Modeling Higher-order Human Beliefs Using the Justified Perspective Model

Wanchun Li*
The University of Melbourne
Melbourne, Australia
wanchun@student.unimelb.edu.au

Chenyuan Zhang*
Monash University
Melbourne, Australia
Faculty of Engineering and IT
The University of Melbourne
Melbourne, Australia
chenyuan.zhang@monash.edu

Weijia Li
Faculty of Engineering and IT
The University of Melbourne
Melbourne, Australia
weijia3@student.unimelb.edu.au

Guang Hu
Faculty of Engineering and IT
The University of Melbourne
Melbourne, Australia
ghu1@student.unimelb.edu.au

Yangmengfei Xu
Department of Mechanical Engineering
The University of Melbourne
Melbourne, Australia
yangmengfeix@student.unimelb.edu.au

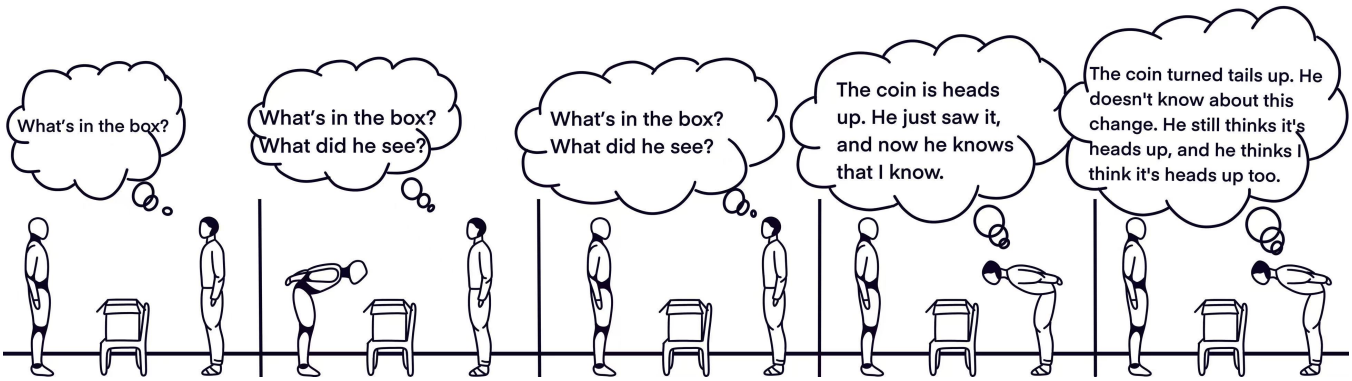

Figure 1: An example on coin problem.

Abstract

This study explores the feasibility of modeling higher-order human beliefs using a generalizable formalization based on the epistemic planning framework, the Justified Perspective (JP) model. Specifically, it investigates (a) whether individuals exhibit consistent belief reasoning abilities and (b) whether these abilities can be inferred from their nesting capabilities within the JP model framework. To address these questions, we propose a novel processing algorithm inspired by Item Response Theory to estimate reasoning abilities based on participants' responses to diverse reasoning scenarios. A pilot experiment was conducted to validate the methodology and refine the experimental design for future studies. While the small sample size limits the statistical significance of the findings, preliminary results suggest the JP model's potential to capture human higher-order beliefs. This work demonstrates the promise of integrating epistemic planning frameworks with human-centered applications, advancing the development of Human Computer Interaction systems capable of understanding and anticipating human cognition.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; • **Theory of computation** → *Semantics and reasoning*; • **Computing methodologies** → Planning and scheduling.

Keywords

Epistemic Planning, Theory of Mind, Item Response Theory

*Both authors contributed equally to this research.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720223

ACM Reference Format:
Wanchun Li, Chenyuan Zhang, Weijia Li, Guang Hu, and Yangmengfei Xu. 2025. Modeling Higher-order Human Beliefs Using the Justified Perspective Model. In *Extended Abstracts of the CHI Conference on Human Factors in*

Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3720223

1 Introduction

Over the past few decades, the development of robotics technologies has broadly expanded human-robot interaction across various applications, such as social robots [22, 30], collaborative supernumerary limbs [27], prosthetics [31], and rehabilitation devices [29]. However, much of the current interaction relies on external measurements (e.g. physiological data [6]) rather than achieving a deeper cognitive understanding. With cognitive understanding, it becomes possible to capture underlying mechanisms and predict human behavior. Thus, effectively understanding human reasoning becomes a key challenge in enhancing both performance and safety in human-robot interactions.

