A Metacognitive Architecture for ToM Revision in AI Agents

Jisu Kim, Mahimul Islam, Ashok Goel

Design Intelligence Laboratory, Georgia Institute of Technology
jisu.kim@gatech.edu, mahimul@gatech.edu, ashok.goel@cc.gatech.edu

Abstract

This paper presents a metacognitive architecture for revising an AI agent's Theory of Mind (ToM) to address misinterpretations in human–AI interaction. The ability to revise an agent's interpretations of users' mental states and characteristics is critical for maintaining trust and positive perceptions, especially in AI-mediated social interactions. To enable ToM revision, we introduce a two-level metacognitive architecture that integrates knowledge-based AI (KBAI) with LLMs. The architecture comprises a cognitive layer that performs the agent's core tasks, and a metacognitive layer that introspects on the cognitive layer using a Task–Method–Knowledge (TMK) model of the agent. The metacognitive layer (1) revises its interpretation of the user in response to user feedback and (2) communicates the revision process to the user.

Motivation

Artificial intelligence (AI) agents often face situations in which revising their interpretations of users is crucial. For example, social AI agents that use large language models (LLMs) are inherently prone to misinterpreting users' intentions, preferences, or characteristics in AI-mediated social interactions (Wang 2024). Such mistakes in user interpretation erode trust in human–AI interactions and harm perceptions of the agent's intelligence and likability (Honig and Oron-Gilad 2018; Lee et al. 2024; Salem et al. 2015).

In human–AI communication, after users recognize the agent's misinterpretation, they may provide feedback to help the agent revise its interpretation. This brings us to the Theory of Mind (ToM) revision (Wang 2024). During ToM revision, the agent introspects, revises its interpretation of the user based on user feedback, and communicates the revision to the user. ToM revision can shape users' perceptions of the agent and enhance trust when misinterpretations arise (Ashktorab et al. 2019; Wang 2024).

A promising approach to supporting ToM revision is to provide AI agents with a metacognitive structure. Metacognition is the process of "reasoning about one's own reasoning" (Cox 2005). It provides a higher-level mechanism for agents to reflect on and adapt their behavior (Cox and Raja 2007). Our prior work on metacognition in AI focused on endowing an agent with a theory of its own mind to support self-explanation (da Miranda et al. 2025). We extend this line of research by enabling the agent not only to explain, but also to revise its interpretation of the user in response to feedback.

In this work, we introduce a metacognitive architecture for ToM revision. The architecture integrates knowledge-based AI (KBAI) with LLMs. It comprises two layers: a cognitive layer that performs the agent's core tasks, and a metacognitive layer that uses the Task–Method–Knowledge (TMK) model to introspect on the cognitive layer. The metacognitive layer responds to user feedback by identifying misinterpretations, revising its interpretation, and communicating the revision via a step-by-step explanation.

ToM Revision in SAMI

SAMI is an AI social agent deployed in Georgia Tech's Online Master of Science in Computer Science program for ten semesters, serving over 11,000 users. It recommends social connections among users based on shared interests and characteristics extracted from their online posts (Kakar et al. 2024). During deployment, users frequently requested revisions to the agent's knowledge base. These requests aimed either to (1) correct LLM-induced misinterpretations, in which the agent failed to infer the contextual meaning of extracted entities, or to (2) update user information due to external changes.

For example, the agent may misinterpret a user's current location when the user mentions both a prior and a current location in their self-introduction post: *"I am an AI researcher in Atlanta, but I was born in Seoul."* In this example, the agent misinterprets "Seoul" as the current location and generates recommendations accordingly. The user then provides feedback indicating the misinterpretation: *"I am not in Seoul. I am in Atlanta."* This feedback triggers the agent's ToM revision. The agent extracts relevant entities from the feedback and localizes the source of misinterpretation. It updates the knowledge base by changing the user's primary location from "Seoul" to "Atlanta". Finally, the agent communicates the revision process to the user through a step-by-step explanation as follows:

- **Extract entities**: I am analyzing your feedback, and it seems that I mistook your location as "Seoul" instead of "Atlanta."
- **Locate relevant task**: I traced the issue back to my

entity-extraction step for your introduction post.

- **Identify misinterpretation**: I misinterpreted "Seoul" in "I was born in Seoul" as your current location.
- **Revise knowledge base**: I updated my knowledge base by replacing "Seoul" with "Atlanta" as your current location.

Metacognitive Architecture for ToM Revision

We present a metacognitive architecture for ToM revision (Figure 1). The architecture comprises two layers: a cognitive layer (Level 1) and a metacognitive layer (Level 2). At Level 1 (the cognitive layer), the agent performs its core tasks. At Level 2 (the metacognitive layer), it introspects on Level 1 to identify the causes of misinterpretations about the user, revise its interpretation, and generate a revision message. For a detailed description of the architecture, see our full paper (Kim, Islam, and Goel 2026).

Level 1 Reasoning

At Level 1, the agent generates initial social recommendations based on users' introduction posts. The agent first classifies the type of post. If the post is an introduction, the agent extracts entities such as hobbies, locations, and academic interests. The agent stores the extracted information in a knowledge base. It then applies a matchmaking algorithm over the knowledge base to identify users with shared attributes. Finally, it generates personalized recommendation messages (Kakar et al. 2024).

