Scientific Electronic Archives
Issue ID: Sci. Elec. Arch. Vol. 17 (3)
Mai/Jun 2024
DOI: http://dx.doi.org/10.36560/17320241939
Article link: https://sea.ufr.edu.br/SEA/article/view/1939

ISSN 2316-9281

Strengthening AI via ToM and MC dimensions

Victoria Bamicha
Net Media Lab Mind - Brain R&D IIT - N.C.S.R. "Demokritos", Athens, Greece

Corresponding author
Athanasios Drigas
N.C.S.R. Demokritos
dr@iit.demokritos.gr

Abstract. Theory of Mind (ToM) highlights the social-cognitive ability of the individual to communicate and interact effectively with the members of each social group. Essentially, it is the cornerstone of social knowledge that allows the recognition and understanding of the thoughts, intentions, and feelings of all involved, promoting social interaction and engagement. Metacognition (MC) is a higher mental ability of the biological mind and is characterized by the observation, control, evaluation, differentiation, and readjustment of the cognitive mechanism, aiming at its optimal performance and maintaining the homeostasis of mental, social, and emotional becoming of an organism. The rapid development of technology in recent decades has promoted the development of Artificial Intelligence (AI) intertwined with the need to integrate ToM and MC capabilities, enriching human communication. This paper investigates how the above-described human cognitive functions are involved in the conception and development of an artificial agent and their influence on human society. The conclusions suggest the importance of being able to read beliefs, emotions, and other factors, but also introspection by an intelligent system for social benefit, including the necessary ethical constraints.
Keywords:artificial intelligence, theory of mind, metacognition, computational theory of mind, autonomous systems

Introduction

AI consists of a collection of software and hardware infrastructures created by humans, which operate in the physical or digital dimension. They collect data that contribute to the perception of their environment and process information given the best choice of action to complete a goal. In addition, artificial systems use symbolic rules or mathematical models, where several times through data analysis, they become adaptive, evaluating the effect of their previous actions on the environment (Samoili et al., 2020). Essentially, artificial intelligence seeks to increase and improve human capacities for activities involving the reconstruction of nature and governing society. It employs intelligent machines to establish a harmonious society between humans and machines (Liu et al., 2018).

Man's complex and flexible cognitive mechanism has prompted him to create artificial intelligence based on a deep understanding of human cognition and its processes. Mastering the conceptual models and their respective applications is considered essential. Expectations in the technology industry have increased mainly with the introduction of computing machines, given it was beneficial for creating, implementing, and executing programs based on a predetermined rational process. Among the cognitive abilities that pose a challenge for artificial intelligence in modeling human cognition is ToM (Erb, 2016).

The ability to read minds occupies a privileged position in the development of human communication and is associated with prominent social, emotional, and cognitive skills (Brock et al., 2018; Bamicha & Drigas, 2022a). Its evolutionary course involves reasoning processes, specific brain connections, and higher cognitive processes, such as executive function, with particular emphasis on working memory, inhibitory control, attention, and

cognitive flexibility (Frith & Happé, 1999; Samson, 2009; Bamicha & Drigas, 2022b).

One of the dominant goals of AI is to mimic human cognition. Enriching it with ToM skills would provide a machine system with the ability to reason, solve problems, make decisions, interact linguistically, and perform other cognitive processes (Garcia-Lopez, 2024). Central to the AI evolution is incorporating the thought of an intelligent being influenced by an emotional state. Comprehension of human thinking, motivations, and goals that dictate the individual's response to various situations is highly significant (Cuzzolin et al., 2020). The advanced form of Artificial Intelligence could improve machine thinking, significantly approaching human thinking (Bakola et al., 2022).

Numerous studies report that executive functions, Cognition, and Metacognition constitute higher mental processes, which interact and are involved in each other's functioning, to achieve a goal. Effectual attribution of mental states to self and others is consciously controlled and evaluated by metacognition (Sodian & Frith, 2008; Bamicha & Drigas, 2023a, b). Metacognition, as a higher cognitive process, allows the person to be aware of their cognitive functions to observe them while they are working, to control them, allowing their differentiation and readjustment when required. Three key processes that pervade the metacognitive mechanism are observation, regulation, and adaptation, or "consciousness", a dynamic process that constitutes a dominant pillar in the cognitive pyramid (Drigas & Papas, 2017; Drigas, Kokkalia, & Economou, 2021).

In artificial intelligence, metacognition has been linked to introspection, allowing the machine to form beliefs about its internal states besides examining the environment in which it operates. In this sense, we could distinguish Metacognitive knowledge, Metacognitive knowledge, Metacognitive regulation, and Metacognitive experience. According to the first function, the system as a cognitive processor knows itself. The second could be about the system's knowledge of what it knows and doesn't know. The latter is associated with a process that incorporates the system's previous experiences related to the goal it will manage. Evaluating the data it has gathered leads to formulating the system's potential forecasts for the result (Ribeiro et al., 2024). Creating AI systems with metacognition would allow systems to think, learn, and adapt to real-world conditions (Johnson, 2022).

Perceived hardware/software complexity of IT systems pushed IBM to Autonomic Computing. The driving force behind the execution of the new perspective was the observation and finding of the more general emerging self-organization and self-awareness of an agent in nature, whereby, by specific processes, it reveals and interprets a complex behavior through unobservable causes. Autonomic Computing intends to foster the growth of various functions related to self-control, self-management, and self-organization in an IT system by increasing its level of autonomy. Indicatively, we mention the cases of autonomous Multi-Processor System-on-Chip platforms deployed in CPS and IoT applications and use self-awareness and self-organization functions, to improve system design (Sadighi et al., 2018).

The effective inclusion and integration of AI into human society requires providing it with ToM capabilities as a necessary condition. Therefore, artificial systems develop better levels of social interaction with humans and other agents. That will improve the understanding of intelligent systems for social cognition. Equally essential is the metacognitive capability in AI since it allows the assessment of system operation by limiting the chances of errors and external interventions, providing the best choice of actions in response to environmental challenges. Consequently, the study of the topic concerning the enhancement of AI with aspects of ToM and MC is worthy of attention, both for human-artificial system communication and the efficiency of AI applications.

Methods and materials

The current literature review relates to the empowerment of AI with dimensions of ToM and Metacognition and the expected consequences on the entire society. Methodologically, narrative review was utilized, as it provides a multifaceted and flexible approach to a research topic. Furthermore, the gathering and synthesis of earlier studies lead to an advancement in knowledge (Collins & Fauser, 2005; Snyder, 2019). The review was conducted in the following international bibliographic databases like Google Scholar and Research Gate, using as search phrases: artificial intelligence, theory of mind, metacognition, computational theory of mind, and autonomous systems. The research included the following stages: search for sources based on keywords, selection of articles according to the subject of the research under study, categorization of the articles according to their content, and writing of the study. This review contained 85 articles, the findings of which led to conclusions indicating the necessity of integrating aspects of ToM and MC into AI to achieve harmonious human-machine coexistence and interaction. The exclusion criteria constituted the research studies were not directly related to the individual research topics and presented deficiencies regarding the clear interpretation and analysis of the information. While English-language sources published in reputable scientific journals were selection criteria, covering the period from 1979 to 2024 and focusing mainly on the last decade from 2014 to 2024.

