Extending the Theory of Mind Framework to Embodied Artificial Agents

A Systematic Literature Review

Aleksandra Maria Jach [1]

Supervisor(s): Chirag Raman[1] **Ojas Shirekar**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Aleksandra Maria Jach
Final project course: CSE3000 Research Project
Thesis committee: Chirag Raman, Ojas Shirekar, Willem-Paul Brinkman

Abstract

This research paper aims to present how Theory of Mind (ToM) - the ability that allows humans to attribute mental states to others - can be used in the context of physically and virtually embodied computational agents. The focus is on using ToM for perspective-taking in environments with multiple computational agents interacting together. A Systematic Literature Review (SLR) was conducted providing a structured search process for collecting the literature. The findings from this review were obtained from 38 selected papers found in the literature since 2018. The review presented how computational ToM was inspired by the human ToM. Further, it summarised the current applications for perspective-taking for multi-agent settings, and it collected insights on high-level implementations of ToM agents. The findings highlight the complexity of ToM agents and the importance of ToM for agent interactions.

1 Introduction

The ability of intelligent agents to navigate social settings has become a significant area of interest in recent years. Unlike humans, who naturally navigate complex social interactions, computational agents often struggle in these environments. To address this, researchers have studied cognitive frameworks such as Theory of Mind (ToM), which refers to the ability to attribute mental states—including intentions, beliefs, and desires—to oneself and others [1]. This ability allows humans to empathize, recognize deception, and adopt different perspectives, making ToM a critical component of social interaction and communication [1]–[3].

Theory of Mind is a concept that initially gained significance in the field of developmental psychology, particularly through studies on children's cognitive development [3], [4]. Crucial experiments in this domain tested whether children understand that others can hold false beliefs about the world, which is a fundamental aspect of ToM [4]. Typically, children around the age of five begin to succeed in these tasks, indicating the development of ToM [1]. In addition to psychology, ToM has been the focus of research in neuroscience and cognitive science, where researchers have identified specific brain regions and correlated cognitive mechanisms associated with ToM [4], [5].

In recent years, ToM has been extensively studied in the context of Human-Agent Interaction (HAI). The goal of ToM in HAI research is to achieve better collaboration and mutual understanding between humans and either physical or virtual embodied agents by combining insights from psychology, neuroscience, and cognitive science [1]. Physical embodiment refers to intelligent agents that exist in tangible spaces, such as robots, while virtual embodiment refers to agents that exist within digital environments, such as chatbots [6].

Although ToM has been studied in HAI, there exists a gap in research concerning the integration of ToM in environments where multiple *embodied agents* interact *with one another*. This gap leads to the primary research question of this paper:

RQ: *How has the framework of Theory of Mind been incorporated to virtually and physically embodied agents with the ability to take perspectives of each other's points of view?*

While ToM takes into account various aspects such as emotion recognition or understanding false beliefs, this research paper focuses on perspective-taking. The following research questions have been formulated to help in answering the main research question:

- **RQ1:** In what ways does Theory Of Mind differ when applied to computational agents and when applied to humans?
- **RQ2:** What are the applications of Theory of Mind in multi-agent systems in which agents can take on each others' perspectives?
- **RQ3:** What are the ways of implementing Theory Of Mind for multi-agent environments?

The first sub-question dives into how Theory Of Mind is understood in the research community in three cases: when applied to humans, when applied to computational agents interacting with humans, and when applied to computational agents interacting with other agents. For the sake of clarity, in this paper "agents" will refer exclusively to computational agents and not humans. Thus, a "multi-agent" environment will imply agent-to-agent interaction. Through answering this question, I aim to understand differences and similarities between how ToM has been conceptualised for humans, and for agents.

The second sub-question explores practical applications of perspective-taking in multi-agent environments. Identifying these applications provides a more detailed motivation for researching this topic.

The third sub-question reviews existing approaches for implementing Theory of Mind in multi-agent systems, including proposed architectures and algorithms. While this research paper does not go in-depth into specific implementations, it provides an overview of the current state of the field.

By answering these questions, this study aims to clarify the application of ToM in computational multi-agent systems.

The remainder of this paper is organized as follows: Section 2 explains the methodology used for conducting this systematic literature review. Section 3 presents the results in relation to the research questions. Section 4 provides a discussion on the findings together with limitations. Section 5 reflects on the ethical aspects of the research.

2 Methodology

This Systematic Review was performed by following steps proposed in "Doing a systematic review: a student's guide" [7], and by complying with PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [8]. After deciding upon a research question, its sub-questions and screening through initial papers, researchers can begin

searching the databases for relevant literature. Systematic literature reviews offer an advantage of having predefined search strategies and criteria, which minimize bias in literature searches and ensure greater transparency and reproducibility. Section 2.1 identifies the chosen databases and outlines the formulated search query. Section 2.2 talks about the next step, which was the screening process, and identifies the resulting records. Section 2.3 clearly defines the criteria used to exclude certain papers during the screening process.

2.1 Selecting Databases and Formulating the Search Query

To ensure a systematic approach, I predefined the databases and keywords for the search query. Below, I outline these choices.

Databases

The aim was to collect comprehensive and recent papers from various fields like Psychology, Computer Science, Neuroscience and Biology in order to deeply explore the topic. The following databases cover all desired fields and are widely known, and thus they were used in this study: Web of Science, Scopus, and IEEE Xplore databases and Cell Press.

Search query

In order to develop a suitable search query, key words and their synonyms had to be identified. This was done through analysing the sub-questions and considering the scope of the study.

The main focus of this study is on Theory of Mind and how it has been applied to intelligent agents. To gain a full understanding of this topic, the paper is not only restricted to multi-agent systems and how they take on each others perspectives, but also covers autonomous agents and their ability to empathise with humans and take their point of view. Furthermore, since computational models are often inspired by the human brain, this study had to dive into how humans make use of Theory of Mind. To restrict the search in the field of neuroscience, the intention was to focus only on theoretical concepts regarding human brain architecture.

Figure 1 presents the resulting query which was used across the chosen databases.

("Theory of Mind")
AND (cognition* OR cognitive)
AND ("decision making") AND
(perspective OR "point of view" OR "frame of reference" OR empathy OR perception)
AND (predict* OR learn* OR understand* OR mimic* OR interpret OR awareness)
AND ("multi-agent" OR "brain-inspired" OR "autonomous agent" OR "intelligent agents" OR "autonomous agents"
OR (("mental state" OR "mental states" OR "belief state" OR "belief states" OR "internal state" OR "internal states") AND architecture))

Figure 1: Search query used for searching databases

The search was performed on 18/05/2024 and it was limited to only include papers in English language that were published in past 6 years (2018-2024) and that were neither a book or a book chapter.

2.2 Selection Process

The selection process included three steps: identification, screening, and eligibility. Figure 2 displays each step clearly and provides the number of articles included in each step.

After conducting the search query, the initial search results were recorded using Zotero, which facilitates the detection of duplicates. Initially, 625 records were identified from the databases; 583 were from the Scopus database, 40 were from Cell databases, 1 was from IIEE and 1 was from Web Of Science database. Excluding duplicated results, 611 records remained for the screening process.

