# Robot, did you read my mind? Modelling Human Mental States to Facilitate Transparency and Mitigate False Beliefs in Human-Robot Collaboration

ANGELOPOULOS, Georgios <http://orcid.org/0000-0001-9866-8719>, HELLOU, Mehdi <http://orcid.org/0000-0002-7502-3130>, VINANZI, Samuele <http://orcid.org/0000-0003-0241-9983>, ROSSI, Alessandra <http://orcid.org/0000-0003-1362-8799>, ROSSI, Silvia <http://orcid.org/0000-0002-3379-1756> and CANGELOSI, Angelo <http://orcid.org/0000-0002-4709-2243>

**Citation:**

ANGELOPOULOS, Georgios, HELLOU, Mehdi, VINANZI, Samuele, ROSSI, Alessandra, ROSSI, Silvia and CANGELOSI, Angelo (2025). Robot, did you read my mind? Modelling Human Mental States to Facilitate Transparency and Mitigate False Beliefs in Human-Robot Collaboration. ACM Transactions on Human-Robot Interaction, 15 (1): 1, 1-29. [Article]

# Robot, Did You Read My Mind? Modelling Human Mental States to Facilitate Transparency and Mitigate False Beliefs in Human–Robot Collaboration

GEORGIOS ANGELOPOULOS, Interdepartmental Centre for Advances in Robotic Surgery, University of Naples Federico II, Napoli, Italy
MEHDI HELLOU, Manchester Centre for Robotics and AI, University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland
SAMUELE VINANZI, School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield, United Kingdom of Great Britain and Northern Ireland
ALESSANDRA ROSSI and SILVIA ROSSI, Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Napoli, Italy
ANGELO CANGELOSI, Manchester Centre for Robotics and AI, University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland

Providing a robot with the capabilities of understanding and effectively adapting its behaviour based on human mental states is a critical challenge in Human–Robot Interaction, since it can significantly improve the quality of interaction between humans and robots. In this work, we investigate whether considering human mental states in the decision-making process of a robot improves the transparency of its behaviours and mitigates potential human's false beliefs about the environment during collaborative scenarios. We used Bayesian inference within a Hierarchical Reinforcement Learning algorithm to include human desires and beliefs into the decision-making processes of the robot, and to monitor the robot's decisions. This approach, which we refer to as Hierarchical Bayesian Theory of Mind, represents an upgraded version of the initial Bayesian Theory of Mind, a probabilistic model capable of reasoning about a rational agent's actions. The model enabled us to track the mental states of a human observer, even when the observer held false beliefs, thereby benefiting the collaboration in a multi-goal task and the interaction with the robot. In addition to a qualitative evaluation, we conducted a between-subjects study (110 participants) to evaluate the robot's perceived Theory of Mind and its effects on transparency and false beliefs in different settings. Results indicate that a robot

Georgios Angelopoulos and Mehdi Hellou contributed equally to this research.

This study was supported by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Actions (Grant No. 955778).

Authors' Contact Information: Georgios Angelopoulos (corresponding author), Interdepartmental Centre for Advances in Robotic Surgery, University of Naples Federico II, Napoli, Italy; e-mail: georgios.angelopoulos@unina.it; Mehdi Hellou, Manchester Centre for Robotics and AI, University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland; e-mail: mehdi.hellou@manchester.ac.uk; Samuele Vinanzi, School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield, United Kingdom of Great Britain and Northern Ireland; e-mail: s.vinanzi@shu.ac.uk; Alessandra Rossi, Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Napoli, Italy; e-mail: alessandra.rossi@unina.it; Silvia Rossi, Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Napoli, Italy; e-mail: silvia.rossi@unina.it; Angelo Cangelosi, Manchester Centre for Robotics and AI, University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland; e-mail: angelo.cangelosi@manchester.ac.uk.

ACM 2573-9522/2025/8-ART1
https://doi.org/10.1145/3737890

which considers human desires and beliefs increases its transparency and reduces misunderstandings. These findings show the importance of endowing Theory of Mind capabilities in robots and demonstrate how these skills can enhance their behaviours, particularly in human–robot collaboration, paving the way for more effective robotic applications.

CCS Concepts: • **Human-centered computing** → **User centered design**; **User studies**; • **Computer systems organization** → **Robotic autonomy**;

Additional Key Words and Phrases: Theory of Mind, Transparency, HRI, False Beliefs, Cognitive Robotics

**ACM Reference format:**
Georgios Angelopoulos, Mehdi Hellou, Samuele Vinanzi, Alessandra Rossi, Silvia Rossi, and Angelo Cangelosi. 2025. Robot, Did You Read My Mind? Modelling Human Mental States to Facilitate Transparency and Mitigate False Beliefs in Human–Robot Collaboration. *ACM Trans. Hum.-Robot Interact.* 15, 1, Article 1 (August 2025), 29 pages.
https://doi.org/10.1145/3737890

## 1 Introduction

As robots become increasingly part of our daily lives, from manufacturing to personal assistance, there's a growing imperative for these machines to understand and respond to human behaviours more intuitively and human-likely [23]. A robot that acts like a human is easier to interact with, as humans are naturally good at communicating and interacting with other humans [34, 57]. A key challenge in this endeavour is the integration of human mental states into robotic behaviour, a concept often referred to as **Theory of Mind (ToM)**. ToM is a cognitive ability that allows individuals to understand and attribute mental states to themselves and others, involving beliefs and desires [21, 51]. When robots can understand and mimic these human mental states, they can act and behave more appropriately and intuitively to humans [33]. This has several benefits, such as fostering more effective and natural interactions, improving the user experience, and expanding the range of tasks that robots can perform [44]. This integration is particularly beneficial in complex environments because it allows robots to adapt to the situation, and to predict human behaviours better and vice versa, leading to a more efficient collaboration in shared tasks [17]. Predicting and understanding others' behaviours is crucial in many robotics applications, such as assistive technology, autonomous vehicles and social robotics [17].

The majority of the studies analysing ToM ability consider the '**False-Belief Understanding**' (***FBu***), which is the ability of people to infer when others hold beliefs that contradict the reality [10, 19, 20, 29, 61]. It is also an appropriate measure for evaluating children's ability to make a distinction between the environment and the human mind [61]. This ability reflects the understanding that humans, in general, act according to how they represent the world, rather than how the world actually is [29]. Comprehension of FB situations can also highlight the children's capacity to adapt their behaviour, for instance, adopting assistance behaviour towards individuals holding FBs [19, 20]. Based on the insights gained from those studies, psychologists established a fundamental experiment called the **Sally–Anne (*SA*)** test to measure *FB* in children [10]. This test involves a scenario where Sally (i.e., one player) places a ball in her basket and then leaves the room. Subsequently, Anne (i.e., a second player) moves the ball from the basket to a box. The child participant, who has been observing these actions, is then asked to predict where Sally will search for the ball upon her return. Children with a well-developed ToM system are expected to anticipate that Sally will look in the basket, adhering to her original belief. Conversely, children with a less developed ToM system may respond based on their knowledge of the ball's location. This discrepancy is an example of a false belief, a misconception or misunderstanding one might have about the beliefs or intentions of another person [61]. In **Human–Robot Interaction (HRI)**, the presence of possible false beliefs can lead to miscommunication or misinterpretation of the robot's actions, thereby affecting the overall interaction. Understanding and addressing false beliefs is critical to developing effective HRIs [17]. Just as the *SA* test demonstrates the importance of understanding and predicting others' beliefs in human–human interactions, the same principles apply to HRIs.

ToM and transparency are inherently connected in HRI, as a robot's ability to understand and reason about human mental states directly impacts how clearly its actions can be interpreted. When robots can accurately model human mental states and adjust their behaviour accordingly, they create a foundation for transparent interaction. This brings us to the concept of transparency in HRI, which is closely related to the robot's ability to understand and communicate about beliefs and intentions. In HRI, transparency contributes to the clarity with which the human user understands a robot's intentions, capabilities and actions. This understanding is enhanced when robots can demonstrate ToM capabilities, as they can better explain their actions in terms of their understanding of human mental states and expectations. Despite its recognised significance, a universally accepted definition of transparency in HRI is still evolving. Current theorising suggests a multi-faceted approach, integrating elements of Explainability, Legibility and Predictability [1, 3, 27]. Importantly, transparency also emerges from how these dimensions help the human perceiver build an accurate mental model of the robot's functioning [2, 41]. This mental model, in turn, shapes the Expectability of the robot's behaviour. Therefore, for a behaviour to be transparent, the robot should not only be perceived as legible, predictable and explainable but also align with the user's expectations, which are rooted in their mental model.

The importance of aligning robot behaviour with human mental models is further emphasised in recent literature. Matthews et al. [43] suggest that transparent information should be designed for compatibility with the operator's mental model. Similarly, a recent paper from Williams et al. [63] discussed that a robot with ToM could potentially provide a higher level of transparency, as it can communicate its intentions and decision-making processes more effectively. This leads us to consider the concept of transparency in the context of HRI since the ability to reason about false beliefs can increase transparency, thereby reducing the likelihood of miscommunication or misinterpretation. Some papers reveal that transparency can enable users to see through the robot and understand its action, making communication with humans more efficient and increasing trust [5]. The robot can also mitigate its errors during a task through explanations, showing its integrity and recognising its imperfection [4]. When combined with ToM, this approach can considerably improve the relationship between humans and robots, especially in collaborating settings [46]. This work investigates transparency by conducting an HRI study, leading to our **Research Question (RQ)**: 'Can users comprehend the advanced cognitive skills of a robot equipped with ToM, and does this improve their interactions with it?'. However, existing methods often fall short in accurately capturing and adapting to the complexity and variability of human mental states, and usually, they are based on predefined models or assumptions about human behaviour.

The contributions of this work are (1) the development of a novel method for incorporating human mental states into robotic behaviour using Bayesian inference within a **Hierarchical Reinforcement Learning (HRL)** algorithm, capable of adapting the robot's behaviour in a multi-goal collaborative scenario. This model, referred to as **Hierarchical Bayesian ToM (HBToM)**, enables the robot to collect and deliver multiple items while considering the users' mental states, including their beliefs about items' location and their desires regarding the sequence of item collection and delivery and (2) the execution of an extensive between-subjects study involving 110 participants, providing empirical evidence for the effectiveness of the method and evaluating the impact of a robot's ToM on transparency and false beliefs across various settings.

## 2 Related Work

The field of psychology has extensively explored ToM [11, 29, 30, 32, 38, 45], but its inclusion in robots and its application in the context of HRI has been recognised as one of the 'grand challenges' in robotics research [65].

### 2.1 ToM in Psychology

Psychologist researchers are particularly interested in the development of ToM in children and their understanding of people's mental states [10, 11, 29, 45, 61]. As already mentioned, an experiment widely used as a baseline measurement for *FBu* is the *SA* test [10]. Inspired by this experiment, researchers implemented a modified version of the story, playing with the ability of the children to participate in the scenario actively. Gopnik et al. [31] integrated a series of exercises to evaluate children's performance based on their verbal feedback. There were various dynamic tasks, including a pretence task where the children were asked the purpose of objects when it was used in different contexts [42]. Similar to the *SA* test, Bartsch and Wellman [11] investigated the ability of children to detect *FB* in storytelling. The questions were more complex and aimed at the intentions behind the actions performed by the story's characters. The purpose was to invoke the children's capability to indicate the mental state behind the character's actions, such as beliefs.

State-of-the-art research does not limit its evaluation to the verbal skills of children but also their interactions within the environment [19, 20, 52]. Buttelman et al. [19, 20] investigated the helping behaviours of infants when this one detects *FB*. Similarly to *SA*, children watched a toy in a box being swapped into another box while an adult witnessed or did not witness the swap. Then, the adult unsuccessfully attempted to open the box where the toy originally was. The infant was then asked to help the adult to open the box. In response to this request, the infant chose to open either the box with the toy or the other box. Priewasser et al. [52] conducted a similar study wherein they exposed the infants to a more complex setting. Other studies focus on participants' ability to understand *FB* and manipulate it in specific game scenarios [59].

While these scenarios have provided valuable insights into the development and understanding of ToM in children, their direct application in HRI scenarios may not be effective and not translate well to the capabilities and constraints of collaborative human-robotic systems. Especially when the robot's design can have an impact on the interactions [40] or the transition from human–human to human–robot environment, which might add some complexity to the experiment.

### 2.2 ToM in HRI

Given the diverse range of studies conducted, roboticists have encountered significant difficulties when attempting to implement suitable experiments for robotic ToM. Ruocco et al. [55] decided to study the correlation between ToM and trust in HRI through the Investment Game [14]. The participant had the opportunity to assess the robot's cognitive abilities by observing its performance on the *SA* test. Considering the prior knowledge of ToM capability of the robot, the participant played the game with the robot by investing coins in the robot during several rounds. The results revealed that, on average, people invested more money in a robot, demonstrating a higher level of ToM. However, there was no significant difference in trust over the different rounds. A study conducted by Romeo et al. [54] provides an interesting perspective by investigating how a robot, through a collaborative task that mimics ToM, influences users' behaviours and trust. The study was conducted in a cooperative maze navigation task with three robot personas: one neutral, one that explains its reasoning in technical terms and one that mimics ToM. Despite their intriguing findings, the use of the Wizard of Oz technique to simulate a ToM-like behaviour in the robot can be seen as a limitation [53]. In another work, Bara et al. [9] ventured into the domain of ToM modelling in situated language communication. Their focus was on collaborative tasks in the 3D virtual block world of Minecraft. They introduced a dataset including several videos and dialogues of collaborative tasks performed by pairs of human subjects in Minecraft. This dataset provides information capturing partners' beliefs of the world and each other as an interaction unfolds. However, the use of a Feed-Forward Network to infer fellow players' mental states raises questions about the applicability of such models in real-world scenarios (e.g., does not incorporate temporal context in mental state inference) [12, 48].

