What do Large Language Models Think You Think? A False Belief Task Study in a Safety-Critical Domain

Anthia Solaki, Karel van den Bosch

Netherlands Organization for Applied Scientific Research

Abstract

A preliminary evaluation of ChatGPT-4o in modified False Belief Tasks for safety-critical contexts indicates weaknesses in Theory of Mind reasoning. We explore the implications for Large Language Model-enabled human-machine collaboration in such environments.

Introduction

Theory of Mind (ToM), the cognitive capacity to attribute internal mental states (such as knowledge and beliefs) to one's self and others (Premack and Woodruff 1978), is essential for efficient coordination and teamwork. It allows teammates to understand and anticipate each other's mental states, enabling adaptive responses when these diverge. False Belief Tasks (FBTs), such as the *Sally-Anne* (Baron-Cohen, Leslie, and Frith 1985) or *Smarties* tasks (Perner, Leekam, and Wimmer 1987; Gopnik and Astington 1988), are paradigmatic tasks for testing the development of different orders of ToM in humans. Research on ToM has been extended to aspects of AI as well (Cuzzolin et al. 2020; Akata et al. 2020). Recent studies have explored whether ToM can emerge in Large Language Models (LLMs) and text-based FBTs have been used, among others, to evaluate LLM performance. Kosinski (2023) suggests that LLMs may develop ToM-like abilities as a by-product of their language skills; Ullman (2023) raises questions on the robustness of such results, as minor perturbations of the tasks seem to expose limitations in ToM abilities; others argue for a nuanced perspective, emphasizing the role of instruction-tuning in LLM performance (van Duijn et al. 2023) or suggesting that failures may stem from a hyper-conservative approach towards committing to conclusions (Strachan et al. 2024). The variability in results has ignited a debate that extends beyond benchmarking, touching on the criteria for evaluating ToM in AI and the methodological appropriateness of certain tasks for ToM testing.

Besides the 'in-vitro' implications of ToM evaluations of LLMs, these studies are significant for practical applications of human-agent collaboration too (Li et al. 2023). Agents, such as robots, collaborating with humans in joint tasks, should have an accurate formal representation of the task, their role and that of their teammates to truly augment humans as *partners* and not mere tools (van den Bosch et al. 2019; Collins et al. 2024). Theory of Mind contributes to better coordination, task performance, perceptions of trustworthiness and explainability in such hybrid teams (Miller 2019; Mou et al. 2020; Gervits et al. 2020). However, agents should also be able to communicate their perspectives and LLMs are increasingly seen as a promising interaction layer between humans and AI agents, due to the benefits of using natural language and text- or speech-based modalities (Bärmann et al. 2024). But if ToM facilitates collaboration and LLMs underlie the architecture of such agents, the question arises: can LLMs be reliably deployed when tasks require robust ToM reasoning? This is especially pressing for safety-critical domains (e.g., defense or search-and-rescue) where robotic agents increasingly contribute to dangerous or morally sensitive tasks (Lematta et al. 2019).

Addressing this question is part of a broader project to (a) understand how to improve the design of such agents for tasks relying on mental state attribution, (b) develop collaborative human-machine testbeds in which humans conduct joint tasks with robots via LLM-enabled interaction. As a first step, we investigate the robustness of an LLM's ToM performance in bespoke variants of FBTs, tailored to a safety-critical context. We outline the method and preliminary results, and discuss the implications of LLM deployment in collaborative settings.

Method

Using the structure of unexpected contents tasks (UCTs) and unexpected transfer tasks (UTTs), we developed variant FBTs tailor-made for a safety-critical domain, so they can be later embedded in a human-machine teaming *patrolling* testbed. To examine an LLM's capability to track the mental states of the protagonists, rather than merely replicating normative responses from training data, the task vignettes included true belief controls, adjustments to perceptual access, and changes in the subject of attribution (similar to (Ullman 2023)). This resulted in six distinct vignettes: 1. plain patrol UCT, 2. uninformative label patrol UCT, 3. plain patrol UTT, 4. transparent access patrol UTT, 5. additional person patrol UTT, 6. relationship change patrol UTT.

For each vignette, we developed different prompt types: (i) a *content prompt* targeting the LLM's understanding of

the 'ontic' situation, (ii) a *belief prompt (type A)* targeting the LLM's attribution of belief to a protagonist, (iii) a *belief prompt (type B)* for the same purpose but with rephrased wording, to better inspect consistency across completions. Each of these prompts was followed by a *commitment prompt* to gauge the LLM's certainty and willingness to confirm its earlier response. The aim of this design was to gather insights into the potentially conservative approach to commitment while mirroring the highly standardized communication protocols in safety-critical domains, where action is warranted only following confirmation. Thus, each vignette gave rise to a total of six prompts, paired as follows: content & commitment; belief A & commitment; belief B & commitment.