Theoretical background

Artificial Intelligence (AI)

In their quest to comprehend and investigate artificial intelligence, numerous scholars have developed some definitions. AI is a technological

discipline that aims to make machines intelligent to understand, interpret, and predict the environment. It is directly related to the cognitive branch of computing that deals with solving problems, which cognitively concerns human learning, imitation, memory, and pattern recognition. In addition, it combines the theory and development of computational systems related to human intelligence functions (Chassignol et al., 2018).

According to McCarthy, 2007 AI utilizes computer mechanisms and programs to understand human intelligence, which are highly effective at human tasks and named as intelligent. It is worth noting that artificial intelligence (AI) employs methodologies and approaches that are not always biologically observable.

Next, we list crucial branches of AI.

Machine learning in which programmers, leveraging complex mathematical expertise, design machine learning algorithms, creating a complete ML system. That enables machines to categorize, decipher, explore, and process data to solve real-world problems (Tyagi, 2021).

Fuzzy logic is a method of dealing with problems that aren't settled by using common sense, which relies on binary values. Since the specific troubles are expressions that are not true or false, making decisions is essential with more information, where their representation requires an intermediate value between absolute truth and absolute falsehood (García et al., 2019).

Artificial intelligence integrating cognitive science and neurobiology creates **Artificial Neural Networks** (ANNs), which simulate and try to copy the function of the human brain and the communication of its neurons. It is a set of algorithms that aims to discover elementary relationships in a set of data through the imitative process of the human brain (Tyagi, 2021).

Natural Language Processing is also designated as computational linguistics and aims to understand natural language. It allows users to communicate with their machines quickly and efficiently, using natural language, reducing the communication gap between humans and machines (García et al., 2019).

Designing, building, using, and operating robots is the primary focus of **Robotics** applications in various scientific disciplines such as medicine, education, and others. It leverages machine learning to develop social interaction in diverse interactivities (Tyagi, 2021).

Computer Vision is a branch of AI that, according to some techniques, enables computers to "learn" to recognize an image and its features. Leveraging machine learning models on images allows the computer to discern elements from the image, distinguish what it is, and separate it from another (García et al., 2019).

Expert systems were among the first successful artificial intelligence software models. An expert system is a computer system that mimics the decision-making intelligence of a human expert. That is achieved depending on the database's most recent updates, from which it draws information, applying rules of reasoning and knowledge related to user questions (Tyagi, 2021).

Theory of Mind (ToM)

Theory of Mind constitutes the fundamental component of social cognition, involving those processes that promote the attribution of mental states to others so that successful social interactions between people develop. The ability to read the mind enables people to recognize that others have different knowledge, beliefs, and desires and act accordingly. It is a decisive factor in the child's cognitive growth, who, growing up, experiences various interactions. As a result, his experiences prompt him to distinguish important behaviors, understanding which of them have positive or negative consequences (Williams et al., 2022).

The conceptual basis of ToM is meta-representation, the individual's ability to represent the world according to one's perspective. At the same time, meta-representation allows one to perceive how others act according to their desires, thoughts, and feelings (Rakoczy, 2022). Representation is considered a primary performance function of mental states. The English philosopher John Locke mentioned the importance of representation, stressing that it makes things present in the mind. Italian philosopher Thomas Aquinas called "species" the property of mental states to represent objects. He even pointed out that the representation resulting from sensory images gives the mind semantic content (Calì, 2020).

Researchers distinguish two different types of ToM, the explicit and implicit process. The first process is rapid and develops early, using implicit, automatic, and unconscious procedures, based on preexisting beliefs and patterns of response. In addition, it utilizes heuristics and biases, which limit direct control and intervention. While the slow, explicit process is controlled, it is distinct from conscious awareness and occurs later in life, as cognitive development is required (Roth et al., 2022).

Metacognition (MC)

An advanced mental process known as metacognition observes, controls, and evaluates a person's behavior and mental processes. Therefore, through applying his metacognitive abilities, man becomes aware of his capabilities and limitations, including all factors that may affect his cognitive performance. Utilizing the metacognitive function, the individual gradually moves towards self-awareness (Drigas & Mitsea, 2020a, b, c; Ribeiro, et al., 2024).

The study of metacognitive theories is concerned with modeling intelligence and higher-order reasoning. It has been demonstrated that selecting an efficient strategic solution and flexibly

treating an issue depends critically on metacognitive attention. At the same time, the meticulous observation of a person's cognitive functions promotes the mental mechanism's control, the search and finding of a solution depending on the conditions that arise (Cox, 2005). Notably, what exactly defines metacognition is awareness, evaluation, and regulation of thought (Drigas, Kokkalia, & Economou, 2021; Flavell, 1979).

Three key characteristics—purposefulness, self-consciousness, and self-awareness—distinguish metacognition from knowledge, according to Worley (2018). Assuming that a cognitive system develops a behavior, the metacognitive system observes, evaluates, and improves the system's performance by changing its behavior. That is achieved by controlling and varying its parameters, mainly by improving thinking. Therefore, metacognition as a higher mental process separates humans from simple reinforcement machines (Cox et al., 2022).

Artificial Intelligence approaches aspects of Theory of Mind

According to the European Commission's High-Level Expert Group (HLEG) on Artificial Intelligence, Artificial intelligence (AI) refers to software systems that perceive their environment by collecting and interpreting data. The ultimate goal is to process information to act to achieve a specific goal. AI systems adapt their behavior depending on the influence of previous experiences based on symbolic rules or learning numerical models (Nebreda et al., 2024).

Schossau& Hintze, 2023 state the evolution of human-level intelligence follows four crucial developmental milestones, including the system's ability to develop representation. First, he starts without representations, then creates representations for his environment, followed by representations of self and others, ending with self-representations related to the evolution of his consciousness state. Regarding artificial intelligence, representations are the information the machine collects about the environment or itself. The performance of machines in complex and dynamic environments requires combining and preserving data derived from these environments so that future uses for them are possible.

Several researchers in the computational theory of mind and artificial intelligence argue that ignoring consciousness can account for the description of mental representation. Since consciousness is not a passive observer of representations but an active process that affects the causal parameter of mental representations, the computational theory of mind lacks key aspects of ToM (Swiatczak, 2011).

Emotion, a central factor in interpersonal communication, transmits information about the emotional state and is the basis for interpreting complex psychological processes and behavioral motivations. Therefore, it is considered a necessary capability of machines to develop intelligence. The evaluation of subjective emotional changes and the gradual building of knowledge constitute data that feeds the artificial intelligence engine by incorporating human attitudes, preferences, and emotional experiences. AI combining human psychological knowledge simulates the person's reasonable thought process creating an emotional interaction between humans and machines, machines and machines. That facilitates communication with the human factor and results in the identification and comprehension of emotions, thereby enhancing the empathy and ToM dimensions of the artificial system (Zhao et al., 2022).

The foundation of effective human-machine interaction is learning through imitation, where machines essentially learn from humans. Previous attempts to integrate ToM into machines did not include the learning factor, failing to capture the accurate working of the human mind. It was evident from most models that they could not capture the dynamic process of experience-based learning. At the same time, they relied on reasoning processes for processing the human linguistic code without considering how the human brain represents and organizes knowledge (Cuzzolin et al., 2020).

Understanding other people's intentions is a complex process. It requires correct decoding of the social information received by the individual and the transfer of his social messages to achieve mutual social communication. Agents possessing ToM skills should be aware of the social and moral norms moreover, the tacit knowledge associated with any social situation. Furthermore, prior knowledge should be utilized to draw conclusions about the intentions and emotions of people and forecast their conduct (Williams et al., 2022).