Consider a simple scenario where robot a and human b are observing a coin c placed in a box. The coin has two possible states: "head" and "tail". The robot or human can only know the state of the coin if they actively peek into the box. Additionally, they can see if the other is peeking. Their actions include "peek", "return" and "flip". The flip action and its outcome are only visible to the one currently peeking, while the one not peeking remains unaware. An example is shown in Figure 1. Initially, neither the robot nor the human has peeked, and the coin is in the "head" state. The robot peeks first and then returns. Subsequently, the human peeks and observes the coin flip. As a result, the robot forms a false belief about the coin's state. Specifically, in the final state, the coin is actually "tail", the human believes it is "tail", the robot believes it is "head", and the human believes that the robot believes it is "head". This difference in beliefs arises from the complex interaction between belief and observation, and was known as nested high-order belief. For example, a second-order belief usually represents the cognition on others' beliefs about their own beliefs [9, 24]. Modeling how humans form these beliefs, especially when involving nested beliefs, remains a significant challenge in human-robot interaction [3, 10, 28, 33].

Building on this complexity, the Theory of Mind (ToM) studies examine the cognitive ability to infer the beliefs, intentions, and emotions of others from their actions or social signals. A classic and widely recognized example in this field is the study by Baron-Cohen et al. [1], which introduced the "Sally-Anne" task. This seminal experiment demonstrated that children with autism have distinct challenges with ToM, specifically in recognizing and understanding the beliefs and intentions of others. In the context of artificial intelligence (AI), ToM provides a framework for predicting and explaining agents' behavior. Current ToM research in AI focuses on computational models for multi-robot simulations and decision-making inference. For example, Winfield and Jirotka [28] examined how robots infer human belief states for autonomous decision-making, and Zhao et al. [33] combined probabilistic programming with symbolic inference to analyze human beliefs from videos. However, these methods lack validation through comprehensive human experiments, limiting real-world application. Similarly, Buehler and Weisswange [3] modeled the human as rational agent acting in partially observable Markov decision process in simulations, but such settings do not fully capture natural human cognition. Gurney and Pynadath [10] highlighted the need for unified frameworks and benchmarking in ToM models to better align with human reasoning. Additionally, Rabinowitz et al. [23]'s learning-based ToM model showed potential in game theory for mimicking human strategic reasoning, yet its applicability remains confined to simulations. The existing body of work suggests that ToM models can handle specific belief reasoning and decision-making tasks; however, their capacity to generalize across different scenarios remains limited.

To address this gap, a systematic formalization dealing with knowledge and beliefs of an agent (either a robot or a human) was proposed and is known as epistemic planning (EP) [2]. EP combines epistemic logic with automatic planning, allowing agents to anticipate and respond to the knowledge and beliefs of other agents. For instance, Hansen and Bolander [11] used it to enable humanoid robots to perform ToM reasoning by updating first- and higher-order beliefs. Similarly, Shvo et al. [26] integrated EP with ToM to improve human-robot interaction by predicting human beliefs and resolving discrepancies. In another example, Shekhar et al. [25] proposed a human-aware task planning framework to predict and adjust belief divergences in human-robot collaboration. An innovative EP approach employs the Justified Perspective (JP) Model [12] for analyzing an agent's epistemic logic, drawing inspiration from two intuitions of human reasoning: human believes what they see; and, for the parts they could not see, human believes what they have seen in the past unless they saw evidence to suggest otherwise. However, these studies lack experiments involving human subjects and higher-order beliefs, which may limit the applicability of their findings to real-world human behaviors and responses.

This pilot study is designed to explore the feasibility of using the JP model to understand higher-order human beliefs. It is important to note that we do not seek to demonstrate that human belief reasoning operates through the same mechanism as the JP model. Rather, we aim to show that the JP model can serve as a useful tool for understanding human belief reasoning across different scenarios. To this end, we developed a series of belief reasoning questions to verify (a) whether the reasoning ability quantified under the JP model framework is consistent, and (b) whether belief reasoning ability directly results from nesting ability as suggested in the JP model. While no definitive conclusions can be drawn due to the limited sample size (5 subjects), our preliminary findings indicate that humans exhibit consistent reasoning ability, with a positive correlation observed between nesting ability and reasoning ability. These results suggest the JP model's potential to effectively capture human higher-order beliefs.

To summarize, the key contributions of this work are as follows: First, we propose an experimental paradigm and a novel processing algorithm, inspired by Item Response Theory (IRT), to measure individuals' belief reasoning abilities under the JP model framework. Second, we conduct a pilot study to validate the proposed paradigm, with preliminary results suggesting that the JP model can be utilized to understand human belief reasoning across different scenarios. Finally, we identify the limitations of the current pilot study and provide insights for refining future research.

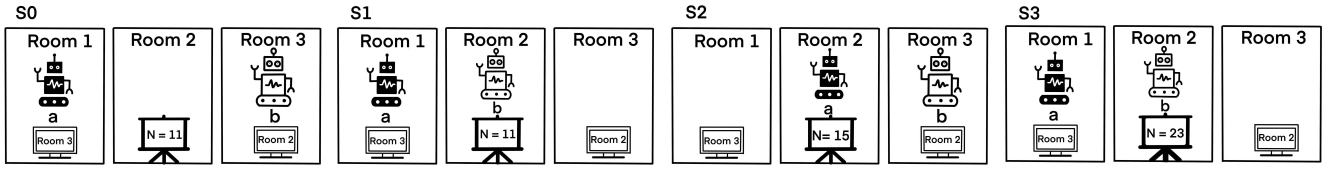

Figure 2: An example of the nesting test (A detailed analysis of this example, along with the belief reasoning results generated from the JP model, can be found in the Appendix A).

