A Review of Theory of Mind and Robotics: Mind Reading in Human-Robot Interaction for Proactive Social Robots.

HELLOU, Mehdi, VINANZI, Samuele and CANGELOSI, Angelo

A Review of Theory of Mind and Robotics: Mind Reading in Human-Robot Interaction for Proactive Social Robots

Mehdi Hellou[1], Samuele Vinanzi[2], and Angelo Cangelosi[1]

[1] Manchester Centre for Robotics and AI,
University of Manchester, Manchester, United Kingdom,
[2] Department of Computing, Sheffield Hallam University,
Sheffield, United Kingdom

Abstract. Social robots are becoming more prevalent in our daily life and will soon become part of our society. To answer this challenge, several studies focused on how we could integrate those technologies into the current world to help humans in their daily tasks. Many areas are involved in those challenging issues, and there is a particular focus on taking inspiration from psychology studies to develop autonomous machines. One of them is interested in using Theory of Mind(ToM), a human ability that enables us to infer our own mental states and other people's mental states. However, it is a challenging task to integrate this ability in robotics. Through an analysis of the literature, we aim to provide some examples and guidelines on applying ToM in robotics to improve their behaviours. This review gives researchers a general view of the domain and insights on how they can contribute to its development.

Keywords: Theory of Mind, Human-Robot Interaction, Social Robotics, Artificial Intelligence

1 Introduction

Over the years, new technologies have appeared in our daily lives, and robots have become increasingly present in our society, indicating that interacting with them will soon be usual. In that case, it is crucial to have technologies that can personalise and adapt their behaviours to our personalities, preferences and needs. Indeed, several studies have discussed that personalisation requires fluid interactions between humans and robots, as it leads to increase user engagement during the interactions [1–3].

In response to this need, many researchers have focused on exploring the intersection of child development and psychology by analysing the notion of the ToM. ToM is a cognitive ability that humans develop during childhood to infer our own and other people's mental states, such as beliefs, desires and intentions [4]. Numerous experiments have been conducted to analyse children's social interaction, with a key focus on false-belief understanding. This exploration investigates the child's ability to recognise the moments when individuals hold beliefs that contradict reality [5, 6]. In that case, researchers introduced an experiment called the "Sally-Anne"test to evaluate ToM in false belief scenario [5], which has been adapted for use in robotics studies.

In this paper, we reviewed key literature that explicitly focus on ToM for Human-Robot Interaction(HRI) and autonomous agents. Specifically, we would like to consider papers developing cognitive models replicating ToM skills in social settings with humans. We decided to separate those papers in terms of Artificial Intelligence methods they exploit in three main classes: (1) "Probabilistic models" encompassing methods using Bayesian inference and Bayesian Networks (BNs); (2) "Deep and Reinforcement Learning models" containing methods that use Deep Learning (DL) models with Neural Networks and Reinforcement Learning (RL) models; and (3) "Hybrid models" which includes models using various methods going from probabilistic to DL models to implement ToM in their robotics systems. To conclude, we provide pros and cons for each main idea and discuss the points to improve in the future.

2 Probabilistic models

The first models designed to approach ToM used BNs, a graphical model for data analysis and a popular representation for encoding uncertain expert knowledge in expert systems [7]. It was suggested that these models could effectively represent the knowledge and learning of infants, wherein it was argued that they represent the world by using a causal map[8]. Among the first attempts, Goodman et al. [9] proposed two causal Bayesian models to illustrate false belief understanding from children. The first model, known as copy theorist (CP), suggested that only the world affects beliefs. The second model, perspective theory (PT), proposes that perception and the world jointly influence the causal relation to change beliefs, which can sometimes be false. The authors compared their models' predictions to children's predictions and explanations in the Sally-Anne test with an exchange-item-location task. Following these ideas, Asakura et al. [10] decided to introduce a new BN by taking into account the concept of simulation theory [11], arguing that children make predictions through a simulating process by imagining being in the position of someone else. As a test, the authors used a determined Band-Aid box's content experiment, where the child needs to predict whether the box will include Band-Aids or another item. The BN is built with three types of variables: the World W representing the true box content's state, the Visual access V indicating whether the child can observe what is inside of the box or not, and the Belief B of the child. The particularity of their models was to design the causal relation between the mental states of the child and others in order to illustrate the children's ability to detect false beliefs and represent their formations during the unexpected content task. Figure 1 displays their proposed BN, referred to as Model (b). They drew inspiration from Model (a) in [9], incorporating two internal mental states: the children's own mental states (V_S and B_S), and the mental states of others (V_O and B_O). These papers serve as