ToM significantly affects human-human and agent-agent interaction. As such, it contributes to understanding human communication, enabling modeling and personalization of user experiences. Simultaneously, it offers the chance to create collaborative human-computer systems work effectively. The computer keeps track of data regarding the user's intentions, beliefs, goals, and behavior and then makes assumptions that lead to conclusions. Specifically, the user models include statistics and machine learning methods, which achieve the direct processing, interpretation, and generalization of the data resulting from the interaction process. Additionally, a form of ToM modeling in deep learning uses meta-learning to achieve action prediction through a sequence of observations (Çelikok et al., 2019).

Computational ToM allows agents to reason about other agents, which are more interpretable, and promote human-agent interaction. Consequently, an agent possessing principal aspects of ToM is informed about other agents' beliefs, recognizing those that are false (Zaroukian,

2022). In the work by Rabinowitz et al. (2018), models underwent training to identify various kinds of agents, as they predicted their later behavior based on their previous. Specifically, they proposed the creation of a ToM-net neural network, which can create models of the agents it meets, study their behavior, and ultimately advance human-machine interaction through applying meta-learning.

It has been established the processing of non-verbal communication presents several difficulties in understanding social signals and behavioral cues. However, an agent with dimensions of social competence should be able to interact with the human agent, collect data regarding their intentions, beliefs, and goals, and use them to conclude. General social AI should process social data, incorporating verbal and non-verbal cues (Williams et al., 2022).

Computational Theory of Mind (CTM) could be considered the revised form of Representational Theory of Mind. CTM treats the brain as a computer's kind and mental processes as computations. Moreover, it considers cognitive states as a set of computational relations governed by a sequence of processes involving various mental representations. At the same time, CTM develops models of cognitive processes that can be applied to artificial information processing systems, attempting to decode the mental processes of the human brain (Erb, 2016; Rescorla, 2015; Pitt, 2022).

Computational models allow the simulation of behavior in various tasks and evaluate a model's performance to interpret human behavior in a broad range of contexts. Subsequently, they provide the possibility of combining the representations of specific mental processes with neural recording techniques. That contributes to gathering information about the corresponding neural circuit, facilitating the understanding of the functioning of ToM (González & Chang, 2021).

Computational models of ToM fall into the following categories.

Bayesian ToM relates to modeling inherent uncertainty that arises from inferring unobservable mental states and can capture participants' judgments. However, the applications of this specific computational paradigm are limited to simple settings, with implications for inference and generalization of results (Langley et al., 2022b).

Game ToM is related to game theory, which has been used to model the representation of others' beliefs, thoughts, intentions, emotions, and desires in interactive economic games. Since the structure of these games is simple, they allow the use of neuroimaging, mapping the mental functions activated in the brain when representing the mental states of others. Probability distributions concerning actions, states, or beliefs of other players function to present a player's perspectives. Modeling gaming behavior involves examining players' reasoning, goals, and motivations (González & Chang, 2021; Cheong et al., 2017).

In addition, **RL** and **IRL** models offer state-of-the-art results in scalable real-world tasks. However, a substantial quantity of data processing or access to a simulator is required, including interpretation limitations. In particular, **Reinforcement Learning** models enable the understanding and automation of goal-driven learning and decision-making. Observing the interaction of an agent with its environment provides the possibility of predicting errors in the outcome. In a typical **RL** environment, learning results from optimal actions aimed at behavior that maximizes a predetermined reward function. Whereas **Inverse Reinforcement Learning** tries to recover the reward function from the observed behavior of the agent. The latter algorithm lends itself to modeling beliefs, goals, and desires from observing the actions of others (Langley et al., 2022b; González & Chang, 2021).

Computational methods make it easier to comprehend how humans adapt to and incorporate the beliefs of others in a social environment that is constantly changing. Studies report that the use of analogical models for the understanding and interpretation of social information by humans identifies the effect of the individual's actions on the beliefs of others and the feedback of these assessments for optimal decision-making (Cheong et al., 2017).

Nguyen & Gonzalez, 2020 developed a Bayesian ToM (BToM) model, which uses Bayesian probabilities and human rationality to identify mental states. In particular, relying on the observation of other agents' actions concludes their beliefs and desires. It is a decision-making algorithm and a set of cognitive mechanisms that develop computational models. The algorithm integrates knowledge of past experiences. Making decisions is aided by the algorithm's integration of prior experiences and representation of those that are pertinent to the current circumstance. The computational process follows inductive learning processes without using large volumes of data or complex models.

The development of artificial agents that can successfully communicate with humans may benefit from ToM research utilizing deep learning. It is enough to consider the importance of ToM in mutual communication and language development. Specifically, advanced deep learning tools can contribute to understanding how ToM works by allowing precise manipulation of words and phrases fed into models. In addition, the possibility of intervention is given to individual structural elements of deep learning, such as specific artificial neurons that model other factors and related components related to ToM processes. A prerequisite is that deep learning models of ToM approximate human ToM ability (Aru et al., 2023).

Several studies report that AI constructs that incorporate ToM dimensions would help interact with individuals diagnosed with neurological disorders such as autism, depression, Alzheimer's, and

schizophrenia by providing empathic healthcare. A consequence would be the reinforcement of psychological treatments, such as cognitive behavioral therapy or mindfulness, enabling robots to understand and express emotions in their communication with humans (Cuzzolin et al., 2020). In addition, artificial intelligence acts effectively, mainly through machine learning, in the diagnosis of various neurodevelopmental disorders, allowing the organization, analysis, and classification of data (Anagnostopoulou et al., 2020; Sideraki & Drigas, 2021; Fotoglou et al., 2022; Kyriakaki et al., 2023; Moraiti& Drigas, 2023; Chaidi & Drigas, 2023). While at the same time AI, utilizing digital technology, facilitates the access of people with special needs to education, significantly reducing discrimination and exclusion (Vrettaros et al., 2006; Karyotaki & Drigas, 2015; Garg & Sharma, 2020; Macpherson et al., 2021; Pappas & Drigas, 2016; Tourimpampa et al., 2018; Chaidi & Drigas, 2022; Papanastasiou et al., 2022; Moraiti et al., 2023; Bamicha & Salapata, 2024).

Artificial Intelligence approaches aspects of Metacognition

The quick growth of AI creates the imperative need to approach artificial systems from a metacognitive perspective, contributing to their self-awareness, self-management, and self-healing. Since using AI systems often leads to critical choices with expected high consequences, security is an indisputable condition of their construction. Metacognition is a capability that can be incorporated into artificial intelligence systems, providing monitoring and understanding of their external and internal operating environment. As a result, it enables systems to control and evaluate their performance, identifying and restoring possible errors. The primary sources of failure of an artificial intelligence system may be due to deficiencies in the design and pre-development process of its engineering and problems arising from its operational use (Johnson, 2022).

Intelligent systems have developed their autonomy by utilizing metacognition. The use of the metacognitive mechanism provides the ability to observe and control their learning and reasoning, which is why metacognition in AI is often referred to as meta-reasoning, introspective monitoring, and meta-level control. However, the enrichment of artificial systems with metacognitive features presents troubles due to the complexity of the individual processes that make up the metacognitive process (Caro et al., 2015).

According to Schmill et al., 2008, an artificial intelligence system characterized by the ability to reason and evaluate its processes has the meta-reasoning ability. Systems that can develop self-models by assessing their internal representations and processes have metacognitive capabilities. Therefore, if an invalidation or alteration of their expected cognitive processes is detected, meta-level error diagnosis and assessment might aid their readjustment. A fact that strengthens their performance and facilitates their application.