2 Methodology

This experiment serves as a pilot study to evaluate the feasibility of the methodology and aims to refine it for subsequent large-scale experiments. A total of five participants were involved in this experiment.

To improve readability, we first define a few key preliminaries. The belief of Agent i on the number N is represented as B_iN in the following sections. *Nesting ability* refers to the depth of belief about others' beliefs (arbitrary nesting) that an individual can understand and infer in an omniscient setting (e.g., without memory limitations). *Reasoning ability* refers to the depth of belief about others' beliefs (arbitrary nesting) that an individual can understand and infer based on available evidence, considering limited information, cognitive load, and other constraints in a specific context. As such, reasoning ability offers a more practical reflection of human belief reasoning in real-world scenarios. For simplicity, this pilot study focuses solely on investigating reasoning ability under memory load.

To investigate the feasibility of using the JP model to anticipate higher-order human beliefs, we will test the following two hypotheses within the JP model framework: (a) **reasoning ability of individuals is consistent across various scenarios**; and, (b) **human belief reasoning abilities are positively correlated to their nesting abilities**.

2.1 Task

The experiment was structured as a computerized test to evaluate the memory, nesting, and reasoning abilities of the participants. Each participant was individually seated in front of a monitor and presented with a scenario involving two robots, a and b, illustrated in Figure 2. These robots could navigate freely among three rooms: Room 1, Room 2, and Room 3. In Room 2, a board presented a two-digit number which either robot could modify. Rooms 1 and 3 each had a monitor showing the occupation of another room. For example, in state 0 (s_0), Robot a is in Room 1, and Robot b is in Room 3 with the board in Room 2 displaying the number "11". Robot a can see that Robot b is in Room 3, while Robot b knows Room 2 is empty and thus infers that Robot a is in Room 1. Before testing, participants were thoroughly briefed on the action sequences and setup to ensure comprehension.

As shown below, the test comprised eight questions, designed based on the cognitive limit of four nesting levels as suggested by the literature [5], using all questions for each level (e.g., 2 for 2 robots) to comprehensively assess this cognitive function. Questions were formatted to evaluate the participants' understanding of nested beliefs, such as B_aB_bN represents Robot a's belief about Robot b's belief regarding the number N. The answers to each question were determined using the JP model at the corresponding nesting level, and participants' responses were compared to these answers to assess their nesting abilities.

- Level 1: B_aN =?, B_bN =?
- Level 2: B_aB_bN =?, B_bB_aN =?
- Level 3: $B_aB_bB_aN$ =?, $B_bB_aB_bN$ =?
- Level 4: $B_aB_bB_aB_bN$ =?, $B_bB_aB_bB_aN$ =?

2.2 Protocol

Given the requirement for reasoning ability, which involves generating beliefs based on limited information, this experiment necessitates participants' capability to remember three two-digit numbers. Therefore, a memory test is conducted during the preliminary phase to screen valid subjects. Following this, a nesting test and a reasoning test are sequentially conducted in Phase 1 and Phase 2, respectively, to assess the hypotheses.

2.2.1 Pre-phase: Memory Test. In the memory test, subjects are presented with sequences of two-digit numbers displayed one at a time on the screen. After each sequence, participants must type the numbers in the correct order. The sequence length is adjusted using an adaptive design [16]: it increases by one after a correct response and decreases by one following an incorrect response. This process continues until the participant accurately recalls sequences of the same length three consecutive times, determining their maximum recall capacity. Participants who successfully recall more than three numbers qualify for subsequent tests. Based on the pioneer literature on human memory ability (4 ± 1 theory) [4], the initial sequence consists of four numbers.

2.2.2 Phase 1: Nesting Test. The nesting test begins with a demonstration designed to familiarize participants with the concept of nesting, the interface, and the notation B_iN, using the coin example as shown in Figure 1. The demonstration includes three scenarios, each featuring different action sequences. After the initial familiarization, as outlined in Section 2.1, the nesting test is conducted with a scenario encompassing four states illustrated in Figure 2. In this test, the two-digit numbers on the board in Room 2 are randomly generated before the experiment and the same for different participants.