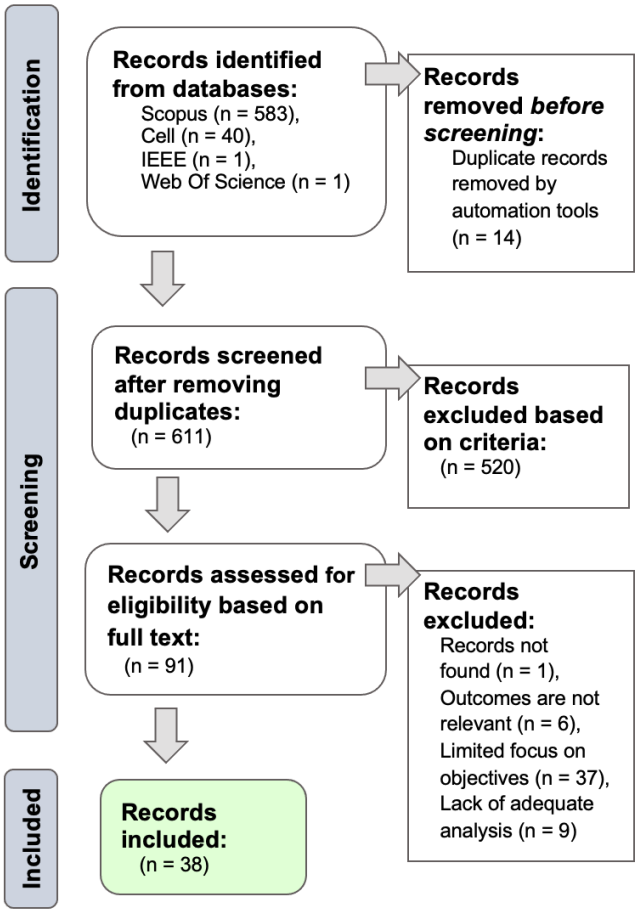

Figure 2: PRISMA flow diagram used to demonstrate the process of finding relevant papers for this systematic review.

For the screening step, I first predefined inclusion and exclusion criteria, described in Section 2.3. The screening process began with reviewing the titles and abstracts of all papers and discarding those that did not match the search criteria. To ease this process, I made use of ASReview LAB, a software that is designed to help with screening abstracts.

Following that, I conducted a full-text screening to further exclude papers. Papers that initially appeared relevant based on their abstracts but did not meet the criteria upon full-text review were marked as "Outcomes not relevant". Papers that met the criteria but offered only surface-level or brief analysis were marked as "Lack of adequate analysis". Papers that focused on supportive topics but had different primary objectives were marked as "Limited focus on objectives". In the end, 38 papers were included in this report.

2.3 Defining Inclusion and Exclusion Criteria

Inclusion and exclusion criteria had to be defined so that the records collected through the database search, and the records discarded through the screening process were reproducible. The following criteria were chosen to limit the scope.

Inclusion Criteria:

- Papers focusing on Theory of Mind in a setting with two or more computational agents
- Papers describing Theory of Mind for robots interacting with humans, where the implementation is described
- Biological perspectives on Theory of Mind
- Articles, conference papers, and reviews written in English
- Papers published in the past 6 years (2018-2024)

Exclusion Criteria:

- Research on how humans perceive robots with Theory of Mind
- Studies on the limitations of Theory of Mind applications in HAI settings
- Papers focusing solely on comparing the efficiency of agents with and without Theory of Mind
- Papers that only mention Theory of Mind in one sentence without focusing on it
- Specific applications of Theory of Mind in HAI settings without detailed implementation
- Papers on agents and understanding human emotion
- Papers about measuring Theory of Mind in agents
- Papers that are too general about cognition and only briefly mention Theory of Mind
- Empirical studies on how humans apply Theory of Mind
- Papers not written in English
- Papers that are not found

3 Results

The results of the review are analysed and described in this section. There are three subsections, each corresponding to a sub-research question defined in Section 1.

3.1 In what ways does Theory Of Mind differ when applied to computational agents and when applied to humans?

In the literature, there is a lack of consensus on what Theory of Mind (ToM) entails. This subsection analyses different outlooks on ToM and shows how computational ToM was inspired by cognitive ToM. Additionally, it describes the differences in how ToM is conceptualised in human-to-agent interaction (HAI) and multi-agent interaction.

Biological Outlook on ToM: Theory of Mind is not yet fully understood from a neuroscientific standpoint. Certain regions of the brain, such as Temporal Parietal Junction (TPJ), precuneus, Superior Temporal Sulcus (STS), Inferior Parietal Lobule (IPL), Inferior Frontal Gyrus (IFG), Premotor Cortex (PMC), and medial Prefrontal Cortex (mPFC) are activated during ToM-related tasks. These regions are associated with belief reasoning, self-other distinction, motor responses, and visual perception [5].

Further, in cognitive sciences, according to others, a distinction is made between "hot" and "cold" processes. Hot cognition relates to social and emotional cognition, while cold processes are non-emotional [1]. Based on these, ToM can be separated into cognitive ToM and affective ToM. Cognitive ToM involves the capacity to understand and represent thoughts, intentions, and beliefs, while affective ToM is concerned with recognizing and representing emotional states and feelings [1]. In the literature, there are various characterisations of mental states in computational and cognitive ToM, and it is not decided whether emotional states should be considered. Some even classify expectations, preferences, or "bodily feelings" as part of ToM [1], [9].

Theory Theory and Simulation Theory of social cognition: Two philosophical theories regarding ToM have been developed: Theory Theory (TT) and Simulation Theory (ST). TT proposes that human gather abstract principles and rules about behaviour from early childhood and use them to explain and predict the behaviour of others [10], [11]. This theory argues that humans create probabilistic reasoning models that stabilise over time [1], [4].

ST involves imitating others' mental states and imagining oneself in someone's situation to simulate their subjective point of view and likely behaviours [4], [12]. ST is more associated with empathy and perspective-taking [4], [12]. There exists a correlation between ToM and empathy, and it is difficult to quantify the overlap [13]. Some argue that TT relies more on cold, intellectual process, while ST relies more on hot cognition [1]. Some sources argue that activity found in the ToM-related regions in the brains seems to be more linked with TT, rather than with ST [4]. In contrast, others suggest that ST is more consistent with findings from neuroscience [5].

In the literature, there are examples of models based on TT, ST, or both [11], [14], [15]. The Bayesian approach, further explained in subsection 3.3, is often used to implement TT. These models typically neglect emotion, representing Cognitive ToM rather than Affective ToM [1], [10], [14]. Bayesian approach is inspired by the fact that human decision making

has been often modeled as Bayesian Inference [16], [17].

