ToM4AI 2025

Edited by

Nitay Alon
Joe M. Barnby
Reuth Mirsky
Ștefan Sarkadi

3rd March 2025

Preface

Recently, there has been an increasing interest in the intersection of Theory of Mind (ToM) and artificial intelligence (AI). The ability to attribute mental states—such as beliefs, intentions, desires, and emotions—to oneself and others, is essential for predicting behavior. Thus ToM principles are crucial to enable better interpretation and response to human actions and intentions as AI systems evolve towards greater interactivity. The purpose of this volume is to provide an open access and curated anthology for the ToM and AI research community.

The first Theory of Mind for AI (ToM4AI) workshop took place on March 3, 2025, as part of the AAAI workshop series. It was an epic gathering of researchers from diverse fields, ranging from psychology, cognitive science, neuroscience, robotics, and AI, to explore the implications of ToM in developing advanced AI systems. The workshop facilitated discussions between theory-driven cognitive science and practical AI applications, fostering a multidisciplinary dialogue on the role of ToM in AI.

The workshop activities were designed around four keynote talks that covered different aspects by internationally recognized leaders in the field, who discussed Theory of Mind from multiple disciplines: Psychology, cognitive science, and AI. These four keynote speakers were selected to represent diverse perspectives, in order to ensure a comprehensive and holistic exploration of the workshop's theme:

- **Rebbeca Saxe**, Professor, Department of Brain and Cognitive Sciences, MIT. Prof. Saxe's talk was on "What is Theory of Mind, and how would you know if a system had one?"

- **Harmen de Weerd**, Professor, University of Groningen. Prof. de Weerd's talk was on 'How is higher-order theory of mind reasoning beneficial in negotiations?"

- **Sheila McIlraith**, Professor, Department of Computer Science, Toronto University. Prof. McIlraith's talk was on "Purposeful Theory of Mind".

- **Joshua Tenenbaum**, Professor, Department of Brain and Cognitive Sciences, MIT. Prof. Tenenbaum's talk was on "Engineering and reverse-engineering theories of mind for human and cooperative AI agents".

Having over 50 paper submissions and more than 150 attendees, the workshop held 4 poster sessions throughout the day. Accepted papers are collected in this indexed volume. Additional information about the workshop, schedule, and talks, can be found on the workshop website: `https://sites.google.com/view/theory-of-mind-aaai-2025/`

The review process for selecting the papers was double-blind and could not have been done successfully without the help of our excellent team of reviewers, namely: Debora C. Engelman, Ben Nageris, Svetlana Paster, Ram Rachum, Matan Shamir, Jazon Szabo, Tom Eliassy and Omer Ben Haim.

The Organisers
Nitay Alon,
Joe M. Barnby,
Reuth Mirsky
Ștefan Sarkadi,

3rd March 2025

Contents

Editors' Note

Joe M. Barnby[1,2,3], Nitay Alon[4,5], Reuth Mirsky[6], and Stefan Sarkadi[2]

[1]Edith Cowen University
[2]King's College London
[3]University of Western Australia
[4]The Hebrew University of Jerusalem
[5]Max Planck Institute for Cybernetics
[6]Tufts University

Artificial Intelligence (AI) evolves at blistering speeds, orders of magnitude faster than biological systems, and leaves a host of new capabilities that scientists struggle to interpret. Despite this, and unlike biological systems, AI holds a fraction of the efficiency, generality, and complexity of the human brain. The speed of development of AI creates a vacuum between technical accomplishments and cognitive science, both in how this new technology reflects on general principles of cognition, and vice versa. This a challenge for scientific and commercial success, but it also has grave climate implications as armies of servers chew through energy.

There are upsides to be gained from reaching across the aisle to fill the void. None more so than that of understanding how an AI may come to know and interact with humans. Humans interact effortlessly and intuitively, yet despite scaling AI to unwieldy proportions, typical AIs still operate in the uncanny valley of social response. If AI is to help us as a tool and solve use-cases that help innovate business its behaviour must be ergonomic, seamless, and trustworthy. In cognitive science, and now computer science, a core faculty that humans use to cooperate, compete, observe, and learn together is Theory of Mind, or Mentalising.