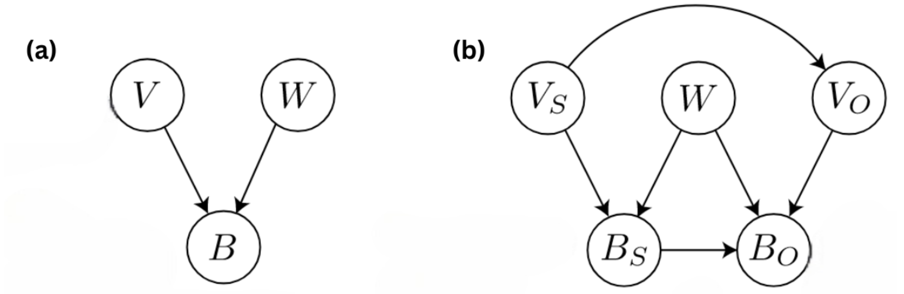

Fig. 1. BNs models from [9](a) and [10](b). Each model depicts the causal relation between the world W, visual access V, and belief B. The model (b) from [10] took inspiration from model (a) which follows the PT's assumptions, by indicating two internal models: the S for the mental states of "self", and the O for the mental states of "others".

valuable resources for understanding the design of computation models for ToM and their applications in child developmental studies.

Probabilistic models utilised for robotics are less present compared to other models. Vinanzi et al. [12] investigated using a BN to build a robot's ToM model for trust in HRI. Inspired by an experiment conducted by Vanderbilt et al. [13], the robot has to ask the help of a participant to find a sticker hidden in two possible positions. As part of the experiment, the informants could either indicate the correct position or the wrong one, and the BN then learns to build trustworthiness about the informants. After a familiarisation phase to mature the robot's ToM model, the robot's cognitive model was put to the test in two phases. In the decision-making phase, the robot has to trust the informant according to the previous interactions when indicating a position to find the stickers. The last phase, the belief estimation phase, was introduced to assess the model's ability to track the informants' beliefs about the sticker's location and predict their suggestions in case the robot asks for help. In addition to their BN, the authors introduced an episodic memory module to apprehend surprises or betrayals from the users.

At the frontier of robotics and autonomous agents, researchers have explored the ability to understand the mental states of virtual agents using the principles of ToM. More specifically, they have studied the intentions of goal-oriented agents to explain their behaviours. Pantelis et al. 2014 [14] built a model to infer the mental states of autonomous agents based on their observable actions in a two-dimensional virtual environment. These agents, called Independent Mobile Personalities (IMPS), navigated in a survival environment with their limited sensory capabilities and were driven by their beliefs, intentions, and one of four distinct goals: attacking, exploring, fleeing, and gathering food. These goals were considered the agent's mental state. The authors implemented a Dynamic Bayesian Network (DBN), a BN that tracks and predicts uncertainty over time, portraying how agents' goals and environments could impact their actions. With Bayesian inference and inverting the entities' plans, the model could predict the potential goals of the agents based on their actions. In a test, human participants were asked to observe different scenes with four virtual agents and estimate their goals. Afterwards, they compared the human judgements to the model predictions to identify analogies between both inferences.

Zhi-Xuan et al. [15] proposed a Bayesian model for boundedly rational planners capable of inferring an agent's goals through optimal and non-optimal sequences of actions. In doing so, they developed a Sequential Inverse Plan Search (SIPS), a probabilistic program performing Bayesian inference over the agent's plan to determine its goal. The model uses Inverse Reinforcement Learning (IRL) to observe the agent's behaviour and invert its plans to understand its intention. Both successful and failed plans were considered for online inference of the agent's intentions. Their experiments took problems from four domains in simulated environments, one of which involved an agent navigating in a maze with doors, keys, and gems. The agent aimed to collect one of the gems in the environment, each of which was locked behind one or more doors that could be opened with the keys. The model was able to predict the agent's interest regarding the gems as it moved through the maze, even when the agent could not reach its goal. Similar to [14], the authors compared the model's results to human inferences in order to investigate the similarities. The idea was to evaluate whether the model's ToM embodiment exhibited a similar process to human reasoning in those scenarios.

Those probabilistic methods accurately describe how the human mind works, especially in inferring the mental states of entities by observing their behaviours. However, those models are limited in their ability to generalise to situations beyond those they have been trained on, which is a known challenge in modern AI studies.