Systems that model and represent belief-generating processes, displaying metacognitive function, are distinguished between those that reason about what action to follow and those that look for the cause of an error, interpreting a failed action. In the first case, the systems choose an action according to the knowledge of the mental mechanism available to the system. In the second instance, systems undertake a feedback process of the reasoning process, providing interpretation and understanding of their operation (Cox, 2005).

Introspective monitoring of an agent's reasoning for effectiveness involves the perceptual process and a form of internal feedback to perform better, but also an evaluation of its meta-deliberative data. More generally, it is necessary to understand and process various events, situations, and actions of other actors in an environment to interact and respond satisfactorily in a social context. Using expressive language abilities is very beneficial for the meta-reflective process (Cox & Raja, 2011).

To be more precise, introspection is a metacognitive process that entails assessing meta-level data obtained at the object level. Finding errors in reasoning at the object level is the primary aim so that the intelligent system, through sufficient information, can make effective corrective decisions at the meta-level. However, it seems that despite the efforts to integrate introspective operations into the systems, they do not have a model of the knowledge they possess (Caro, Gomez, & Giraldo, 2017).

Per an earlier report, AI incorporates metacognition to build robust systems through two basic metacognitive processes: introspective monitoring and meta-level control. Caro et al., 2015 proposed a new Domain-Specific Visual Language (DSVL) for modeling metacognition in an intelligent system. They called it M++ and it includes the two functions of metacognition mentioned above. It provides precision in metacognitive concepts and a visual framework for the software engineering of such systems. It can also support the rapid prototyping of metacognitive architectures and enhance the analog system's design, testing, and updating.

The ability of intelligent systems to observe and control the processes of learning and integration of information has led to an increase in their autonomy, mainly in the choice of decisions. The main metacognitive processes of a system include metamemory, self-regulation, and metacomprehension. Metamemory constitutes the mechanism of control and observation of memory processes. Self-regulation is directly related to the adaptive action of the system regarding its learning processes. Finally, metacomprehension as a metacognitive component concerns the degree of understanding of the information received by the system (Caro Piñeres& Jiménez Builes, 2013).

The meta-reasoning process is characterized by self-adjustment, seeking to improve the performance of an autonomous agent. As a result, it can leverage algorithms to process various information it receives through sensors to plan tasks and make decisions that make it efficient. In addition, it can understand the environment, determining its actions. A multi-agent system may include additional reasoning algorithms such as coordination and clustering. Moreover, agents should consider the actions of multiple agents as they interact dynamically in the environment. In this case, system performance can be affected by meta-reasoning's effects (Langlois et al., 2020).

M'Balé& Josyula, 2013 point out that agents should adapt to various contingencies to correct failures and errors. If not, they are deemed fragile systems. Therefore, the capacity to employ metacognitive elements enables performance management and monitoring, enabling remedial interventions. Artificial systems show limited adaptability and flexibility, as any deviation from their specifications makes their operation difficult.

An intelligent agent possessing metacognitive abilities perceives stimuli from the environment and acts rationally, choosing actions that will lead him to achieve his goal. Meta reasoning processes presuppose the perception of reasoning and its control, aiming to improve the quality of its decision-making, and distinguishing the mental actions in which it excels and lags. Consequently, it preserves equilibrium within the computational process and the behavior it will manifest (Cox & Raja, 2011).

Some agents have episodic memory, which enhances their performance by supporting their cognitive ability. M'Balé& Josyula, 2013 presented the design of a metacognitive agent that can be connected to any cognitive agent, aiming to improve the adaptability of the cognitive system. In particular, the metacognitive agent continuously observes the performance of the cognitive agent and gradually becomes aware of its behavior and expectations about the environment. In addition, the metacognitive system forms its expectations by evaluating what it observes while identifying indications of violations of cognitive system expectations by suggesting corrective solutions. The communication interface between the two agents uses messages, and the metacognitive agent operates externally to the cognitive agent without necessarily sharing the same resources.

Computational metacognition includes the ability of Intelligent Systems (IS) to monitor and control their own learning and reasoning processes, which in human intelligence are related to higher cognitive functions. Metacognition allows an intelligent system to display metacognitive capabilities from at least two cognitive fields (object level and meta-level). In the first metacognitive ability, the intelligent agent has a reasoning model for its environment allowing problem-solving. While the second concerns a level of representation of the agent's reasoning (Caro, Gomez, & Giraldo, 2017). Essentially, computational metacognition aims to harness knowledge from the operational process of human metacognition and metacognitive approaches to artificial intelligence. Declarative representation and monitoring of cognitive processes in an intelligent system constitute dominant processes for self-management and performance of its cognitive function (Cox et al., 2022).

Jackson, 2020 wants to give another dimension to metascience, stating that it is possibly closely related to metacognition in human intelligence and human-level artificial intelligence. He suggests that the representation and processing that could support the metacognition of an AI system could also enhance an AI system that reasoned meta-scientifically about various fields of science. This view rests on the reasoning that scientific reasoning can be considered a subcategory of cognition in general, and metascientific reasoning can be considered a subcategory of metacognition. Metascience is considered the organized and procedurally unified acquisition of knowledge for systematic methods of knowledge acquisition, the science concerned with the understanding and formation of science in all scientific fields.

Especially significant for the effectiveness of an AI system is its ability to create from its experiences a database of knowledge resulting from monitoring performance, error rates, and prediction results of previous situations. It is the metacognitive memory of the system, an essential tool for evaluating its capabilities, processing data, and making decisions, improving its functionality (Johnson, 2022).

Metamemory, according to cognitive psychology is a component of metacognition and includes self-observation and control of memory processes by the human factor. Researchers Yamato et al., 2020 studied an advanced neural network that has a metamemory function based on the self-report of memory and analyzed the mechanism of metamemory. They developed neural networks utilizing neuromodulatory neurons, which can dynamically alter the plasticity of a neuron's connection. In particular, they looked at the neural network's structure, dynamics, and behavior, in which two modulatory neurons regulate certain connections from standard neurons to another modulatory neuron. The modulatory neuron could influence the network circuit according to the result of monitoring the memory state in the choice phase, allowing the network to respond accordingly to the experiment task.

Crowder & Shelli Friess, 2011 argue that an artificial system has cognitive self-awareness when it incorporates an artificial cognitive neural framework by evaluating its cognitive relationships within the artificial intelligence system. It could be a neural processing system that uses a modular

artificial neural architecture, providing flexibility and diversity in system capabilities. Particularly useful for the intelligent system would be mastering the concepts of emotions assisting in the information processing depending on the environment and immediate response in real-time. In addition, the system with metamemory features would allow access to cognitive data processing, providing their analysis and storage for later use.

Intelligence estimation of agents could be attributed to their efficiency in solving multiple and innovative tasks, using knowledge and models derived from past experiences. In this sense, meta-learning and knowledge transfer are the criteria of his intelligence. Langdon et al., 2022 point out that using models that reinforce and guide behavior and learning can help improve meta-learning, social cognition, and consciousness in AI as well as humans. In particular, meta-learning in artificial intelligence systems, say, the learning of learning algorithms and the selection and use of models and knowledge, is crucial for solving new situations. Behaviorally, it is characterized by the integration of experience into pre-existing knowledge. Consequently, the possibility of adaptation, the flexibility of artificial systems, and successful cooperation with humans is promoted.