Complexity of Human ToM: Some argue that traditional probabilistic models do not fully capture human reasoning, and when they attempt to do so, they become highly complex [18]. Other ways of representing human thinking have been studied, such as the Mirror Neuron System (MNS). According to some sources, MNS represents a low-level architecture of Human ToM, but to fully represent it, more complex models need to be designed, combining imagination and memory [5]. In cognitive sciences, MNS is responsible for mirroring behaviour of others [4], [5]. Others argue that, similar to empathy, MNS is not a sub-component of ToM, and the concepts are not mutually exclusive in terms of cognitive functions and brain regions associated with them [4], as represented in Figure 3. Models like MNS are complex, and are implemented via Spiking Neural Networks and Reinforcement Learning, which are often used on its own to model ToM [5]. More on this can be found in Section 3.3.

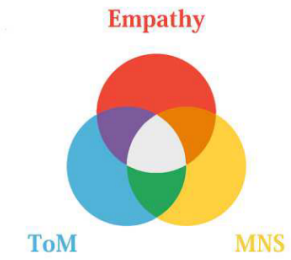

Figure 3: There is an overlap between Theory of Mind (ToM), empathy, and Mirror Neuron Systems (MNS). The overlap is difficult to quantify, thus there is no general agreement on what ToM consists of. *Source:* Adapted from [4]

When humans make decisions, they do not rely solely on ToM but also on other cognitive methods and biases. Examples of these include making decisions based on "gut feeling." or having a selective bias which allows humans to reduce complexity of perception [18]. Humans are not always rational, which further increases complexity [19]. To model human irrational thinking, some propose quantum-based methods [18]. Another quality humans possess is introspection, which is related to ToM, although it remains undecided how introspection fits into ToM's definition. Nevertheless, some include it in ToM implementations as further discussed in Section 3.3 [20].

Zero, first, second and higher orders of ToM, and strategic reasoning: There exist multiple orders of ToM that go up to infinity [12]. Zero-order ToM agents cannot reason about mental content of another agent, but they can make predictions about the behaviour of others based only on previously observed behaviour [21], [22]. First-order of ToM corresponds to "I think that you think". Second-order of ToM represents "I think that you think that I think" [22]. The same recursive pattern can be constructed in the higher orders. On average, humans are hypothesised to reach a 1.6th level with their reasoning, which could suggest that in some cases ToM agents might not need to use higher-order implementations [17]. Apart from ToM, there is also the concept of strategic reasoning, which some consider a component of ToM [10], [20]. In the computational world, strategies seem to be modeled in such a way that they make predictions without the use of internal states [21], [23]. Humans may use those predictive models to anticipate previously encountered situations, but switch to ToM when they try to predict someone's behaviour in a novel situation with infinite possibilities. However, sometimes it is hard to distinguish if they are using ToM or simple behaviour-based strategies [21], [24]. These concepts from human ToM inspire the implementation ideas discussed further in Section 3.3.

Inspirations taken from human-to-human collaboration and HAI for multi-agent systems: According to research, when humans interact with agents, they use less cognitive energy than when interacting with other humans [25]. This relates to the complexity of human thinking, as discussed earlier in this paper. Humans' use of cognitive functions increases as agents become more human-like and as humans better understand the hidden states of these agents [20], [26]. This could suggest that when an agent interacts with a human, it will also need to use more energy than when interacting with another agent.

In HAI, there is a strong focus on making agents human-like, so that humans can trust them, develop feelings of closeness, or empathize with them [27]. Similarly, in competitive settings such as negotiations involving agents, interpersonal relationships should be taken into consideration [14]. In many multi-agent environments, these factors do not need to be acknowledged at all, or to a limited extent since agents often do not have emotional needs. Emotional needs are considered when modelling a simulation with aim that an agent will be human like [27].

Furthermore, when modelling agents that interact with humans, it is important to consider social rules, norms, ethics, morality, values, emotion and appraisal. These factors help agents select actions that best suit the current situation from a range of possible actions [19], [27]–[29]. Social norms and rules do not need to be taken into account when building multi-agent environment, however, unless it is a task that specifically requires it, such as a simulation [27].

One of the ways to tackle the complexity of implementing ToM for HAI or multi-agent systems is by introducing other cognitive ideas to help model it. Since the focus is on the multi-agent environment, inspiration can be taken from human social settings. People tend to orientate themselves towards others who are like-minded, and through that they collectively form social groups of individuals who think similarly [20]. In such cases, people assume that to some degree everyone thinks the same as they do, so they put less cognitive energy into trying to understand and interpret others [20]. Some authors have proposed to use this social-ability, referred to as intra-group homophily, to implement agents [20], [28]. By assuming that agents think similarly and adding constraints to their thinking, agents can narrow the range of possible actions of other agents. Agents are considered to be more constrained than humans, who display unpredictable behaviours [10]. Yet, even in HAI setting, in the topic of

ToM, the concept of bounded rationality has been introduced. Bounded rationality suggests that there are some limits to human rationality when making decisions, which often leads humans to reach satisfactory solutions over the most optimal ones [18]. Thus, in HAI, there are examples where the implementation of ToM includes constraints on human thinking [20].

It is worth noting that humans have ways of reducing their cognitive load when interacting together by actions that are outside of the scope for computational agents. An example of that would be communicating through the use of physical objects like cards during collaboration games like Hanabi, where humans tend to outperform agents [30]. Hanabi is a game in which players work together to arrange a deck of cards into sequences by color and number. In this game players can see everyone's cards, but not their own, so they often rely on clues from others to know what card to play [15], [30], [31].

General and task-specific ToM: Most ToM implementations in both HAI and multi-agent settings are task-specific. However, especially in the context of HAI, some argue that it is necessary for agent to be very adaptive [10], [12]. According to Rudrauf, adaptive agents should be capable of simulating, predicting, and evaluating behaviors like joint attention, emotional expressions, and, more broadly, navigation within a three-dimensional environment [12].

3.2 What are the applications of Theory of Mind in multi-agent systems in which agents can take on each others' perspectives?

Multi-agent interactions are not as often observed as human-agent interactions. Below is the list of use cases found in literature.

List of use cases: classic game playing, chess, poker, simulated robot soccer, dialog systems, autonomous military systems, autonomous vehicles, automated negotiation, deception and scepticism, human user modelling, commercial video games, multi-agent learning, prisoner dilemma, ultimatum games, hide and seek simulation, casino task simulation, sender-receiver games, Virtual Reality, robots with incomplete self knowledge, planning and navigation tasks [11], [12], [17], [18], [20], [24], [31]–[33].

In the list above, there are examples of both virtual agents and physically-embodied agents. Multi-agent interaction seems to be most often studied in the context of Game Theory, which inspired an area of ToM that is referred to as Game ToM.

Uses for a collaborative setting: There are examples when researchers used ToM to achieve *collective intelligence*, an idea that acting together leads to achieving better results than acting individually [34]. When trying to achieve collaboration, modeling other agents is essential when there are no coordination or communication protocols [32]. ToM allows agents to achieve social perceptiveness, which is a necessary component of collaboration. Together with Goal Alignment, ToM has been studied to achieve better results [34]. It should be noted, however, that on its own, ToM might not suffice for achieving collaboration and instead it could be combined with other cognitive principles such as active inference [35].