3 Deep Learning and Reinforcement Learning models

In this section, we explore data-driven models, including DL and RL methods, which have become a significant part of AI sectors utilised in various domains such as industries, healthcare or even public settings such as museums and restaurants. Researchers have investigated different ways to implement ToM by using those methodologies, which require training on a consequent amount of data if they want to learn. Rabinowitz et al. [16] established one of the well-known ToM neural networks called *ToMnet*, which aims to predict agents' next behaviours according to a set of their previous behavioural actions. The authors argue that ToM is essentially an issue of "meta-learning" or "learning to learn" [17], which can be summarised as the ability of an agent to choose the appropriate algorithm to learn according to the issues it's facing. They described *ToMnet* with two central concepts: (1) the ability of the network to estimate common behaviours on a set of agents by learning the weights of the network, and (2) the learning of an agent-specific character and mental state that makes it distinct from others. The model was then evaluated in several simulations that involved designing agents moving in a grid world with distinct behaviours. For example, they might have different environmental goals, or the environment might change during the process. The latter was executed to test the model's ability to detect when agents held false beliefs. The agents could differ from their observation spaces or the algorithms they used, going from random policies to policies learned with deep RL. In their results, the authors demonstrated that with meta-learning, the *ToMnet* could flexibly learn behaviour from different agents while generalising the processes controlling the agents' decision-making. In the future, they would like to adapt the cognitive model to more complex settings, such as 3D environments, and add real design to the observers' visibility, which was open in their situations.

Following this methodology, Oguntola et al. [18] developed a similar cognitive ToM model based on the belief, desires and intentions (BDI) model, a concept well-defined in folk psychology [19]. This model was designed to infer the mental states of an observed entity, similar to human reason and behaviour. In doing so, the authors divided their models into three main modules: a "Belief Model" to illustrate the agent's beliefs and update them according to the observation space, a "Desire Model" to compute the agent's intent given the updated belief, and an "Action Model" which is responsible for generating or predicting the agent's actions given the beliefs and intents. They exploited a rule-based model to interpret and update the belief states according to the observations. On the other hand, they used encoder-decoder neural networks for the desires and inverse-action models inspired by U-Nets [20]. Additionally, they incorporated the "concept whitening" to modify the neural network's layers [21] and increase its interpretability. To evaluate their model, they built a simulated search and rescue task in a Minecraft environment, wherein human participants were asked to rescue as many victims as possible. The model predicted the participants' actions regarding different concepts, such as their intent to rescue the victims present in the field of view and their injuries (critical or non-critical).

Another big field in Machine Learning methods is the study of RL models utilised in different issues, especially focusing on making multi-agent systems to learn how to perform complex tasks in a specific environment, usually grid-types. Researchers investigated these methods to demonstrate ToM-like behaviours with those agents. Freire et al. [22] proposed a model-free RL called Control-based Reinforcement Learning (CRL) implemented with two agents competing against each other in a 2D environment to obtain rewards with different values. The agents were simulated as mobile robots that can move and sense the elements within the environment, including the presence of different rewards. The baseline ToM model consists of two layers: a "Reactive Layer", working at a low level to make the agent learn how to use its sensors and actuators to detect the rewards and escape from the opponent. A second layer incorporating a Temporal-Difference RL algorithm (TD-learning) [23], to learn a policy strategy for maximising the rewards given the environmental state. The interaction between these layers formed the model's foundation, enabling the agents to select the best course of action. Moreover, the authors extended this basic model by implementing two additional versions. The first variant is responsible for inferring the other agent's actions, while the second version has a different strategy. Specifically, it predicts the outcome of the other agent by simulating its behaviours. Although not explicitly discussed in the paper, it could be argued that these versions try to mimic the main theories introduced by psychologists in ToM. Specifically, the second model's version can be linked to the "simulation theory" principles. Mentioned in Section 2, this theory states that ToM is carried out through a process of role-taking to determine the mental state of someone. By contrast, the first model follows the idea in "theory-theory", suggesting that ToM resides in applying causal relations in the world with human psychology, enabling someone to predict the mental states of others [24]. More precisely, how a person perceives the world and the actors interacting and moving within the environment greatly influence their attributions of mental states to these different actors. Following these two theoretical approaches, the authors conducted an experiment comprising two 2D robots in a simulated environment to catch a reward out of two options with different values. With the help of their sensors, they could detect whether the reward was high or low within a certain range. If both robots reached the same goal, no points were awarded. On the other hand, if one robot reached a reward, it received the points attached to it, while the other received the points attached to the alternative reward.