While artificial intelligence has contributed significantly to solving various problems, implementing multiple processes in a single system and flexibly managing and coordinating them presents limitations. Dehaene et al., 2021 point out that when an artificial system has access to a set of information in its cognitive system, which it can recall, process, and act upon, it develops conscious functions. In addition, the intelligent system that can monitor its cognitive function process of data processing and analysis and collecting information about its performance has a form of introspection. Essentially, he creates internal representations of his knowledge and abilities related to metacognition. Therefore, the mentioned skills would cause an AI system to behave as though it were conscious. According to Kralik et al., 2018 consciousness includes metacognitive aspects that contribute to its effective functioning and lead to decision-making.

Then follows a scheme that summarizes the utilization of the cognitive functions of Tom and MC by the biological and artificial mind.

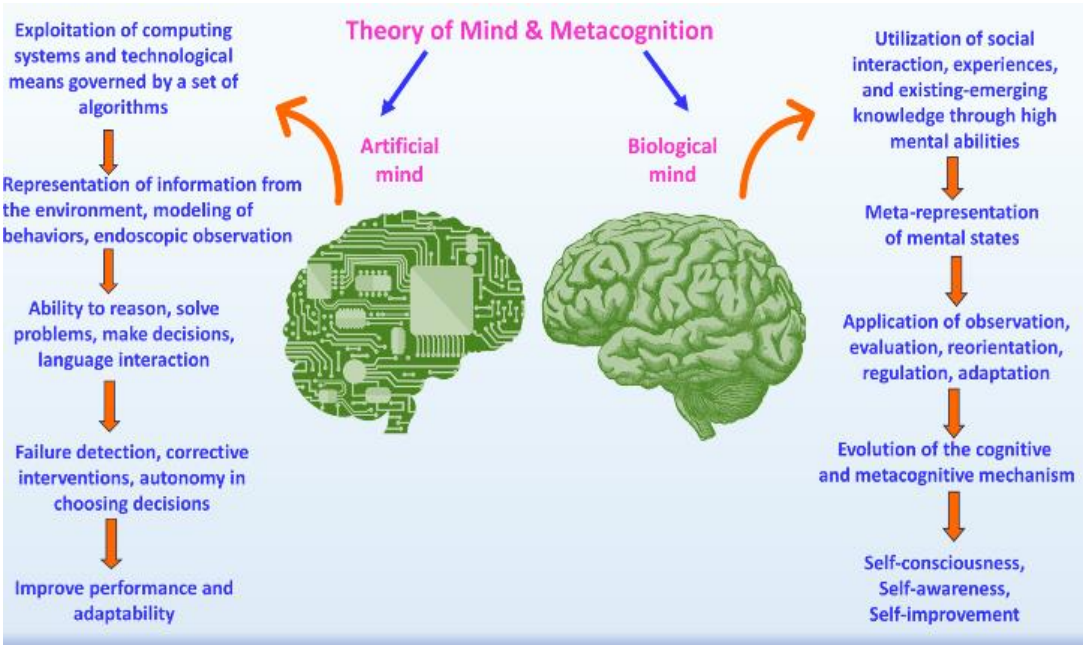

Figure 1. Theory of Mind and Metacognition in Biological and Artificial Mind

Authors Bamicha and Drigas, considering the findings of studies (Frith & Happé, 1999; Samson, 2009; Bamicha & Drigas, 2022a, b; Williams et al., 2022; Bamicha & Drigas, 2023a, b; Garcia-Lopez, 2024; Drigas & Papas, 2017; Cox & Raja, 2011; Johnson, 2022; Drigas & Mitsea, 2020a, b, c; Ribeiro et al., 2024; Schossau& Hintze, 2023; Nebreda et al., 2024) summarize the use of higher cognitive functions of ToM and MC by humans and artificial intelligence.

Results and Discussion

Artificial intelligence is characterized by scientificity because it includes intriguingfindings from different fields of knowledge, such as logic, statistics, engineering, image processing, linguistics, philosophy, psychology, and neurology. It has various digital tools with accessible user interfaces, covering separate scientific fields and requiring their appropriate selection and utilization (Ertel, 2011).

The dominant role of artificial intelligence should be to enhance humanity and respect human autonomy, while at the same time, it should be characterized by transparency and ethics. Also, the creation of technical systems must aim at increasing human efficiency while respecting human dignity and preserving cultural diversity. Moreover, its design concept is to protect personal information and maintain privacy. However, it should undergo an algorithmic process that allows the human agent to undo the inadvertent error. Finally, it is necessary to avoid bias in conjunction with appropriate and representative research (García et al., 2019).

Artificial intelligence (AI) is a fundamental evolution in the field of technology, including Machine Learning (ML), which is about the ability of machines to learn from data. Deep learning (DL), which is based on artificial neural networks and aims to facilitate learning effectively, occupies a prominent role. The collaboration of deep learning with reinforcement learning, where an agent acquires knowledge through interaction with the real world, receives the corresponding consequence, has spectacular results (Cuzzolin, 2020).

Utilizing smart devices with the ability to adopt flexible strategies often creates the impression of knowing and awareness of a situation, as well as intentions and beliefs (Erb, 2016). Engineering social knowledge and the emergence of an agent's social intelligence requires the cognitive base of interactions and an understanding of how messages are integrated through cues to support developing agents with social intelligence. Additionally, modeling elements of culture according to social norms can enhance understanding of social relationships (Williams et al., 2022).

Data modeling processes in the agent use reinforcement learning and imitation learning, replicating present behavior without considering internal mental states. In addition, a significant factor of successful human-agent interaction is the trust of the individual towards the agent, especially when the way agents are pushed to conclusions is not perceived. Consequently, the two-way exchange of information and the training of humans in factors that affect the functionality of an AI system would enhance the development of Artificial ToM (Williams et al., 2022).

Regardless of the theory used by research efforts to create Computational ToM (CToM), they all converge on the idea of an artificial brain whose mental processes are analogous to a computer, making decisions through mental algorithms. CToM offers a framework for artificially simulating human cognition and behavior. It enables the creation of artificial models that exhibit intelligent behavior, including problem-solving, learning, and decision-making, not just in people but also in many systems. In addition, CToM serves as a valuable resource for the design of intelligent systems by providing insights into the fundamental cognitive mechanisms that govern human behavior (Garcia-Lopez, 2024).

Several researchers approaching metacognition in the research field of artificial intelligence describe the term metareasoning computationally in terms of specific programs and algorithms. Various studies analyzing metacognition have focused on data from human experience and behavior (Cox & Raja, 2011). The additional cognitive function of metacognition in artificial systems indicates yet another push in the evolution of AI. Metacognition gives the system knowledge about itself and its knowledge, providing an understanding of processes. Therefore, self-diagnosis and observation of system internal indications by the system itself, recording errors, and adopting desired behavior, depending on existing situations, would enhance its adaptability and efficiency (Johnson, 2022).

Computational metacognition provides autonomy and awareness to Intelligent Systems by observing and controlling their learning and reasoning processes. Modeling metacognition in an artificial system presents difficulties due to the complex components involved. Especially when it requires the integration of many aspects of metacognition, such as metamemory, meta-understanding, and self-regulation (Caro, 2014).

MC in AI includes the process of Self-Analysis, or Introspection, creating the conditions for observing its reasoning. Metamemory relates to the system's memory capabilities and strategies that help represent, maintain, retrieve, and self-monitor the memory so that the system evaluates the data gathering and takes appropriate action. In addition, the system's generation of assumptions through validity checks and determination is a prerequisite of the self-assessment process (Crowder & Shelli Friess MA, 2011).