2.2.3 Phase 2: Reasoning Test. Augmenting the nesting test with limited information, the reasoning test requires the subjects to remember previously displayed numbers, making it more reflective of real-world scenarios and human reasoning processes. In this test, the two-digit number on the board in Room 2 in the previous state disappeared when a new state was introduced. For example, as

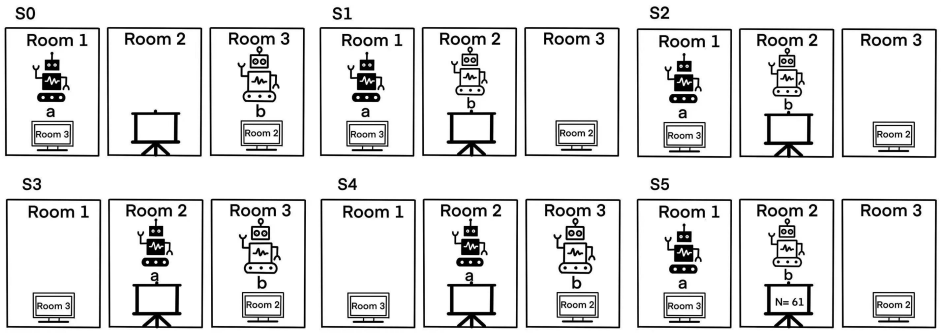

Figure 3: An example of reasoning test.

illustrated in Figure 3, when s_5 is displayed, the number on the board in s_4 disappeared. In the reasoning test, only the numbers need to be memorized, as the action sequence remains visible throughout the trial to control for distracting factors.

The reasoning test includes four different scenarios. Each scenario in the experiment consists of nine states, starting from state 0 (s_0) to state 8 (s_8). Each scenario is structured into seven trials that progressively reveal the states to the participants. In Trial 1, s_0 to s_2 are displayed simultaneously. Trial 2 then extends the display to include State 3, showing s_0 to s_3 together. Subsequent trials continue this pattern of adding one state per trial, so that by Trial 3, States 0 to 4 are displayed, continuing until Trial 7, which displays all states from 0 to 8.

In each of these trials, participants need to answer eight questions that test their understanding and reasoning based on the currently displayed states, as introduced in Section 2.1. The two-digit numbers are randomly generated before the experiment and are different in each scenario, yet consistent across different participants. The first four states in the first scenario are the same as those in the nesting test to validate the efficacy of the test, and the second scenario changes the action sequences of the robots from the first trial. The third scenario differs from the first scenario with different seeing rules (the information on the monitor in Room 1 and 3), and the fourth scenario differs from the first scenario with both action sequences and seeing rules.

2.3 Data Analysis

The binary responses, indicating whether each answer is correct or not (i.e. consistent with the answer generated from the JP model), are recorded for all questions from both tests. In the nesting test, each subject answers 8 questions, while in the reasoning test, each subject answers 56 questions in each scenario; therefore, 224 questions in total.

2.3.1 Nesting Test. A participant's nesting ability is evaluated based on their consistent ability to answer questions correctly at the assessed level and all preceding levels. In other words, to be classified as possessing a specific level of nesting ability, a participant must answer all questions correctly at that level and at all lower levels.

2.3.2 Reasoning Test. As the reasoning test requires subjects to remember numbers, there is potential noise from guessing or errors. To analyze this noisy data, Item Response Theory (IRT) [7], a probabilistic model that describes the relationship between a subject's latent ability (θ) and their test item responses, was employed. In this study, reasoning ability is estimated using the two-parameter logistic model (2PL) of the IRT, defined as $P(x|\theta, a, b) = \frac{1}{1+e^{-a(\theta-b)}}$. Here, x is the subject's response to the question, and 1 for correct, 0 for incorrect; $P(x|\theta, a, b)$ represents the probability of a subject correctly/incorrectly answering an item; θ denotes the subject's reasoning ability inferred from the test, constrained to $0 \leq \theta \leq 4$; a is the discrimination parameter of the item, set to 1 to indicate uniform discrimination, meaning each item is considered equally effective at distinguishing between different levels of subject ability; b is the difficulty parameter, aligned with the nesting levels of questions in the reasoning test and set at $b = 1, 2, 3, 4$ for levels 1 through 4, respectively.

To estimate the reasoning ability (θ) of each subject, the Maximum Likelihood Estimation (MLE) method is used. The likelihood function for a subject's responses is calculated as $L(\theta) = \prod_{i=1}^{n} P(x_i|\theta, a_i, b_i)^{x_i}(1 - P(x_i|\theta, a_i, b_i))^{1-x_i}$, where n is the total number of questions. In this experiment, there are 56 questions in each scenario. The reasoning ability $\hat{\theta}$ is estimated by maximizing the likelihood function: $\hat{\theta} = \arg\max_\theta L(\theta)$. In this experiment, $\hat{\theta}$ is calculated in each scenario j, denoted $\hat{\theta}_j$ for each subject.

3 Preliminary Results

Potential statistical analysis methods for future human experiments are discussed in Section 4.4.

The results of the pre-phase (memory test) are shown in the second column in Table 1. All participants had a memory span of three to five numbers, supporting the 4 ± 1 theory of short-term memory in [4]. Some participants were observed to be able to remember all the numbers but struggled to recall them in the right order. This suggests that retrieval can be challenging even when memory capacity is enough.