In the field of HRI, Zeng et al. [25] proposed a Brain-inspired Model of Theory of Mind (Brain-ToM model), which was implemented in a humanoid robot performing two false belief tasks. The model uses a Spiking Neural Network (SNN), an Artificial Neural Network where neurons process temporal inputs and outputs. More specifically, the authors were inspired by the Voltage-driven Plasticity-centric SNN [26], a methodology to feed-forward SNN. As a test-bed, the authors conducted an HRI experiment with two NAO robots from Aldebaran United Robotic Group, a small humanoid robot predominantly used for HRI research. Two false belief tasks were conducted during the experiment: "Opaque-and-Transparent Blindfold Test"(OTBT) and "Turn Around Test"(TAT). The settings of both experiments were the same, with slight differences between them. It included two robots facing each other across with a ladybird and two boxes (green and yellow). One robot played the actor role, and the other one played the role of the participant. During the experiment, the ladybird was hidden in the yellow box. The actor robot was then blindfolded with a transparent or opaque blindfold, and an experimenter changed the ladybird's location to the green box. Finally, the blindfold was removed, and the participant robot was asked "in which box the actor robot thinks the ladybird is" and "where the real location of the ladybird is". Using the Brain-ToM model to detect and recognise the object, the model predicted the robot's action according to the question and the environmental inputs. The authors then tested the model with the same setting but changed the action of the actor robot turning its head around to avoid seeing the table. This second experiment considered whether the robot turned its head (False Belief condition) or not (True Belief condition). The results from both experiments were compared with children's performance in similar trials.

In summary, the literature demonstrates that only a few data-driven models specifically address ToM's challenge. Substantial ToM cognitive models combine multiple independent architectures, a topic we will explore in the next section.

4 Hybrid models

As pointed out in the last paragraph, an essential part of ToM computational models is built with different modules, leading to the emergence of hybrid models with distinct functions to infer particular mental states and their interactions. Patacchiola et al. [27] created a cognitive architecture for trust and ToM in humanoid robots. This model integrates an Actor-Critic framework to interpret biological observations combined with a model called ERA [28], an architecture gathering self-organising maps (SOM) as functions approximator, and a BN to represent the intrinsic environment's values. The model is used to compute trust in object-naming learning, assisting the robot in learning new objects with human assistance. The model is divided into three main parts:

- **Actor**: the ERA model used the SOMs to cluster the information from the speech and vision received from the robot's sensor and convert them into beliefs.
- **Intrinsic environment**: The BN represents the cognitive model's internal module to define trust. It uses the beliefs and computes a negative, positive, or null cost to bootstrap the RL mechanisms and build an internal reward model when external rewards are unavailable.
- **Critic**: The Critic uses the cost from the BN to update functions and the Actor's functions essential to generate the appropriate actions for the robot.

In their proposed methods, the model does not utilise external rewards from the environments. Instead, the reward is computed from the BN, which outputs a cost based on the predicted beliefs. Two experiments were set up to evaluate the application of the model. The first included a simulation of a trust experiment from Vanderbilt et al. [13] building by using Python. This involved a scenario where a child learned to trust an adult according to their answers in a hiding-sticker game. The child observed an adult consistently pointing to the correct location of a sticker and then observed another adult purposely pointing to the wrong location. This process helped build trust in the reliable adult and distrust in the unreliable adult. The same method was used in the simulation to build the trust module for the robot. In the second experiment, the researcher implemented an object name learning task for the robot with the help of two humans: one who gave the right name for the object and one who did not. The robot had to learn which one of the humans to trust through a training trial (familiarisation) and make its decision in the test trial (*explicit judgement* and *endorsement*). During the *endorsement* trial, two informants introduced a new object to the robot by proposing new names. Based on the level of trust for each informant built during the training trial, the robot chose the appropriate name for the object.

As an extension of their previous work [12], Vinanzi et al. [29] proposed a model for intention reading and trust in human-robot collaboration. They implemented an agent possessing the ability to infer the non-verbal intentions of others and to evaluate how likely they are to achieve their goals, jointly understanding what kind and which degree of collaboration they require. The model proposed is divided into two main modules. One is responsible for reading the intentions of people, and the other module is for the robot's trustworthiness toward the human when participating in the collaborative task. For the evaluation, the authors set up a collaborative game wherein the robot and the user must work together to align blocks with different colours in specific settings. In their setting, a goal sequence represents a full line of 4 coloured blocks. During the task, if the human was trusted or the partial sequence was valid, the robot would collect the next predicted blocks and hand them over to the user. In the case of an unreliable collaborator, the robot would position the blocks on the building area in what it considered the correct order, attempting to rectify the errors that had been committed. The robot would also explain why the partial sequence was invalid in the latter case. This experiment is among the few in the literature that embodies a ToM computational model into a robot and evaluates it in a human-robot collaboration task.