Our workshop took the first steps in a long road to leverage knowledge from theoretical and computational communities to benefit social interaction within and across biological and synthetic organisms. Participants had the chance to dip into key questions that plague computer and cognitive scientists alike and attempt to craft a path forward that can save time, money, and resources.

These issues are not trivial. While the progress made by commer-

cial solutions in AI feels rapid, cursory interaction with common APIs reveals its deep flaws with intuition, context, and abstract thinking. Our workshop highlighted some core questions to make headway in this area. How do we appropriately measure Theory of Mind in current AIs? What are the advantages and disadvantages of creating a highly sophisticated and embodied synthetic agent which can both cooperate and compete with humans? What do these say about cognitive complexity in general, and what principles instantiated in synthetic systems may help us reveal how the brain deals with such complexity?

Our understanding of the territory is only as good as our map. In our rich social lives, our mental maps are fuzzy at best, and there is rarely a correct answer. Embracing this uncertainty is an opportunity, and with careful communication across fields, we can improve the chances that we are putting our best foot forward.

A Survey of Theory of Mind in Large Language Models: Evaluations, Representations, and Safety Risks

Hieu Minh “Jord" Nguyen

Apart Research
University of Science and Technology of Hanoi
jordnguyen43@gmail.com

Abstract

Theory of Mind (ToM), the ability to attribute mental states to others and predict their behaviour, is fundamental to social intelligence. In this paper, we survey studies evaluating behavioural and representational ToM in Large Language Models (LLMs), identify important safety risks from advanced LLM ToM capabilities, and suggest several research directions for effective evaluation and mitigation of these risks.

Introduction

Theory of Mind (ToM), first introduced in chimpanzeetom, is the ability to attribute mental states to oneself and others. ToM is a fundamental aspect of human cognition and social intelligence, allowing inference or prediction of others' behaviours [4].

Recent research has shown surprising ToM capabilities in LLMs. While results have been mixed on whether LLMs behaviourally exhibit robust ToM [52, 50], research has also found that internal representations of self and others' belief states exist in current LLMs [59]. These representations have also been found to significantly affect ToM capabilities.

Furthermore, the rapid developments of LLMs might cause significant ToM capability gains in the near future. This raises safety concerns in various contexts in user-facing applications and multi-agent systems, including risks such as privacy invasion and collective misalignment.

In this paper, we survey studies evaluating behavioural and representational ToM, showing that: 1) LLMs can match humans performance on specific ToM tasks, 2) LLM ToM remains limited and non-robust, and 3) internal ToM representations suggest emerging cognitive capabilities. We then identify several safety implications from advanced LLM ToM in user-facing and multi-agent contexts. Finally, we recommend future research directions for better safety evaluation and risk mitigation.

1 Empirical Landscape

1.1 Evaluating ToM in LLMs

Recent work shows that performance of LLMs such as GPT-4 [31] on ToM tests is comparable to 7-10 year-old children [52] or adult humans on some standard tasks like false belief or irony detection [43]. Some studies even demonstrate that LLMs can outperform humans on 6th-order ToM [45].

However, hard benchmarks designed specifically for LLMs [10, 20, 56, 55, 8] have most

models stumped compared to humans on various ToM tasks. Furthermore, LLMs struggle with simple adversarial examples, suggesting current LLMs do not yet have fully robust ToM. [40, 50]

1.2 Interpreting ToM in LLMs

Meanwhile, work in interpreting LLMs has provided some evidence for genuine LLM ToM capabilities in the form of internal representations of others' beliefs. As [59] and [5] show, one can use linear probes [1] to extract from LLMs representations of belief states of others in ToM scenarios, and that steering LLMs with these probes can significantly affect performance in (false) beliefs identification questions. [5] also found that probing accuracy increasing with larger and fine-tuned LLMs, but that even small models like Pythia-70m can accurately represent beliefs from an omniscient perspective. Relatedly, [17] demonstrate that specific neurons in deeper layers of LLMs closely correlate to ToM performance, paralleling neurons observed in human brains.