Uses for a competitive setting: Besides collaborative settings, there are also *competitive settings* and mixed settings, where there is also proof that ToM agents achieve better performance in comparison to agents without ToM [27]. An example of a competitive setting is a simple rock-paper-scissors game, where players compete directly to win each round [22]. In a mixed setting like robot soccer, two teams play against each other [11]. Humans themselves are also more likely to make use of ToM in mixed-motivated settings where both cooperative and competitive aspects play a role [21]. For a competitive environment, it has been studied that even if ToM does not accurately model beliefs of opponents, when higher orders of ToM are used it still holds an advantage [21]. In some cases, agents could rely on simple strategic reasoning or the zeroth order of ToM. However, when the complexity of environment increases, there is a need for higher ToM to allow for perspective-taking [21], [23].

Modelling simulations: ToM is essential when modelling simulations. For example, when testing attack and defense systems such that the agents resemble humans who do not make optimal decisions [36]. When modelling a simulation of humans, factors like emotion or feature extraction have to be taken into account, which makes the tasks more difficult [12].

Uncertainty reduction: It is worth noticing that ToM is used to reduce uncertainty about agents' environment [27], [32], [34]. Agents are usually limited to partial observability, meaning that they do not get all the information about the environment [15]. Further, other agents are often considered to be the most complex parts of the environment [31]. It has been shown that in some settings, ToM accounts for 15% information gain [15].

An example from literature can be found where there were 2 agents - one Strong agent that was very perceptive about the environment, and one Weak agent that was "blind" and could not see its environment (it was very uncertain about the environment), but it had been gifted with the Theory of Mind [34]. The idea of the experiment was for the Strong agent to guide the Weak agent through the environment. The Weak agent was able to sense the environment through the lens of the Strong agent, and through that navigate through it. In this specific paper, the agents made decisions using Free Energy Principle, implemented with Markov blankets. At each step, the agents selected the action that minimised their free energy relative to the target they wanted to achieve, and they updated their beliefs in such a way that they matched their sensory (external environment) states as best as possible. The Strong agent could "easily" achieve their target as they had a lot of knowledge about the environment. The Weak agent had to learn about its sensory environment from the Strong agent, by creating partner's mental model of the world. At the end, the Weak agent combined both the mental representation of the Strong agents' model and their own model to form a decision. Additionally, this model introduced a "second-order theory of mind", signifying that the Weak agent would build a model of what the Strong agent thought the Weak agent's

model was. The Weak agent used it further as part of the Free Energy Principle to make a decision. So ToM here is used in such a way that the Strong agent's model influenced the Weak agent's model [34].

3.3 What are the ways of implementing Theory Of Mind for multi-agent environments?

This section does not go into detail about possible implementations but rather presents themes regarding modelling ToM seen in literature. There are different ways of thinking about artificial Theory of Mind, and some divide ToM models into categories like Game ToM, Observational Reinforcement Learning, and Bayesian ToM [1].

The general idea: Typically there are two or more agents in a defined environment. Depending on the task, the agents can be modeled differently. In general, an intelligent agent will always be able to perform some action and will have some target that it wants to reach. It also has a position in a physical or virtual space of the environment and a way of reasoning about which action to select. Agents also tend to store a history of observed interactions (e.g., past actions, states, etc.). If storing the entire history is unfeasible, agents can compute factors based on past interactions [32]. In some cases, models consider only the most recent actions [21]. In the context of ToM agents, there exist internal states which are responsible for implementing beliefs and desires [21]. Researches have developed various layered architectures for agents, which define different components of agent models. For example, one study created an Adaptive Layer that contained internal states, and a Reactive Layer that interacted with the outside environment [23]. However, a detailed analysis of different layered architectures is outside of the scope of this research.

Stochastic vs deterministic actions: When creating an agent architecture, it has to be decided whether actions are stochastic or deterministic. According to Albrecht's paper, deterministic actions reduce the complexity of the problems, while stochastic actions account for randomness. Yet, even with the introduction of deterministic actions, trying to take on the perspective of another agent is difficult. Considering a scenario in which an agent models another agent's decision-making as a deterministic finite automaton and tries predicting its actions, the learning process is considered to be NP-complete in both the exact and approximate cases [32]. Thus, various approaches focus on reducing the complexity of problems.

Use of probabilistic models: The most commonly used models for ToM are probabilistic models [37]. Nevertheless, it is worth acknowledging that in this review, there is an even split between studies that implement probabilistic and machine learning methods, as depicted in Figure 4.

The Bayesian approach, to define shortly, is about computing the likelihood of observing someone's action and combining it with prior knowledge of beliefs of the observed agent, through which a prediction of someone's approximate behaviour, beliefs, or goals can be made [28], [37], [38]. To implement decision-making in ToM, partially observable Markov decision processes (POMDPs) are often used. In an

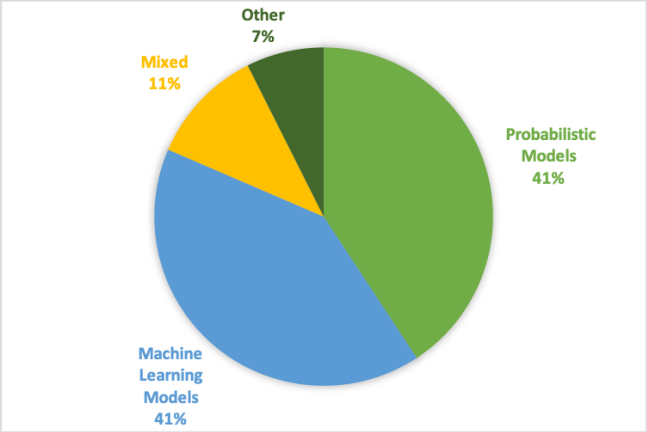

Figure 4: A chart illustrating the distribution of Theory of Mind implementations used in the selected studies. The following presents which models correspond to which studies. The studies that do not provided detailed implementation are not given here. **Probabilistic models**: [2], [4], [14], [16], [21], [22], [26], [33]–[35], [37]. **Machine learning models:** [3], [5], [12], [15], [23], [24], [30], [36], [38]–[40]. **Both:** [17], [20], [31]. **Other:** [13], [29].

environment modeled by an MDP, the agent's task involves observing the current state, evaluating the potential rewards for each available action, and selecting the optimal action based on this information. Additionally, in POMDPs, states are not directly observable, so the agent must collect information after making an action, and based on that update models of itself and of other agents [28]. Furthermore, there have been extensions developed like Interactive POMDPs (IPOMDPs), which are meant to account for other agents beliefs, desires and intentions. IPOMDPs were presented in literature in the context of deception [17]. Moreover, some propose POMDPs in the context of human group decision-making with the idea of modelling average group member's mind [16].

MDPs are also often used in Game ToM where all agents are homogeneous [41]. In this context, MDPs value function can be defined over the combined state spaces of all agents in the environment, with the assumption that the reward is the same and known to all agents [38]. In the context of Game ToM, it should be noted that game theory specific concepts like Nash Equilibrium can get integrated in ToM implementation [20], [21]. It is also worth noting that predictive models can be used to implement strategic reasoning, where there is no need for an implementation of internal states [21], [24].