In another domain, Baker et al. [30, 31] introduced a well-known model for ToM called the Bayesian Theory of Mind (BToM). As a continuation of their study in understanding human action using Bayesian inverse planning [32], BToM is a probabilistic model used to infer people's goals by observing their behaviours. By conceiving their problem as a partial observable Markov decision process (POMDP) and combining it with Bayesian inference, the model aims to explain the behaviour of autonomous agents by predicting their beliefs and preferences. The model is based on the "principal of rational action" arguing that agents choose their actions to maximise a reward function regarding their desires. In doing so, they are using their observations about the environment and, hence, what they believe is the true state of the environment. These ideas are formulated as POMDPs to represent the agents' planning, incorporating Bayesian inference to capture agents' beliefs and to update them. By inverting the agent's plans, BToM succeeds in extracting its beliefs and preferences. For example, the authors built a 2D-grid environment with different food trucks in which an agent was placed. The agent's goal was to find a place to eat based on their preferences and the availability of foods among three options: Korean, Lebanese and Mexican. Additionally, the agent could only observe a partial section of the environment, which increased the complexity of the problem. In that setting, BToM was utilised to track the agent's mental state according to (1) the agent's preference regarding the three food options and (2) its belief about the food trucks' locations. Subsequently, they compared the model's result to human participants on the predictions of those mental states in order to evaluate the similarities between human judgement and the model's inferences. They also tested another scenario to demonstrate the versatility of their model, illustrated by a similar scenario. Unlike the first experiment, the foods' preferences were known, and their locations were unknown during the interaction. To add further complexity, certain food trucks could be closed or open. While the agent moved in the environment, human participants and the BToM predicted the carts' location. Some researchers have pursued their proposed ideas, exploiting BToM to different challenges where ToM is a key point to solve their issues.

For instance, Poppel et al. [33] proposed the hierarchical active inference for collaborative agents (HAICA), a combination of BToM processes with a perception-action system based on predictive processing and active inference. With this model, the authors aimed to allow an agent to infer other agents' mental states, such as intentions and goals, which they called "belief resonance", and adapt their behaviour to collaborate and achieve a common task. The collaborative scenario is from the Overcooked game, a 2D environment where multiple agents work together to prepare as many ordered meals as possible in a limited time. Utilising ToM to enhance collaboration between the chef agents seems appropriate given the setting. To optimise the efficiency of an agent's task, HAICA is divided into two main parts, illustrating the mental states of the concerned agent and the mental of the other agent. Each main structure comprises two hierarchical layers that display the agent's intention and goal. The goal's layer determines the agent's main task that it wants to achieve, such as preparing the soup. This part is crucial to deciding on the necessary steps to achieve the goal, which are specified in the low-level layer called the intention layer. Observation inputs continuously update the agent's intention and goal, computing predictions to produce the agent's actions. Complementing the first part, a mentalising component based on the BToM process provides information about the other agent's goal and intention through observations.

In another study, Hellou et al. [34] experienced BToM in HRI situations to improve the behaviours of social robots by integrating advanced cognitive skills. They were particularly interested in developing helping behaviours in false belief scenarios, highlighting the importance of using ToM to enhance the robot's skills. The researchers designed a cognitive architecture suitable for humanoid robots, which comprised different modules they could use to interpret users' mental states, such as vision. Their experiment replicated a psychology study by Buttelam et al. [35], evaluating the ability of children to help adults when they were holding false beliefs in a swapping toys' location situation. In their experiment, the researchers used BToM to infer the preferences and beliefs of a human interacting with different toys and placing them in specific boxes. By intentionally changing the toys' locations, the authors wanted to create false beliefs in the person's mind and analyse the robot's answer. The initial experiment was realised by simulation to validate the performance of the computational ToM model in such situations and its use in real-time experiments. Finally, a demonstration with the humanoid robot Pepper from Aldebaran United Robotics Group confirmed the robot's cognitive skills to assist users in this particular scenario. The authors claimed that this model could be operated in more complex environments.