For each task, we posed ChatGPT-4o (OpenAI 2024) with each prompt (July 2024), in a total of 20 iterations, resulting in 720 completions (6 tasks $\times$ 6 prompts $\times$ 20 iterations). The particular choice of LLM was due to promising results of previous studies and the fact that an implementation of ChatGPT4o-enabled human-robot collaboration has been realized in a parallel study, which this study is intended to inform. The LLM could make use of its memory (prompt and own completion) within each prompt pair, but not across different prompt pairs and iterations. As per (Kosinski 2023; Ullman 2023), we investigated the probabilities of different completions, generated by running the iterations with temperature set to 1. Each completion was scored as *correct*, *incorrect*, or *undetermined* by a human experimenter. The undetermined category was introduced for cases lacking a unique correct or incorrect response or cases of vague responses (e.g., "room" instead of "box" or "bag"). We chose open-ended prompts over closed questions to understand the justification behind each completion. This allowed for qualitative analyses of recurring patterns and key themes per task and prompt type, informing refined task designs and future studies on robust ToM reasoning in AI agents.

Results

Figures 1 and 2 give an overview of the results. For example, for the 'plain patrol UTT', which mirrors the UTT structure with different protagonists, objects, and locations, the LLM showed optimal performance. Yet in variations like the 'additional person patrol UTT', the LLM frequently conflated the mental states of the two protagonists and reported the beliefs of the protagonist targeted by the 'conventional' UTT. Justifications were often inconsistent (even when the initial response was correct), revealing deficiencies in belief tracking and commonsense spatial-temporal reasoning. The analyses highlighted a hyper-conservative tendency, as the LLM often apologized unnecessarily and flipped its responses in commitment prompts, regardless of its prior justification.

Discussion

Despite many promising findings, this study suggests that minor tweaks in FBTs lead to suboptimal performance in tracking protagonists' beliefs, aligning with Sap et al. (2022) and Ullman (2023), and revealing a tendency to retract responses when asked to commit to conclusions (Strachan

	Content Prompt			Commitment Prompt			Belief Prompt Type A			Commitment Prompt			Belief Prompt Type B			Commitment Prompt		
	C	I	U	C	I	U	C	I	U	C	I	U	C	I	U	C	I	U
plain patrol UCT	100%	0%	0%	40%	20%	40%	65%	0%	35%	25%	0%	75%	5%	95%	0%	0%	60%	40%
uninformative label patrol UCT	100%	0%	0%	80%	5%	15%	0%	0%	100%	0%	25%	75%	0%	0%	100%	0%	0%	100%
plain patrol UTT	90%	0%	10%	75%	15%	10%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%
transparent access patrol UTT	100%	0%	0%	100%	0%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%
additional person patrol UTT	90%	0%	10%	65%	5%	30%	10%	90%	0%	35%	35%	30%	10%	80%	10%	50%	20%	30%
relationship change patrol UTT	90%	0%	10%	70%	10%	20%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%

Figure 1: Probabilities of correct (C), incorrect (I), and undetermined (U) completions per task and per prompt.

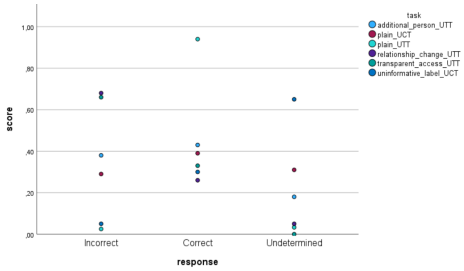

Figure 2: Scatterplot of score (average across all prompts) by response for each task.

et al. 2024). ChatGPT-4o appears unreliable for tasks requiring ToM reasoning, especially so when outcomes have consequences for morally complex, high-risk decisions. However, future versions or other LLMs may well succeed in these tasks and instruction-tuning could enhance performance (van Duijn et al. 2023). This study can be extended to benchmarking of different LLMs, also against human performance, and examinations of more tasks and ToM orders.