Another class of approaches in the field of HRI employs **Inverse Reinforcement Learning (IRL)**. This technique is used to learn an underlying reward function, which is assumed to guide the human behaviour. This can be seen as a form of ToM, where the robot tries to interpret human behaviours based on inferred goals. Jara-Ettinger [37] shows that certain relationships exist between ToM and IRL. For example, it allows interpreting an agent's reward as its preference and the policy as its intentions. Sadigh et al. [56] apply IRL by learning the reward function of a human to model the human's next action in an autonomous driving scenario. However, this approach assumes that the agent is a rational planner and, therefore, acts optimally given its reward function. Moreover, IRL algorithms require numerous demonstrations, and they are often considered a black box to humans [16]. Despite those limitations, IRL points out that humans tend to act according to plans, which supports many studies concentrated on the definition of intention in human psychology [24].

In a different approach, Vinanzi et al. [60] developed a robot learning architecture based on **Bayesian Networks (BNs)**. This architecture, supported by an episodic memory system, is capable of estimating the trustworthiness of human partners. The results of their study showed that their architecture achieved the same results obtained by the children participating in the study. This study focused on integrating ToM for social robots in collaborative settings [28, 49, 54, 55]. Furthering the discussion on Bayesian models, Baker et al. [7] proposed a **Bayesian ToM (BToM)** model in **Partially Observable Markov Decision Processes (POMDPs)** and Bayesian inference. The study acknowledges that goals and beliefs need to be inferred simultaneously, which is a significant insight into the complexity of implementing ToM in robots. Similarly, Hellou et al. [35], building upon Baker's work, applied BToM to a simulated dynamic HRI scenario involving false beliefs. Despite the absence of experiments with humans, their results show that robots can better assist and collaborate with them when they track the mental states of humans. Those two papers demonstrated the utility of BNs and POMDPs to study the roles and impact of beliefs and desires in the decision-making process of users. This can be essential when designing autonomous systems which are intended to interact with humans.

Involving false-belief situations, Zeng et al. [66] 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, an Artificial Neural Network where neurons process temporal inputs and outputs [68]. The tasks included two humanoid robots interacting in a replicated *SA* experiment where one robot would develop false belief regarding the possible location of a ladybird in two boxes. The other robot observing the scene was asked the critical false-belief question where the initial robot will look after the place of the ladybird was exchanged. The results from experiments were compared with children's performance in similar trials to evaluate the model's ToM. In another context, Pöppel and Kopp [50] focused their work on explaining the uncertainties of social agents navigating in an unknown maze environment. In doing so, they implemented a modified version of the BToM to predict an agent's behaviour by observing its actions. More specifically, the model was able to provide explanations on the agent's goal when its location or the environment's settings were unknown. To tackle false beliefs in human–robot collaboration tasks, Favier et al. [28] introduced a Task Planner [18] in a collaborative cooking scenario between a human and robot. The goal of the task was to prepare pasta within a shared environment between the human and the robot, where several situations were simulated. More specifically, the authors introduced situations where the human's beliefs might be different from the actual state of the environment, which could result to a task's failure. In response to these issues, the authors extended their original model to detect false beliefs and enable proactive behaviour for the robot to help the human to achieve the task.

Buehler et al. [17] designed a Human–Robot sushi-making task to evaluate the influence of a human-centric communication concept on performance. Their approach integrated ToM into a communication assistant to support humans in a cooperative setting. By considering the human's beliefs, the robot guided them to select the correct ingredient in cases of false beliefs or suggested the appropriate actions to follow in order to prepare the correct meal. The study found that compared to conditions without information exchange, participants assisted by the robot could recover from unawareness much earlier. Despite their positive results, the use of the Wizard of Oz, where a human operator covertly controls the robot's actions, presents a limitation since a controlled robot is serving more as a proxy for a human and less as an independent entity [53].

Given the aforementioned studies and their limitations as discussed above (the reliance on the Wizard of Oz and the use of simulations devoid of human participants), it becomes evident that there is a need for further research in this area. Especially in dynamic environments where false beliefs may arise, it is crucial for the robot to be transparent and provide clear reasoning for its actions. This attitude can help users to have a better understanding of the current environment's state, which is crucial when both humans and robots are working together [28]. Additionally, this transparency could help build trust between the human and the robot and aid in better collaboration by aligning the human's understanding with the robot's actions [47]. Therefore, developing a ToM cognitive model for robots should incorporate these aspects. This work aims to explore the development of ToM cognitive model for robots to enhance their interactions with humans, particularly during false-belief situations. As highlighted in the papers reviewed above, addressing these specific scenarios is crucial, especially in the context of collaborative tasks. Through this study, we hope to gain insights into how humans perceive this robot behaviour in which it exhibits advanced cognitive abilities enabled by ToM.

## 3 Technical Approach

Inspired by the BToM concept put forth by Baker et al. [6–8], this work proposes to combine HRL with a human in the loop by observing the interaction with a modified version of the BToM, a probabilistic model used to infer a human's joint belief-desires (see Figure 1). We refer to this model as HBToM. It combines Bayesian updates and POMDPs to reason about an observer's mental states. However, applying POMDPs can become computationally intractable in complex environments due to the exponential growth of the state space. This complexity arises from the need to maintain and update belief states over numerous possible world states, which makes it challenging to scale POMDPs to more complex, real-world scenarios. To address this, we adopt HRL to solve the POMDPs. With a human in the loop, the interaction is observed through the HBToM. HRL offers several advantages, such as the ability to learn solutions to sub-problems that can be reused in solving other related problems. It also structures the problem hierarchically, enabling the robot to learn over temporally extended transitions and explore over extended periods. Furthermore, the robot can reason about the instructions given by users based on their initial beliefs and desires. The agent receives instructions from the observer, who works with it to accomplish particular objectives. Given the agent's task and instructions, the model does the following: (1) updates observers' beliefs along the collaborative task, (2) learns various possible policies that the agent could follow based on human desires, (3) jointly infers the beliefs and desires of the observer all along the collaborative task and (4) extracts these mental states over the entire sequence.

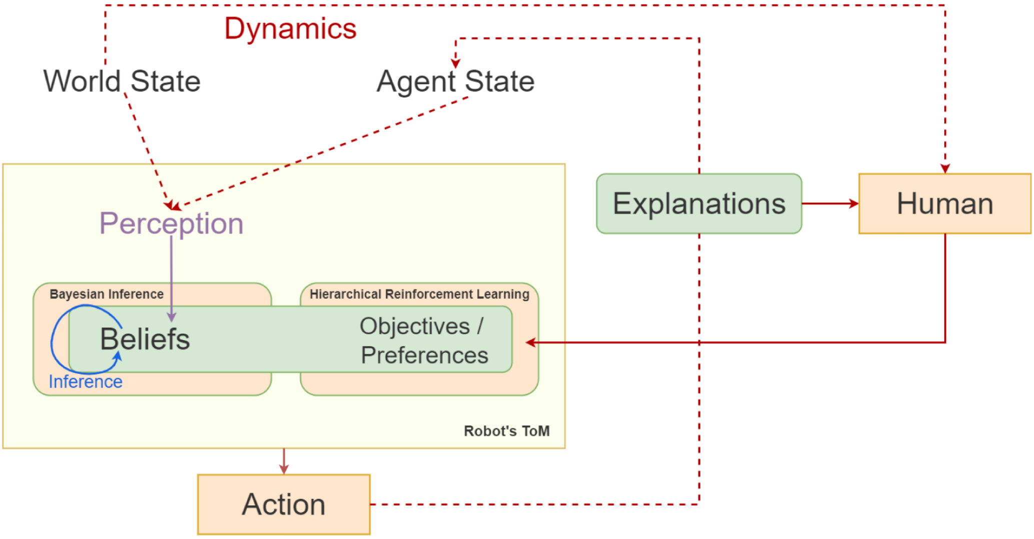


Fig. 1. The proposed architecture inspired by [6]. The cognitive model uses the interactions with the human to extract their desires and the environment with the agent's position to update their beliefs. The inference of both mental states influences the robot's decisions by generating the appropriate explanations.

### 3.1 Update the Observer's Belief

In the context of inferring the mental states of a human observer interacting with an autonomous robot, this approach employs the Bayesian inference process. The robot infers the current human's belief $B$ and desire $D$ based on its action $A$ and the human's observation [50]. This can be represented using Bayes' theorem in Equation (1):

$$P(B, D|A, O) = \frac{P(A, O|B, D)P(B, D)}{P(A, O)}. \tag{1}$$

Here, $P(A, O|B, D)$ is the likelihood of the human having the observation $O$ when the robot takes an action $A$ given the human's belief $B$ and desire $D$. $P(B, D)$ is the marginal probability of the belief and desire based on prior beliefs and desires, and $P(A, O)$ is the normalisation probability over all possible actions and observations. Based on these inferred mental states, the robot provides explanations to collaborate with humans.

The model updates the observers' beliefs regarding their observation, defined through an observation model $O$. This model considers the robot's action $a \in A$, where $A$ is the robot's action space, the world $w \in W$, where $W$ represents the possible worlds, and the position of the robot $p \in P$, where $P$ represent the possible positions of the robot in an environment. Considering the possible observations and the previous observer's beliefs, the model determines and updates their beliefs regarding a specific world $w$, using Bayesian update. For instance, to determine the new belief $b_t(w_t)$ of the observer at a time $t$ regarding the world $w_t$, considering the action $a_{t-1}$ to go from position $p_{t-1}$ to $p_t$, the new observation $o_t$, and the prior belief $b_{t-1}$, we update the new belief $b_t$ as follows in Equation (2):

$$b_t(w_t) \propto P(o_t \mid w_t, p_t, a_{t-1}, b_{t-1})\, P(w_t, p_t \mid w_{t-1}, p_{t-1}, a_{t-1})\, b_{t-1}(w_{t-1}). \tag{2}$$

In Equation (2), $b_t(w_t)$ denotes the likelihood that the world $w$ is true at $t$. $P(o_t \mid w_t, p_t, a_{t-1}, b_{t-1})$ is the likelihood of having the observation $o_t$ regarding the robot's position $p_t$, the current world $w_t$, the action taken by the robot and the prior belief $b_{t-1}$. $P(w_t, p_t \mid w_{t-1}, p_{t-1}, a_{t-1})$ is the likelihood to have the world $w_t$ and the robot's position $p_t$ according to the previous world $w_{t-1}$, robot's position $p_{t-1}$, and the action taken by the robot $a_{t-1}$. The robot updates the observers' beliefs regarding previous information to align their beliefs with the current state of the world and what they visually perceive.

### 3.2 Policy Learning with ToM

The proposed model leverages POMDPs to represent the robot's plan in an environment, taking into account the observers' mental states (beliefs and desires). The state space $S = \{P, B\}$ comprises the robot's position $P$ in the environment, and the observers' belief set $B$ about the world. The action space is denoted by $A$, and the reward function $R(s, a, s')$, where $s$ is the current state, $a$ is the action taken and $s'$ is the subsequent state. The transition probability, denoted as $P(s'|s, a)$, represents the probability of transitioning to state $s'$ given the current state $s$ and action $a$.

However, in real-world scenarios, the complexity of collaborative tasks often necessitates the decomposition of the main goal into composite and primitive tasks. This is where HRL comes into play in this proposed approach, specifically using the MAXQ algorithm [25, 26]. By structuring the problem hierarchically, MAXQ enables the robot to learn solutions to sub-problems that can be reused in solving other related problems.

In a generic scenario, the composite tasks could be any high-level tasks that the robot needs to accomplish, and the primitive actions could be any low-level actions that the robot can perform. For instance, consider a scenario where the task is to clean an environment by collecting items and placing them in their corresponding locations. The composite tasks are two-fold: (1) *collect*, which involves going to the item's location and picking it up and (2) *deliver*, which involves delivering the item to its target location. The primitive actions describe the robot's navigation in the environment: $\{Down, Top, Right, Left\}$; and its interaction with items: $\{PickUp, Deliver\}$.

Within our framework, we define these composite tasks as sub-tasks, that enables the robot to achieve the main goal—collecting and delivering all items to their targeted locations. Additionally, we compute a complete function $C(s, i)$ to evaluate the value of completing a sub-task in relation to the remaining sub-tasks. By following the hierarchy, the HRL updates the Q-value for a sub-task $i \in \{deliver, collect\}$ as follows:

$$Q(i, s, a) = V(i, s) + C(i, s, a), \tag{3}$$

where the value function $V(i, s)$ is the actual value in state $s$ during the sub-task $i$, while $C(i, s, a)$ is indicating the quality to complete the sub-task $i$ in state $s$ when executing the primitive task $a$. At each training step, we establish the new values of the two main functions as follows:

$$V(i, s) = \sum_{s'} P(s' \mid s, a)\ (R(i, s, a, s') + \gamma \cdot V(i, s')). \tag{4}$$

$$C(i, s, a) = \gamma \cdot \sum_{s'} P(s' \mid s, a) \max_{a'} Q(j, s', a'). \tag{5}$$

In Equation (4), we compute the expected reward when following the sub-task $i$ in the next state $s'$ with $V(i, s')$ discounted with a factor $\gamma$. Using the complete function $C(i, s, a)$ in Equation (5), we determine the maximum Q-value for the next sub-task $j$ in state $s'$, which represents the future reward when completing the sub-task $i$ and moving to the next one. Both Equations (4) and (5) rely also on the transition function $P(s' \mid s, a)$, which indicates the probability of transitioning from state $s$ to $s'$ after taking the action $a$.