The results of the nesting test are shown in the third column in Table 1 and the results agreed with the cognitive limit of four nesting levels from the literature [5]. The preliminary results of the reasoning test are shown in the last five columns in Table 1.

Table 1: Preliminary results in memory test, nesting test, and reasoning test.

	Memory	Nesting	Reasoning test				
Subject	test	test	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	Variance
Subject 1	4	2	2.44	2.45	2.44	2.46	0.0001
Subject 2	5	3	3.00	3.00	3.00	3.00	0.0000
Subject 3	4	3	2.94	2.96	3.00	3.00	0.0009
Subject 4	3	2	2.54	2.56	2.54	2.57	0.0002
Subject 5	4	3	2.88	2.86	3.00	3.00	0.0057

The average variance among the five participants in the reasoning test was 0.0014, which is notably smaller than the total variance of all $\hat{\theta}$ values (0.0556). This suggests that individual reasoning ability remains consistent across scenarios within the JP framework (**Hypothesis a**). Additionally, participants with a nesting ability assessed at level 3 outperformed those with a nesting ability at level 2 in the reasoning test. This indicates a positive correlation between belief reasoning ability and nesting ability (**Hypothesis b**).

4 Discussion

This pilot study aimed to explore the feasibility of using the JP model to understand higher-order human belief reasoning. Although the small sample size and other limitations prevent definitive conclusions, the preliminary results suggest that human reasoning ability is consistent within the JP model framework, and that a positive correlation exists between reasoning ability and nesting ability, as proposed by the model. In this section, we discuss the limitations of the current study and outline directions for future research.

4.1 Scenario Design

In epistemic planning, it is important to test the generalization of the approach on diverse domains. In the current reasoning test, although four scenarios were developed with changes to seeing rules and action sequences in the reasoning test, they are still within the same "Number" domain [18]. This limitation may reduce the practical relevance of our findings and their applicability to real-world contexts. Additionally, similar scenarios could induce more pronounced learning effects, and participants may develop task-specific strategies that do not reflect general reasoning abilities, further diminishing the reliability of the results. In subsequent experiments, the scenarios in the "Number" domain should be expanded to include more numbers of agents, rooms, and different seeing rules to enhance generalizability. Furthermore, introducing new domains such as "Corridor" [15] and "Grapevine" [20] would enhance the diversity of the experimental conditions. In the Corridor domain, robots are able to share knowledge with nearby robots, while in the Grapevine domain, they can share or distort their beliefs. These scenarios introduce additional complexity and could be more effective in differentiating participants' reasoning abilities across a broader range of contexts.

The single-trial design of the nesting test shares similar issues with the reasoning test, reducing its reliability and robustness in identifying meaningful nesting ability. To improve the nesting test, future experiments could incorporate diverse domains and tasks across multiple scenarios, as outlined above.

4.2 Participant Engagement

The current task lacks reward and/or feedback, and the inclusion of an excessive number of number-filling questions has led to reduced participant engagement. Offering performance feedback after each scenario could promote a sense of accomplishment and competence, helping to maintain participants' motivation and focus throughout the experiment. In future experiments, external rewards, such as points or incentives, could be introduced to further enhance motivation, as suggested by existing literature [14].

Additionally, the pilot study took approximately 1 hour and 30 minutes to complete, which was significantly longer than anticipated. This extended duration resulted in participant fatigue, which affected concentration and response time. To mitigate these issues, future experiments should be divided into multiple stages with breaks in between.

4.3 Learning Effect

Unsurprisingly, a notable learning effect was observed, and most participants (4 out of 5) demonstrated improved performance in later scenarios, as they became familiar with task structures, rules, and strategies. This learning effect potentially introduces bias when analyzing inter-scenario data, and may lead to overestimating participants' reasoning ability, leading to a less reliable conclusion. As mentioned in Section 4.1, introducing diverse tasks/domains could be helpful to reduce the learning effect. With the similar setup of nesting and reasoning tests, the scenarios among these two tests could be mixed in a random or reverse order for different subjects to statistically control for any biases that might arise from the order.

4.4 Data Analysis

The pilot study included only five participants, which is not sufficient to detect meaningful effects or group differences with any statistical tool. The results of this study mainly validated the feasibility and rationale of the methodology, rather than providing strong evidence to accept or reject the hypotheses. In the subsequent experiment, a larger number of participants will be recruited. Tools like G*Power and Bayesian Factor Design Analysis will be used to calculate the necessary sample size [8] based on key factors such as expected effect size, significance level, statistical power, and statistical method.

With a large sample size, we can run statistical analysis to verify our hypothesis as follows.

