Establishing the Cooperative Game Wavelength as a Testbed to Explore Mutual Theory of Mind

Katelyn Morrison[1], Zahra Ashktorab[2], Djallel Bouneffouf[2], Gabriel Enrique Gonzalez[2], and Justin D. Weisz[2]

[1]Carnegie Mellon University
[2]IBM Research

Abstract

Machine learning (ML) and human-centered AI (HCAI) researchers have considered numerous methods to evaluate Theory of Mind (ToM)-like capabilities in artificial intelligence (AI). These methods have independently captured multiple aspects of ToM capabilities (*i.e.*, beliefs, knowledge). Recent research has proposed exploring Mutual Theory of Mind (MToM) as a way to understand how a human's mental model and an AI's user model can be mutually shaped to benefit future interactions. However, there is a lack of methods for understanding the development and impact of MToM-like capabilities in human-AI teams. We propose using a collaborative party game called Wavelength as a testbed to explore the complexities of MToM-like capabilities in human-AI teams. We compare Wavelength to other methods (*i.e.*, Overcooked, Hanabi) and discuss how game mechanics help players mutually construct, recognize, and revise their models of their teammates. Lastly, we briefly suggest how future work can explore MToM with Wavelength.

1 Introduction

Tasks and communications among humans have been effective due to humans' Theory of Mind (ToM) capabilities: the ability to make conjectures about the thoughts, feelings, and intentions of others [20, 1]. Recently, this concept has been applied to AI systems, such as large language models, to evaluate their ability to solve problems that require the types of perspective-taking provided by having a ToM-like capability [12, 21]. With the increasing use of artificial intelligence (AI) in countless user-facing applications, it is essential to understand if AI has ToM-like capabilities [26]. Similarly, research has captured the importance of humans accurately developing a mental model of the AI system to improve their collaboration [1, 13].

While ToM is unidirectional [25], recent research has proposed exploring bidirectional dynamics in human-AI teams to understand how a human's mental model and AI's user model may be mutually shaped through interactions with each other [25]. This process, also known as Mutual Theory of Mind [25], is particularly important during cooperative tasks where continuous behavior adaption and recursive mentalizing are necessary [25]. ML and HCAI researchers have leveraged various methods to assess the Theory of Mind and Mutual Theory of Mind capabilities in human-human and human-AI collaborations [16]. Despite these approaches, enabling MToM can lead to insights on how to enhance collaborations and design

interactions with collaborative AI systems [26].

We present a case for the cooperative party game Wavelength [6] as a unique space to study MToM-like capabilities in human-AI teams due to its mechanics requiring the team members to (1) find common ground [9], (2) coordinate perspectives [26], and (3) adapt behaviors [27]. Finding common ground is an important aspect of human-AI symbiosis [8] that many cooperative games do not require and thus cannot observe. While Wavelength does not cover all contexts where MToM-like capabilities can exist and flourish, it provides an interesting space for cooperative teaming contexts where adapting behaviors are necessary.

2 Related Work

Cooperative games have long served as a critical testbed for understanding how humans collaborate with AI, offering structured environments where shared goals between humans and AI can be examined. Research in this area has highlighted the ability of cooperative games to explore user perceptions of teammates [3, 2] and decision-making strategies [18] within human-AI teams. Additionally, numerous recent works in the human-computer interaction community have investigated mental models and Theory of Mind with cooperative games, such as Hanabi [15, 22, 11, 5, 14, 4, 17], Taboo [19, 28], CodeNames [23, 24], and purpose-built games [13, 3, 2, 29, 7]. Despite research being increasingly conducted in this area, previous works have not considered nor analyzed game mechanics from an MToM perspective.

Hanabi. Hanabi has been increasingly popular in the space of mental/user modeling. In Hanabi, players need to interpret their teammates' clues and predict how their teammates will act based on what they tell them, relying strongly on their beliefs to align with their teammate's beliefs. The strict communication (interpreting other players' intentions) mirrors human intentionality strategies, and subtle signaling [10]. These elements—cooperative play, information asymmetry, and limited structured communication—make Hanabi a compelling test case for AI research, as it presents unique challenges in modeling teamwork and intention inference [5].

Despite the growing interest in Hanabi, recent work has pointed out the difficulties of studying Theory of Mind-like capabilities in Hanabi with human-AI teams due to AI's current limitations and difficulties in getting AI to play competently [22]. [22] found that human superiority in Hanabi stems from the manipulation of physical game pieces, rule negotiation, and social coordination, where AI agents still fall short. With information asymmetry being core to Hanabi, this presents a challenging environment to investigate human-AI teams' ability to find common ground and coordinate their behaviors based on ongoing gameplay.