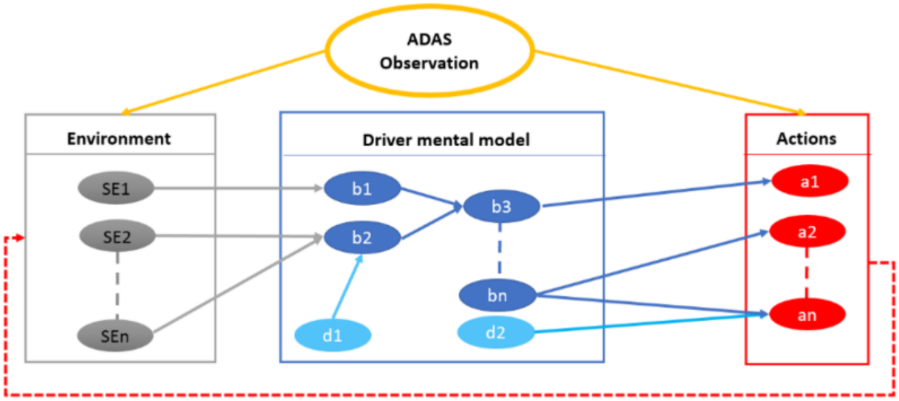

Fig. 2. The cognitive model's architecture proposed in [36] to interpret a driver's mental state by using BToM. The central part (in blue) determines the mental states, including the beliefs and preferences. The mental states influence the actions of the drivers (in red), which directly impact the environment's states (in grey). Here, the ADAS uses the information on the drivers to improve their driving experiences according to their beliefs and preferences.

In a more realistic scenario, Darwish et al. [36] took advantage of BToM and POMDPs to learn the behaviours of drivers and assist them in improving their driving experiences through an advanced driver assistance system (ADAS). Their main model, illustrated in Figure 2, predicts a specific driver mental model through the BToM (in blue) according to the environment's state (in grey) and the driver's actions (in red), having a direct impact on the environment. Combined with an artificial neural network (ANN), the BToM aims to determine the driver's preferences or driving styles, which could be aggressive or conservative, and their knowledge (beliefs) about acceptable actions regarding the situations. The authors experienced their modified BToM with the driving simulator CARLA [37], including a deep RL method to adapt the speed control system of the car. The proposed cognitive model is used to modify the speed control system by suggesting personalised speed based on drivers' preferences while ensuring their safety. To evaluate the versatility of their model, the authors altered the environment by adjusting the weather. As a future direction, they plan on adapting more elements from the environment, such as the type of road or the traffic.

Lee et al. [38] exploited the BToM to adapt the nonverbal behaviours of a social robot in a storytelling scenario. They conducted a user-robot experiment where children and parent participants were recruited to tell a story to a robot. During the interaction, the cognitive model adapts the robot's attention while listening to the children to manipulate their beliefs. The author's idea was to evaluate the robot's influence on the storytellers' behaviours, whose role was to catch their attention. Furthermore, they assessed the participants' awareness of the robot's behaviours in terms of human likeness and intelligence. The results showed that the BToM model, which adjusted the robot's attention, was perceived as more human-like and acceptable by the parents compared to a model with a consistently high level of attention throughout the interactions.

5 Discussion

For many years, ToM has been the subject of numerous studies and hypotheses from psychologists in cognitive development who seek to understand its process in the human mind. These studies have led to the emergence of different theories, such as "theory-theory" [24] or "simulation theory" [11], which provide explanations for the cognitive components involved in ToM. As a result, we have gained a deeper understanding of the functioning of the human mind and fundamental mental states, such as desires, perceptions, beliefs, knowledge, thoughts, intentions, and feelings. They also contributed to the development of several experiments, which are still relevant nowadays, such as false belief understanding experiments, and have also been helpful in the creation of cognitive computational models for autonomous systems.

In this current literature review, we explored some of these methods to replicate ToM for autonomous systems such as robots. Among these methods, probabilistic models are the first ones exploited, and a large number of them are based on the different techniques proposed by Bayesian inference. This was also the technique that psychologists used to explain the functioning of ToM in children, making it an important starting point for researchers in the field. It also goes hand in hand with how the human mind works to represent the belief, which can intuitively be seen as a Bayesian problem. BN was the main model that the investigators used due to their versatility, ease of creation, and good performance, especially in tracking people's beliefs. It can integrate time dependence with DBN and demonstrate a high level of cognition, such as trust in HRI. However, their structures can only be designed for specific problems, showing their limitations in terms of generalisation to more significant topics. To address this limitation, BN can be combined with other AI models, which can bring additional abilities to build complete cognitive architectures. Those cognitive architectures can resolve the issue of generalisation by using data-driven models, including DL and RL with neural networks, which have been popular in the field of AI in the past decades.