4.4.1 Hypothesis a: reasoning ability of individuals is consistent across various scenarios. This hypothesis can be examined using a Bayesian approach by comparing two different parameterizations of IRT models. The first model assumes that each individual has a consistent reasoning ability across all scenarios, while the second model allows reasoning ability to vary across scenarios. Given the observed data, we can estimate the posterior distributions for both models. Model comparison will be conducted using the Bayes factor, with the model assuming consistency in reasoning ability placed in the numerator. A larger Bayes factor would indicate stronger

evidence in favor of the hypothesis that reasoning ability remains consistent across different scenarios.

4.4.2 Hypothesis b: human belief reasoning abilities are positively correlated to their nesting abilities. Once we obtain human reasoning ability and nesting ability using the processing algorithm proposed in this paper, a correlation analysis can be performed. If the correlation is found to be significant, we can accept the hypothesis.

4.5 Future Works

If both hypotheses are verified by future experiments, it means the JP model can be used to understand humans' high-order belief reasoning.

By integrating human goal recognition models [32], autonomous agents could better anticipate human objectives, enabling more efficient assistance. Consider a scenario where a human interacts with a supernumerary limb, such as a wearable robotic arm. Prior research has demonstrated the effectiveness of a third arm controlled via physiological signals (e.g., foot movements) for completing tasks that typically require three hands [21]. However, this approach may impose additional cognitive load on users. With the personalized JP model, the supernumerary limb could simulate the user's higher-order beliefs about the environment. By incorporating human goal recognition models, the robotic arm could proactively anticipate user intentions and assist in task completion without requiring direct control inputs.

Another promising future direction is applying the proposed model to understanding trust in human-AI interaction. As a specialized form of belief, trust plays a crucial role in ensuring safety in human-robot collaboration. The proposed model allows for the nesting of trust models across multiple agents, providing a valuable framework for studying trust in multi-agent interactions [13, 17, 19].

5 Conclusion

This work presents a pilot study involving five participants to explore the feasibility of anticipating higher-order human beliefs through a generalizable formalization using the Justified Perspective (JP) model. We introduce a novel processing algorithm inspired by Item Response Theory (IRT) to effectively estimate human belief reasoning ability from their responses across different scenarios. The preliminary results suggest that human reasoning ability is consistent within the JP model framework, and is positively correlated with nesting ability. The limitations of the current study have been identified and discussed, along with potential statistical analysis methods for future empirical research. These findings highlight the potential of bridging theoretical epistemic frameworks with practical human-centered applications, providing valuable insights for the design of intelligent human-computer interaction systems that more effectively comprehend human cognition.

References

[1] 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.

[2] Thomas Bolander and Mikkel Birkegaard Andersen. 2011. Epistemic planning for single-and multi-agent systems. *Journal of Applied Non-Classical Logics* 21, 1 (2011), 9–34.

[3] Moritz C. Buehler and Thomas H. Weisswange. 2018. Online inference of human belief for cooperative robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*. IEEE, Madrid, Spain, 409–415. doi:10.1109/IROS.2018.8594076

[4] Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences* 24, 1 (2001), 87–114.

[5] Robin IM Dunbar. 2003. The social brain: mind, language, and society in evolutionary perspective. *Annual review of Anthropology* 32, 1 (2003), 163–181.

[6] Jonathan Eden, Mario Bräcklein, Jaime Ibáñez, Deren Yusuf Barsakcioglu, Giovanni Di Pino, Dario Farina, Etienne Burdet, and Carsten Mehring. 2022. Principles of human movement augmentation and the challenges in making it a reality. *Nature Communications* 13, 1 (2022), 1345.

[7] Susan E Embretson and Steven P Reise. 2013. *Item response theory for psychologists*. Psychology Press, New York, USA.

[8] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[9] Liesbeth Flobbe, Rineke Verbrugge, Petra Hendriks, and Irene Krämer. 2008. Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information* 17 (2008), 417–442.

[10] Nikolos Gurney and David V. Pynadath. 2022. Robots with Theory of Mind for Humans: A Survey*. In *31st IEEE International Conference on Robot and Human Interactive Communication, RO-MAN*. IEEE, Napoli, Italy, 993–1000. doi:10.1109/RO-MAN53752.2022.9900662

[11] Lasse Dissing Hansen and Thomas Bolander. 2020. Implementing theory of mind on a robot using dynamic epistemic logic. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence Organization, Yokohama, Japan, 1615–1621.

[12] Guang Hu, Tim Miller, and Nir Lipovetzky. 2023. Planning with Multi-Agent Belief Using Justified Perspectives. In *Proceedings of the Thirty-Third International Conference on Automated Planning and Scheduling*, Sven Koenig, Roni Stern, and Mauro Vallati (Eds.). AAAI Press, Prague, Czech Republic, 180–188. doi:10.1609/ICAPS.V33I1.27193

[13] Carolina Centeio Jorge, Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams. In *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021) (CEUR Workshop Proceedings, Vol. 3022)*, Dongxia Wang, Rino Falcone, and Jie Zhang (Eds.). CEUR-WS.org, London, UK.