Other Approaches. Numerous other approaches have been used to understand, simulate, and capture the Theory of Mind-like capabilities in AI [16]. For example, one cooperative game, Taboo, has been considered in a ToM-based experiment [28]. However, research has found that people without shared knowledge could still perform well when playing Taboo as players do not need to rely on hidden information or be able to predict their teammate's next move[19]. Therefore, constructing a model of your teammate may not be necessary for game success, and the process of MToM may never be observed. The potential to play the game without both teammates needing to construct models of each other raises concerns about the suitability of the game as a testbed to study MToM.

Aside from Taboo, other work explores ToM-like capabilities in human-AI teams with an environment derived from the game Overcooked [29, 7]. ANother thread of work explore AI's ToM-like capabilities with the cooperative game CodeNames [23]. CodeNames presents similar game mechanics to Taboo and requires more agents/humans to play, making exploring these capabilities in dyads difficult. And, while

Overcooked may be ideal for exploring MToM processes in human-AI teams, the task environment can get overly complex, making it difficult to test different interventions that help human-AI teams leverage MToM. Additionally, it is more difficult to capture if and how team members find common ground due to Overcooked being task-based instead of reliant on knowledge.

3 MToM with Wavelength

Wavelength is a cooperative subjective rating game released in 2019 by CMYK Games [6]. Wavelength is designed to assess how well players can read each other's thoughts. Players work together to score points by predicting the location of a hidden target on a spectrum based on a clue. The game consists of two rounds: the clue-giving round and the clue-guessing round.

During the clue-giving round, one player aims to provide a clue that conceptually maps to "where the provided target range is located on the given spectrum" [6]. For example, if the clue-giver was given the spectrum of *simple vs. difficult* and a target region closer to the *difficult* side, a good clue for their teammate might be *space travel*.

During the clue-guessing round, the challenge is not only in selecting the position of the target but also in understanding "the psychic's" thought process, which requires considering how the psychic views the spectrum in relation to their teammate, the clue they provided, and where the target falls on the spectrum. The game's outcome is successful when the players' perceptions align, making it an attractive tool to assess Mutual Theory of Mind.

Imagine the following gameplay example with an AI teammate: in the first round, the AI has the spectrum *underrated skill vs. overrated skill* and gives the human the clue *prompt engineering* as an *overrated skill* based on its assumption that the human does not know about the struggles and difficulties associated with prompt engineering. However, the human guesses the target region is closer to *underrated skill* as they have themselves struggled with prompt engineering, and they imagine that the AI would level with them on such a skill being difficult.

Informed by our own play of Wavelength and observation of strangers playing the game together, we argue that a team's success at Wavelength is strongly tied to knowing about your teammate and being able to recognize what your teammate knows about you. This awareness occurs during the review of the guessing round when you and your teammate see the clue, target location, and guessed location, helping players naturally recognize their [incorrect] models of their teammates and sparking conversations to find common ground and "revise" those models of each other. By finding common ground through gameplay, players coordinate their perspectives by [mis]aligning clues/guesses and adapt their behaviors by tailoring future clues, strengthening their team's performance. While these can be captured through Wavelength, other games (*i.e.*, Hanabi, CodeNames, Overcooked) cannot due to the game mecahnics.

4 Conclusion

Large language models have been shown to be capable of playing games [**?**, 24], which is why we believe AI can play Wavelength as well. Wavelength can allow for the exploration of MToM-like capabilities within human-human, AI-AI, and human-AI teams. Wavelength can help address challenges such as if a certain architecture has ToM- and MToM-like capabilities. Additionally, the game presents a unique design space for methods that help human-AI teams more effectively construct, recognize, revise, and maintain their models of each other throughout collaborations. Wavelength game mechanics allow for easy measurement of subjective and behavioral observations of a team's MToM through game score, alignment of clue/guess rationalizations, and per-

ception of clue personalization. Ultimately, we hope this initiates engaging conversations and studies across the AI and HCAI communities to progress human-AI collaborations through novel approaches for Mutual Theory of Mind-like capabilities.

Acknowledgements

Grammarly and Writeful were used to help rephrase content to improve clarity and check for spelling or grammar mistakes. Any content generated with AI was further modified to align with the author's writing style.