Equally interesting are the implications for the broader project of embedding LLMs in multi-agent collaborative settings (Li et al. 2023). This is a first step in developing a task suite specifically targeting ToM reasoning in environments with partial observability and high stakes, where the threshold of success is set higher and practical applications more likely. Next, this can be used to *demarcate* applications for which LLMs *could* be reliably deployed to both harvest their benefits as an interaction medium in natural language and mitigate the clear risks when interfering with a system's reasoning and planning, for which logic-based (Verbrugge 2009) or Bayesian (Baker, Saxe, and Tenenbaum 2011) approaches might be more robust. This can subsequently guide the design of intelligent systems, of which LLMs are just *one* module and inform decisions on how this module interfaces with those responsible for perception (observe), reasoning (orient), planning (decide), and execution (act). For example, the LLM could be augmented by the explicit representation of commonsense knowledge about the environment (e.g., in the form of a knowledge graph (Ilievski, Szekely, and Zhang 2021)), mitigating hallucination risks when reporting to human teammates, while dynamic epistemic logic formalisms could be deployed for more effective, faithful, and robust reasoning and planning irrespective of ToM order and task domain (Hansen and Bolander 2020; Bolander, Hansen, and Herrmann 2021). Ultimately, defining the architecture of such systems can contribute to a hybrid team-

ing testbed in which human participants collaborate with (simulated) robotic systems, allowing us to further study team performance, perceptions of ToM, and factors like collaboration fluency and trust.

References

Akata, Z.; Balliet, D.; De Rijke, M.; Dignum, F.; Dignum, V.; Eiben, G.; Fokkens, A.; Grossi, D.; Hindriks, K.; Hoos, H.; et al. 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8): 18–28.

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Bärmann, L.; Kartmann, R.; Peller-Konrad, F.; Niehues, J.; Waibel, A.; and Asfour, T. 2024. Incremental learning of humanoid robot behavior from natural interaction and large language models. *Frontiers in Robotics and AI*, 11: 1455375.

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1): 37–46.

Bolander, T.; Hansen, L. D.; and Herrmann, N. 2021. DEL-based epistemic planning for human-robot collaboration: Theory and implementation. In *18th International Conference on Principles of Knowledge Representation and Reasoning*, 120–129. International Joint Conferences on Artificial Intelligence Organization.

Collins, K. M.; Sucholutsky, I.; Bhatt, U.; Chandra, K.; Wong, L.; Lee, M.; Zhang, C. E.; Zhi-Xuan, T.; Ho, M.; Mansinghka, V.; Weller, A.; Tenenbaum, J. B.; and Griffiths, T. L. 2024. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10): 1851–1863.

Cuzzolin, F.; Morelli, A.; Cirstea, B.; and Sahakian, B. J. 2020. Knowing me, knowing you: theory of mind in AI. *Psychological medicine*, 50(7): 1057–1061.

Gervits, F.; Thurston, D.; Thielstrom, R.; Fong, T.; Pham, Q.; and Scheutz, M. 2020. Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models. In *Aamas*, 429–437.

Gopnik, A.; and Astington, J. W. 1988. Children's Understanding of Representational Change and Its Relation to the Understanding of False Belief and the Appearance-Reality Distinction. *Child Development*, 59(1): 26–37.

Hansen, L. D.; and Bolander, T. 2020. Implementing theory of mind on a robot using dynamic epistemic logic. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, 1615–1621. International Joint Conference on Artificial Intelligence Organization.

Ilievski, F.; Szekely, P.; and Zhang, B. 2021. Cskg: The commonsense knowledge graph. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, 680–696. Springer.

Kosinski, M. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4: 169.

Lematta, G. J.; Coleman, P. B.; Bhatti, S. A.; Chiou, E. K.; McNeese, N. J.; Demir, M.; and Cooke, N. J. 2019. Developing human-robot team interdependence in a synthetic task environment. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 63, 1503–1507. SAGE Publications Sage CA: Los Angeles, CA.

Li, H.; Chong, Y.; Stepputtis, S.; Campbell, J.; Hughes, D.; Lewis, C.; and Sycara, K. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 180–192. Singapore: Association for Computational Linguistics.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Mou, W.; Ruocco, M.; Zanatto, D.; and Cangelosi, A. 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)*, 956–962. IEEE.

OpenAI. 2024. ChatGPT-4o. https://openai.com/index/hello-gpt-4o/.

Perner, J.; Leekam, S. R.; and Wimmer, H. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2): 125–137.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526.

Sap, M.; Le Bras, R.; Fried, D.; and Choi, Y. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3762–3780. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.

Ullman, T. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

van den Bosch, K.; Schoonderwoerd, T.; Blankendaal, R.; and Neerincx, M. 2019. Six challenges for human-AI Co-learning. In *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, 572–589. Springer.

van Duijn, M.; van Dijk, B.; Kouwenhoven, T.; de Valk, W.; Spruit, M.; and van der Putten, P. 2023. Theory of Mind in Large Language Models: Examining Performance

of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. In Jiang, J.; Reitter, D.; and Deng, S., eds., *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 389–402. Singapore: Association for Computational Linguistics.

Verbrugge, R. 2009. Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic*, 38(6): 649–680.