However, DL models are limited due to the lack of data, especially in scenarios involving false belief understanding. RL models can use simulation to overcome data limitations. Additionally, many problems have been designed as POMDPs to represent the BDI of an agent and extract those mental states to reason on the agent's behaviour. It was especially efficient with hybrid models that use strong points from the different categories pointed out earlier. For instance, BToM successfully uses POMDPs combined with Bayesian inference to track the beliefs and desires of an autonomous agent. By successfully replicating human-like cognitive processes, BToM has been subject to different applications, particularly HRI for social robots. However, the models mentioned above have mainly been learnt in particular settings, such as 2D grid environments, challenging their integration into the real world. Some attempts have been made to embody ToM skills in an autonomous machine, specifically with humanoid robots, to investigate whether the user of ToM could improve their behaviours in social interactions with humans. These studies included integrating BToM into HRI experiments to produce proactive behaviours for social robots in complex situations, including false belief situations, and exploring the use of ToM for trust in human-robot collaboration. It can be particularly useful when robots and humans are working together, such as industrial robotic arms operating on lines with the help of human workers, or social robots helping nurseries or hospitals. Some researchers decided to investigate the impact on real user-case scenarios such as the development of social robots for educational purposes with Lee et al. [38], or the improvement of driving skills with personalised advanced driver assistance system [36].

While there have been some efforts to apply ToM in real-life scenarios, especially for robotics, more studies are needed to assess how these cognitive skills can be effectively integrated with machines and their impact on human perspectives.

6 Conclusion

In this concise literature review, we outlined several helpful examples for researchers interested in taking inspiration from psychological studies, particularly in child developmental psychology, to improve machine behaviour. We provided details about the different techniques used in the AI domain, categorising them into three main sections to clarify their pros and cons. We hope this might assist researchers in choosing appropriate methods for their problems and motivate them to create new models replicating ToM abilities. Finally, we discussed the limitations encountered when attempting to connect ToM and Robotics, making those issues challenging for the following years.

Acknowledgment

The PERSEO project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778.

References

1. B. Irfan, A. Ramachandran, S. Spaulding, D. F. Glas, I. Leite, and K. L. Koay, "Personalization in long-term human-robot interaction," in *2019 14th ACM/IEEE*

International Conference on Human-Robot Interaction (HRI), pp. 685–686, IEEE, 2019.

2. M. Hellou, N. Gasteiger, J. Y. Lim, M. Jang, and H. S. Ahn, "Personalization and localization in human-robot interaction: A review of technical methods," *Robotics*, vol. 10, no. 4, p. 120, 2021.
3. N. Gasteiger, M. Hellou, and H. S. Ahn, "Factors for personalization and localization to optimize human–robot interaction: A literature review," *International Journal of Social Robotics*, vol. 15, no. 4, pp. 689–701, 2023.
4. D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behavioral and Brain Sciences*, vol. 1, no. 4, p. 515–526, 1978.
5. S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind" ?," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.
6. H. M. Wellman, D. Cross, and J. Watson, "Meta-analysis of theory-of-mind development: The truth about false belief," *Child development*, vol. 72, no. 3, pp. 655–684, 2001.
7. D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," in *Uncertainty Proceedings 1994* (R. L. de Mantaras and D. Poole, eds.), pp. 293–301, San Francisco (CA): Morgan Kaufmann, 1994.
8. A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks, "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets.," 2004.
9. N. D. Goodman, C. L. Baker, E. B. Bonawitz, V. K. Mansinghka, A. Gopnik, H. Wellman, L. Schulz, and J. B. Tenenbaum, "Intuitive theories of mind: a rational approach to false belief," in *Proceedings of the Twenty-Eigth Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum*, 2006.
10. N. Asakura and T. Inui, "A bayesian framework for false belief reasoning in children: a rational integration of theory-theory and simulation theory," *Frontiers in Psychology*, vol. 7, p. 221010, 2016.
11. P. L. Harris, "From simulation to folk psychology: the case for development.," *Mind & Language*, 1992.
12. S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi, "Would a robot trust you? developmental robotics model of trust and theory of mind," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, p. 20180032, 2019.
13. K. E. Vanderbilt, D. Liu, and G. D. Heyman, "The development of distrust," *Child development*, vol. 82, no. 5, pp. 1372–1380, 2011.
14. P. C. Pantelis, C. L. Baker, S. A. Cholewiak, K. Sanik, A. Weinstein, C.-C. Wu, J. B. Tenenbaum, and J. Feldman, "Inferring the intentional states of autonomous virtual agents," *Cognition*, vol. 130, no. 3, pp. 360–379, 2014.
15. T. Zhi-Xuan, J. Mann, T. Silver, J. Tenenbaum, and V. Mansinghka, "Online bayesian goal inference for boundedly rational planning agents," *Advances in neural information processing systems*, vol. 33, pp. 19238–19250, 2020.
16. N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, "Machine theory of mind," in *International conference on machine learning*, pp. 4218–4227, PMLR, 2018.
17. J. Schmidhuber, J. Zhao, and M. Wiering, "Simple principles of metalearning," 1996.
18. I. Oguntola, D. Hughes, and K. Sycara, "Deep interpretable models of theory of mind," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 657–664, IEEE, 2021.

