as "emotion recognition technology" and it is advancing in theory, but its application is still infeasible.

**ToM in Cognitive Science**

Theory of mind is a branch of cognitive science that investigates the cognitive ability of

individuals in understanding others' mental states and affirming differences in beliefs, wants, intentions, and internal feelings or emotional behaviors (Hopcroft, 2025). ToM is a psychological skill acquired throughout childhood to allow humans to navigate social interactions and explain or predict their actions or reactions in reference to their personal experiences and drive (Rakoczy, 2022).

## ToM in AI

In artificial intelligence, ToM refers to the ability of an AI system to understand, explain, and predict others' mental states, taking into consideration their backgrounds. In this context, others include individuals and AI systems (Mao et al., 2024). Duplicating natural cognitive capabilities in computer science enables achieving human-like AI. Although ToM is still under development, theoretically, it is expected to revolutionize interactions between humans and systems in terms of naturalness and effectiveness (El Qasemy, 2025).

### *ToM Aspects*

Key aspects of ToM in AI involve beliefs, desires, intentions, and emotions (Bamicha & Drigas, 2022). Others may believe in things that are outside the AI's scope of knowledge; this could include cultural customs, religious beliefs, and more. ToM in AI would not only indicate that the AI system will have the capability of understanding that others hold different beliefs from its own but also the capability of recognizing that others have different desires, e.g., goals and drives (Tesar et al., 2020).

From a theoretical perspective, an intelligent system with ToM would be able to predict actions or reactions based on one's intentions in life. In autonomous vehicles, ToM AI could anticipate the intentions of both humans and machines in its environment (Nebreda et al., 2024). ToM AI could also understand emotions and respond accordingly; thus, a deep connection to a human's emotional state would occur in an interaction between the human and the ToM AI system (Bamicha & Drigas, 2022).

### *Traditional AI Technology and Genuine Comprehension*

Traditional AI technology, whether it interacts with humans or not, generates output according to predefined rules and patterns (Aggarwal et al., 2025). For instance, algorithmic trading (AT) is a form of traditional AI focused on automation rather than adaptive learning. It operates through a rule-based system that relies on predefined algorithms and quantitative models to execute profitable trades based on market data, e.g., time, stock price, and volume (Ben Zidane, 2025). In contrast, from a theoretical perspective, ToM AI, while still in progress, is expected to generate output based on a genuine comprehension of its users' psychology: thoughts and emotions, which is a high cognitive process that involves brain connections beyond reasoning (Bamicha & Drigas, 2022).

#### ToM Instances.

In theory, a ToM robotic aide could detect signs of user fatigue and respond with empathy. Such fatigue could be inferred from non-verbal cues, such as actions and posture, rather than verbal indicators. Another example of ToM in practice could involve a virtual assistant adapting its speech to its interlocutor and enhancing explanations when it notices confusion, showing signs of social awareness (Williams et al., 2022). Building on these examples, intelligent agents with ToM could revolutionize many fields, including educational technology, healthcare, and collaborative robotics (Williams et al., 2022).

#### *Educational Technology.*

In the field of education, ToM AI would assess learners' engagement and motivation through multiple reliable educational tools such as emotion-aware learning environments, eye-tracking systems, and physiological sensors. After a full assessment, ToM AI could theoretically interpret behavioral cues as indicators of internal mental states: curiosity, confusion, or frustration (Rosenberg-Kima & Thomas, 2022)

***Hypothetical Example.***

Hypothetically, a future ToM, an AI-enabled intelligent tutoring system, might use a combination of eye-tracking data, facial expression analysis, and interaction logs from a learning platform to infer that a student is showing signs of disengagement. This includes but is not limited to frequent gaze shifts, reduced interaction time, and expressions of boredom (Wang et al., 2021; Aly, 2025). Drawing on these inputs, the ToM AI could model the learner's internal state as unmotivated or cognitively overloaded and then adapt the instructional content in real time by simplifying the material, offering encouragement, or switching to a more interactive activity (Sharma et al., 2020). In this scenario, the eye-tracking device, facial recognition software, and platform usage analytics serve as reliable educational tools that feed into the ToM AI's inference system.