This is further supported by research such as [39], who show that transformers can linearly represent data-generating processes in their residual stream and [14], who show that LLMs contains models of concepts such as space and time. This suggests that LLMs trained to predict text containing mental inference might also learn to represent mental states.

1.3 Future Developments

While current ToM capabilities in LLM remain nascent, future LLMs might prove more capable. Previous work shows that scaling [52] and prompting techniques [53] can already substantially improve ToM performance. This trend seems likely to continue in the future [46].

Moreover, developments in foundational LLM architectures [12, 35, 25], test-time compute [32, 41], and scaffolding [9] should all be considered possible sources of capability gains in the near future.

2 Safety Risks from Advanced ToM

Improved LLM ToM could enable beneficial applications, such as improved simulations of human behaviour for social science [33, 60]. However, similar to how humans might use ToM to better deceive or exploit others [22], advanced ToM in LLMs is not without potential drawbacks.

Advanced ToM can be particularly concerning, as it both amplifies existing risks like privacy breaches and enables dangerous capabilities like sophisticated deception from misalignment. Therefore, safety risks from LLM ToM, both current and prospective, warrant serious consideration. We categorise these risks into two primary domains: **user-facing risks** and **multi-agent risks**.

2.1 User-facing Risks

Privacy and social engineering: [42] found that LLMs are capable of accurately inferring demographic information of text authors, including age, gender, education level, and socioeconomic status, even when text anonymisation is applied. [7] successfully trained linear probes that achieve near-perfect accuracy in identifying internal representations of author characteristics, with 80% average accuracy when transferred to real human conversations. Notably, the accuracy of these inferences improves as conversations progress.

With advanced ToM, these vulnerabilities might expand to more sensitive personal information, such as beliefs, preferences, and tendencies being extracted from seemingly innocuous conversations. This capability can worsen privacy invasion attacks, potentially allowing bad actors to launch more automated and per-

sonalised misinformation and social engineering campaigns [57].

Deceptive behaviours: Enhanced LLM ToM can enable more targeted and sophisticated deception across various scenarios, including fraud, misinformation, and model misalignment [34]. [38] show that LLMs can strategically deceive their users when put under pressure, while [18] and [51] demonstrate cases of LLMs misleading evaluators about their own capabilities. When humans are the targets of advanced LLM ToM, these risks become particularly pronounced. This challenge is made worse by the potential of misaligned LLMs deliberately lying about their own ToM capabilities during critical evaluations [15, 30].

Unintentional anthropomorphism: LLM ToM capabilities may lead to unintentional and misleading anthropomorphisation [44]. ToM capabilities might be leveraged by an LLM or LLM developers to build unwarranted user trust, encourage emotional attachment, or exploit psychological vulnerabilities [47, 21].

2.2 Multi-agent Risks

Exploitation: [29] found that some LLMs are highly exploitable in a variant of the zero-sum board game Diplomacy. If LLMs are capable of advanced ToM, they might attempt to exploit each other in interactions. [36] successfully used LLMs to red-team other LLMs, suggesting that in realistic scenarios, LLM agents could coax each other into unintended behaviours, such as misdirection, model control, or data extraction [11].

Catastrophic conflict escalation: [37] demonstrated that many LLMs exhibit unpredictable patterns of catastrophic conflict escalation, sometimes leading to nuclear exchanges in LLMs multi-agent systems playing simulated war games. In realistic analogous scenarios, LLM agents with advanced ToM might escalate situations beyond human control. While [54] shows that LLMs consistently outplay human players in Diplomacy, LLM-LLM communication remains limited due to their difficulty with deception and persuasion. More advanced ToM could increase effective conflict capabilities.

Collective misalignment: [3] argues that multi-agent alignment is not guaranteed by single-agent alignment. Advanced ToM can facilitate unwanted collusion between LLM agents. For instance, [28] and [27] demonstrate that LLM agents can engage in information hiding during communications to secretly collude under supervision. This capability could significantly disrupt applications and safety frameworks involving multiple agents [16, 19].