Bayesian models are mainly used to handle simple situations with an uncertain environment where agents are assumed to act quite rationally [1], [37]. Langley presents some ways of further reducing the uncertainty in Bayesian environment [1].

Use of machine learning models: Some alternatives to Bayesian models include (Deep or Spiking) Neural Networks or (Inverse) Reinforcement Learning [12], [15], [37]. Neural networks require vast amounts of training data, which is be very expensive. These networks also hard to interpret due to their black-box nature [5], [12], [40]. However, they can be

trained on agents' own beliefs and on other agents' beliefs [5], [39]. The advantage of Neural Networks, is that they are often inspired by human brain architecture [5], [39].

In reinforcement learning (RL), agents make decisions based on a policy that is a combination of their world model and a reward function. The idea of inverse reinforcement learning (IRL) is to learn another agent's reward function. IRL is also considered computationally expensive [18], [23], [28], [38].

Machine learning methods are commonly combined with the probabilistic methods, as also depicted in Figure 4. For example, together with POMDP framework, there is a use made of Inverse Reinforcement Learning, with the purpose of inferring the reward function from observed behaviour [38].

Use of constraints: A theme worth discussing is the constraints on beliefs. Modelling all possible beliefs about the environment would be intractable, since often there are too many beliefs to model. Therefore, various sources introduce constraints, abstractions, or "structured representations" to simplify the belief space [20], [24], [27], [34]. Harre illustrates this concept using chess as an example: even though there are many possible combinations of chess moves in the early stages of a game, there is a limited number of chess openings used by intermediate and professional players [20]. This aligns with the idea of bounded rationality of humans [18]. Through the use of abstractions, the beliefs can be mapped and reduced to simpler representations, which can be further reduced to final ones [27].

Use of logic: There are examples in literature where logic was used to implement dynamic beliefs that are dropped when they are no longer relevant [15]. Logic can also be used to implement clauses that reduce the state space of its own actions or possible beliefs of another agent [15], [27].

Validation of predictions: Another important theme is validation of predictions. When a Bayesian agent makes a hypothesis about another agent, they might evaluate that hypothesis once the other agent completes its action. This may lead to the agent updating its parameters [28]. This idea of parameter adjustment is also seen in other literature, for example, in the context of introspection. ToM can be applied by an agent also to oneself. The prediction of the ToM output can be compared with the actual action made by the agent to adjust ToM reasoning [20]. It can be noticed, that a common way of comparing predictions to actual results is through Kullback-Leibler (KL) divergence, which measures how much information is lost when one probability distribution is used to approximate another one [12], [23], [34].

More than 2 agents in an environment: In environments where multiple agents interact and they are dependent on each other, it might be useful to consider the joint probabilities of observed agents [32]. Furthermore, there is often an assumption that when each agent models observed agents, agents who observe the same agent will have a similar model of that observed agent [16].

Parameter adjustment: In some cases, beliefs are updated with an adjustable learning rate [21]. In the example of the Strong and Weak agents described in Section 3.2, the Strong agent's influence might not always be the most optimal [34]. Thus, research proposes an idea of "degree of alterity", which means the extent to which the Strong agent's model will influence the Weak agent's model. Depending on the chosen alterity, there is a risk of over-fitting on early established states, which leads to a situation "the blind leading the blind". Alternatively, there is a risk of the Weak agent missing relevant information about the environment found by the other agent [34]. ToM can be effectively used to stabilise or reduce computational uncertainty, but the parameters have to be adjusted correctly to the specific environment that the agents are in [34]. The same idea is generalised in the situations when there is a leader and a follower [35].

Switching between ToM and strategic reasoning: Even though ToM agents can reason at higher-order levels of ToM, it might not always be necessary [22]. In some cases, it might be enough to just follow simple strategic reasoning [21]. In literature it has been proposed that agents can adjust their ToM level so that it matches the level of the opponent in competitive settings [33]. If agents decide that their ToM representation is not accurate, then they can switch to simple strategic reasoning [21], [33]. There exist complex layered architectures that allow for integrating different cognitive mechanisms, strategic reasoning and ToM [23]. It should be also noted that a higher order of ToM does not necessarily mean better results [37]. De Weerd argues that performance of higher-order ToM agents is not task-specific, but it depends on how unpredictable the environment is. Thus, depending on the complexity of the environment, an agent will achieve better results with a lower order of ToM [22].

4 Discussion and Concluding Remarks

In this paper I have shown how computational agents can make use of ToM with the purpose of perspective-taking in cooperative, competitive or mixed settings or when attempting to simulate humans. The inspirations for computational ToM have be taken from biology and neuroscience, as well as from research done in the areas of human-agent interaction. Various ideas for modelling ToM were presented tackling the issue of high computational costs. With the work done in this field, there is a hope to create agents that will be collaborative and solving more complex problems in uncertain environments.

4.1 Discussion of results

This subsection provides thorough reflection on the given sub-questions, with the aim of answering the primary research question. The first sub-question explored themes in neuroscience, biology, sociology and HAI. The second sub-question illustrated current use case of ToM for perspective-taking found in the literature. Lastly, the third sub-question discussed computational implementation of the inspirations presented in the first question. Together, the results provide an overview of how the framework of Theory of Mind has been incorporated to virtually and physically embodied agents with the ability to take perspectives of each other's points of view. Below I discuss the notable findings.

Philosophical challenges and considerations in applying ToM to multi-agent systems: Simulation Theory outlook on ToM, fits perspective-taking idea more than the Theory Theory outlook, yet Theory Theory is still used to get an understanding of someone's internal states. Further, perspective-taking seems to rely on empathy, which takes into consideration emotions and feelings. I argue that incorporating emotion recognition for the purpose of ToM in a multi-agent setting seems mostly relevant when creating human simulations, which is a task specific situation, otherwise it could be ignored. If the emotional aspect of ToM is not a focus of ToM agents, then it could be argued that ToM agents rely on cold (non-emotional) cognitive processes which, according to literature, seem to be associated more with Theory Theory. Thus, it remains ambiguous which philosophy to adopt. For multi-agent environments, further research is needed to determine what perspective-taking entails for non-emotional agents. While Simulation Theory appears preferable, in practice, Theory Theory or a combination of both (ST and TT) might be more suitable, especially given that TT is less computationally expensive.

Task-specific vs. General approach for ToM implementation: For the purposes of efficiency, researchers should consider withholding from creating general implementations and focus on task-specific ones. Task-specific implementations seem to be already computationally expensive, so quality generalised models that would encompass many different cases seem out of reach.

Integrating ToM with other cognitive inspirations: Theory of Mind should not be used on its own but in the combination with other cognitive ideas inspired by how humans think and act in social settings. Examples include: intra-group homophily, bounded rationality, goal alignment, active inference etc. Those could be applied both in the context of achieving efficiency, but also better results.