Autonomous systems based on artificial intelligence (AI) and machine learning (ML) are used in various fields, including healthcare, transportation, finance, industrial automation, etc. However, their increasing use raises concerns about their reliability and safety. Seshia, 2019 brings forward the concept of simulating an autonomous system's surroundings. Specifically, it refers to the introspection of the system for the modeling of the environment that utilizes the presumptions of the algorithm about the surroundings, recognizes its weak points, and shapes its safe operation.

Kuchling et al., 2022 state that metacognition involves regulatory processes embedded in the functioning of a great system and acts as a metacognitive model that regulates individual components of the system. Therefore, self-observation and self-regulation are essential to the internal regulators and post-processors of the system that perform metacognition. They consider that the regulatory capacity of a system increases as the processor's ability to access more data that it encodes increases.

Sadighi et al., 2018 characteristically emphasize that the adoption of self-awareness of a

computer system, the ability to recognize its state depending on the conditions, to identify possible actions and their effect on the environment, constitute an essential treatment of its complexity. Furthermore, Anderson et al., 2008 point out that the metacognitive aspects of an AI system provide automation and flexibility to deal with the unexpected. Detecting a fault by the system is equivalent to finding a mismatch between the expected and the observed result. Intelligent systems are characterized by fragility when they are incapable of handling new contingencies with changes or failures, as they are ineffective in the predetermined goals. Incorporating metacognitive processing into the system could improve its performance by dealing with contingency through three fundamental processes, identifying the damage, analyzing and evaluating the cause, and choosing the best solution to the problem.

Conclusions

In conclusion, AI excels compared to human intelligence in the speed and efficiency of processing large amounts of information, recognizing patterns, and predicting outcomes based on data. However, it lags in responding immediately to new situations as it has not developed heuristics and intuitive abilities, limiting its flexibility and adaptability. The human mind can sufficiently handle environmental complexity and unpredictability, especially with the contribution of AI, which acts flexibly in various and different ways in the diversified conditions of the external and internal world. Even though it has advanced, AI's conquest of ToM and MC is still in its infancy compared to its complete integration into intelligent systems.

The uniqueness of the human being might theoretically be included in future research when designing and creating models, specifically, the complex and different treatment and attitude of the human factor in emotional and social situations. Additionally, the ability of humans to decode and understand AI actions would be a bridge of communication and trust between them. Also, as intelligent systems include aspects of the Theory of Mind and Metacognition, they come closer to human intelligence. A fact that makes it vital to include ethical criteria in all stages of their evolution, so that their use is primarily beneficial to humans.

References

Anagnostopoulou, P., Alexandropoulou, V., Lorentzou, G., Lykothanasi, A., Ntaountaki, P., & Drigas, A. (2020). Artificial intelligence in autism assessment. International Journal of Emerging Technologies in Learning (iJET), 15(6), 95-107. https://doi.org/10.3991/ijet.v15i06.11231

Anderson, M. L., Fults, S., Josyula, D. P., Oates, T., Perlis, D., Wilson, S., & Wright, D. (2008). A Self-Help Guide For Autonomous Systems. AI Magazine, 29(2), 67. https://doi.org/10.1609/aimag.v29i2.212

Aru, J., Labash, A., Corcoll, O., & Vicente, R. (2023). Mind the gap: challenges of deep learning approaches to Theory of Mind. Artificial Intelligence Review, 1-16. https://doi.org/10.1007/s10462-023-10401-x

Bakola, L. N., Drigas, A., & Skianis, C. (2022). Emotional Intelligence vs. Artificial Intelligence: The interaction of human intelligence in evolutionary robotics. Research, Society and Development, 11(16). http://dx.doi.org/10.33448/rsd-v11i16.38057

Bamicha, V., & Drigas, A. (2022a). The Evolutionary Course of Theory of Mind-Factors That Facilitate or Inhibit Its Operation & the Role of ICTs. Technium Soc. Sci. J., 30, 138-158. https://doi.org/10.47577/tssj.v30i1.6220

Bamicha, V., & Drigas, A. (2022b). ToM & ASD: The interconnection of Theory of Mind with the social-emotional, cognitive development of children with Autism Spectrum Disorder. The use of ICTs as an alternative form of intervention in ASD. Technium Social Sciences Journal, 33, 42-72. https://orcid.org/0000-0001-5637-9601

Bamicha, V., & Drigas, A. (2023a). Consciousness influences in ToM and Metacognition functioning-An artificial intelligence perspective. Res earch, Society and Development, 12(3). https://doi.org/10.33448/rsd-v12i3.40420

Bamicha, V., & Drigas, A. (2023b). Theory of Mind in relation to Metacognition and ICTs. A metacognitive approach to ToM. Scientific Electronic Archives, 16(4). https://doi.org/10.36560/16420231711

Bamicha, V., & Salapata, Y. (2024). LLLT applications may enhance ASD aspects related to disturbances in the gut microbiome, mitochondrial activity, and neural network function. Brazilian Journal of Science, 3(1), 140-158. https://doi.org/10.14295/bjs.v3i1.457

Brock, L. L., Kim, H., Gutshall, C. C., &Grissmer, D. W. (2018). The development of theory of mind: Predictors and moderators of improvement in kindergarten. Early Child Development and Care. https://doi.org/10.1080/03004430.2017.1423481

Calì, C. (2020). Representation, Internal. In: Vercellone, F., Tedesco, S. (eds) Glossary of Morphology. Lecture Notes in Morphogenesis. Springer, Cham. https://doi.org/10.1007/978-3-030-51324-5_104

Caro Piñeres, M. F., & Jiménez Builes, J. A. (2013). Analysis of models and metacognitive architectures in intelligent systems. Dyna, 80(180), 50-59. Retrieved from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0012-73532013000400007&lng=en&tlng=en]

Caro, M. F., Josyula, D. P., Cox, M. T., & Jiménez, J. A. (2014). Design and validation of a metamodel

for metacognition support in artificial intelligent systems. Biologically Inspired Cognitive Architectures, 9, 82-104. https://doi.org/10.1016/j.bica.2014.07.002

Caro, M. F., Josyula, D. P., Jiménez, J. A., Kennedy, C. M., & Cox, M. T. (2015). A domain-specific visual language for modeling metacognition in intelligent systems. Biologically Inspired Cognitive Architectures, 13, 75-90. https://doi.org/10.1016/j.bica.2015.06.004

Caro, M. F., Gomez, A. A., & Giraldo, J. C. (2017). Algorithmic knowledge profiles for introspective monitoring in artificial cognitive agents. In 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). doi: 10.1109/ICCI-CC.2017.8109792

Çelikok, M. M., Peltola, T., Daee, P., & Kaski, S. (2019). Interactive AI with a Theory of Mind. arXiv preprint arXiv:1912.05284. https://doi.org/10.48550/arXiv.1912.05284

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: a narrative overview. Procedia Computer Science, 136, 16-24. https://doi.org/10.1016/j.procs.2018.08.233

Cheong, J. H., Jolly, E., Sul, S., & Chang, L. J. (2017). Computational models in social neuroscience. Computational models of brain and behavior, 229-244. https://doi.org/10.1002/9781119159193.ch17

Chaidi, I., & Drigas, A. (2022). Digital games & special education. Technium Soc. Sci. J., 34, 214. http://dx.doi.org/10.47577/tssj.v34i1.7054