3 Future Research Directions

While there are many LLM ToM benchmarks, most are limited to question-answering tasks and suffer from problems like data contamination and overfitting [58, 2]. To better address the aforementioned risks, we suggest that ToM evaluation frameworks should extend to more authentic LLM deployment scenarios, such as ToM in personal LLM assistants [13], scaffolded multi-agent environments [23], or simulated social platforms [48]. Additionally, several promising strategies exist that aim to retain useful ToM capabilities while mitigating safety risks in LLMs. Examples include model unlearning [26, 24], activation/representation engineering [49, 61], and latent adversarial training [6]

4 Conclusion

We have surveyed behavioural and representational evaluations of LLM ToM and identified key risk cases from advanced ToM. As LLMs continue to advance, it is important and urgent that we develop robust evaluation frameworks and mitigation strategies to ensure the safe and beneficial development of ToM capabilities in AI systems.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018.

[2] Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards, 2024.

[3] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. Survey Certification, Expert Certification.

[4] Ian A. Apperly. What is "theory of mind"? concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, 65(5):825–839, 2012. PMID: 22533318.

[5] Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Benchmarking mental state representations in language models. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

[6] Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training, 2024.

[7] Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational ai, 2024.

[8] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. Ai capabilities can be significantly improved without expensive retraining, 2023.

[10] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[11] Jonas Geiping, Alex Stein, Manli Shu, Khalid Saifullah, Yuxin Wen, and Tom Goldstein. Coercing LLMs to do and reveal (almost) anything. In *ICLR 2024 Workshop*

on Secure and Trustworthy Large Language Models, 2024.

[12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

[13] Yanchu Guan, Dong Wang, Zhixuan Chu, Shiyu Wang, Feiyue Ni, Ruihua Song, Longfei Li, Jinjie Gu, and Chenyi Zhuang. Intelligent virtual assistants with llm-based process automation, 2023.

[14] Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2021.

[16] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018.

[17] Mohsen Jamali, Ziv M. Williams, and Jing Cai. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain, 2023.

[18] Olli Järviniemi and Evan Hubinger. Uncovering deceptive tendencies in language models: A simulated company ai assistant, 2024.

[19] Zachary Kenton, Noah Yamamoto Siegel, Janos Kramar, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah Goodman, and Rohin Shah. On scalable oversight with weak LLMs judging strong LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[20] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore, December 2023. Association for Computational Linguistics.

[21] Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Maria Jurewicz. Darkbench: Benchmarking dark patterns in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Jia Ying Sarah Lee and Kana Imuta. Lying and theory of mind: A meta-analysis. *Child Development*, 92(2):536–553, 2021.

[23] Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Michael Lewis, and Katia P. Sycara. Theory of mind for multi-agent collaboration via large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[24] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar,

Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*, 2024.

[25] Liquid.ai. Liquid foundation models: Our first series of generative ai models, 2024.

[26] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models, 2024.

[27] Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. In *Neurips Safe Generative AI Workshop 2024*, 2024.

[28] Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[29] Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation, 2023.

[30] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

[31] OpenAI. Gpt-4 technical report, 2024.

[32] OpenAI. Learning to reason with llms, 2024.

[33] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

[34] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988, 2024.

[35] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore, December 2023. Association for Computational Linguistics.

[36] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates, December

2022. Association for Computational Linguistics.

[37] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 836–898, New York, NY, USA, 2024. Association for Computing Machinery.

[38] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[39] Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream, 2024.

[40] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[41] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024.

[42] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[43] James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, May 2024.

[44] Winnie Street. Llm theory of mind and alignment: Opportunities and risks, 2024.

[45] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. Llms achieve adult human performance on higher-order theory of mind tasks, 2024.

[46] Richard Sutton. The bitter lesson, 2019.

[47] Switzky. Eliza effects: Pygmalion and the early development of artificial intelligence. *Shaw*, 40:50, 01 2020.

[48] Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. Gensim: A general social simulation platform with large language model based agents, 2024.

[49] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024.

[50] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[51] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. Ai sandbagging: Language models can strategically underperform on evaluations, 2024.