Optimal use of ToM orders in agent design: Based on the reviewed literature, it seems like most use-cases of ToM go up to the second ToM order. Even though, in general, ToM should support recursive methods, in might be unnecessary to invest in such costly recursion to go up to higher-orders of ToM. Humans themselves rarely make use of higher-order ToM, so why would agents do it? In predictive environments simple strategies seem to be enough, so switching between simple strategies and ToM seems like an implementation worth considering when designing agent architecture. Even if a task increases in uncertainty, there could be an implementation which switches between different orders of ToM.

Probabilistic models vs. Machine Learning implementations: In terms of implementation of ToM, there are ways of applying probabilistic models, or more expensive machine learning models. Currently, the probabilistic models (sometimes combined with some reinforcement learning) seem to be most commonly used, since many problems described in Section 3.2 are of low complexity. More research needs to be performed about advantages and disadvantages of different approaches.

4.2 Limitations and Future Work

Despite the comprehensive nature of this literature review, limitations must be acknowledged. Limitations regarding reproducibility and validity of this research are mentioned in Section 5.

Ambiguity in terminology regarding perspective-taking
Perspective-taking is a skill correlated with empathy, which makes it unclearly defined in the context of multi-agent systems. While a comprehensive linguistic analysis of the term and a precise definition could have been beneficial, the time constraints of this research did not allow for such an exploration. For the purposes of this paper, "perspective-taking" was interpreted as an ability to understand the view point of another agent, which is highly correlated with ToM itself. Thus, any form of ToM, whether it is TT or ST could align with this definition of perspective-taking. I would also argue that in general, the definitions applicable to humans may not clearly align with those used for computational agents. It should be noted, that for the future, more accurate distinctions between ToM and perspective-taking should be acknowledged in the multi-agent environment.

Lack of detailed description for the use cases of ToM
Even though the use cases of ToM are listed, this paper does not provide detailed explanation for the given examples. This is due to the fact that some examples need a handful of explanations. If there is a need for further explanations, the citations next to listed examples can be consulted.

Lack of in-depth analysis of probabilistic and machine learning approaches
This paper does not go in detail with different implementations of Bayesian, Inverse Reinforcement Learning, Neural Networks, Reinforcement Learning Theory of Mind. There is a need for detailed explanation of those methods, and for analysis on how those probabilistic and machine learning methods can be combined. For more information regarding different implementations, refer to the sources cited in Figure 4 in Section 3.3.

Lack of performance analysis
Multiple papers regarding ToM efficiency were excluded from this study. However, while some papers included in this research offered brief insights into performance, it would be beneficial to have an exhaustive numerical analysis of ToM performance across different scenarios in the future studies.

5 Responsible Research

This section provides a reflection on the ethical aspects of this report. When writing a report it is always essential to ensure that methods and results are reproducible, transparent and valid.

5.1 Reproducibility of this report

As mentioned in Section 2, this paper was written following PRISMA guidelines. Thus, Section 2 documents the entire process of acquiring papers for this review. I believe that given the query, the date of the search, and the clearly defined

inclusion and exclusion criteria, another researcher would analyse and extract data from the same papers and achieve similar results.

The methodology section does not provide a detailed explanation of how the data were extracted from the papers. In practice, this involved rereading the papers multiple times and summarising the content for each sub-question.

5.2 Validity and transparency of the results

After filtering abstracts, the full texts of the papers were obtained from reliable databases. Papers for which a full text could not be found were marked as not found in the process.

There is, however, a limitation regarding the search process related to the search query. For the word "interpret," a wildcard could have been used to detect more keywords like "interpretation" or "interpreting." The initial reason for not including that was due to some databases limiting the number of wildcards that can be used. In hindsight, there was no need to include that keyword at all.

Furthermore, it is worth considering that restricting the query to just the Theory of Mind might have been limiting. There might be papers that discuss ways of integrating perspective-taking for collaboration without strictly implementing ToM. Nevertheless, those papers could be useful for providing ideas on how to implement ToM for perspective-taking.

Another limitation is that some important keywords were only identified after conducting the research. For example the word "mentalizing" is used often instead of Theory of Mind [4]. In the conducted research, if paper's abstract was on mentalizing but did not directly refer to Theory Of Mind, then it was filtered out. Thus, some essential papers might have not been acknowledged. The same goes for more specific concepts like Mirror Neuron Systems, whose significance to ToM was only conducted after gathering the results.

References

[1] C. Langley, B. Cirstea, F. Cuzzolin, and B. Sahakian, "Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review," *Frontiers in Artificial Intelligence*, vol. 5, 2022. DOI: 10 . 3389 / frai . 2022 . 778852. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85128846514 & doi = 10 . 3389 % 2ffrai . 2022 . 778852 & partnerID = 40 & md5 = ed45f44108eace3efd094de8eff4a3e8.

[2] M. Persiani, "Towards we-intentional human-robot interaction using theory of mind and hierarchical task network," in *International Conference on Computer-Human Interaction Research and Applications, CHIRA - Proceedings*, vol. 2021-October, 2021, pp. 291–299. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85146197180 & partnerID = 40 & md5 = d7589484368b4f26f1ebc87fd1a98e7d.

[3] J. Williams, S. Fiore, and F. Jentsch, "Supporting artificial social intelligence with theory of mind," *Frontiers in Artificial Intelligence*, vol. 5, 2022. DOI: 10 . 3389 / frai . 2022 . 750763. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85126656853 & doi = 10 . 3389 % 2ffrai . 2022 . 750763 & partnerID = 40 & md5 = dfa16d9e237b454daaee1485b4c2e6fe.

[4] D. Alcalá-López, K. Vogeley, F. Binkofski, and D. Bzdok, "Building blocks of social cognition: Mirror, mentalize, share?" *Cortex*, vol. 118, pp. 4–18, 2019. DOI: 10.1016/j.cortex.2018.05.006. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85048345192 & doi = 10 . 1016 % 2fj . cortex . 2018 . 05 . 006 & partnerID = 40 & md5 = ae927e95a7f3765aac18c896018fc0dd.

[5] A. Gorgan Mohammadi and M. Ganjtabesh, "On computational models of theory of mind and the imitative reinforcement learning in spiking neural networks," *Scientific Reports*, vol. 14, no. 1, 2024. DOI: 10 . 1038 / s41598 - 024 - 52299 - 7. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85182809127 & doi = 10 . 1038 % 2fs41598 - 024 - 52299 - 7 & partnerID = 40 & md5 = f9609b820173c30b03965bbe350f8078.

[6] B. Lugrin, "Introduction to socially interactive agents," in *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, 1st ed. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–20, ISBN: 9781450387200. [Online]. Available: https://doi.org/10.1145/3477322.3477324.

[7] A. Boland, M. G. Cherry, and R. Dickson, Eds., *Doing a Systematic Review: A Student's Guide*. SAGE, 2014.

[8] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and T. P. Group, "Preferred reporting items for systematic reviews and meta-analyses: The prisma statement," *Annals of internal medicine*, vol. 151, no. 4, pp. 264–269, 2009.