For the reward function $R(i, s, a, s')$, the model computes a reward based on the robot's distance to the preferred goal and its maximum distance to the other goals, which are goals of the same nature (*collect* or *deliver*). In our scenario, the robot must collect and deliver items to designated locations. Specifically, the robot handles two distinct collection tasks, $collect\{n\}$, and two distinct delivering tasks, $deliver\{n\}$, with $n \in \{1, 2\}$ representing the item to be collected and delivered. As a result, the reward function needs to consider these sub-tasks, and the preferences of the agent. We calculate the reward when the robot goes from a state $s$ to $s'$ as follows:

$$R(i, s, a, s') = R(i, a, s') + f(i, s, s', g), \tag{6}$$

where $R(i, a, s')$ is the immediate reward when the agent is performing the primitive task $a$ in the state $s$, e.g., collecting an item. Additionally, we introduced a shaping reward function $f$, considering the agent's current location to the optimal goal's position $g$ considering the user's preference and the nature of the sub-task $i$. By doing so, we designed the shaping function to prioritise the user's preferences:

$$f(i, s, s', g) = \gamma \cdot \sigma(g, s') - \sigma(g, s). \tag{7}$$

Here, $\sigma(g, s') = 1 - \frac{d(g,s')}{w+h}$, where $d(g, s')$ denotes the Manhattan distance between the state $s'$ and the goal $g$, and $w$ and $h$ respectively stand for the dimension of the environment and $\gamma$ is a discount factor. This reward function is designed to guide the robot in learning the correct action sequence that aligns with the user's preferences. It was also proven to be powerful means to improve reinforcement learning performance [62], which is crucial to ensure smooth and efficient human–robot collaboration.

An essential aspect of this context is that the human is involved in the decision-making process, selecting the order of sub-tasks based on their preferences. This serves as the foundation for the model to predict the human's preferences, indicating the robot the item they would like collected first, based on their beliefs. As mentioned earlier, the sub-tasks require the robot to collect and deliver two items (a ball and a teddy bear), executed in a sequence that reflects the user's preferences in the collaborative task with the robot (more details about the task provided in Section 4).

### 3.3 Inference of the Mental States

The policies learned in relation to the observers' mental states are integrated with Bayesian updates to jointly infer their beliefs and desires. This approach enables the model to track these internal states during the interaction. The model utilises a **Dynamic BN (DBN)** to capture the dependencies of external and internal factors, such as the robot's location, observations, desires and beliefs. A visualisation of these dependencies is illustrated in Figure 2, with the key feature that the Action variable influences the state of the world, compared to the initial version in [6]. Indeed, the robot agent can modify the environment by collecting and delivering items. The forward-backward algorithm [67] is employed to compute the probabilistic values of the beliefs and desires. These probabilistic predictions are important for tracking the cognitive states of the observers interacting with the robot, which is a crucial step in this study.

To compute the joint probabilities of the beliefs and desires over a full trajectory sequence up $T$, at any time $t \le T$, the model synthesises the outcomes of the forward and backward algorithms:

$$P\left(b_t, d_t \mid w_t, s_t\right) \propto P\left(b_t, d_t \mid w_{1:t}, s_{1:t}\right) \cdot P\left(s_{t+1:T} \mid b_t, d_t, w_{t+1:T}\right),$$

where $P\left(b_t, d_t \mid w_{1:t}, s_{1:t}\right)$ is the forward distribution and $P\left(s_{t+1:T} \mid b_t, d_t, w_{t+1:T}\right)$ is the backward distribution. The variables $b$ and $d$ represent the beliefs and desires of the observers, while $w$ and $s$ represent the real world over time and the robot's state. This prediction, informed by both past and anticipated future information, enables the robot to reason about the mental states of the observers and provides explanations regarding their instructions. These scripted explanations varied depending on the mental states considered by the model and whether the human held a false belief or not.

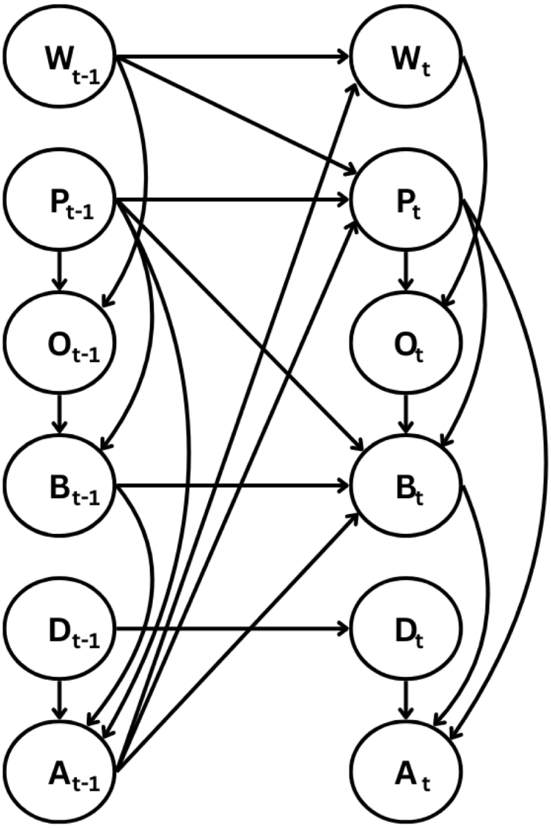


Fig. 2. A visualisation of the DBN is used to represent the dependencies between each element, including the mental states (beliefs and desires) and external factors.

### 3.4 The Task

These hypotheses form the basis of our investigation, and to explore them, this work used the *SA* test, which serves as a foundation for our study [10]. In the presented study, a variant of the *SA* test was adapted to create and fit in a more collaborative HRI scenario. In our scenario, a human and a robot share the task of cleaning a room and organising two toys—a teddy bear and a ball—in their respective designated locations. The human instructs the robot to commence cleaning from a specific preferable first location or item (as depicted in Figure 3 on the left is the teddy bear, and on the right is the ball). We used Unity Engine to create the simulation, chosen for its ease of use and extensive customisation options for the environment. Specifically, we built a pleasing environment to encourage participants to collaborate with the robot. The environment was built by using the '3D Tilemap' system, available on GitHub[1]. This technology enabled us to develop the 3D grid environment for the collaborative task and integrate various elements, including obstacles and toys (ball and teddy bear). For the robot model, we decided to use an appealing, freely available design in Unity. As a result, we exploited the 3D model of Wall-E, available for free on Sketchfab[2]. Through this system and the integration of the various elements, we could develop our own environment

[1] https://github.com/michaelsgamelab/MGL-3D-Rule-Tiles.

[2] https://sketchfab.com/3d-models/wall-e-free-download-undetailed-f40a3cb97ce24465973d9a726d0d4463.

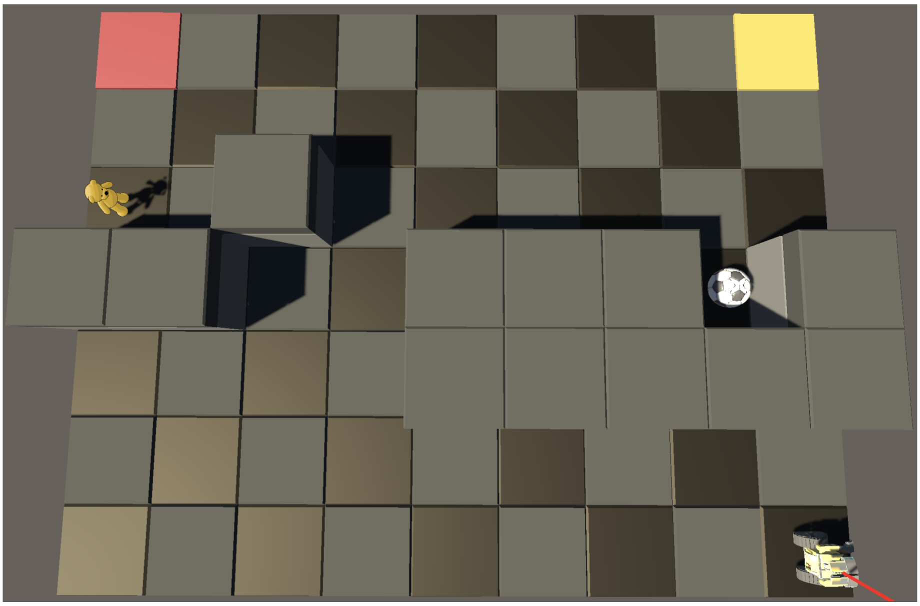


Fig. 3. The environment comprises a robot and two toys, each positioned in specific locations and with specific targeted location.

for the experiment. Furthermore, the simulation is publicly available on GitHub under an open source license[3].

To introduce a false-belief scenario, we simulated fog in the environment and swapped the positions of the items (as depicted in Figure 4). A robot with a ToM system would adhere to the human's initial preference and start cleaning from the originally specified location or item. Upon reaching the item, it would recognise that it is not the item initially indicated by the human due to the swap. The robot always gives explanations of its actions, and since we have a dynamic map with dynamic preferences, it splits the whole cleaning task into sub-tasks.

## 4 Hypotheses

To investigate our RQ and the influence of user desires and beliefs on ToM and Transparency during a collaborative task, and in light of the existing literature [43, 63], we formulated the following hypotheses:

- *Hypothesis 1 (H1)*: A robot that takes into account the desires and beliefs of the human will be perceived as displaying the most advanced ToM skill.
- *Hypothesis 2 (H2)*: The robot with the most advanced perceived ToM will also be perceived as displaying the most transparent robotic behaviour.
- *Hypothesis 3 (H3)*: The robot with the most transparent behaviour will also be perceived by the humans with the highest levels of trust in the robot.

To investigate our approach and explore the hypotheses, we created the following conditions:

- *Condition 1 (C1)*: The robot considers both the human's desire (the chosen toy) and beliefs. When the human instructs the robot to start cleaning from a specific location or item

[3] https://github.com/Mehdi-Hellou/Collaborative-Robot-Interaction-Simulation.

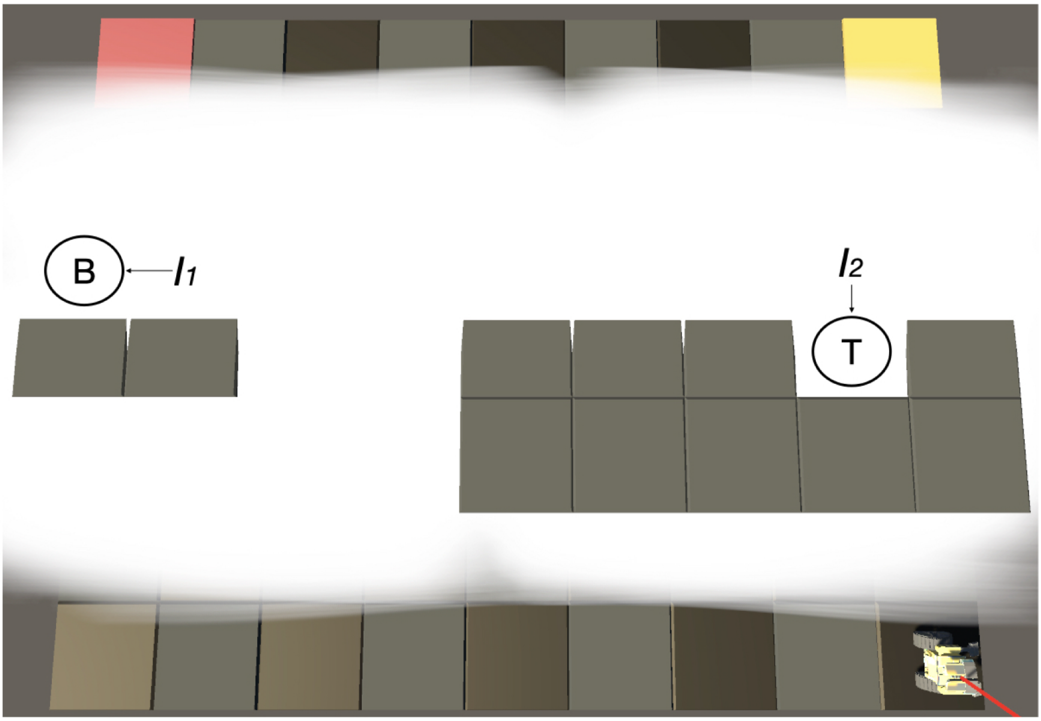


Fig. 4. The experimental environment with added fog, where objects $\mathcal{B}$ and $\mathcal{T}$ are rearranged to test participants' beliefs according to the initial setting in Figure 3. $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the location wherein the items are located.