Task, Method, Knowledge Representation

To support Level 2 introspection over Level 1, we use TMK to represent the agent's Level 1 reasoning process. TMK enables ToM reasoning by providing the agent with a structured, interpretable self-model of its internal processes. It encodes three components: Tasks denote the goals the agent tries to achieve, Methods specify how tasks are executed, and Knowledge refers to the information the agent uses (Goel and Rugaber 2014). For ToM revision, we focus on the Task model. It decomposes the agent's Level 1 process into interpretable units, such as *"Identify and Extract Entities from the post."* This expressiveness provides the structural basis for localizing where a misinterpretation has happened (Goel and Rugaber 2017).

Level 2 Reasoning

At Level 2, the metacognitive layer leverages TMK to introspect on Level 1. When a user's post is classified as feedback requesting a revision, the agent executes Level 2 to identify the cause of the misinterpretation and revise its interpretation of the user. First, the agent extracts task-relevant entities, the specific information the user is requesting to revise, from the user feedback. It then identifies the Task in the TMK that led to the misinterpretation. Using the identified task, the agent performs a dictionary lookup in a solution library to retrieve the associated revision function. We constructed this solution library from misinterpretations observed during deployment and their corresponding revision functions. The agent then applies the retrieved function to update the user data stored in the knowledge base. At each stage, the agent generates intermediate natural-language messages describing its reasoning and actions. These messages are compiled into a single step-by-step revision message and presented to the user.

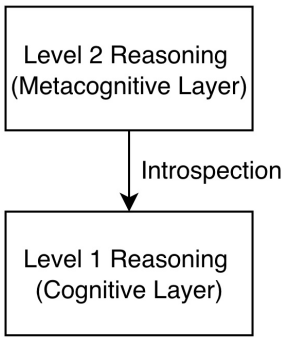

Figure 1: Overview of the two-level metacognitive architecture for ToM revision. Level 1 executes the agent's primary tasks, and Level 2 introspects on Level 1 to revise the agent's interpretation of the user.

Discussion

To examine the behavior of the proposed architecture across different misinterpretation scenarios, we employed a controlled synthetic data check (Nauta et al. 2023). We used 20 cases derived from real user data and misinterpretations observed during SAMI's deployment. Of the 20 cases, the metacognitive architecture completed revisions in 15 cases. In these cases, the agent identified the task-relevant misinterpretation, revised its knowledge base for the user, and generated an explanation of the revision. These results suggest that the architecture can improve the interpretability and transparency of an AI agent's behavior by revising and communicating its interpretation of the user.

More broadly, this work offers a generalizable architecture for developing trustworthy AI agents. It demonstrates how a metacognitive layer can support introspection for ToM revision. Furthermore, our work highlights the complementary benefits of combining KBAI and LLMs. KBAI, including TMK, the solution library, and the knowledge base, supports more interpretable and controlled agent behavior (Pan et al. 2024; Gaur and Sheth 2024). At the same time, LLMs provide the semantic flexibility needed to generate natural-language reasoning and responses. Our work points to new directions for designing AI agents that foster trust in human–AI interaction through transparent ToM revision and communication.

Acknowledgments

This research was supported by a grant from the US NSF (#2247790) to the National AI Institute of AI for Adult Learning and Online Education (aialoe.org).

References

Ashktorab, Z.; Jain, M.; Liao, Q. V.; and Weisz, J. D. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Cox, M.; and Raja, A. 2007. Metareasoning: A manifesto. *BBN Technical*.

Cox, M. T. 2005. Metacognition in computation: A selected research review. *Artificial Intelligence*, 169(2): 104–141.

da Miranda, M. P.; Islam, M.; Basappa, R.; Taylor, T.; and Goel, A. 2025. Adaptable Social AI Agents. In *Proceedings of the 1st Workshop on Advancing Artificial Intelligence through Theory of Mind (AAAI-25)*.

Gaur, M.; and Sheth, A. 2024. Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety. *AI Magazine*, 45(1): 139–155.

Goel, A. K.; and Rugaber, S. 2014. Interactive metareasoning: Towards a CAD-like environment for designing game-playing agents. In *Computational creativity research: Towards creative machines*, 347–370. Springer.

Goel, A. K.; and Rugaber, S. 2017. GAIA: A CAD-Like Environment for Designing Game-Playing Agents. *IEEE Intelligent Systems*, 32(3): 60–67.

Honig, S.; and Oron-Gilad, T. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, 9: 861.

Kakar, S.; Basappa, R.; Camacho, I.; Griswold, C.; Houk, A.; Leung, C.; Tekman, M.; Westervelt, P.; Wang, Q.; and Goel, A. K. 2024. SAMI: an AI actor for fostering social interactions in online classrooms. In *International Conference on Intelligent Tutoring Systems*, 149–161. Springer.

Kim, J.; Islam, M.; and Goel, A. 2026. A Metacognitive Architecture for Correcting LLM Errors in AI Agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lee, Y.; Son, K.; Kim, T. S.; Kim, J.; Chung, J. J. Y.; Adar, E.; and Kim, J. 2024. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent ai generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2518–2531.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; Van Keulen, M.; and Seifert, C. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s): 1–42.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 3580–3599.

Salem, M.; Lakatos, G.; Amirabdollahian, F.; and Dautenhahn, K. 2015. Would you trust a (faulty) robot? Effects of error, task type, and personality on human-robot cooperation and trust. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 141–148.

Wang, Q. 2024. *Mutual theory of mind for human-AI communication in AI-mediated social interaction*. Ph.D. thesis, Ph. D. Dissertation. Georgia Institute of Technology.