Chaidi, I., & Drigas, A. (2023). Digital Gaming and Autistic Spectrum Disorder. International Journal of Emerging Technologies in Learning (iJET), 18(22), 4-23. . DOI:10.3991/ijet.v18i22.34497

Collins, J. A., & Fauser, B. C. (2005). Balancing the strengths of systematic and narrative reviews. Human reproduction update, 11(2), 103-104. https://doi.org/10.1093/humupd/dmh058

Cox, M. T. (2005). Metacognition in computation: A selected research review. Artificial intelligence, 169(2), 104-141. https://doi.org/10.1016/j.artint.2005.10.009

Cox, M. T., & Raja, A. (2011). Metareasoning: An Introduction. Proc. Metareasoning, pp. 3-14. https://doi.org/10.7551/mitpress/9780262014809.001.0001

Cox, M., Mohammad, Z., Kondrakunta, S., Gogineni, V. R., Dannenhauer, D., &Larue, O. (2022). Computational metacognition. arXiv preprint arXiv:2201.12885. https://ui.adsabs.harvard.edu/link_gateway/2022arXiv220112885C/doi:10.48550/arXiv.2201.12885

Crowder, J. A., & Shelli Friess MA, N. C. C. (2011). Metacognition and metamemory concepts for AI systems. In Proceedings on the International Conference on Artificial Intelligence (ICAI) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). Retrieved from https://www.proquest.com/openview/7f3758944801234b137b25813441813d/1?pq-origsite=gscholar&cbl=1976349

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%2FS0033291720000835

Dehaene, S., Lau, H., Kouider, S. (2021). What Is Consciousness, and Could Machines Have It? In: von Braun, J., S. Archer, M., Reichberg, G.M., Sánchez Sorondo, M. (eds) Robotics, AI, and Humanity. Springer, Cham, 43-56. https://doi.org/10.1007/978-3-030-54173-6_4

Drigas, A. S. & Papas, M. A. (2017). The Consciousness-Intelligence-Knowledge Pyramid: An 8x8 Layer Model. International Journal of Recent Contributions from Engineering Science & IT, 5(3), 14-25. https://doi.org/10.3991/ijes.v5i3.7680

Drigas, A. & Mitsea, E. (2020a). The Triangle of Spiritual Intelligence, Metacognition and Consciousness. International Journal of Recent Contributions from Engineering Science & IT, 8(1), 4-23. https://doi.org/10.3991/ijes.v8i1.12503

Drigas, A. & Mitsea, E. (2020b). A Metacognition Based 8 Pillars Mindfulness Model and Training Strategies. International Journal of Recent Contributions from Engineering Science & IT, 8(4), 4-17. https://doi.org/10.3991/ijes.v8i4.17419

Drigas, A. & Mitsea, E. (2020c). The 8 Pillars of Metacognition. International Journal of Recent Contributions from Engineering Science & IT, 15(21), 162-178. https://doi.org/10.3991/ijet.v15i21.14907

Drigas, A., Kokkalia, G. & Economou, A. (2021). An 8-Layer Model for Metacognitive Skills in Kindergarten. NEUROLOGY AND NEUROBIOLOGY, 4(1), 2-10. http://dx.doi.org/10.31487/j.NNB.2021.01.01

Erb, B. (2016). Artificial Intelligence & Theory of Mind. Ulm University (2016), 1-11. Retrieved from https://www.researchgate.net/publication/308608903_Artificial_Intelligence_Theory_of_Mind

Ertel, W. (2011). Introduction. In: Introduction to Artificial Intelligence. Undergraduate Topics in Computer Science. Springer, London. https://doi.org/10.1007/978-0-85729-299-5_1

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American psychologist, 34(10), 906.

https://psycnet.apa.org/doi/10.1037/0003-066X.34.10.906

Frith, U. & Happé, F.G.E (1999). Theory of Mind and Self-Consciousness: What Is It Like to Be Autistic? Mind & Language, 14 (1), 1–22. https://doi.org/10.1111/1468-0017.00100

Fotoglou, A., Moraiti, I., Dona, K., Katsimperi, A., Tsionakas, K., Karabatzaki, Z., & Drigas, A. (2022). IoT Applications help people with Autism. Technium Soc. Sci. J., 31, 115. DOI: 10.47577/tssj.v31i1.6422

García, C. G., Valdez, E. R. N., Díaz, V. G., Bustelo, B. C. P. G., & Lovelle, J. M. C. (2019). A Review of Artificial Intelligence in the Internet of Things. IJIMAI, 5(4), 9-20. DOI: 10.9781/ijimai.2018.03.004

Garcia-Lopez, A. (2024). Theory of Mind in Artificial Intelligence Applications. In The Theory of Mind Under Scrutiny: Psychopathology, Neuroscience, Philosophy of Mind and Artificial Intelligence (pp. 723-750). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-46742-4_23

Garg, S., & Sharma, S. (2020). Impact of artificial intelligence in special need education to promote inclusive pedagogy. International Journal of Information and Education Technology, 10(7), 523-527. doi: 10.18178/ijiet.2020.10.7.1418

González, B., & Chang, L.J. (2021). Computational Models of Mentalizing. In: Gilead, M., Ochsner, K.N. (eds) The Neural Basis of Mentalizing. Springer, Cham. https://doi.org/10.1007/978-3-030-51890-5_15

Jackson, P. (2020). Toward metascience via human-level AI with metacognition. Procedia Computer Science, 169, 527-534. https://doi.org/10.1016/j.procs.2020.02.214

Johnson, B. (2022). Metacognition for artificial intelligence system safety–An approach to safe and desired behavior. Safety Science, 151, 105743. https://doi.org/10.1016/j.ssci.2022.105743

Karyotaki, M., & Drigas, A. (2015). Online and other ICT Applications for Cognitive Training and Assessment. International Journal of Online Engineering, 11(2). http://dx.doi.org/10.3991/ijoe.v11i2.4360

Kralik, J. D., Lee, J. H., Rosenbloom, P. S., Jackson Jr, P. C., Epstein, S. L., Romero, O. J., ... & McGreggor, K. (2018). Metacognition for a common model of cognition. Procedia computer science, 145, 730-739. https://doi.org/10.1016/j.procs.2018.11.046

Kuchling, F., Fields, C., & Levin, M. (2022). Metacognition as a Consequence of Competing Evolutionary Time Scales. Entropy (Basel, Switzerland), 24(5), 601. https://doi.org/10.3390/e24050601

Kyriakaki, E., Karabatzaki, Z., & Salapata, Y. (2023). Mobile applications for autism. Eximia, 8, 51-66. Retrieved from https://eximiajournal.com/index.php/eximia/article/view/241

Langdon, A., Botvinick, M., Nakahara, H., Tanaka, K., Matsumoto, M., & Kanai, R. (2022). Meta-learning, social cognition and consciousness in brains and machines. Neural Networks, 145, 80-89. https://doi.org/10.1016/j.neunet.2021.10.004

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, 62. https://doi.org/10.3389/frai.2022.778852

Langlois, S. T., Akoroda, O., Carrillo, E., Herrmann, J. W., Azarm, S., Xu, H., & Otte, M. (2020). Metareasoning structures, problems, and modes for multiagent systems: A survey. IEEE Access, 8, 183080-183089. https://doi.org/10.1109/ACCESS.2020.3028751

Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., & Lee, I. (2018). Artificial Intelligence in the 21st Century. IEEE Access, 6, 34403-34421. https://doi.org/10.1109/ACCESS.2018.2819688