References

[1] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175, 2023.

[2] Zahra Ashktorab, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. Effects of communication directionality and ai agent differences in human-ai interaction. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15, 2021.

[3] Zahra Ashktorab, Q Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–20, 2020.

[4] Christiane Attig, Patricia Wollstadt, Tim Schrills, Thomas Franke, and Christiane B Wiebel-Herboth. More than task performance: Developing new criteria for successful human-ai teaming using the cooperative card game hanabi. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2024.

[5] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

[6] BoardGameGeek. Wavelength, 2019.

[7] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

[8] Janghee Cho and Emilee Rader. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–28, 2020.

[9] HH Clark. Grounding in communication. *Perspectives on socially shared cognition/American Psychological Association*, 1991.

[10] Markus Eger, Chris Martens, Pablo Sauma Chacón, Marcela Alfaro Córdoba, and Jeisson Hidalgo-Cespedes. Operationalizing intentionality to play hanabi with human players. *IEEE Transactions on Games*, 13(4):388–397, 2020.

[11] Andrew Fuchs, Michael Walton, Theresa Chadwick, and Doug Lange. Theory of mind for deep reinforcement learning in hanabi. *arXiv preprint arXiv:2101.09328*, 2021.

[12] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–12, 2020.

[14] Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7187–7194, 2020.

[15] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[16] Yuanyuan Mao, Shuang Liu, Qin Ni, Xin Lin, and Liang He. A review on machine theory of mind. *IEEE Transactions on Computational Social Systems*, 2024.

[17] Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, and Carles Sierra. Combining theory of mind and abductive reasoning in agent-oriented programming. *Autonomous Agents and Multi-Agent Systems*, 37(2):36, 2023.

[18] Imani Munyaka, Zahra Ashktorab, Casey Dugan, James Johnson, and Qian Pan. Decision making strategies and team efficacy in human-ai teams. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–24, 2023.

[19] Monique MH Pollmann and Emiel J Krahmer. How do friends and strangers play the game taboo? a study of accuracy, efficiency, motivation, and the use of shared knowledge. *Journal of language and social psychology*, 37(4):497–517, 2018.

[20] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[21] Jonan Richards and Mairieli Wessel. What you need is what you get: Theory of mind for an llm-based code understanding assistant. *arXiv preprint arXiv:2408.04477*, 2024.

[22] Matthew Sidji, Wally Smith, and Melissa J Rogerson. The hidden rules of hanabi: How humans outperform ai agents. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[23] Matthew Sidji, Wally Smith, and Melissa J Rogerson. Human-ai collaboration in cooperative games: A study of playing codenames with an llm assistant. *Proceedings of the ACM on Human-Computer Interaction*, 8(CHI PLAY):1–25, 2024.

[24] Matthew Stephenson, Matthew Sidji, and Benoît Ronval. Codenames as a benchmark for large language models. *arXiv preprint arXiv:2412.11373*, 2024.

[25] Qiaosi Wang. *MUTUAL THEORY OF MIND FOR HUMAN-AI COMMUNICATION IN AI-MEDIATED SOCIAL INTERACTION*. Phd thesis, Georgia Institute of Technology, December 2024. Available at `https://www.researchgate.net/publication/387522283_MUTUAL_THEORY_OF_MIND_FOR_HUMAN-AI_COMMUNICATION_IN_AI-MEDIATED_SOCIAL_INTERACTION`.

[26] Qiaosi Wang and Ashok K Goel. Mutual theory of mind for human-ai communication. *arXiv preprint arXiv:2210.03842*, 2022.

[27] Justin D Weisz, Michael Muller, Arielle Goldberg, and Dario Andres Silva Moran. Expedient assistance and consequential misunderstanding: Envisioning an operationalized mutual theory of mind. *arXiv preprint arXiv:2406.11946*, 2024.

[28] Yuan Yao, Haoxi Zhong, Zhengyan Zhang, Xu Han, Xiaozhi Wang, Kai Zhang, Chaojun Xiao, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. Adversarial language games for advanced natural language intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14248–14256, 2021.

[29] Shao Zhang, Xihuai Wang, Wenhao Zhang, Yongshan Chen, Landi Gao, Dakuo Wang, Weinan Zhang, Xinbing Wang, and Ying Wen. Mutual theory of mind in human-ai collaboration: An empirical study with llm-driven ai agents in a real-time shared workspace task. *arXiv preprint arXiv:2409.08811*, 2024.