***ToM in AI: The Design***

ToM in AI is still under development; thus, its current applications are either in research or experimental phases. Yet, below is a brief overview of how AI with acquired ToM would proceed:

Step 1: Observing people's behaviors and communication, specifically nuances of interactions.

Step 2: Saving the information observed and beginning to recognize patterns in thoughts and feelings (Andrews et al., 2023).

The intelligent agent uses its observations and stored information regarding patterns to infer a person's thoughts and feelings in a specific situation. If its conclusions are inaccurate, then it learns and improves. This process is called machine learning (El Qasemy, 2025).

For consistency and clarity purposes, let's follow up on the example of the virtual assistant previously shared. In an interaction between a ToM virtual assistant and a person, the ToM virtual assistant could notice short sentences, assume that the person is either busy or not in the mood for long conversations, and would adapt its responses accordingly, e.g., by replacing details with conciseness.

***ToM AI Versus Traditional AI***

Traditional AI acts or reacts according to the predefined rules and patterns (Aggarwal et al., 2025), while ToM AI is intended to react according to its understanding of its interlocutor's thoughts and emotions. As of now, ToM AI is still under development; thus, this statement is more of an aim than a reality.

***Challenges***

Achieving a fully-fledged ToM AI requires major advancements in existing technologies. Genuinely grasping human intelligence might require utilizing neural networks (NNs), which are fundamentally different from the NNs used in limited memory AI. NNs used in limited memory AI rely primarily on historical data to decide the next steps or predict an outcome. These AI systems do not store the data they come across for future development or long-term learning (Aru et al., 2023).

ToM AI developers are encountering many other challenges. At the outset, humans have a mind that's complex enough to understand, and even if understood, it is changeable. Building a ToM AI that not only understands the human mind but also grasps that it can change based on many factors is not an easy task (Cuzzolin et al., 2020; Wang et al., 2021).

The AI system must also be capable of differentiating between emotions, beliefs, and needs. Not to overlook the fact that the AI system would be expected to interact with a variety of human beings who might be on the spectrum or facing neurological diseases and psychiatric disorders such as dementia and schizophrenia, respectively (Cuzzolin et al., 2020).

**Understanding the Human Mind**

Perception of emotions and beliefs comes to humans naturally, although it is not always accurate, as affected by personal background and cognitive health status (Tesar et al., 2020). One of the main challenges is developing AI

systems that are capable of interpreting cues accurately, whether verbal or non-verbal (Cuzzolin et al., 2020; Wang et al., 2021). This capability is hard to reach even in humans who often times misjudge each other and misread signals. This could be due to differences in maturity levels, emotional intelligence, or social experiences (Tesar et al., 2020). Not only is it challenging to create a ToM AI that understands the human mind, but it also interprets situations accurately in a complex environment where the exact same signals could mean different things to different people.

**Constructing Mental Models**

Constructing mental models with ToM capabilities requires crafting precise representations of existing intelligent entities, whether artificial or human. The difficulty is in the 'how' (Andrews et al., 2023). How can meta-learning be utilized to construct mental models with ToM? For clarification purposes, meta-learning is a subfield of machine learning focused on enabling models to improve their learning process over time by learning from previous experiences. Thus, meta-learning is also called learning to learn (Vettoruzzo et al., 2024). Instead of training a model just to perform a specific task, meta-learning trains a model to generalize across tasks so it can quickly adapt to new, unseen tasks with minimal data (Vettoruzzo et al., 2024).

***Types of Meta-Learning Approaches***

There are three types of meta-learning approaches: model-based, metric-based, and optimization-based.

The model-based meta-learning approach designs networks with internal memory, such as long short-term memory networks (LSTMs), to enable rapid adaptation to new data. In contrast, the metric-based approach learns to compare new examples with known ones, as in Siamese networks (Tian et al., 2022), while the optimization-based approach focuses on improving the efficiency of learning better or faster, e.g., model-agnostic meta-learning (MAML) (Vanschoren, 2019).