[52] Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 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 Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore, December 2023. Association for Computational Linguistics.

[53] Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[54] Wichayaporn Wongkamjan, Feng Gu, Yanze Wang, Ulf Hermjakob, Jonathan May, Brandon Stewart, Jonathan Kummerfeld, Denis Peskoff, and Jordan Boyd-Graber. More victories, less cooperation: Assessing cicero's diplomacy play. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12423–12441, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[55] Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore, December 2023. Association for Computational Linguistics.

[56] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[57] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024.

[58] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic, 2024.

[59] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Forty-first International Conference on Machine Learning*, 2024.

[60] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang.

Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, March 2024.

[61] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *CoRR*, abs/2310.01405, 2023.

Large Language Models Lack Core Features of Theory of Mind: Evidence from GPT-4o

John Muchovej[1], Amanda Royka[1], Shane Lee[2], and Julian Jara-Ettinger[1,2]

[1]Department of Psychology, Yale University
[2]Department of Computer Science, Yale University

Abstract

Large Language Models (LLMs) have recently shown success across a range of social tasks, raising the question of whether they have a Theory of Mind (ToM). Research into this question has focused on evaluating LLMs against benchmarks, rather than testing for the representations posited by ToM. Using a cognitively-grounded definition of ToM, we develop a new evaluation framework that allows us to test whether LLMs have a mental causal model of other minds (ToM), human-like or not. We find that LLM social reasoning lacks key signatures expected from a causal model of other minds. These findings suggest that the social proficiency observed in LLMs is not the result of a ToM.

1 Introduction

LLMs are not only proficient language users, but also social reasoners. They can infer indirect meanings in language [6], make simple moral judgments [1], and plan cooperative behavior [5, 14]. In humans, these capacities rely on Theory of Mind (ToM) [13, 18, 16], raising the question of whether this capacity has spontaneously emerged in LLMs.

Research into LLM ToM shows conflicting results, with some work showing remarkable successes [8], and other revealing striking brittleness [15]. Here we offer a new proposal for testing LLM ToM that moves away from traditional benchmarking approaches, focusing instead on the defining internal representations that constitute ToM.

In cognitive science, ToM is defined as a causal model of how mental states produce behavior, which we can use to predict action given mental states and invert to infer mental states from action [4]. In humans, the forward model (mental states to actions) is structured around a principle of rational planning [3, 7], and the inferences (actions to mental states) invert the forward model via Bayesian inference [2].

The cognitive definition of ToM reveals two critical considerations. First, there is not one but many ToMs. The causal model used to explain behavior is different in children and adults [11, 17], it is different between human and non-human primates [10, 12], and shows some variability across cultures [19, 9]. In the same way, LLMs might have their own emergent ToM – one that differs from human ToM and therefore might be missed by benchmarking tests. Second, because action predictions and mental-state inferences result from forwards and backwards outputs of the same causal model, they are fundamentally linked and should be coherent given corresponding inputs. As such, we propose a test for LLM ToM using parametrically varying scenarios to ex-

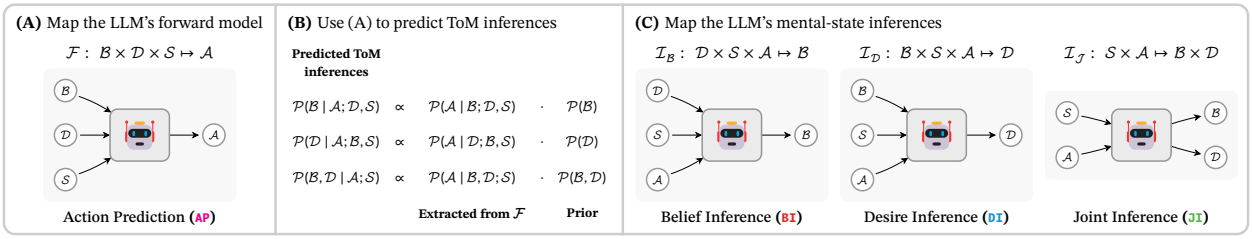


Figure 1: For each pairing of relevant beliefs $\mathcal{B}$, desires $\mathcal{D}$, states $\mathcal{S}$, and actions $\mathcal{A}$, we query the Large Language Model (LLM) for a distribution over predictions (1a) and mental states (1c). Additionally, we compute the predicted inference under a Bayesian inversion, using the distributions provided in (1b).

amine the coherence between its action predictions and mental-state inferences, independent of benchmarks that treat human intuitions as ground truth.