[9] F. Bianco and D. Ognibene, "Transferring adaptive theory of mind to social robots: Insights from developmental psychology to robotics," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11876 LNAI, 2019, pp. 77–87. DOI: 10 . 1007 / 978 - 3 - 030 - 35888 - 4_8. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85076567912 & doi = 10 . 1007 % 2f978 - 3 - 030 - 35888 - 4_8 & partnerID = 40 & md5 = 642c09899284f338f44d1b58c44db8cd.

[10] N. Gurney and D. Pynadath, "Robots with theory of mind for humans: A survey," in *RO-MAN 2022 - 31st IEEE International Conference on Robot and Human Interactive Communication: Social, Asocial, and Antisocial Robots*, 2022, pp. 993–1000. DOI: 10.1109/RO-MAN53752.2022.9900662. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=

2 - s2 . 0 - 85140740397 & doi = 10 . 1109 % 2fRO - MAN53752 . 2022 . 9900662 & partnerID = 40 & md5 = 0c456b488df751e7feb5bc14af5ef11b.

[11] A. Winfield, "Experiments in artificial theory of mind: From safety to story-telling," *Frontiers Robotics AI*, vol. 5, JUN 2018. DOI: 10 . 3389 / frobt . 2018 . 00075. [Online]. Available: https://www.scopus.com/inward/record . uri ? eid = 2 - s2 . 0 - 85050078775 & doi = 10 . 3389 % 2ffrobt . 2018 . 00075 & partnerID = 40 & md5 = 027045b4e7c2699565506e837975b2c6.

[12] D. Rudrauf, G. Sergeant-Perhtuis, Y. Tisserand, T. Monnor, V. De Gevigney, and O. Belli, "Combining the projective consciousness model and virtual humans for immersive psychological research: A proof-of-concept simulating a ToM assessment," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 2, 2023. DOI: 10 . 1145 / 3583886. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85171730285 & doi = 10 . 1145 % 2f3583886 & partnerID = 40 & md5 = 2ba23b9d6844e4fcf5d41523d5416fa3.

[13] M. Schurz, J. Radua, M. Tholen, *et al.*, "Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind.," *Psychological Bulletin*, vol. 147, no. 3, pp. 293–327, 2021. DOI: 10.1037/bul0000303. [Online]. Available: https://www.scopus.com/inward/record . uri ? eid = 2 - s2 . 0 - 85101886820 & doi = 10 . 1037 % 2fbul0000303 & partnerID = 40 & md5 = 40bc8066a15ae36163e0b34274786259.

[14] L. Ouali, N. Sabouret, and C. Rich, "Guess my power: A computational model to simulate a partner's behavior in the context of collaborative negotiation," in *Advances in Intelligent Systems and Computing*, vol. 868, 2018, pp. 1317–1337. DOI: 10 . 1007 / 978 - 3 - 030 - 01054-6_92. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057128090&doi=10.1007%2f978-3-030-01054-6_92&partnerID=40&md5=c7c311c550601bcbea16faffe8bf7e92.

[15] N. Montes, M. Luck, N. Osman, O. Rodrigues, and C. Sierra, "Combining theory of mind and abductive reasoning in agent-oriented programming," *Autonomous Agents and Multi-Agent Systems*, vol. 37, no. 2, 2023. DOI: 10.1007/s10458-023-09613-w. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85168313041 & doi = 10 . 1007 % 2fs10458 - 023 - 09613 - w & partnerID = 40 & md5 = a8225f2dc3a14be08d0178dab0fb651f.

[16] K. Khalvati, S. Park, S. Mirbagheri, *et al.*, "Modeling other minds: Bayesian inference explains human choices in group decision-making," *Science Advances*, vol. 5, no. 11, 2019. DOI: 10.1126/sciadv.aax8783. [Online]. Available: https://www.scopus.com/inward/record . uri ? eid = 2 - s2 . 0 - 85075440998 & doi = 10 . 1126 % 2fsciadv . aax8783 & partnerID = 40 & md5 = 6285304b30772c161148e1649889c229.

[17] N. Alon, L. Schulz, J. Rosenschein, and P. Dayan, "A (dis-)information theory of revealed and unrevealed preferences: Emerging deception and skepticism via theory of mind," *Open Mind*, vol. 7, pp. 608–624, 2023. DOI: 10 . 1162 / opmi_a_00097. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85172471162 & doi = 10 . 1162 % 2fopmi_a_00097 & partnerID = 40 & md5 = 37d4e93725d1d0e1b4560cc1a80c37a6.

[18] A. Fuchs, A. Passarella, and M. Conti, "Modeling, replicating, and predicting human behavior: A survey," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 18, no. 2, 2023. DOI: 10 . 1145 / 3580492. [Online]. Available: https : / / www . scopus . com/inward/record.uri?eid=2-s2.0-85152768694&doi=10.1145%2f3580492&partnerID=40&md5=d3b3d845d2047752543b0974f4dc73c3.

[19] M. Ho and T. Griffiths, "Cognitive science as a source of forward and inverse models of human decisions for robotics and control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 33–53, 2022. DOI: 10.1146/annurev-control-042920-015547. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129831031&doi=10.1146%2fannurev-control-042920-015547&partnerID=40&md5=ddaf504845f1389889cf759ee3434180.

[20] M. Harré, "What can game theory tell us about an AI 'theory of mind'?" *Games*, vol. 13, no. 3, 2022. DOI: 10 . 3390 / g13030046. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85133340046 & doi = 10 . 3390 % 2fg13030046 & partnerID = 40 & md5 = 27b8b1319e0b123ec0083414c578eeac.

[21] H. De Weerd, D. DIepgrond, and R. Verbrugge, "Estimating the use of higher-order theory of mind using computational agents," *B.E. Journal of Theoretical Economics*, vol. 18, 2018. DOI: 10.1515/bejte-2016-0184. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050935461&doi=10.1515%2fbejte-2016-0184&partnerID=40&md5=c6df118fcafd70dda8915f96d5df0455.

[22] H. de Weerd, R. Verbrugge, and B. Verheij, "Higher-order theory of mind is especially useful in unpredictable negotiations," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 2, 2022. DOI: 10 . 1007 / s10458 - 022 - 09558 - 6. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85129757300 & doi = 10 . 1007 % 2fs10458 - 022 - 09558 - 6 & partnerID = 40 & md5 = 33193d1d659ce73a8c34dc63865cea4b.

[23] I. Freire, X. Arsiwalla, J.-Y. Puigbò, and P. Verschure, "Modeling theory of mind in dyadic games using adaptive feedback control," *Information (Switzerland)*, vol. 14, no. 8, 2023. DOI: 10.3390/info14080441. [Online]. Available: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85168770847 & doi = 10 . 3390 % 2finfo14080441 & partnerID = 40 & md5 = 9a0b1a1dbdd4ef08fcb51caee8211060.

[24] M. K. Ho, R. Saxe, and F. Cushman, "Planning with theory of mind," *Trends in Cognitive Sciences*, vol. 26,

no. 11, pp. 959–971, 2022, ISSN: 1364-6613. DOI: 10.1016/j.tics.2022.08.003. [Online]. Available: https://doi.org/10.1016/j.tics.2022.08.003 (visited on 2024-05-14).