[14] Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin* 119, 2 (1996), 254.

[15] Filippos Kominis and Hector Geffner. 2015. Beliefs In Multiagent Planning: From One Agent to Many. In *Proceedings of the Twenty-Fifth International Conference on Automated Planning and Scheduling, ICAPS*, Ronen I. Brafman, Carmel Domshlak, Patrik Haslum, and Shlomo Zilberstein (Eds.). AAAI Press, Jerusalem, Israel, 147–155. http://www.aaai.org/ocs/index.php/ICAPS/ICAPS15/paper/view/10617

[16] Marjorie R Leek. 2001. Adaptive procedures in psychophysical research. *Perception & psychophysics* 63, 8 (2001), 1279–1292.

[17] Roy J Lewicki and Chad Brinsfield. 2015. Trust research: measuring trust beliefs and behaviours. In *Handbook of research methods on trust*. Edward Elgar Publishing, Cheltenham, UK, 46–64.

[18] Weijia Li, Guang Hu, and Yangmengfei Xu. 2024. Beyond Static Assumptions: the Predictive Justified Perspective Model for Epistemic Planning. *CoRR* abs/2412.07941 (2024). doi:10.48550/ARXIV.2412.07941 arXiv:2412.07941

[19] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. 2024. Integrity-based explanations for fostering appropriate trust in AI agents. *ACM Transactions on Interactive Intelligent Systems* 14, 1 (2024), 1–36.

[20] Christian J. Muise, Vaishak Belle, Paolo Felli, Sheila A. McIlraith, Tim Miller, Adrian R. Pearce, and Liz Sonenberg. 2015. Planning Over Multi-Agent Epistemic States: A Classical Planning Approach. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Blai Bonet and Sven Koenig (Eds.). AAAI Press, Austin, Texas, 3327–3334. doi:10.1609/AAAI.V29I1.9665

[21] A Noccaro, J Eden, G Di Pino, D Formica, and E Burdet. 2021. Human performance in three-hands tasks. *Scientific reports* 11, 1 (2021), 9511.

[22] Nami Ogawa, Jun Baba, and Junya Nakanishi. 2024. Investigating Effect of Altered Auditory Feedback on Self-Representation, Subjective Operator Experience, and Task Performance in Teleoperation of a Social Robot. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 588, 18 pages. doi:10.1145/3613904.3642561

[23] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew M. Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning, ICML (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.).

PMLR, Stockholmsmässan, Stockholm, Sweden, 4215–4224. http://proceedings.mlr.press/v80/rabinowitz18a.html
[24] Fernanda Rubio, Catherine Neira, César Villacura-Herrera, and Ramon Castillo Guevara. 2022. First and Second-Order Theory of Mind as Predictors of Cooperative Behaviors in Preschool and School Children. *Psykhe* 31, SI 1 (2022), 1–15.
[25] Shashank Shekhar, Anthony Favier, and Rachid Alami. 2024. An Epistemic Human-Aware Task Planner which Anticipates Human Beliefs and Decisions. *CoRR* abs/2409.18545 (2024). doi:10.48550/ARXIV.2409.18545 arXiv:2409.18545
[26] Maayan Shvo, Ruthrash Hari, Ziggy O'Reilly, Sophia Abolore, Sze-Yuh Nina Wang, and Sheila A. McIlraith. 2022. Proactive Robotic Assistance via Theory of Mind. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*. IEEE, Kyoto, Japan, 9148–9155. doi:10.1109/IROS47612.2022.9981627
[27] Kohei Umezawa, Yuta Suzuki, Gowrishankar Ganesh, and Yoichi Miyawaki. 2022. Bodily ownership of an independent supernumerary limb: an exploratory study. *Scientific reports* 12, 1 (2022), 2339.
[28] Alan FT Winfield and Marina Jirotka. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180085.
[29] Yangmengfei Xu, Vincent Crocher, Justin Fong, Ying Tan, and Denny Oetomo. 2021. Inducing human motor adaptation without explicit error feedback: A motor cost approach. *IEEE transactions on neural systems and rehabilitation engineering* 29 (2021), 1403–1412.
[30] Yangmengfei Xu, Suxuan Tian, Guojing Wang, and Bin Tang. 2025. From Isolation to Connection: Community Service Robots for Social Cohesion and Sustainability. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25)*. IEEE Press, Melbourne, Australia, 1953–1956.
[31] Tianshi Yu, Alireza Mohammadi, Ying Tan, Peter Choong, and Denny Oetomo. 2024. Discrete-Target Prosthesis Control Using Uncertainty-Aware Classification for Smooth and Efficient Gross Arm Movement. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 32 (2024), 3210–3221. doi:10.1109/TNSRE.2024.3450973
[32] Chenyuan Zhang, Charles Kemp, and Nir Lipovetzky. 2024. Human Goal Recognition as Bayesian Inference: Investigating the Impact of Actions, Timing, and Goal Solvability. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, Auckland, New Zealand, 2066–2074. doi:10.5555/3635637.3663071
[33] Yibiao Zhao, Steven Holtzen, Tao Gao, and Song-Chun Zhu. 2015. Represent and Infer Human Theory of Mind for Human-Robot Interaction. In *2015 AAAI Fall Symposia*. AAAI Press, USA, Arlington, Virginia, 158. http://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11719