In theory, meta-learning can help an AI system quickly infer the beliefs or goals of a new agent it has not seen before, just by observing a small amount of behavior; essentially learning how to model minds faster.

**Ethical Implications of ToM AI**

Theory of mind capabilities could raise important ethical and societal concerns. Intelligent systems that model human cognition could compromise privacy, introduce bias, and unintentionally or deliberately manipulate user behavior (Langley et al., 2022).

Privacy could be compromised through data collection, non-encrypted storage, and interpretation of users' sensitive personal information as a means to make behavioral predictions. While biases could be introduced through perpetuation of societal or cultural inequities. Introduction of biases would be the result of feeding the ToM AI subjective data for training purposes (Langley et al., 2022); thus, the importance of carefully and ethically designing intelligent systems (Edwards, 2024).

Manipulation represents another important ethical implication associated with intelligent systems and can arise when behavior prediction is used to influence decisions of ToM AI users without transparent disclosure (Langley et al., 2022). Manipulation is discussed in detail in the societal implications section of this literature review.

**Societal Implications of ToM AI**

From a societal perspective, intelligent systems that model human cognition may shape choices, alter the nature of social engagement, and expand machine authority.

***Shaping Choices***

Prediction of intentions or preferences could affect human decisions in a very subtle manner, shaping choices in many areas, including but not limited to education, healthcare, or consumer decisions, without the user's conscious awareness (Matta, 2026).

Impacts on decision-making processes could either be beneficial or manipulative, depending on how the ToM AI model is designed and how transparent it is. For instance, a positive impact of ToM AI on decision-making in the educational field could involve suggesting interesting resources that would help the user make faster, well-informed decisions. While a negative impact could involve manipulation through neuromarketing techniques and strategic nudging towards products or services, shaping choices in ways that might not be immediately perceptible to users (Chokshi, 2025).

### *Altering the Nature of Social Engagement*

ToM AI could influence human communication and human-to-human collaboration. Equipped with functions such as anticipating needs, providing suggestions, and potentially acting as a social partner, ToM AI might seem all-encompassing or even self-contained. Thus, humans might feel enticed to fully rely on it. Besides full reliance, humans might also replace their human-to-human interactions with human-to-AI interactions or at least mediate or guide existing human-to-human interactions via ToM AI, resulting in a decrease in direct human-to-human interactions and eventually altering the nature of social engagement (Freund, 2023).

### *Expanding Machine Authority*

In theory, a ToM AI could appear socially intelligent, although it may not possess genuine social intelligence. This behavior, although pseudo-social and simulated by predictions of mental states through pattern recognition and probabilistic inference, could overly appeal to humans (Deel et al., 2023). Humans might assume that the said socially intelligent system understands context perfectly and over-trust its recommendations, resulting in full reliance, especially when the performance of the ToM AI aligns with users' expectations. In contrast, a detection of bias or failure from the ToM AI's side could erode human trust and affect potential adoption of general automated systems (Freund, 2023).

## Fully-Fledged ToM AI: Feasibility

The analyses of ToM AI feasibility present conflicting conclusions. On one hand, certain studies suggest that despite ToM being a prominent concept, its feasibility is not confirmed and if achieved, ways to verify its reliability must be identified (Wang et al., 2024). On the other hand, researchers, scientists, and technology developers are hinting at ToM being achievable, especially after the recent breakthroughs in technology. Some believe that progress is shaping up, exceptionally after the significant advances observed in large language models such as GPT-4 and LLaMA2. Although these large language models' responses are similar to those of humans, one cannot safely consider human-like responses as an achieved ToM capability. Moreover, although GPT-4 and LLaMA2 seem to understand irony or hints, it is unclear whether this is the route to ToM (Wang, 2023). This conflict in conclusions emphasizes the need for caution, scrutiny, and further testing prior to large-scale implementation.