(left for the teddy bear, right for the ball), the robot complies. If the item has been swapped, it informs the human, 'It appears that someone has relocated the teddy bear and the ball. Would you like me to pick up this item, or should I proceed to the other location?' The robot then follows the subsequent desires of the human.

- *Condition 2 (C2)*: The robot considers only the human's beliefs. It proceeds to the first location and, upon encountering a swapped item, informs the human, 'It seems that someone has changed the location of the teddy bear with the ball'. The robot picks the item up, places it in the specified location and continues with the remaining item.
- *Condition 3 (C3)*: The robot considers only the human's desire (the chosen toy). When the human instructs the robot to begin cleaning from a specific location or item (left for the teddy bear, right for the ball), the robot adheres to the human's choice. It informs the human, 'I am picking up the object that you suggested to pick up first'. The robot picks the item up, places it in the specified location and continues with the remaining item.
- *Condition 4 (C4)*: The robot disregards both the human's desire (the chosen toy) and beliefs. It proceeds to the first location (not selected by the human) and informs the human, 'I am picking up the object'. The robot picks the item up, places it in the specified location and continues with the other remaining item.

## 5 Experimental Method

An online between-subject study was conducted to assess the proposed approach. It included a prior pilot experiment that evaluated task clarity prior to the main investigation. The study was advertised via relevant e-mail lists and several social media platforms. Snowball sampling was also used, with participants being asked to share the study information with interested friends and colleagues. The ethical committee of the University of Naples Federico II granted approval for this experimentation.

## 5.1 Procedure and Measurements

The study was run using the Qualtrics software platform and a virtual environment developed in Unity 2021. At the commencement of the experimental session, participants were provided with a comprehensive informed consent document outlining the research's objectives and methodology. Upon granting their consent, participants provided responses to demographic inquiries, which included age, gender, education level, English proficiency, country of residence, prior exposure to robots and their predisposition towards robots. Then, participants were introduced to the experimental virtual environment (as depicted in Figure 3).

Following this introduction, an attention verification check was introduced to ensure participants' engagement in the study. Participants were tasked with indicating the positions of the toys within the virtual environment. It is worth noting that individuals who did not meet the age requirement (under 18 years old), lacked proficiency in the English language, or failed to pass the attention verification were excluded from the study. Subsequently, a simulated fog was introduced into the virtual environment (see Figure 4), marking the beginning of the interaction between the participant and the virtual robot. The experimental conditions were randomly assigned to each participant, with each participant exposed to only one condition. To mitigate bias, irrespective of the condition, the robot posed a question about human preferences regarding the picking of the first item ('Should I pick the object on the left or on the right?'), even if it did not consider this input for maintaining internal validity [36]. The desires (the choice of the first item to be cleaned by the robot) were communicated through the user's selection on the virtual environment. Finally, the experimental session had an approximate duration of 10 minutes.

After the interaction, to evaluate the transparency of robot behaviours, participants' responses were collected through various questions. More specifically, participants were asked to answer 7-point Likert scale questions regarding perceived Legibility (understanding immediate intent), perceived Predictability (anticipating future actions) and perceived Expectability (alignment with contextual norms or prior expectations), paired with previous work [2], which impact Transparency as explained previously:

- *Legibility*: 'To what extent did you understand what the robot was doing?';
- *Predictability*: 'To what extent could you predict the robot's next movement or behaviour?';
- *Expectability*: 'To what extent did the robot behave as you expected?'.

To assess whether participants believed the robot exhibited a ToM-like behaviour, participants were presented with four additional questions. These questions were adapted from a prior study [13] but included an additional inquiry regarding the robot's ability to communicate environmental changes. These questions aimed to gauge the robot's capacity to recognise alterations in its surroundings and effectively convey these changes to human participants. Such abilities entail both self-awareness and an understanding of what information humans may find pertinent, which are facets of ToM. The questions in this category were as follows:

- To assess participants' perceptions of the robot's ability to understand their knowledge of the environment (henceforth, *Awareness*): 'To what extent did you believe the robot was aware of your knowledge of the map?';

- To measure the robot's success in selecting the intended item that participants initially wanted (henceforth, *Alignment*): 'To what extent did the robot successfully pick up the item you initially intended to select?';
- To evaluate the extent to which the robot accurately grasped participants' intentions and communication (henceforth, *Comprehension*): 'To what extent did the robot accurately understand you?';
- To gauge the robot's effectiveness in conveying information about environmental changes to participants (henceforth, *Communication*): 'To what extent did the robot help you understand the environmental changes?'.

Participants were also instructed to complete the **Robotic Social Attributes Scale (RoSAS)** Questionnaire [22] to assess their perceptions of the robot's social attributes. Finally, the **Multi-Dimensional Measure of Trust (MDMT)** v2 questionnaire [58] was used to evaluate participants' levels of trust in the robot.

### 5.2 Participants

In the conducted online study, a total of 164 participants were initially recruited. After careful screening, one individual was excluded due to their age falling below the threshold of 18; four participants were excluded due to insufficient proficiency in the English language, seven were excluded for failing to pass the attention verification, and an additional 42 were removed from the analysis due to non-completion of the study. The final dataset for analysis comprised 110 participants, consisting of 71 males, 38 females and 1 participant who chose not to disclose their gender. No non-binary or other gender identities were reported. This distribution of participants allowed us to achieve our *a priori* calculation, which aimed for an effect size of $d = 0.25$ with a power of 0.80 at an alpha level of 0.05.

The participants' age range spanned from 18 to 64 years ($M = 30.33$, SD = 9.04). They were drawn from a variety of countries, with the largest representation originating from Italy (34 individuals), France (14 individuals) and the United Kingdom of Great Britain and Northern Ireland (13 individuals). The study encompassed a total of 21 different countries (Greece, Germany, the United States of America, Argentina, Spain, Pakistan, Austria, India, Iran, Netherlands, Japan, Nigeria, Oman, Peru, Portugal, Sweden, Switzerland and Turkey), with respondents hailing from various regions across nearly all continents. This international participation underscores the global scope and inclusivity of the study, reflecting a wide range of cultural and regional backgrounds among the individuals involved.

Participants' educational backgrounds included a high school degree or equivalent (15 individuals), a Bachelor's degree (26 individuals), a Master's degree (45 individuals) and a Doctorate (24 individuals). The majority of respondents (72.7%) also stated having prior experience with robots. Finally, we observed that they did not exhibit negative bias towards robots ($M = 1.92$, SD = 0.92).

## 6 Results

### 6.1 HBToM Model Performance

To evaluate the performance of the model, two types of evaluations were initiated, including (1) the assessment of the HRL to demonstrate the accuracy of the agent's planning in this particular scenario and (2) the analysis of the model to infer both human beliefs and desires. In the first evaluation, the focus was on assessing the HRL to learn optimal policies according to the preferences. In contrast, the second evaluation centres on the model's ability to predict users' mental states in relation to the situation. As mentioned previously, the solutions of the POMDPs are essential to infer human preferences. Following this approach, the scenario was divided into four sub-tasks that the robot

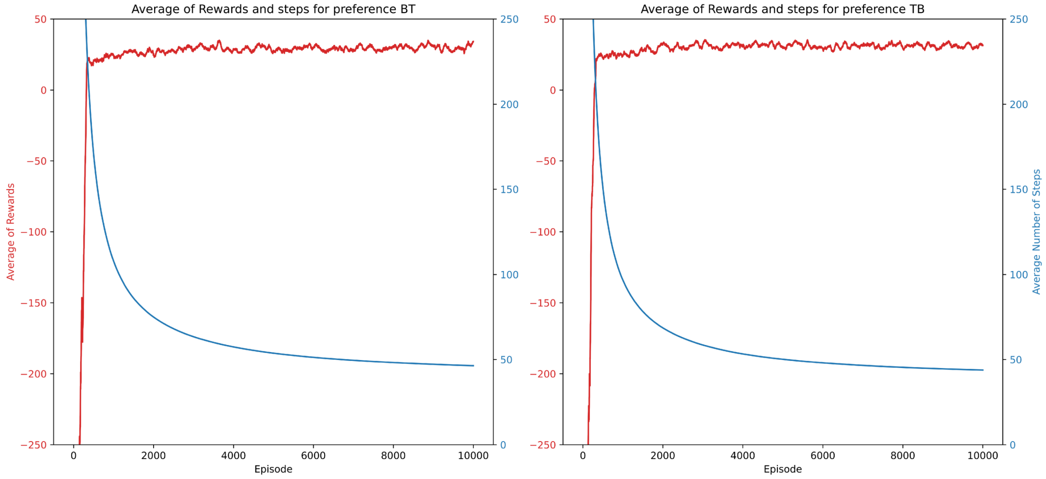


Fig. 5. The performance of the learning models used to illustrate the agent's plans according to the observers' preferences. The left graph depicts the average rewards and steps for the observers preferring the ball over the teddy bear, while the right graph depicts considering that the human prefers the teddy bear over the ball.

needs to achieve while considering the observers' mental states: (1) take the first item, (2) deliver it to its corresponding goal, (3) take the second item and (4) deliver it to its corresponding place.

In Figure 5, we depict the performance of our model through the average rewards received and steps the agent has taken to achieve the four sub-tasks over 10,000 episodes. The two graphs correspond to the POMDPs resolution for each preference: $\mathcal{BT}$ when the agent follows the path for an observer preferring the ball over the teddy bear, and conversely, for the preference $\mathcal{TB}$. The learning is made regarding the following parameters:

- To interpret the observers' initial belief, we initialise the world in two different settings: (1) $\mathcal{B}$ is in $\mathcal{I}_1$ and $\mathcal{T}$ in $\mathcal{I}_2$ and (2) $\mathcal{T}$ is in $\mathcal{I}_1$ and $\mathcal{B}$ in $\mathcal{I}_2$.
- Fifty percent of the time, we swapped the toys' location to create false beliefs for the observer.

Those parameters enable the model to learn about several conditions and to predict the observers' beliefs and desires in different situations. The results are similar for both desires and demonstrate the model's ability to resolve the different sub-tasks considering the desires and beliefs. The model's inferring capacity is analysed in the next paragraphs, considering specific conditions.

Afterwards, different situations were compared to validate the cognitive model's ability to accurately track and infer mental states. Drawing from the first condition (C1) in the online experiment, two contexts were characterised where the robot interacts with users by considering their beliefs and desires. These scenarios fully utilised the cognitive model to demonstrate its effectiveness in specific situations. In both contexts (Figures 6 and 9), the user asked the robot to take $\mathcal{B}$, believing that the ball is in $I_2$ (step 1). However, the robot detected the presence of $\mathcal{T}$ instead and questioned whether the user still wanted this item or preferred to inspect the other location, $I_1$. Based on the user's instruction, the model's prediction is analysed.

More specifically, in Scenario 1 (see Figure 6), the observer asked the robot to check the other location (step 2), indicating a level of trust in the system. The robot's model aligns with this hypothesis by revealing a change in the user's belief about the swapping event (Figure 7). It accurately determined that the user's beliefs now match the actual world setting ($\mathcal{B}$ in $I_1$ and $\mathcal{T}$ in $I_2$). In the final step, the model detected a change in the environment when the robot holds an item

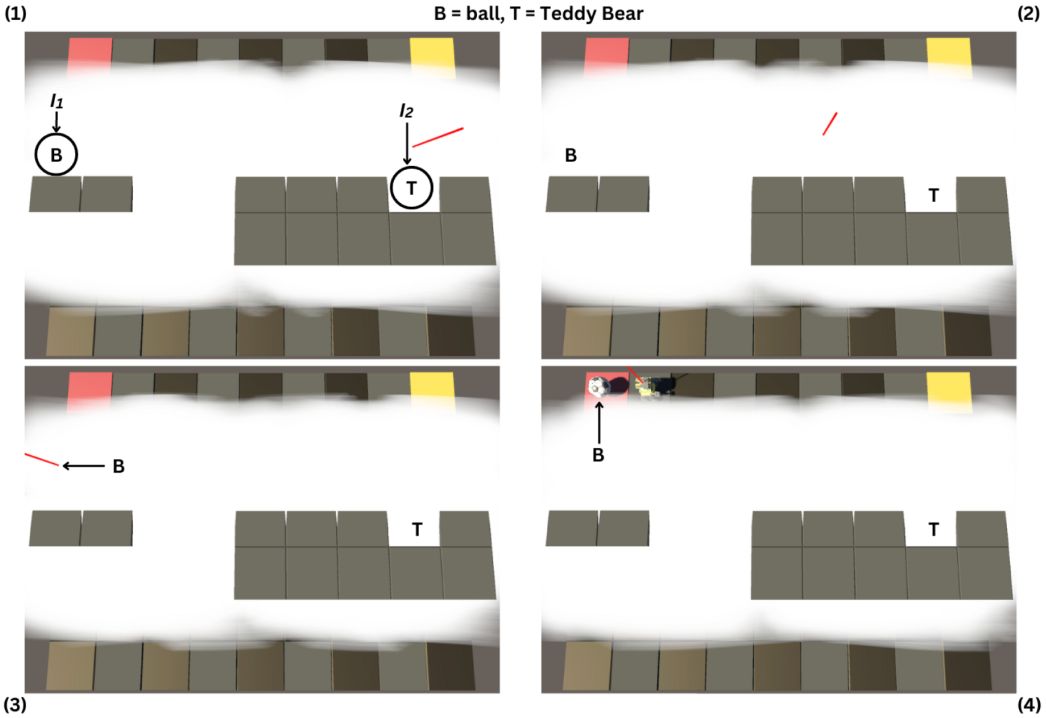


Fig. 6. Scenario 1: The observer asked the robot to check the other location. $I_1$ and $I_2$ correspond to the first and second items' location.