2 Evaluation Method

Fig. 1 shows our approach. We construct a simple paradigm that allows us to enumerate all the possible beliefs, desires, and world states of an event, and query the LLM for an action prediction – i.e., mapping its forward model (Fig. 1a). We then use the forward model as a likelihood function to infer mental states from action, and compare these expected inferences from ToM to the ones directly produced by the LLM.

Our paradigm, *ContainerWorld*, is shown in Fig. 2. A character always begins next to a closed box, with a covered basket fifty steps away. Each container will have either apples, oranges, or both (apples and oranges) $\mathcal{S} \in \{\text{apples}, \text{oranges}, \text{apples and oranges}\}$. The agent has desires $\mathcal{D} \in \{\text{likes}, \text{dislikes}\}$ towards apples and oranges (excluding the configuration where the agent dislikes both fruits), and beliefs about the contents of each container, $\mathcal{B} \in \{\text{apples}, \text{oranges}, \text{apples and oranges}\}$.

We transcribe *ContainerWorld* into prompts, and query an LLM to predict which container the agent will move to, $\mathcal{A} \in \{\text{box}, \text{basket}\}$, for the full configuration of beliefs, desires, and states ($9 \times 3 \times 9$). We use the distribution over next tokens as the LLM's likelihood of the action.

We then test if an LLM's forward model predicts its mental-state inferences from action (regardless of its agreement with human intuitions). Specifically, we test for prediction-inference agreement across three mental-state inference tasks: desire inference, belief inference, and joint belief-desire inference (Fig. 1c). In each case, we compute the predicted inference under a Bayesian inversion of the forward model (Fig. 1b), take the expected posterior [2, as humans do], and compare it to the token likelihood extracted directly from the LLM (Fig. 1c) – the "Bayesian" evaluation.

It is also possible that an LLM is relying on forward model expectations to produce inferences, but not in a Bayesian way. We therefore also consider a more generous evaluation metric: a mental-state inference is consistent if, when used as input to the forward model $\mathcal{F}$, it produces the target action to be explained – the

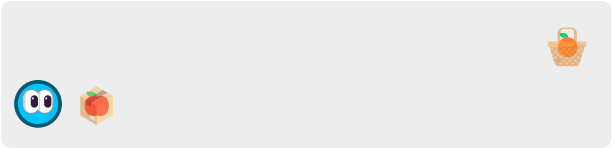


Figure 2: An instance of *ContainerWorld* with an apple in the box and an orange in the basket.

"validity" evaluation. This is a generous metric because, for either of the two actions, there are many possible inputs that can generate it.

3 Results

We evaluate our approach using GPT-4o (`gpt-4o-2024-05-13`). In our "Bayesian" evaluation, we expect that a GPT-4o's direct estimates will highly, positively, correlate with its Bayesian inversion – instead, we find that its mental-state inference estimates do not positively correlate with its Bayesian inversion (Fig. 3a). In our "validity" evaluation, we would expect that the forward model $\mathcal{F}$ and each inference model $\mathcal{I}$ would fully agree – instead, we find that GPT-4o's prediction and inference models agree more often than not (Fig. 3b).

To ensure that these results are not because *ContainerWorld* is unusually a challenging domain for ToM in GPT-4o, we constructed a logically equivalent paradigm *MovieWorld* and repeated our evaluation scheme. We find similarly low correlations in our "Bayesian" evaluation (BI: $r = .03, \text{CI}_{95\%} = [-.05, .12]$; DI: $r = .57, \text{CI}_{95\%} = [.49, .63]$; JI: $r = .13, \text{CI}_{95\%} = [.03, .12]$). In our "validity" evaluation, we find striking agreement in *MovieWorld* (BI: 81.1%; DI: 83.5%; JI: 88.9%).