With increased curiosity and interest, technology enthusiasts are impatiently waiting for what's next, while experts in the field are hitting the brakes and warning against premature machine anthropomorphism (Wang et al., 2024). Several scholars caution that applying human-centered terminology to describe a machine's function could blur the line between human intelligence and artificial intelligence, potentially misleading the public (Wang et al., 2021). The world may not be ready yet for this drastic shift in technology and we should carefully examine when it would be appropriate to make it happen because, beyond the ultimate realization of ToM, there will be self-awareness, the latest and greatest stage of AI development.

## ToM AI Versus Self-Aware AI

Based on theoretical research, ToM AI could model the mental states of other agents, human or artificial, and then use that model to predict or explain the agent's behavior. Mental states involve beliefs, desires, intentions, and knowledge. This is a much more advanced cognitive capability than current AI generally has (Li et al., 2025).

Self-aware AI could recognize its existence as a non-human subject, act or react based on ethics, and could feel empathy (Shi et al., 2025; Khusainova & Filippova, 2022). However, existing theory posits that this form of empathy is expected to be cognitive rather than emotional (Khusainova & Filippova, 2022). Technology facilitating the achievement of self-awareness in AI is not available at the moment and whether it will ever be remains debatable.

## Recommendations

Cross-disciplinary research combining insights from cognitive science, psychology, computer science, and most importantly, ethics is highly recommended to support designing robust, responsible, and socially responsive ToM models in AI. Empirical validation in real-world settings is also recommended to not only assess the effectiveness or reliability but also further identify any additional ethical implications of AI systems with ToM capabilities in practice. Building on ToM, future research may turn to self-awareness, the next frontier in AI development.

## Conclusion

Traditional AI follows fixed rules and provides an already set output for a given input. Thus, there is no genuine comprehension of thoughts or emotions in the process. Contrary to traditional AI, theory of mind AI strives to fully perceive subtle human thoughts and feelings and react accordingly. Although ToM AI is still under development, excellent progress is observed in areas such as emotion recognition, predictive modeling, conversational AI, simulation and cognitive modeling, and multi-agent systems. These breakthroughs are a step towards an AI that does not complete tasks solely but understands the agent it works with. Yet, full practical feasibility remains uncertain, and if achieved, ways to verify reliability must be identified.

Achieving a fully fledged ToM AI requires substantial technology, including sophisticated neural networks and meta-learning approaches capable of modeling the complexity, variability, and evolving nature of the human mind. A fully fledged ToM AI should distinguish among emotions, beliefs, and needs while accurately interpreting verbal and nonverbal cues across diverse individuals and contexts, an ability that even humans find challenging to achieve. Beyond these technical challenges, the development of ToM AI raises significant ethical and societal concerns, including privacy risks, bias, and behavioral manipulation, along with broader impacts on shaping human choices, altering social engagement, and expanding machine authority. The world might not be ready for this drastic shift in technology and experts caution against premature machine anthropomorphism. This research matters because our world still does not fully understand human thoughts and is nonetheless expected to create AI systems that can adapt to our ever-changing minds and/or distinguish between various mental states: thoughts versus feelings, bearing in mind that internal states are unobservable and rather inferred.

## References


Aggarwal, R., Sachan, S., Verma, R., & Dhanda, N. (2025). Traditional AI vs. modern AI. In A. Sharma, Nayancy, & R. Verma (Eds.), *The confluence of cryptography, blockchain and artificial intelligence* (pp. 51–75). CRC Press. https://doi.org/10.1201/9781032711515

Aly, M. (2025). Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model. *Multimedia Tools and Applications, 84*(13), 12575-12614. https://doi.org/10.1007/s11042-024-19392-5

Andrews, R. W., Lilly, J. M., Srivastava, D., & Feigh, K. M. (2023). The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science, 24*(2), 129-175. https://doi.org/10.1080/1463922X.2022.2061080

Aru, J., Labash, A., Corcoll, O., & Vicente, R. (2023). Mind the gap: Challenges of deep learning approaches to theory of mind. *Artificial Intelligence Review, 56*(9), 9141-9156. https://doi.org/10.1007/s10462-023-10401-x