[25] K. McDonald, J. Pearson, and S. Huettel, "Dorsolateral and dorsomedial prefrontal cortex track distinct properties of dynamic social behavior," *Social Cognitive and Affective Neuroscience*, vol. 15, no. 4, pp. 383–393, 2020. DOI: 10.1093/scan/nsaa053. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096661594&doi=10.1093%2fscan%2fnsaa053&partnerID=40&md5=f545dfd09539225ba06f96ece2391afb.

[26] K. Veltman, H. de Weerd, and R. Verbrugge, "Training the use of theory of mind using artificial agents," *Journal on Multimodal User Interfaces*, vol. 13, no. 1, pp. 3–18, 2019. DOI: 10.1007/s12193-018-0287-x. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058847008&doi=10.1007%2fs12193-018-0287-x&partnerID=40&md5=81db2e36fe192919f32024882046a53a.

[27] E. Erdogan, F. Dignum, R. Verbrugge, and P. Yolum, "Abstracting minds: Computational theory of mind for human-agent collaboration," in *Frontiers in Artificial Intelligence and Applications*, vol. 354, 2022, pp. 199–211. DOI: 10.3233/FAIA220199. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142189161&doi=10.3233%2fFAIA220199&partnerID=40&md5=968f5ea4519d951102aeb8acc8e62701.

[28] N. Gurney, S. Marsella, V. Ustun, and D. Pynadath, "Operationalizing theories of theory of mind: A survey," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13775 LNCS, 2022, pp. 3–20. DOI: 10.1007/978-3-031-21671-8_1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147995389&doi=10.1007%2f978-3-031-21671-8_1&partnerID=40&md5=f05298fe6e8e3408c8a296c00dccb2d6.

[29] C. Sirithunge, K. Priyanayana, H. Ravindu, *et al.*, "Tell me more! a robot's struggle to achieve artificial awareness," in *29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020*, 2020, pp. 60–66. DOI: 10.1109/RO-MAN47096.2020.9223458. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095747317&doi=10.1109%2fRO-MAN47096.2020.9223458&partnerID=40&md5=936e64ebff42c67a00846a9c45352df1.

[30] M. Sidji, W. Smith, and M. Rogerson, "The hidden rules of hanabi: How humans outperform AI agents," in *Conference on Human Factors in Computing Systems - Proceedings*, 2023. DOI: 10.1145/3544548.3581550. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160022425&doi=10.1145%2f3544548.3581550&partnerID=40&md5=7293c8aa910ff0108e8f4d9de87147a5.

[31] N. Bard, J. Foerster, S. Chandar, *et al.*, "The hanabi challenge: A new frontier for AI research," *Artificial Intelligence*, vol. 280, 2020. DOI: 10.1016/j.artint.2019.103216. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076276822&doi=10.1016%2fj.artint.2019.103216&partnerID=40&md5=e7086c3f9d2a02bc508d363739fce2f9.

[32] S. Albrecht and P. Stone, "Autonomous agents modelling other agents: A comprehensive survey and open problems," *Artificial Intelligence*, vol. 258, pp. 66–95, 2018. DOI: 10.1016/j.artint.2018.01.002. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042112383&doi=10.1016%2fj.artint.2018.01.002&partnerID=40&md5=95a8d9e9daea384f35c19dd1a0c35229.

[33] D. Kröhling, O. Chiotti, and E. Martínez, "Artificial theory of mind in contextual automated negotiations within peer-to-peer markets," *Engineering Applications of Artificial Intelligence*, vol. 120, 2023. DOI: 10.1016/j.engappai.2023.105887. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147088195&doi=10.1016%2fj.engappai.2023.105887&partnerID=40&md5=6dcef09472c77d29bcf83ce14ea91348.

[34] R. Kaufmann, P. Gupta, and J. Taylor, "An active inference model of collective intelligence," *Entropy*, vol. 23, no. 7, 2021. DOI: 10.3390/e23070830. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85109378686&doi=10.3390%2fe23070830&partnerID=40&md5=48a42a9906b998b74a23c48ad8d7bedd.

[35] J. Pöppel, S. Kahl, and S. Kopp, "Resonating minds—emergent collaboration through hierarchical active inference," *Cognitive Computation*, vol. 14, no. 2, pp. 581–601, 2022. DOI: 10.1007/s12559-021-09960-4. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120333496&doi=10.1007%2fs12559-021-09960-4&partnerID=40&md5=0f4b79cb230dd40fa6480c315a4d3f4a.

[36] T. Malloy and C. Gonzalez, "Learning to defend by attacking (and vice-versa): Transfer of learning in cybersecurity games," in *Proceedings - 8th IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2023*, 2023, pp. 458–464. DOI: 10.1109/EuroSPW59978.2023.00056. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168246847&doi=10.1109%2fEuroSPW59978.2023.00056&partnerID=40&md5=d8277db64b98b9bb38215867bce2da77.

[37] M. Patricio and A. Jamshidnejad, "Dynamic mathematical models of theory of mind for socially assistive robots," *IEEE Access*, vol. 11, pp. 103 956–103 975, 2023. DOI: 10.1109/ACCESS.2023.3316603. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85173037027&doi=10.1109%2fACCESS.2023.3316603&partnerID=40&md5=8e5ee4daafae2f81b4f9f0faa9570092.

[38] J. Ruiz-Serra and M. Harré, "Inverse reinforcement learning as the algorithmic basis for theory of mind: Current methods and open problems," *Algorithms*, vol. 16, no. 2, 2023. DOI: 10.3390/a16020068. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85148711669&doi=10.3390%2fa16020068&partnerID=40&md5=d8fcd7c767ada1ecdfe02c811a1914ee.

[39] Z. Zhao, F. Zhao, Y. Zhao, Y. Zeng, and Y. Sun, "A brain-inspired theory of mind spiking neural network improves multi-agent cooperation and competition," *Patterns*, vol. 4, no. 8, 2023, ISSN: 2666-3899. DOI: 10.1016/j.patter.2023.100775. [Online]. Available: https://doi.org/10.1016/j.patter.2023.100775 (visited on 2024-05-14).

[40] Z. Zhao, E. Lu, F. Zhao, Y. Zeng, and Y. Zhao, "A brain-inspired theory of mind spiking neural network for reducing safety risks of other agents," *Frontiers in Neuroscience*, vol. 16, 2022. DOI: 10.3389/fnins.2022.753900. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133922021&doi=10.3389%2ffnins.2022.753900&partnerID=40&md5=c9fe731811bfe112f75f3405080339f6.

[41] A. Favier, S. Shekhar, and R. Alami, "Models and algorithms for human-aware task planning with integrated theory of mind," in *IEEE International Workshop on Robot and Human Communication, RO-MAN*, 2023, pp. 1279–1286. DOI: 10.1109/RO-MAN57019.2023.10309437. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186984853&doi=10.1109%2fRO-MAN57019.2023.10309437&partnerID=40&md5=61b0c0f0a32e5bc6a4a639478c370e8e.