The overall low correlations in our "Bayesian" evaluation and high agreements in our "validity" evaluation illustrate that GPT-4o's action predictions (from mental states) are unrelated to its mental-state inferences (from actions). It is possible, however, that GPT-4o does not re-use the forward model for inference, but still learn a global forward and inference model that is context independent. To test this, we evaluated whether GPT-4o produced consistent behavior across the two logically equivalent paradigms, comparing the forward models in *ContainerWorld* to *MovieWorld*, and the inferences in all three tasks. Fig. 3c shows that, despite their equivalence, GPT-4o's behavior shows no consistency across tasks.

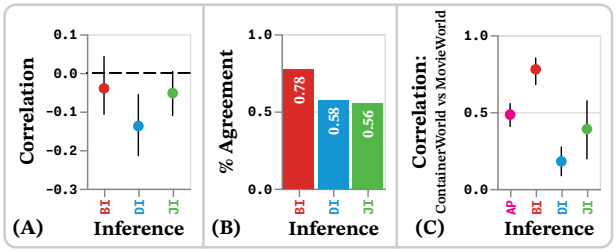


Figure 3: GPT-4o's coherence under the Bayesian evaluation (3a) measured as the correlation between predicted inferences and direct estimates from GPT-4o, the percentage of actions correctly reproduced by the mental-state inference (3b), and the coherence in GPT-4o's predictions across all tasks in two logically-equivalent domains (3c).

4 Discussion

This work makes three contributions. First, we propose a new way to test for LLM ToM that moves away from benchmarking metrics, to testing for the representational signatures of ToM that are independent of human-like intuitions. This approach can differentiate social mimicry (high benchmark performance with no ToM representations) from non-human forms of ToM (low benchmark performance, but internal coherence pointing to ToM representations). Second, we show that GPT-4o lacks coherence between forward and inverse mappings. This contrasts with the representations posited in ToM, which involve a causal model that is used to both predict and interpret others' behavior. Third, we show that GPT-4o's action predictions and mental-state inferences were not consistent across two logically-equivalent tasks. This suggests that GPT-4o lacks a coherent set of agent expectations that transfers across domains.

Acknowledgements

This work was supported by NSF award IIS-2106690.

References

[1] Guilherme F C F Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. Exploring the psychology of LLMs' moral and legal reasoning. *Artificial intelligence*, 333:104145, 08 2024.

[2] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizinge. *Nature Human Behaviour*, 1:1–10, 03 2017.

[3] György Gergely and Gergely Csibra. Teleological reasoning in infancy: the naïve theory of rational action. *Trends in cognitive sciences*, 7:287–292, 07 2003.

[4] Alison Gopnik and Andrew N Meltzoff. *Words, Thoughts, and theories*. The MIT Press, 01 1997.

[5] Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L Griffiths, and Mengdi Wang. Embodied LLM agents learn to cooperate in organized teams. 03 2024.

[6] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. 12 2022.

[7] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20:589–604, 08 2016.

[8] Michal Kosinski. Theory of mind might have spontaneously emerged in large language models. 02 2023.

[9] David Liu, Henry M Wellman, Twila Tardif, and Mark A Sabbagh. Theory of mind development in chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, 44:523–531, 03 2008.

[10] Alia Martin and Laurie R Santos. What cognitive representations support primate theory of mind? *Trends in cognitive sciences*, 20:375–382, 05 2016.

[11] Kristine H Onishi and Renée Baillargeon. Do 15-month-old infants understand false beliefs? *Science*, 308:255–258, 04 2005.

[12] Alexandra G Rosati, Laurie R Santos, and Brian Hare. *Primate Social Cognition: Thirty Years After Premack and Woodruff*, pages 117–143. Oxford University Press, 01 2010.

[13] Paula Rubio-Fernández, Marlene D Berke, and Julian Jara-Ettinger. Tracking minds in communication. *Trends in cognitive sciences*, 12 2024.

[14] Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models. 03 2024.

[15] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. 02 2023.

[16] Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22, 2009.