(step 3) (i.e., a high value for the ball for H, Figure 7) and when the robot successfully delivered $\mathcal{B}$ (step 4). Regarding the user's preferences (as shown in Figure 8), we deliberately selected two scenarios in which the user favours the ball over the teddy bear. As expected, in the first scenario, the robot's model effectively utilised the user's beliefs to make accurate inferences about their preferences, even when they held false beliefs in step 1.

Regarding Scenario 2 (see Figure 9), the observer intentionally ignored the robot's explanations and requested it to pick up the item at its current position $I_2$. In this case, the model captured the phenomenon of humans holding false beliefs (steps 1 and 2) until they could observe that the robot had taken the teddy bear instead of the ball (steps 3 and 4) (Figure 10). Similarly to the first example, the robot's cognitive model appropriately inferred that the user prefers the ball over the teddy bear (Figure 11), even when they had false beliefs (steps 1 and 2), or they notice that the robot is holding the teddy bear instead of the ball (steps 3 and 4).

In both scenarios for the preferences, we notice the cognitive model's inference in the last step (step 4) becomes nearly equal for both items, even though the model assigns the appropriate importance to the ball. We can explain this tendency by the fact that in both policies, when the observer holds true belief and only one item remains, the robot follows the same behaviour: to take the last item and deliver it to its corresponding location.

Overall, these results accurately reflect human intentions based on internal cognitive states, such as beliefs and desires. This highlights the strong connection between beliefs and desires in determining the intentions of a rational agent, such as a human, and underscores the importance of using the BDI model to reason on these agents' mental states [15, 24]. In the next section, we

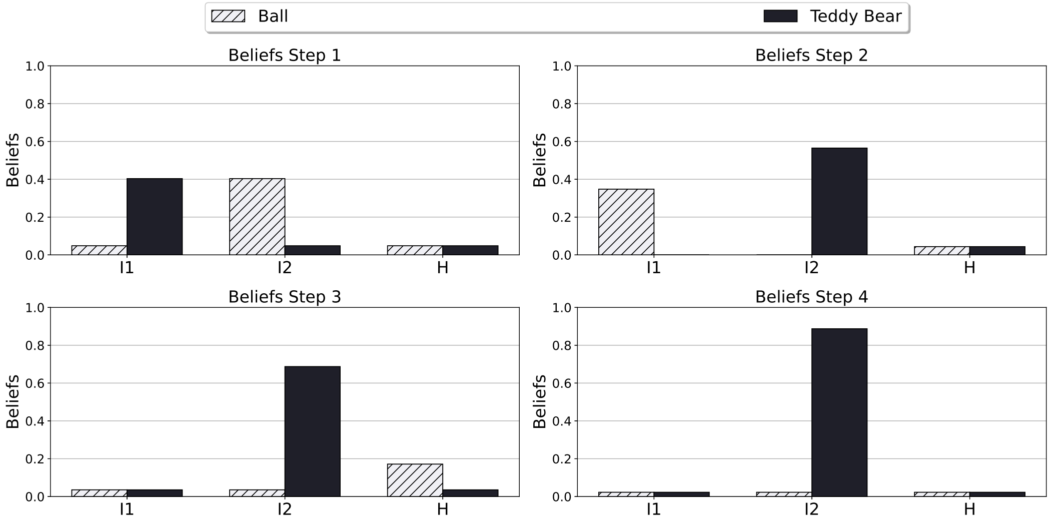


Fig. 7. The results of the beliefs' prediction at each step in Scenario 1 are represented as probability distributions, illustrating how the model adjusts its predictions over time. $I_1$ and $I_2$ are the first and second locations of the items. H indicates whether the robot is holding an item or not.

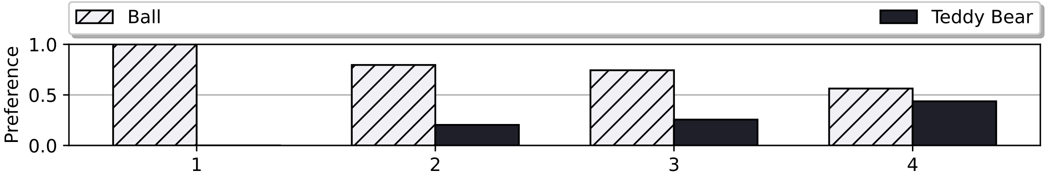


Fig. 8. The results of the preferences' prediction at each step in Scenario 1 are represented as probability distributions, similar to the belief's prediction.

will compare how human participants perceived the robot displaying different levels of ToM and examine the impact on transparency and social skills.

## 6.2 Transparency Ratings

We also assessed the concept of Transparency, incorporating three fundamental components: Legibility, Predictability and Expectability. We began by examining the normality of our data through a Shapiro–Wilk test, which yielded statistically significant p-values for all three factors, indicating non-normal data distribution. Consequently, we employed an Independent-Samples Kruskal–Wallis test to evaluate potential differences among the conditions.

Our analysis revealed a statistically significant difference in terms of Legibility among the conditions, with a test statistic value of $H(3) = 11.305$, p = 0.010. *Post-hoc* pairwise comparisons indicated significant differences between Conditions C3 and C2 (p = 0.049) and between Conditions C3 and C1 (p = 0.036). However, no statistically significant differences were observed between other condition pairs after applying the Bonferroni correction for multiple tests. A statistically significant difference was observed in terms of Predictability among the conditions, with a test statistic value of $H(3) = 14.224$, p = 0.003. The only significant pairwise comparison was between conditions C4 and C1 (p = 0.002), while no other condition pairs showed statistical significance following the Bonferroni correction. Finally, Expectability demonstrated a statistically significant

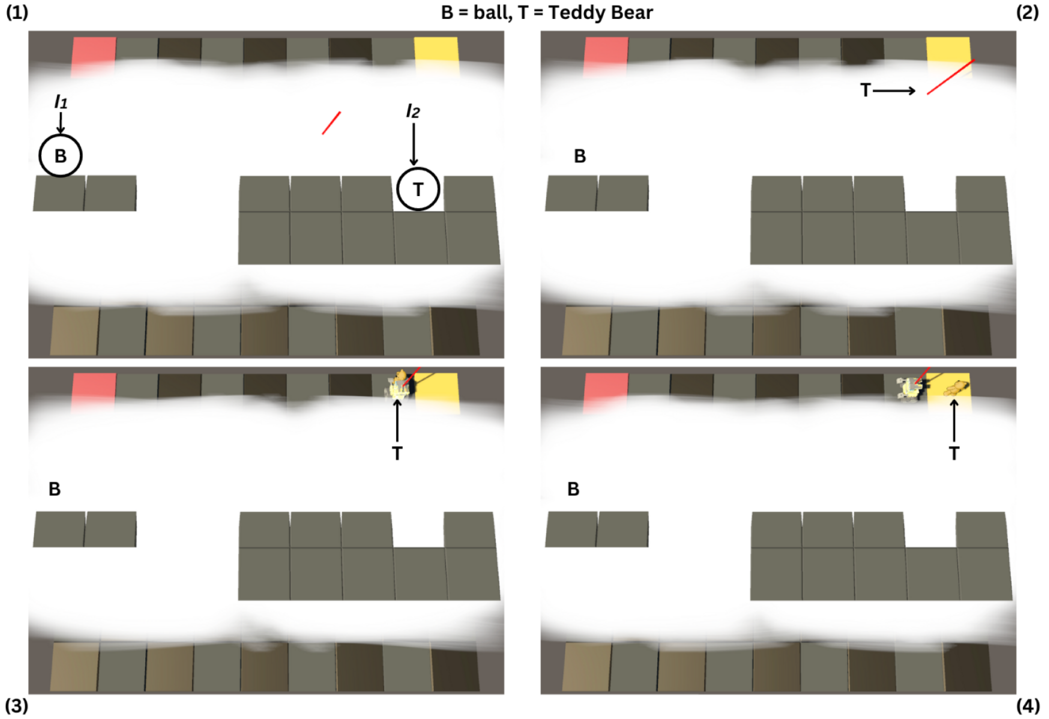


Fig. 9. Scenario 2: The observer ignored the robot's explanations and requested it to pick up the item at its current position. $I_1$ and $I_2$ correspond to the first and second items' location.

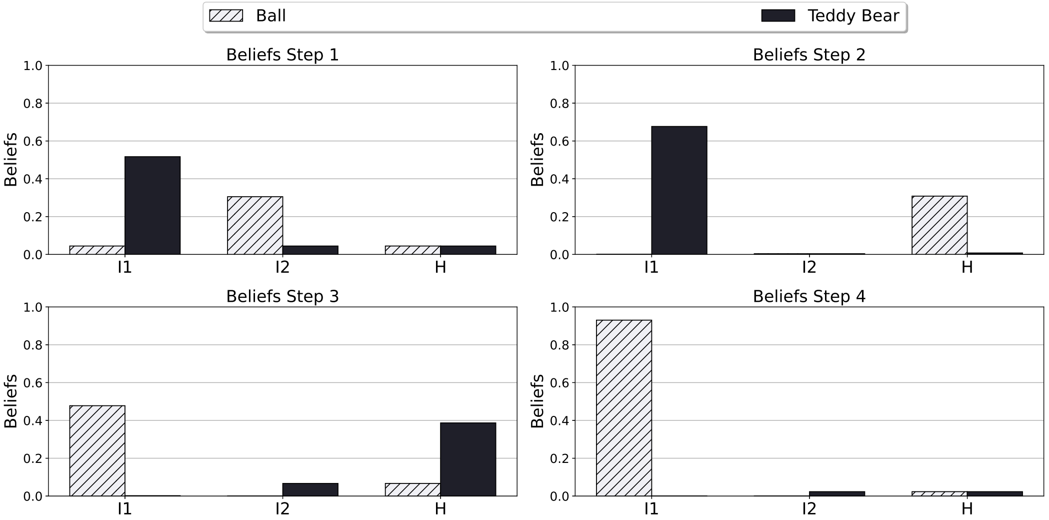


Fig. 10. The results of the beliefs' prediction at each step in Scenario 2 are represented as probability distributions, illustrating how the model adjusts its predictions over time. $I_1$ and $I_2$ are the first and second locations of the items. H indicates whether the robot is holding an item or not.

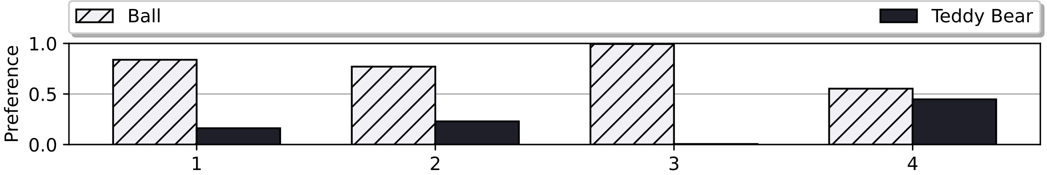


Fig. 11. The results of the preferences' prediction at each step in Scenario 2 are represented as probability distributions, similar to the belief's prediction.

difference among the conditions, with a test statistic value of $H(3) = 24.337$, $p < 0.001$. *Post-hoc* analyses revealed significant differences between Conditions C3 and C2 (p = 0.009), C3 and C1 ($p < 0.001$), C4 and C2 (p = 0.016) and C4 and C1 ($p < 0.001$). No statistically significant differences were found between other condition pairs after the Bonferroni correction. To visually represent the data, we created a Boxplot (Figure 12) comparing the medians and spread of the factors related to Legibility, Predictability and Expectability in each condition.