Bamicha, V., & Drigas, A. (2022). The evolutionary course of theory of mind factors that facilitate or inhibit its operation & the role of ICTs. *Technium Social Sciences Journal. 30,* 138-158. https://doi.org/10.47577/tssj.v30i1.6220

Ben Zidane, M. (2025). Stock price and volume volatility of mergers and acquisitions companies. *Westcliff International Journal of Applied Research, 9*(2), 5-11. https://doi.org/10.47670/wuwijar20252MZ

Bhuiyan, M. M. (2024, June 3). The illusion of boundless AI: Analyzing limitations and ethical concerns. [Preprint]. TechRxiv. https://doi.org/10.36227/techrxiv.171742375.53309794/v1

Chokshi, S. (2025). The ethics of AI nudges: How AI influences decision-making. *Asian Management Insights, 12*(1), 84–91. https://ink.library.smu.edu.sg/ami/278

Cuzzolin, F., Morelli, A., Cirstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: Theory of mind in AI. *Psychological Medicine, 50*(7), 1057-1061. https://doi.org/10.1017/S0033291720000835

Deel, G., Cain, C., & Cain, L. (2023). Ethical considerations for the future of superhuman artificial intelligence: A viewpoint. *Robonomics: The Journal of the Automated Economy, 4*, 41-41. https://journal.robonomics.science/index.php/rj/article/view/41

Edwards, H. (2024). Toward Ethical AI: Relational Dynamics, Theory of Mind, and Human-Compatible Artificial Intelligence. *Contexts, 1*(5).

El Qasemy, H. (2025). Cognitive technologies: Machine learning, artificial intelligence, and convolutional neural networks in computer vision. Westcliff International *Journal of Applied Research, 9*(1), 5-17. https://doi.org/10.47670/wuwijar20251HEQ

Freund, L. (2023). Exploring the Intersection of Rationality, Reality, and Theory of Mind in AI Reasoning: An Analysis of GPT-4's Responses to Paradoxes and ToM Tests [Preprint].PhilArchive. https://philarchive.org/rec/LUCPAT-6

Gordon, J., & Hernandez-Lobato, J. M. (2020). Combining deep generative and discriminative models for Bayesian semi-supervised learning. *Pattern Recognition*, 100, 107156. https://doi.org/10.1016/j.patcog.2019.107156

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5-14. https://doi.org/10.1177/0008125619864925

Hopcroft, R. L. (2025). Neurosociology and theory of mind (ToM). In W. Kalkhoff, J. Dippong, & R. B. Firat (Eds.), *Handbook of neurosociology* (pp. 269–281). *Springer.* https://doi.org/10.1007/978-3-031-95615-7_15

Khusainova, F., & Filippova, V. (2022). Self-awareness of artificial intelligence: Scoping review. *Global and Regional Research, 4*(2), 118–125.

Langley, C., Cirstea, B. I., Cuzzolin, F., & Sahakian, B. J. (2022). Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in artificial intelligence*, *5*, 778852. https://doi.org/10.3389/frai.2022.778852

Li, X., Shi, H., Xu, R., & Xu, W. (2025). AI Awareness [Preprint]. ArXiv. https://doi.org/10.48550/arXiv.2504.20084

Luo, S., Li, Y., Gao, P., Wang, Y., & Serikawa, S. (2022). Meta-seg: A survey of meta-learning for image segmentation. *Pattern Recognition, 126*, 108586. https://doi.org/10.1016/j.patcog.2022.108586

Mao, Y., Liu, S., Ni, Q., Lin, X., & He, L. (2024). A review on machine theory of mind. *IEEE Transactions on Computational Social Systems, 11*(6), 7114–7132. https://doi.org/10.1109/TCSS.2024.3416707

Matta, D. (2026). Artificial intelligence and theory of mind. *Journal of Psychology and AI,* (21), 2628373. https://doi.org/10.1080/29974100.2026.2628373

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of artificial neural networks and deep learning. In O. A. Montesinos López, A. Montesinos López, & J. Crossa (Eds.), *Multivariate statistical machine learning methods for genomic prediction* (pp. 379–425). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_10