19. M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge, “The belief-desire-intention model of agency,” in *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL’98 Paris, France, July 4–7, 1998 Proceedings 5*, pp. 1–10, Springer, 1999.
20. O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.
21. Z. Chen, Y. Bei, and C. Rudin, “Concept whitening for interpretable image recognition,” *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.
22. I. T. Freire, J.-Y. Puigbò, X. D. Arsiwalla, and P. F. Verschure, “Modeling the opponent’s action using control-based reinforcement learning,” in *Biomimetic and Biohybrid Systems: 7th International Conference, Living Machines 2018, Paris, France, July 17–20, 2018, Proceedings 7*, pp. 179–186, Springer, 2018.
23. R. S. Sutton, “Learning to predict by the methods of temporal differences,” *Machine learning*, vol. 3, pp. 9–44, 1988.
24. J. H. Flavell, “Cognitive development: children’s knowledge about the mind.,” *Annual review of psychology*, vol. 50, pp. 21–45, 1999.
25. Y. Zeng, Y. Zhao, T. Zhang, D. Zhao, F. Zhao, and E. Lu, “A brain-inspired model of theory of mind,” *Frontiers in Neurorobotics*, vol. 14, p. 60, 2020.
26. T. Zhang, Y. Zeng, D. Zhao, and M. Shi, “A plasticity-centric approach to train the non-differential spiking neural networks,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
27. M. Patacchiola and A. Cangelosi, “A developmental cognitive architecture for trust and theory of mind in humanoid robots,” *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1947–1959, 2022.
28. A. F. Morse, J. de Greeff, T. Belpeame, and A. Cangelosi, “Epigenetic robotics architecture (era),” *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 325–339, 2010.
29. S. Vinanzi, A. Cangelosi, and C. Goerick, “The collaborative mind: intention reading and trust in human-robot interaction,” *Iscience*, vol. 24, no. 2, 2021.
30. C. L. Baker, R. R. Saxe, and J. B. Tenenbaum, “Bayesian theory of mind: Modeling joint belief-desire attribution,” in *In Proceedings of the Thirtieth Third Annual Conference of the Cognitive Science Society*, pp. 2469–2474, 2011.
31. C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, “Rational quantitative attribution of beliefs, desires and percepts in human mentalizing,” *Nature Human Behaviour*, vol. 1, no. 4, p. 0064, 2017.
32. C. L. Baker, R. Saxe, and J. B. Tenenbaum, “Action understanding as inverse planning,” *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.
33. 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.
34. M. Hellou, S. Vinanzi, and A. Cangelosi, “Bayesian theory of mind for false belief understanding in human-robot interaction,” in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1893–1900, IEEE, 2023.
35. D. Buttelmann, H. Over, M. Carpenter, and M. Tomasello, “Eighteen-month-olds understand false beliefs in an unexpected-contents task,” *Journal of experimental child psychology*, vol. 119, pp. 120–126, 2014.

36. A. Darwish and H. J. Steinhauer, “Learning individual driver’s mental models using pomdps and btom,” in *Proceedings of the 6th International Digital Human Modeling Symposium*, pp. 51–60, 2020.
37. A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “Carla: An open urban driving simulator,” in *Conference on robot learning*, pp. 1–16, PMLR, 2017.
38. J. J. Lee, F. Sha, and C. Breazeal, “A bayesian theory of mind approach to nonverbal communication,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 487–496, IEEE, 2019.