Following the approach of Angelopoulos et al. [2], we conducted a factor analysis to evaluate the construct of Transparency, which integrates Legibility, Predictability and Expectability. The KMO measure of sampling adequacy, with a value of 0.698, supported the suitability of our dataset for factor analysis. Subsequently, our factor analysis yielded factor loadings of 0.815 for Legibility, 0.852 for Predictability and 0.828 for Expectability. These factor loadings underscore the significant contributions of three elements to the overarching construct of Transparency, with Predictability being the most influential factor, closely followed by Expectability and Legibility. To determine the relative importance of each component, we normalised the standardised factor loadings and calculated the following relative importance scores: $R(L)$: 0.327, $R(P)$: 0.341 and $R(E)$: 0.332. Utilising these weights in a weighted sum formula, we obtained the Transparency score as:

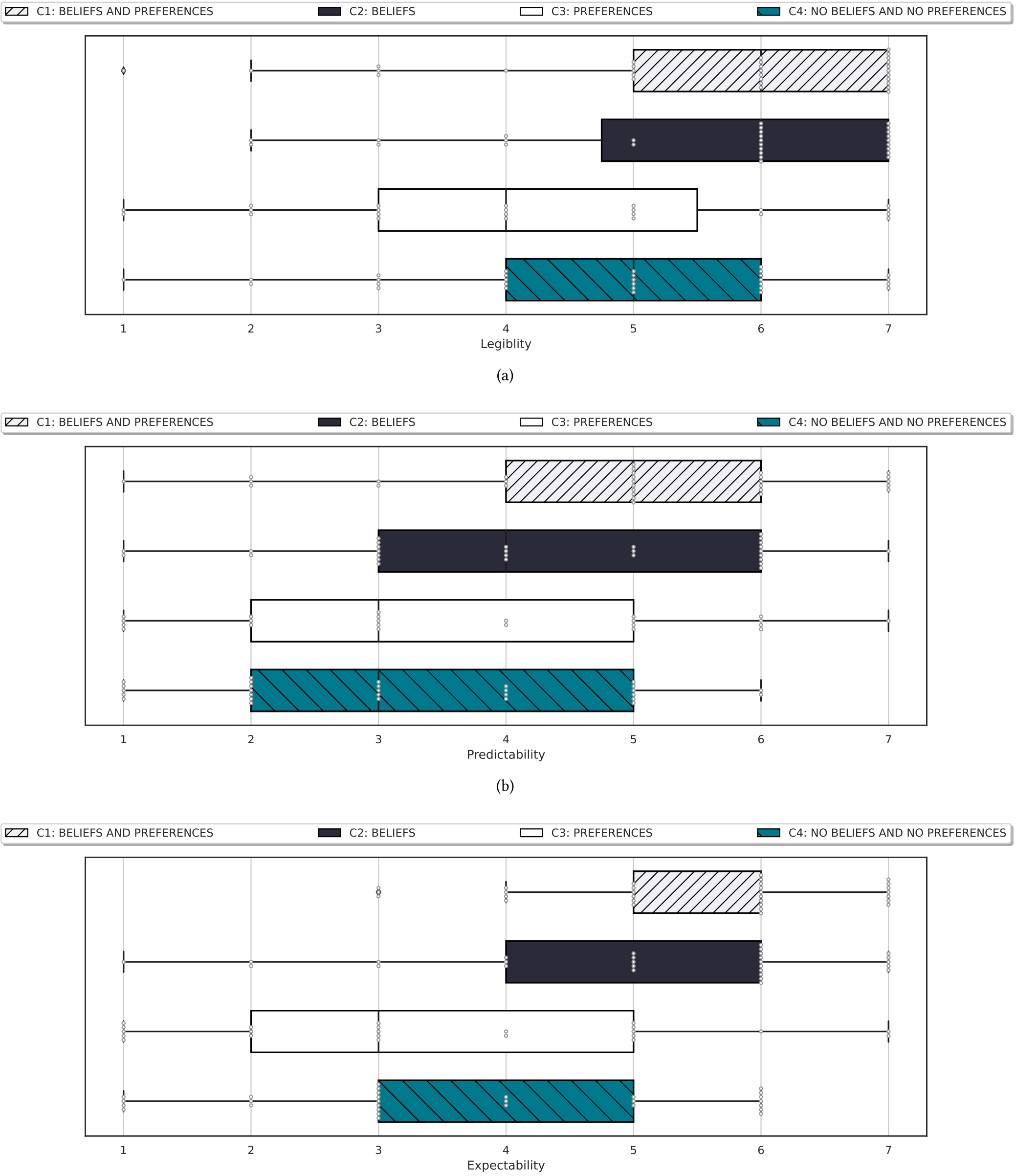
$$\text{Transparency} = 0.327 \cdot \text{Legibility} + 0.341 \cdot \text{Predictability} + 0.332 \cdot \text{Expectability}. \tag{8}$$

Subsequently, a new Shapiro–Wilk test was conducted to examine the normality of the overall Transparency data, which indicated a normal data distribution. A Levene's test showed no strong evidence of significant variance differences between the groups. Therefore, we proceeded with the Independent-Samples $t$-test. A statistically significant difference was observed between C1 and C4 ($t(57) = 4.547, p < 0.001$), between C1 and C3 ($t(51) = 3.918, p < 0.001$) and between C2 and C3 ($t(57) = 4.547, p = 0.008$) and C2 and C4 ($t(57) = 4.547, p = 0.002$). The Transparency scores for each condition are presented in Figure 13.

## 6.3 ToM Ratings

We assessed the perception of ToM encompassing four key aspects: Awareness of human knowledge, Alignment with user preferences, Comprehension and Communicating Environmental Changes. We initiated our analysis by evaluating the normality of our dataset through the Shapiro–Wilk test, which yielded statistically significant p-values for the four factors, signifying a non-normal data distribution. Consequently, we used an Independent-Samples Kruskal–Wallis test to examine potential differences among the conditions.

The analysis revealed a statistically significant difference regarding Awareness of human knowledge among the conditions, with a test statistic value of $H(3) = 12.434$, p = 0.006. *Post-hoc* pairwise comparisons indicated significant differences between Conditions C1 and C4 (p = 0.007). However, no statistically significant differences were observed between other condition pairs following the Bonferroni correction for multiple tests. Similarly, a statistically significant difference was observed in terms of Alignment with user's preferences among the conditions, with a test statistic value



(a)

(b)

(c)

Fig. 12. Boxplot compares the medians and spread of the data by condition for perceived (a) Legibility, (b) Predictability and (c) Expectability.

of $H(3) = 9.690$, p = 0.021. The only significant pairwise comparison was between conditions C1 and C4 (p = 0.014), while no other condition pairs exhibited statistical significance after applying the Bonferroni correction. Regarding Comprehension, our analysis revealed a statistically significant difference with a test statistic value of $H(3) = 38.137$, $p < 0.001$. *Post-hoc* analyses revealed significant differences between Conditions C3 and C1 (p = 0.038), C4 and C2 ($p < 0.001$) and C4

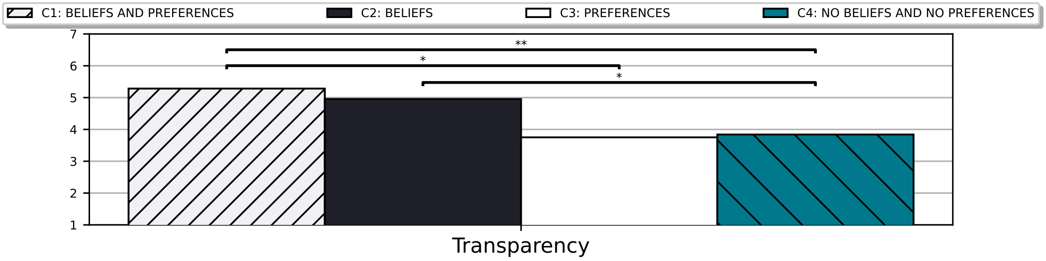
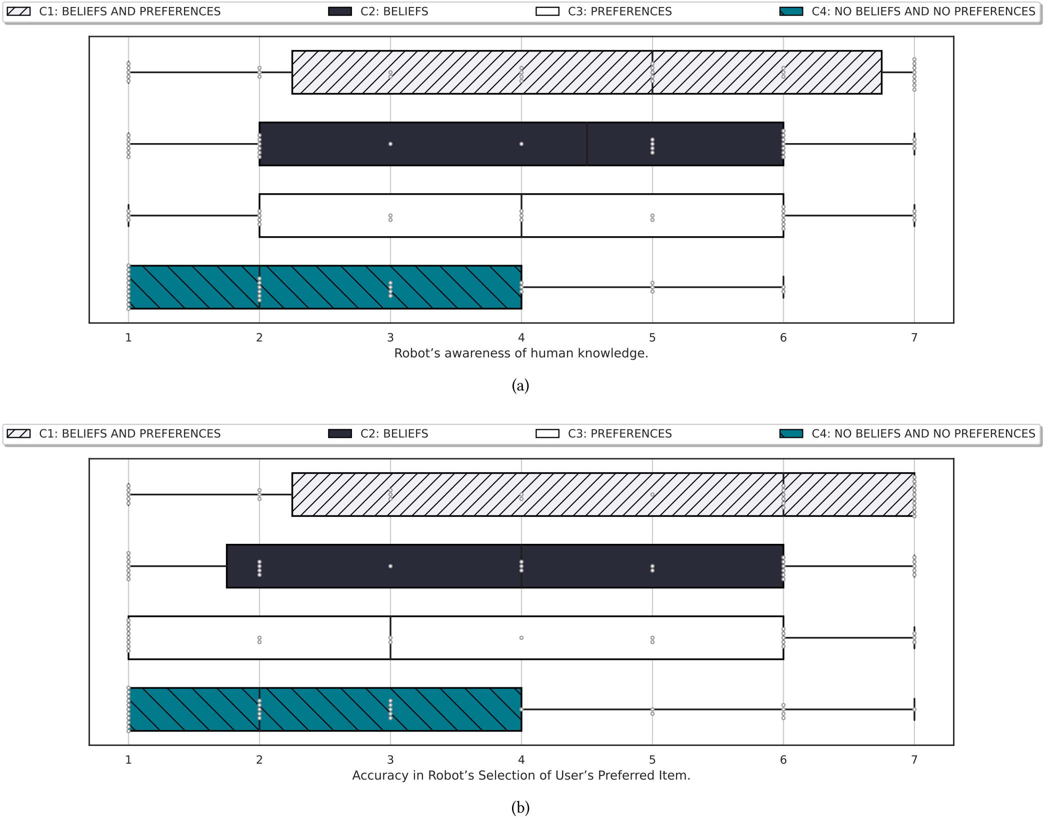


Fig. 13. Transparency for each condition (* for p < 0.05 and ** for p ≤ 0.001).

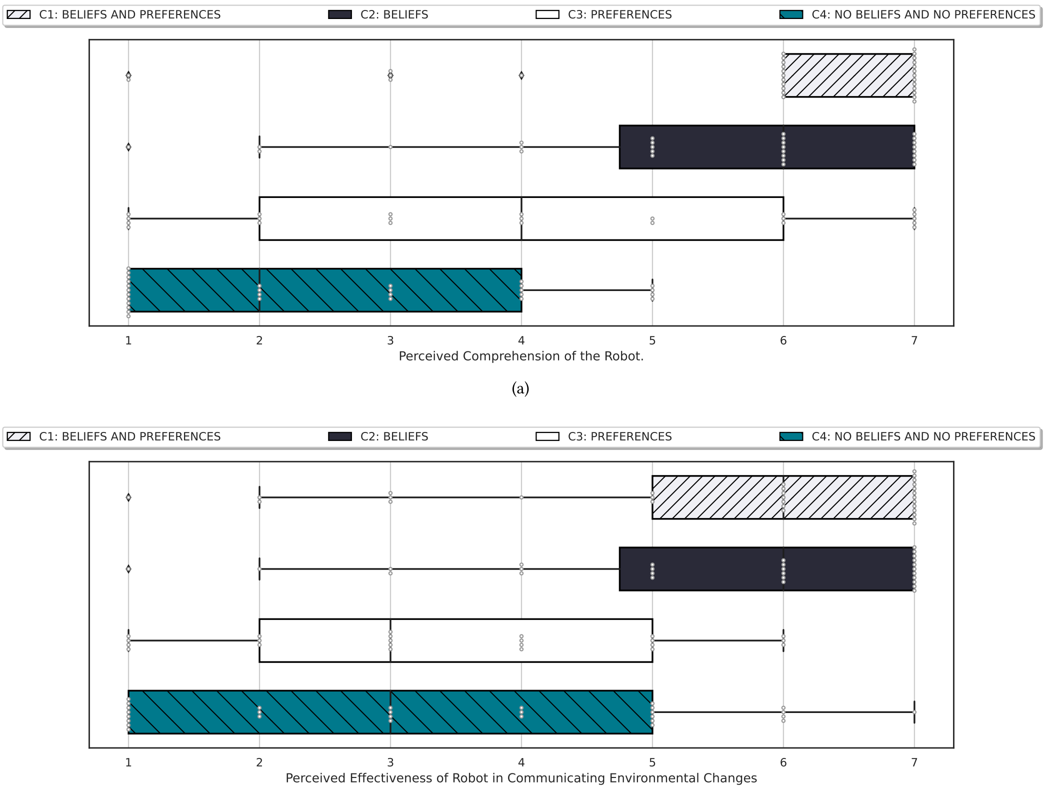


(a)

(b)

Fig. 14. Boxplot compares the medians and spread of the data by condition for (a) Awareness and (b) Alignment.

and C1 ($p < 0.001$). Finally, Communicating Environmental Changes demonstrated a statistically significant difference among the conditions, with a test statistic value of $H(3) = 30.535$, $p < 0.001$. *Post-hoc* analyses revealed significant differences between Conditions C3 and C2 ($p = 0.001$), C3 and C1 ($p = 0.001$), C4 and C2 ($p = 0.001$) and C4 and C1 ($p < 0.001$). No statistically significant differences were found between other condition pairs after the Bonferroni correction. Figures 14 and 15 present Boxplots comparing the central tendencies and variability of the data for Awareness, Alignment, Comprehension and Communication Environmental Changes across each condition, offering a clear overview of the data distribution for each ToM component.



(a)

(b)

Fig. 15. Boxplot compares the medians and spread of the data by condition for (a) Comprehension and (b) Communication.