A An example of JP model

A.1 State 0

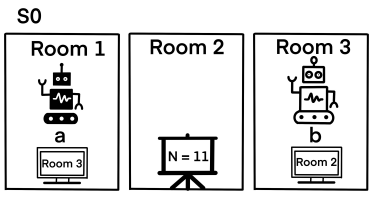

Figure 4: An example of the nesting test s_0.

In the s_0, Robot a is in Room 1, and Robot b is in Room 3. And the board in Room 2 shows that the number is 11. And at this time, neither Robot a nor Robot b knows that the number on the board in room 2, so they have no beliefs.

- Level 1: $B_aN = none$, $B_bN = none$
- Level 2: $B_aB_bN = none$, $B_bB_aN = none$
- Level 3: $B_aB_bB_aN = none$, $B_bB_aB_bN = none$
- Level 4: $B_aB_bB_aB_bN = none$, $B_bB_aB_bB_aN = none$

A.2 State 1

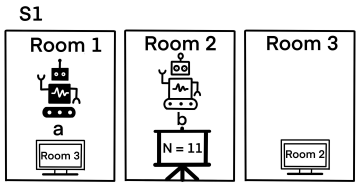

Figure 5: An example of the nesting test s_1.

In the s_1, Robot a remains in the room 1, and Robot b enters the room 2, and sees the number 11. Robot b knows that the number is 11, Robot a still has no beliefs. Also, Robot a knows that Robot b has seen the number, but Robot a doesn't know what it is.

- Level 1: $B_aN = none$, $B_bN = 11$
- Level 2: $B_aB_bN = none$, $B_bB_aN = none$
- Level 3: $B_aB_bB_aN = none$, $B_bB_aB_bN = none$
- Level 4: $B_aB_bB_aB_bN = none$, $B_bB_aB_bB_aN = none$

A.3 State 2

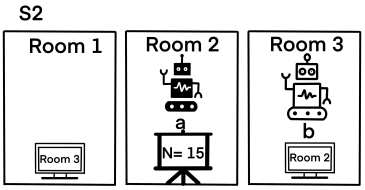

Figure 6: An example of the nesting test s_2.

In the s_2, Robot b leaves the Room 2, and goes back to the Room 3, and Robot a enters into the Room 2 and changes the number on the board in the room 2 from 11 to 15.

- Level 1: $B_aN = 15$, $B_bN = 11$

Robot a enters Room 2 and sees that the number is 11. At this point, Robot a realizes that Robot b must have seen 11 earlier. Then, Robot a changes the number to 15, and Robot b doesn't know that this number change.

- Level 2: $B_aB_bN = 11$, $B_bB_aN = 11$

Robot a changes the number to 15. At this point, a knows that b is unaware of this change and still believes the number is 11. Robot b thinks that Robot a sees the number 11.

- Level 3: $B_aB_bB_aN = 11$, $B_bB_aB_bN = 11$

Robot a knows that Robot b is unaware of the change to the number, so Robot a knows that Robot b still believes the number what Robot a thinks is 11.

- Level 4: $B_aB_bB_aB_bN = 11$, $B_bB_aB_bB_aN = 11$

Similar to the level 3.

A.4 State 3

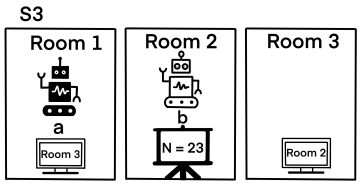

Figure 7: An example of the nesting test s_3.

In the s_3, Robot b goes to the Room 2 and sees the number 15 and changes it to the 23. And Robot a goes back to the Room 1.

- Level 1: $B_a N = 15$, $B_b N = 23$

At this state, Robot a still believes the number is 15 because it doesn't realize the number has been changed. Robot b, on the other hand, believes the number is 23 because it just changed the number.

- Level 2: $B_a B_b N = 15$, $B_b B_a N = 15$

Robot a still believes the number is 15. Robot b believes Robot a believes the number is 15, because it sees the number 15 which is changed by Robot a, so it knows that Robot a believes the number is 15.

- Level 3: $B_a B_b B_a N = 15$, $B_b B_a B_b N = 11$

Robot a still believes the number is 15. And Robot b thinks that Robot a still thinks the belief of Robot b remains in the previous state, which is 11.

- Level 4: $B_a B_b B_a B_b N = 15$, $B_b B_a B_b B_a N = 11$

Similar to the level 3.