Macpherson, T., Churchland, A., Sejnowski, T., DiCarlo, J., Kamitani, Y., Takahashi, H., & Hikida, T. (2021). Natural and Artificial Intelligence: A brief introduction to the interplay between AI and neuroscience research. Neural Networks. https://doi.org/10.1016/j.neunet.2021.09.018

M'Balé, K., & Josyula, D. (2013). Integrating Metacognition into Artificial Agents. Common Model of Cognition Bulletin, 1(1), 55–62. Retrieved from https://ojs.library.carleton.ca/index.php/cmcb/article/view/2677

McCarthy, J. (2007). What is artificial intelligence? http://www-formal.stanford.edu/jmc/

Moraiti, I., & Drigas, A. (2023). AI Tools Like ChatGPT for People with Neurodevelopmental Disorders. International Journal of Online & Biomedical Engineering, 19(16). DOI:10.3991/ijoe.v19i16.43399

Moraiti, I., Fotoglou, A., & Drigas, A. (2023). Digital and Mobile Applications for Autism Inclusion. International Journal of Online & Biomedical Engineering, 19(11). https://doi.org/10.3991/ijoe.v19i11.37895

Nebreda, A., Shpakivska-Bilan, D., Camara, C., & Susi, G. (2024). The Social Machine: Artificial Intelligence (AI) Approaches to Theory of Mind. In The Theory of Mind Under Scrutiny: Psychopathology, Neuroscience, Philosophy of Mind and Artificial Intelligence (pp. 681-722). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-46742-4_22

Nguyen, T. N., & Gonzalez, C. (2020). Cognitive Machine Theory of Mind. In CogSci, 2560-2566 https://cognitivesciencesociety.org/cogsci20/papers/0612/0612.pdf

Papanastasiou, G., Drigas, A., & Skianis, C. (2022). Serious Games: How do they impact special education needs children. Technium Education and Humanities, 2(3), 41-58. https://orcid.org/0000-0001-5637-9601

Pappas, M. A., & Drigas, A. S. (2016). Incorporation of Artificial Intelligence Tutoring Techniques in Mathematics. International Journal of Engineering Pedagogy, 6(4),12-16. https://doi.org/10.3991/ijep.v6i4.6063

Pitt, D. (2022). Mental Representation. The Stanford Encyclopedia of Philosophy, Edward N. Zalta & Uri Nodelman (eds.), URL: https://plato.stanford.edu/archives/fall2022/entries/mental-representation/

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018, July). Machine theory of mind. In International conference on machine learning (pp. 4218-4227). PMLR. https://doi.org/10.48550/arXiv.1802.07740

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

Rescorla, M. (2015). The Computational Theory of Mind. The Stanford Encyclopedia of Philosoph, Edward N. Zalta (ed.), URL: https://plato.stanford.edu/archives/fall2020/entries/computational-mind/

Ribeiro, B.A., Coelho, H., Ferreira, A.E., Branquinho, J. (2024). Metacognition, Accountability and Legal Personhood of AI. In: Sousa Antunes, H., Freitas, P.M., Oliveira, A.L., Martins Pereira, C., Vaz de Sequeira, E., Barreto Xavier, L. (eds) Multidisciplinary Perspectives on Artificial Intelligence and the Law. Law, Governance and Technology Series, vol 58. Springer, Cham. https://doi.org/10.1007/978-3-031-41264-6_9

Roth, M., Marsella, S., & Barsalou, L. (2022). Cutting Corners in Theory of Mind. Retrieved from: https://ceur-ws.org/Vol-3332/paper11.pdf

Sadighi, A., Donyanavard, B., Kadeed, T., Moazzemi, K., Muck, T., Nassar, A., … Kurdahi, F. (2018). Design methodologies for enabling self-awareness in autonomous systems. In 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE). https://doi.org/10.23919/DATE.2018.8342259

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence. EUR 30117 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-17045-7. doi:10.2760/382730, JRC118163

Samson, D. (2009). Reading other people's mind: Insights from neuropsychology. Journal of Neuropsychology, 3(1), 3-16. https://doi.org/10.1348/174866408X377883

Schmill, M. D., Oates, T., Anderson, M. L., Josyula, D., Perlis, D., Wilson, S., & Fults, S. (2008). The role of metacognition in robust AI systems. 163-170. Paper presented at 2008 AAAI Workshop, Chicago, IL, United States. Retrieved from: https://www.researchgate.net/publication/254467928

Schossau, J., & Hintze, A. (2023). Towards a Theory of Mind for Artificial Intelligence Agents. In ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference. MIT Press. https://doi.org/10.1162/isal_a_00605

Seshia, S.A. (2019). Introspective Environment Modeling. In: Finkbeiner, B., Mariani, L. (eds) Runtime Verification. RV 2019. Lecture Notes in Computer Science(), vol 11757. Springer, Cham. https://doi.org/10.1007/978-3-030-32079-9_2

Sideraki, A., & Drigas, A. (2021). Artificial Intelligence (AI) in Autism. Technium Soc. Sci. J., 26, 262. http://dx.doi.org/10.33448/rsd-v11i16.38057

Snyder, H. (2019). "Literature review as a research methodology: An overview and guidelines," Journal of Business Research, Elsevier, 104(C), 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

Sodian, B., & Frith, U. (2008). Metacognition, theory of mind, and self-control: The relevance of high-level cognitive processes in development, neuroscience, and education. Mind, Brain, and Education, 2(3), 111–113. https://doi.org/10.1111/j.1751-228X.2008.00040.x

Swiatczak, B. (2011). Conscious representations: An intractable problem for the computational theory of mind. Minds and Machines, 21(1), 19-32. http://dx.doi.org/10.1007%2Fs11023-010-9214-y

Tourimpampa, A., Drigas, A., Economou, A., & Roussos, P. (2018). Perception and text comprehension. It's a matter of perception! International Journal of Emerging Technologies in Learning (Online), 13(7), 228. https://doi.org/10.3991/ijet.v13i07.7909

Tyagi, N. (2021). 6 Major branches of artificial intelligence (AI). Artificial Intelligence, analyticSteps. Retrieved from https://www.analyticssteps.com/blogs/6-major-branches-artificial-intelligence-ai

Vrettaros, J., Vouros, G., Drigas, A. (2006). An Intelligent System for Solo Taxonomy. In: Shi, Z., Shimohara, K., Feng, D. (eds) Intelligent Information Processing III. IIP 2006. IFIP International Federation for Information Processing, vol 228. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-44641-7_44

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.2022.750763n

Worley, P. (2018). Plato, metacognition and philosophy in schools. Journal of Philosophy in Schools, 5(1). http://doi.org/10.21913/jps.v5i1.1486

Yamato, Y., Suzuki, R., & Arita, T. (2020). Evolution of metamemory based on self-reference in artificial neural network with neuromodulation. Criterion, 1, C1. Retrieved from http://evolinguistics.net/acv/pdf/research_papers/200200325234622840.pdf

Yang, G. Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., ... & Wood, R. (2018). The grand challenges of science robotics. Science robotics, 3(14), eaar7650. https://doi.org/10.1126/scirobotics.aar7650

Zaroukian, E. (2022). Theory of Mind and Metareasoning for Artificial Intelligence: A Review. Retrieved from https://apps.dtic.mil/sti/pdfs/AD1175466.pdf

Zhao, J., Wu, M., Zhou, L., Wang, X., & Jia, J. (2022). Cognitive psychology-based artificial intelligence review. Frontiers in Neuroscience, 16. https://doi.org/10.3389/fnins.2022.1024316