Similar to the Transparency assessment, we used a factor analysis approach to integrate four core components: Awareness of human knowledge, Alignment with user's preferences, Comprehension and Communication of Environmental Changes. The KMO measure of sampling adequacy, with a value of 0.693, indicated the appropriateness of our dataset for factor analysis. The factor loadings from our analysis revealed a significant contribution of Comprehension, with a factor loading of 0.879. The remaining components included Communicating Environmental Changes (0.754), Alignment with user's preferences (0.709) and Awareness of human knowledge (0.644). To establish the relative importance of each component, we normalised the standardised factor loadings and calculated the following relative importance scores: $R(Comprehension)$: 0.293, $R(Communication)$: 0.251, $R(Alignment)$: 0.236 and $R(Awareness)$: 0.215.

ToM score was then computed using these weights in a weighted sum formula:

$$\text{ToM} = 0.215 \cdot \text{Awareness} + 0.236 \cdot \text{Alignment} + 0.293 \cdot \text{Comprehension} + 0.251 \cdot \text{Communication}. \tag{9}$$

Subsequently, a new Shapiro–Wilk test was conducted to assess the normality of the overall ToM data, which indicated a non-normal data distribution. Therefore, we proceeded with the Independent-Samples Kruskal–Wallis test. Our analysis revealed a statistically significant difference in terms of Awareness of human knowledge among the conditions, with a test statistic value of $H(3) = 29.857$, $p < 0.001$. *Post-hoc* pairwise comparisons indicated significant differences between Conditions C1 and

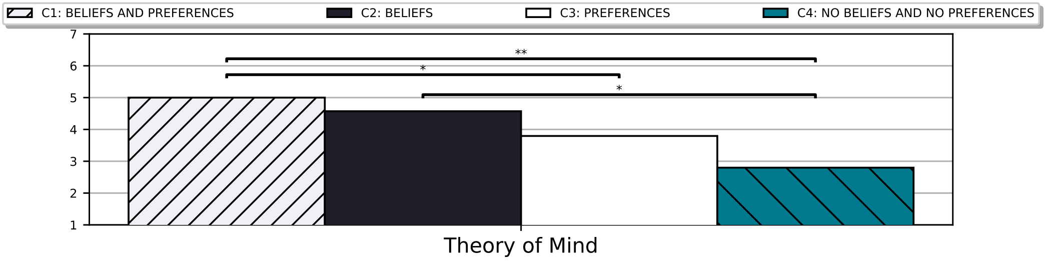


Fig. 16. Perceived ToM for each condition ($^*$ for $p < 0.05$ and $^{**}$ for $p \leq 0.001$).

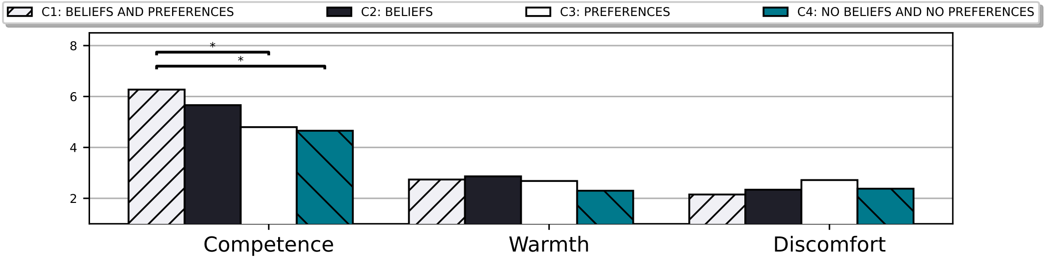


Fig. 17. Average of the RoSAS questionnaire responses for each condition ($^*$ for $p < 0.05$ and $^{**}$ for $p \leq 0.001$).

C4 ($p < 0.001$), C1 and C3 ($p = 0.045$) and C2 and C4 ($p < 0.001$). ToM scores for each condition are presented in Figure 16.

To further understand the relationship between perceived Transparency and perceived ToM, a Pearson correlation analysis was used to measure the relationship between the two variables. The results of this analysis revealed a statistically significant between perceived Transparency and ToM (0.607, $p < 0.01$). This positive correlation suggests that an increase (or decrease) in the perceived Transparency corresponds to an increase (or decrease) in the perceived ToM. This implies that individuals who perceive higher levels of Transparency are also likely to attribute a higher level of ToM to the robot.

### 6.4 RoSAS Ratings

Our initial analysis regarding the RoSAS questionnaire involved assessing the internal reliability across the conditions. We employed Cronbach's alpha test to evaluate the questionnaire's Competence, Warmth and Discomfort factors. The Competence factor demonstrated Cronbach's alpha values of $\alpha_{C1} = 0.70$, $\alpha_{C2} = 0.94$, $\alpha_{C3} = 0.79$ and $\alpha_{C4} = 0.92$. For the Warmth factor, we observed values of $\alpha_{C1} = 0.90$, $\alpha_{C2} = 0.87$, $\alpha_{C3} = 0.80$ and $\alpha_{C4} = 0.91$. The Discomfort factor yielded Cronbach's alpha values of $\alpha_{C1} = 0.83$, $\alpha_{C2} = 0.81$, $\alpha_{C3} = 0.89$ and $\alpha_{C4} = 0.88$.

To verify the normality of our data, we conducted the Shapiro–Wilk test for each condition in the RoSAS questionnaire. The results indicated that our data did not adhere to a normal distribution. As a result, we utilised a non-parametric test, specifically the Independent-Samples Kruskal–Wallis test, to investigate potential differences among the conditions.

Our analysis revealed a statistically significant difference in terms of Competence among the conditions, as evidenced by a test statistic value of $H(3) = 17.790$ and a significance level of $p < 0.001$. *Post-hoc* pairwise comparisons highlighted significant differences between Conditions C1 and C4 ($p = 0.003$) and between Conditions C1 and C3 ($p = 0.005$). However, no significant differences were detected between other conditions on other scales of RoSAS. The scores for each condition are visually represented in Figure 17.

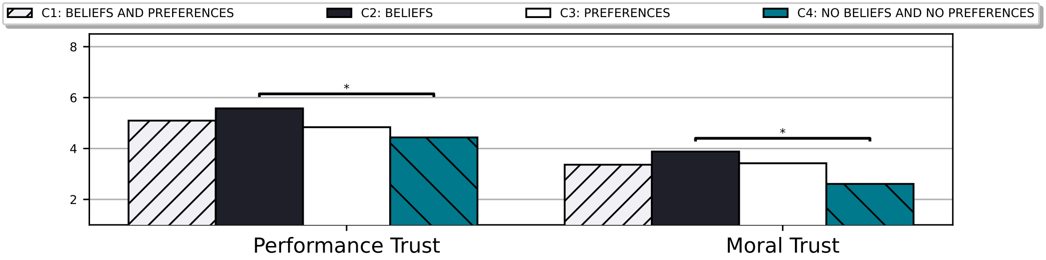


Fig. 18. Average of the MDMT questionnaire responses for each condition (* for $p < 0.05$ and ** for $p \leq 0.001$).

## 6.5 Trust Ratings

As a first step, we evaluated the internal reliability of the MDMT questionnaire for each condition. A Cronbach's alpha test for the Performance Trust factor of the MDMT questionnaire was respectively $\alpha_{C1} = 0.79$, $\alpha_{C2} = 0.91$, $\alpha_{C3} = 0.83$, $\alpha_{C4} = 0.92$. Finally, the Moral Trust factor had a Cronbach's alpha of $\alpha_{C1} = 0.93$, $\alpha_{C2} = 0.87$, $\alpha_{C3} = 0.91$, $\alpha_{C4} = 0.94$.

The variables were tested for normal distribution using the Shapiro–Wilk test. The data were normally distributed, $p > 0.05$. A Levene's test for assessing the equality of variances for our dataset was not statistically significant ($p > 0.05$), suggesting that the assumption of homogeneity of variances was met for our data. Following this, we used an Independent-Samples $t$-test to compare the means of our groups. The $t$-test showed a statistically significant difference between Condition 2 and Condition 4 in the Performance Trust ($t(55) = 2.535, p = 0.014$) and Moral Trust factor ($t(55) = 2.877, p = 0.006$). The MDMT scores for each condition are presented in Figure 18.

Subsequent analyses were conducted to elucidate the relationship between perceived Transparency, operationalised through the variables Legibility, Predictability and Expectability and the Transparency subscale of the MDMT v2. A Pearson correlation analysis yielded a significant positive correlation of 0.253 ($p < 0.01$) between the perceived Transparency and the MDMT Transparency subscale. This indicates that an increase in scores on the MDMT Transparency subscale is associated with a corresponding increase in scores on the perceived Transparency measure. Factor analysis further corroborated this relationship. The KMO measure of sampling adequacy was 0.762, and Bartlett's Test of Sphericity was significant ($\chi^2(21) = 238.639, p < 0.001$), suggesting that the data were suitable for factor analysis. The rotated component matrix revealed that the MDMT Transparency items (transparency, genuine, sincere, candid) loaded strongly onto one factor with loadings ranging from 0.686 to 0.844, while the perceived Transparency items loaded onto a separate factor with loadings ranging from 0.809 to 0.841.

A subsequent regression analysis was performed to predict perceived Transparency based on the MDMT Transparency subscale. The model was statistically significant, with an F-value of 4.981 and a significance level of $p = 0.001$. This suggests that at least one predictor among transparency, genuine, sincere and candid, significantly contributes to predicting perceived Transparency. Notably, transparency (the item from the MDMT Transparency subscale) had an unstandardised coefficient (B) of 0.273. The $t$-value for this item was 3.675, which is statistically significant ($p < 0.001$), indicating its significant predictive power for perceived Transparency. However, the genuine, sincere and candid predictors did not significantly contribute to the model when other variables were controlled for. Moreover, a new factor analysis incorporating Legibility, Predictability, Expectability and Transparency yielded a KMO measure of 0.753. The loadings for Legibility, Predictability and Expectability were 0.781, 0.837 and 0.805, respectively, while the loading for Transparency was lower at 0.587, indicating a moderate positive relationship.

## 7 Discussion

The study results shed light on the formulated hypotheses that explore the impact of ToM on Transparency within the scope of collaborative human–robot tasks. Firstly, concerning *Hypothesis 1*, the data suggest that the robot designed to consider user desires and beliefs (Condition C1) was indeed perceived as possessing a more developed ToM system compared to other conditions. This led to significant improvements in the interactions, particularly in evaluations of social skills and trust. However, it is noteworthy that the difference in perceived ToM between C1 and C2 was not statistically significant. This finding indicates that while integrating user desires and beliefs positively impacts perceived ToM, the extent of this impact may not significantly differ from a condition where only the user's beliefs are considered. As a result, we can affirm that our H1 is partially supported by our results. However, future studies could be conducted to analyse whether significant differences emerge when comparing the consideration of a unique mental state (beliefs or desire) to the consideration of both mental states simultaneously.

Secondly, regarding *Hypothesis 2*, the study results reveal that the robot with the most developed ToM system (Condition C1) is also perceived as displaying the highest Transparency. It is crucial to note that the Transparency ratings for Condition C2, which represents a ToM system without explicit consideration of user desires and beliefs, did not significantly differ from those of Condition C1. This outcome suggests that while ToM positively influences perceived Transparency, there may be a threshold beyond which additional enhancements in ToM do not significantly contribute to increased Transparency, as evidenced by the comparable ratings of C1 and C2. Moreover, Transparency in our experiment is expressed in a particular manner related to the scenario, which may differ from other tasks, potentially leading to different results. Similar to H1, our second hypothesis (H2) is partially supported by our results, indicating that further studies are required to compare the integration of a single mental state to the integration of both mental states simultaneously.

In addition, for *Hypothesis 3*, the study findings indicate that humans perceive the robot with the highest levels of trust to be Condition C2. Notably, Condition C2 significantly differs from Condition C4 in terms of performance and moral trust of the MDMT v2 scale; therefore, *Hypothesis 3* was not confirmed. This suggests that while Condition C1 was perceived as having the most advanced ToM system and high Transparency, it was Condition C2 that garnered the highest trust from humans. This could be due to various factors that are not directly related to Transparency or ToM, and further investigation is needed to understand these dynamics fully.

Finally, the results showed that a robot equipped with ToM capabilities, such as the ability to infer when humans hold incorrect beliefs (e.g., in a false-belief scenario), can adjust its actions and offer explanations that account for the human's misunderstanding. This improves the robot's ability to collaborate in dynamic environments and enhances the user's ability to follow the robot's reasoning. By providing explanations that consider the human's perspective, the robot's actions become more predictable and easier to comprehend, increasing overall transparency. A positive correlation was also observed between perceived transparency and ToM. This indicates that as a robot's ToM capabilities improve, users perceive the robot as more transparent. When the robot can infer and respond to human beliefs, users find it easier to follow the robot's reasoning and predict its behaviour.

### 7.1 Limitations

The study was conducted in an online environment, which inherently lacks the physical presence and real-time interaction that a traditional, in-person setting would provide. The lack of physical presence might limit the participants' ability to fully understand and appreciate the robot's behaviours and actions, which could, in turn, negatively affect their ratings of the robot's trust levels. Indeed, Kumar et al. [39] in their work showed that the enjoyment and trust levels when interacting with an in-person robot were higher than in the online experiment. However, previous work [64] suggested that online experiments have a similar effect on HRI performance as in-person experiments. Nevertheless, further research is needed to validate these results in real-world studies.