Nebreda, A., Shpakivska-Bilan, D., Camara, C., & Susi, G. (2023). The social machine: Artificial intelligence (AI) approaches to theory of mind. In T. Lopez-Soto, A. Garcia-Lopez, & F. J. Salguero-Lamillar (Eds.), *The theory of mind under scrutiny* (pp. 681–722). Springer. https://doi.org/10.1007/978-3-031-46742-4_22https://doi.org/10.1007/978-3-031-46742-4_22

Oluwagbenro, M. B. (2024). Generative AI: Definition, concepts, applications, and future prospects. [Preprint]. TechRxiv. https://doi.org/10.36227/techrxiv.171746875.59016695/v1

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology, 1*(4), 223-235. https://doi.org/10.1038/s44159-022-00037-z

Rosenberg-Kima, R. B., & Thomas, A. (2022). A teacher without a soul? Social-AI, theory of mind, and consciousness of a robot tutor. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 240–244). https://doi.org/10.1007/978-3-031-11647-6_43

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

Sharma, K., Giannakos, M., & Dillenbourg, P. (2020). Eye-tracking and artificial intelligence to enhance motivation and learning. *Smart Learning Environments, 7*(1), 13. https://doi.org/10.1186/s40561-020-00122-x

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: Definition and background. In *Mission AI: The new system technology* (pp. 15–41). Springer International Publishing. https://doi.org/10.1007/978-3-031-21448-6_2

Shi, H., Yin, B., Teng, L., & Ma, B. (2025). Empathetic AI encounters: Pathways to prosocial behavior. *Journal of Service Research*, Advance online publication. https://doi.org/10.1177/109467052513805

Szeliski, R. (2022). *Computer vision: Algorithms and applications* (2nd ed.). Springer Nature.

Tesar, B., Deckert, M., Schmoeger, M., & Willinger, U. (2020). Electrophysiological correlates of basic and higher order cognitive and affective theory of mind processing in emerging and early adulthood - An explorative event-related potentials study to investigate first, second, and third-order theory of mind processing based on visual cues. *Frontiers in Human Neuroscience, 14*, 79. ttps://doi.org/10.3389/fnhum.2020.00079

Tian, Y., Zhao, X., & Huang, W. (2022). Meta-learning approaches for learning-to-learn in deep learning: A survey. *Neurocomputing, 494*, 203-223. https://doi.org/10.1016/j.neucom.2022.04.078

Vettoruzzo, A., Bouguelia, M. R., Vanschoren, J., Rögnvaldsson, T., & Santosh, K. C. (2024). Advances and challenges in meta-learning: A technical review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(7), 4763–4779. https://doi.org/10.1109/TPAMI.2024.33578475

Vettoruzzo, A., Bouguelia, M. R., Vanschoren, J., Rögnvaldsson, T., & Santosh, K. C. (2024). Advances and challenges in meta-learning: A technical review. *IEEE transactions on pattern analysis and machine intelligence, 46*(7), 4763-4779. https://doi.org/10.1109/TPAMI.2024.33578477

Wang, J. (2023). Self-awareness: A singularity of AI. *Philosophy, 1*3(2), 68-77. https://doi.org/10.17265/2159-5313/2023.02.003

Wang, Q., Saha, K., Gregori, E., Joyner, D., & Goel, A. (2021). Towards mutual theory of mind in human-AI interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–14). https://doi.org/10.1145/3411764.3445645

Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J. D., & Goel, A. K. (2024). Theory of mind in human-AI interaction. In *Extended abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–6). https://doi.org/10.1145/3613905.3636308

Williams, J., Fiore, S. M., & Jentsch, F. (2022). Supporting artificial social intelligence with theory of mind. *Frontiers in Artificial Intelligence, 5*, 750763. https://doi.org/10.3389/frai.2022.750763

Zhang, Y., & Gosline, R. (2023). Human favoritism, not AI aversion: People's perceptions and bias toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making, 18*, e41. https://doi.org/10.1017/jdm.2023.37