## 8 Conclusions

In this project, we aimed to explore the connection between Transparency and ToM in HRI, investigating how predicting users' mental states, including beliefs and desires, could enhance the collaboration between the human and the robot, particularly during false-belief contexts.

Our methodology combined HRL with a modified BToM version, which we referred to as HBToM. The model was able to track the beliefs and desires of a human agent, thereby improving collaboration in a multi-objective environment. We analysed four distinct conditions, assessing their impact on beliefs and desires on the robot's ToM for achieving higher transparency and trust levels. The findings of this study suggest that a robot, considering human desires and beliefs, not only enhances transparency but also mitigates false beliefs. This demonstrates the importance of endowing a robot with ToM, which can significantly benefit collaboration by improving transparency and trust while also fostering the robot's social skills. Specifically, it not only illustrates the robot as an effective machine for goal-oriented tasks but also as a personalised assistant. Furthermore, a positive correlation was observed between perceived transparency and ToM. In conclusion, these findings highlight the potential of ToM in enhancing robot transparency and, consequently, fostering successful HRIs.

Future studies will focus on evaluating the robustness of the proposed approach in more complex scenarios and real-world settings. The model will be refined to incorporate more complex human preferences, paving the way for its adoption in real-world settings. We are also planning on analysing the effect of a robot following users' beliefs, even when they have false beliefs. We believe that it will reinforce the impact of robots endowed with ToM.

## References

[1] Victoria Alonso and Paloma De La Puente. 2018. System transparency in shared autonomy: A mini review. *Frontiers in Neurorobotics* 12 (2018), 83.

[2] Georgios Angelopoulos, Pasquale Imparato, Alessandra Rossi, and Silvia Rossi. 2023. Using theory of mind in explanations for fostering transparency in human-robot interaction. In *International Conference on Social Robotics*. Springer, Singapore, 394–405.

[3] Georgios Angelopoulos, Dimitri Lacroix, Ricarda Wullenkord, Alessandra Rossi, Silvia Rossi, and Friederike Eyssel. 2024. Measuring transparency in intelligent robots. arXiv:2408.16865. Retrieved from http://arxiv.org/abs/2408.16865

[4] Zhangyunfan Bai and Ke Chen. 2024. Effects of explanations by robots on trust repair in human-robot collaborations. In *International Conference on Human-Computer Interaction*. Springer, 3–14.

[5] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. 2018. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems* 8, 4 (2018), 1–30.

[6] Chris Lawrence Baker. 2012. *Bayesian Theory of Mind: Modeling Human Reasoning about Beliefs, Desires, Goals, and Social Relations*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[7] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 0064.

[8] Chris L. Baker, Rebecca R. Saxe, and Joshua B. Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. *In 33rd Annual Conference of the Cognitive Science Society*, 2469–2474.

[9] Cristian-Paul Bara, Sky Ch-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. arXiv:2109.06275. Retrieved from https://arxiv.org/abs/2109.06275

[10] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21, 1 (1985), 37–46.

[11] Karen Bartsch and Henry Wellman. 1989. Young children's attribution of action to beliefs and desires. *Child Development* 60, 4 (1989), 946–964.

[12] P. G. Benardos and G.-C. Vosniakos. 2007. Optimizing feedforward artificial neural network architecture. *Engineering Applications of Artificial Intelligence* 20, 3 (2007), 365–382.

[13] Brenda Benninghoff, Philipp Kulms, Laura Hoffmann, and Nicole C. Krämer. 2013. Theory of mind in human-robot-communication: Appreciated or not? *Kognitive Systeme* 2013, 1 (2013), 1–7.

[14] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10, 1 (1995), 122–142. DOI: https://doi.org/10.1006/game.1995.1027

[15] Michael Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

[16] Daniel Sundquist Brown. 2020. *Safe and Efficient Inverse Reinforcement Learning*. Ph.D. Dissertation.

[17] Moritz C. Buehler, Jürgen Adamy, and Thomas H. Weisswange. 2021. Theory of mind based assistive communication in complex human robot cooperation. arXiv:2109.01355. Retrieved from https://arxiv.org/abs/2109.01355

[18] Guilhem Buisan, Anthony Favier, Amandine Mayima, and Rachid Alami. 2022. HATP/EHDA: A robot task planner anticipating and eliciting human decisions and actions. In *2022 International Conference on Robotics and Automation (ICRA)*, 2818–2824. DOI: https://doi.org/10.1109/ICRA46639.2022.9812227

[19] David Buttelmann, Malinda Carpenter, and Michael Tomasello. 2009. Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition* 112, 2 (2009), 337–342.

[20] David Buttelmann, Harriet Over, Malinda Carpenter, and Michael Tomasello. 2014. Eighteen-month-olds understand false beliefs in an unexpected-contents task. *Journal of Experimental Child Psychology* 119 (2014), 120–126.

[21] Stephanie M. Carlson, Melissa A. Koenig, and Madeline B. Harms. 2013. Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science* 4, 4 (2013), 391–402.

[22] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *2017 ACM/IEEE International Conference on Human-Robot Interaction*, 254–262.

[23] Ginevra Castellano. 2020. What kind of human-centric robotics do we need? Investigations from human-robot interactions in socially assistive scenarios. In *8th International Conference on Human-Agent Interaction*, 1–2.

[24] Daniel C. Dennett. 1989. *The Intentional Stance*. MIT Press, Cambridge, MA.

[25] Thomas G. Dietterich. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research* 13 (2000), 227–303.

[26] Thomas G. Dietterich. 1998. The MAXQ method for hierarchical reinforcement learning. In *International Conference on Machine Learning (ICML)*, Vol. 98, 118–126.

[27] Mica R. Endsley. 2017. From here to autonomy: Lessons learned from human–automation research. *Human Factors* 59, 1 (2017), 5–27.

[28] Anthony Favier, Shashank Shekhar, and Rachid Alami. 2023. Anticipating false beliefs and planning pertinent reactions in human-aware task planning with models of theory of mind. In *PlanRob Workshop—International Conference on Automated Planning and Scheduling (ICAPS '23)*. Retrieved from https://hal.science/hal-04163435

[29] John H. Flavell. 1999. Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology* 50, 1 (1999), 21–45.

[30] Philip Gerrans. 2002. The theory of mind module in evolutionary psychology. *Biology and Philosophy* 17 (2002), 305–321.

[31] Alison Gopnik and Virginia Slaughter. 1991. Young children's understanding of changes in their mental states. *Child Development* 62, 1 (1991), 98–110.

[32] Alison Gopnik and Henry M. Wellman. 1994. The theory theory. In *An Earlier Version of This Chapter Was Presented at the Society for Research in Child Development Meeting, 1991.* Cambridge University Press.

[33] Jesse Gray and Cynthia Breazeal. 2005. Toward helpful robot teammates: A simulation-theoretic approach for inferring mental states of others. In *AAAI 2005 Workshop on Modular Construction of Human-Like Intelligence*.

[34] Frank Hegel, Sören Krach, Tilo Kircher, Britta Wrede, and Gerhard Sagerer. 2008. Theory of mind (ToM) on robots: A functional neuroimaging study. In *3rd ACM/IEEE International Conference on Human Robot Interaction*, 335–342.

[35] Mehdi Hellou, Samuele Vinanzi, and Angelo Cangelosi. 2023. 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)*, 1893–1900.

[36] Guy Hoffman and Xuan Zhao. 2020. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction* 10, 1 (2020), 1–31.

[37] Julian Jara-Ettinger. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences* 29 (2019), 105–110.

[38] Christopher Krupenye and Josep Call. 2019. Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science* 10, 6 (2019), e1503.

[39] Shikhar Kumar, Eliran Itzhak, Yael Edan, Galit Nimrod, Vardit Sarne-Fleischmann, and Noam Tractinsky. 2022. Politeness in human–robot interaction: A multi-experiment study with non-humanoid robots. *International Journal of Social Robotics* 14, 8 (2022), 1805–1820.

[40] Sonya S. Kwak. 2014. The impact of the robot appearance types on social interaction with a robot and service evaluation of a robot. *Archives of Design Research* 27, 2 (2014), 81–93.

[41] Minha Lee, Peter Ruijten, Lily Frank, and Wijnand IJsselsteijn. 2023. Here's looking at you, robot: The transparency conundrum in HRI. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2120–2127.

[42] Alan M. Leslie. 1987. Pretense and representation: The origins of "theory of mind". *Psychological Review* 94, 4 (1987), 412–426.

[43] Gerald Matthews, Jinchao Lin, April Rose Panganiban, and Michael D. Long. 2020. Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems* 50, 3 (2019), 234–244.

[44] Peter E. McKenna, Marta Romeo, Jhielson Pimentel, Mohammed Diab, Meriam Moujahid, Helen Hastie, and Yiannis Demiris. 2023. Theory of mind and trust in human-robot navigation. In *1st International Symposium on Trustworthy Autonomous Systems*, 1–5.

[45] Andrew N. Meltzoff. 1995. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31, 5 (1995), 838–850.

[46] Wenxuan Mou, Martina Ruocco, Debora Zanatto, and Angelo Cangelosi. 2020. When would you trust a robot? A study on trust and theory of mind in human-robot interactions. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 956–962.

[47] Birthe Nesset, David A. Robb, José Lopes, and Helen Hastie. 2021. Transparency in HRI: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 313–317.

[48] Erhan Oztop, Daniel Wolpert, and Mitsuo Kawato. 2005. Mental state inference using visual control parameters. *Cognitive Brain Research* 22, 2 (2005), 129–151.

[49] Massimiliano Patacchiola and Angelo Cangelosi. 2022. A developmental cognitive architecture for trust and theory of mind in humanoid robots. *IEEE Transactions on Cybernetics* 52, 3 (2022), 1947–1959. DOI: https://doi.org/10.1109/TCYB.2020.3002892

[50] Jan Pöppel and Stefan Kopp. 2018. Satisficing models of Bayesian theory of mind for explaining behavior of differently uncertain agents: Socially interactive agents track. In *17th International Conference on Autonomous Agents and Multiagent Systems*, 470–478.

[51] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526. DOI: https://doi.org/10.1017/S0140525X00076512

[52] Beate Priewasser, Eva Rafetseder, Carina Gargitter, and Josef Perner. 2018. Helping as an early indicator of a theory of mind: Mentalism or teleology? *Cognitive Development* 46 (2018), 69–78.

[53] Laurel D. Riek. 2012. Wizard of OZ studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

[54] Marta Romeo, Peter E. McKenna, David A. Robb, Gnanathusharan Rajendran, Birthe Nesset, Angelo Cangelosi, and Helen Hastie. 2022. Exploring theory of mind for human-robot collaboration. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 461–468.

[55] Martina Ruocco, Wenxuan Mou, Angelo Cangelosi, Caroline Jay, and Debora Zanatto. 2021. Theory of mind improves human's trust in an iterative human-robot game. In *9th International Conference on Human-Agent Interaction, Virtual Event*. ACM, 227–234. DOI: https://doi.org/10.1145/3472307.3484176

[56] Dorsa Sadigh, Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. 2016. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*. David Hsu, Nancy M. Amato, Spring Berman and Sam Jacobs (Eds.), Vol. 2, MIT Press, Cambridge, MA, 1–9.

[57] Brian Scassellati. 2002. Theory of mind for a humanoid robot. *Autonomous Robots* 12, 1 (2002), 13–24.

[58] Daniel Ullman and Bertram F. Malle. 2019. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. *In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 618–619.

[59] Kimberly E. Vanderbilt, David Liu, and Gail D. Heyman. 2011. The development of distrust. *Child Development* 82, 5 (2011), 1372–1380.

[60] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. 2019. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 20180032.

[61] Henry M. Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72, 3 (2001), 655–684.

[62] Eric Wiewiora. 2011. Potential-based shaping and Q-value initialization are equivalent. arXiv:1106.5267. Retrieved from http://arxiv.org/abs/1106.5267

[63] Jessica Williams, Stephen M. Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in Artificial Intelligence* 5 (2022), 750763.

[64] Sarah N. Woods, Michael L. Walters, Kheng Lee Koay, and Kerstin Dautenhahn. 2006. Methodological issues in HRI: A comparison of live and video-based methods in robot to human approach direction trials. In *15th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN '06)*. IEEE, 51–58.

[65] Guang-Zhong Yang, Jim Bellingham, Pierre E. Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, Vijay Kumar, Marcia McNutt, Robert Merrifield, et al. 2018. The grand challenges of science robotics. *Science Robotics* 3, 14 (2018), eaar7650.

[66] Yi Zeng, Yuxuan Zhao, Tielin Zhang, Dongcheng Zhao, Feifei Zhao, and Enmeng Lu. 2020. A brain-inspired model of theory of mind. *Frontiers in Neurorobotics* 14 (2020), 60.

[67] Luke S. Zettlemoyer, Brian Milch, and Leslie Pack Kaelbling. 2008. Multi-agent filtering with infinitely nested beliefs. In *21st International Conference on Neural Information Processing Systems (NIPS '08)*. Curran Associates Inc., Red Hook, NY, 1905–1912.

[68] Tielin Zhang, Yi Zeng, Dongcheng Zhao, and Mengting Shi. 2018. A plasticity-centric approach to train the non-differential spiking neural networks. In *AAAI Conference on Artificial Intelligence*. AAAI Press, New Orleans, LA, 620–628.

Received 17 January 2024; revised 14 April 2025; accepted 22 May 2025