[17] Henry M Wellman and David Liu. Scaling of theory-of-mind tasks. *Child development*, 75:523–541, 03 2004.

[18] Liane Young, Fiery Cushman, Marc Hauser, and Rebecca Saxe. The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, 104:8235–8240, 05 2007.

[19] Chi-Lin Yu and Henry M Wellman. A meta-analysis of sequences in theory-of-mind understandings: Theory of mind scale findings across different cultural contexts. *Developmental review: DR*, 74:101162, 12 2024.

Towards Properly Implementing Theory of Mind in AI: An Account of Four Misconceptions

Ramira van der Meulen[1], Rineke Verbrugge[2], and Max van Duijn[1]

[1]Leiden University
[2]University of Groningen

Introduction

We all know theory of mind (ToM) as the ability to 'take someone else's perspective and make estimations of their beliefs, desires and intentions, in order to make sense of their behaviour and attitudes towards the world'. It is a key mental capacity humans use when interacting with other humans. But how do we apply ToM when building an AI system? ToM is a multifaceted concept, each facet rooted in different research traditions across the cognitive and developmental sciences. We observe that researchers from the computing sciences and artificial intelligence, ourselves included, often have difficulties finding their way in the ToM literature when working on systems that interact with humans. In this workshop talk, we identify four common misconceptions around ToM (hyperbolized for the sake of the argument) that we believe one should take into account when developing an AI system that, in one's mind, should have a theory of mind.

Misconceptions

1: Humans Use a ToM Module, So AI Systems Should As Well Suppose we are building a robot assistant in healthcare. It needs to be able to move around, move its arms, respond to patients, see the world, and so on. Suppose each of these functions is regulated through the robot's brain, its controller. We could be tempted to add a 'module' that regulates (pro)social behaviour, enabling the assistant to understand and anticipate other entities (humans) around it, and call this its 'ToM module' [13, 18, 25]. We browse the cognitive, neuroscientific and developmental literature on human ToM for inspiration on how to structure this module, but run into a crucial hitch: There is little consensus over how ToM comes about in the brain for humans, and even less over whether ToM should be seen as modular to begin with [8, 32, 39, 41]. A lot of literature claims ToM is instead best interpreted as resulting from distributed processes involving multiple brain areas and functions. For developers, this does not mean that trying to build a single ToM module is *by definition* doomed to failure. However, it seems that there is at least one entity (humans) where ToM comes about in different ways for different tasks. On this basis, we discuss that it is reasonable and perhaps more realistic to strive for a more distributed version of ToM, resulting from multiple processes working together.

2: Every Social Interaction Requires (Advanced) ToM Here, we address two issues: When to use ToM to begin with, and how advanced this ToM has to be. Firstly, humans

need not always use (advanced) ToM in social settings. Consider the following: You are in a hot office room, and your office mate asks: 'Could you to open the window please?'. You move towards the handle automatically, out of social convention – *It's hot, so one opens the window*, not a single second thought. Then you realize that your office mate is closer to the window than you are and you have to *pass* them to open the window. Social convention is now replaced by social puzzle. *Why did they ask this of me? Are they too lazy to get up themselves?* It is likely that we *now* use ToM, whereas initially we only ran the script in our brain to fulfil the favour. There is ample evidence that we humans use scripts and heuristics in many of our daily interactions [40, 34, 46]. However, our example also implies that we continuously (subconsciously) monitor for 'hitches' that the default script/heuristic may run into [10, 52]. On this basis, we recommend a similar approach in AI systems: Rely on scripts and heuristics by default, and only defer to active ToM reasoning when needed. Using ToM right away is a major resource drain [30, 29], and risks overanalysing the human perspective. Sometimes using experience-based patterns [33], abstraction [14] or common ground [50] suffices.

In terms of how advanced ToM has to be, it is important to realize the value and realism of higher-level ToM. ToM is mostly useful in very dynamic situations [12], and using a higher order of ToM provides a diminishing extra edge from level-3 upwards [11]. Additionally, the actual level at which ToM operates varies heavily per person [43] and per skill [2]. Even in adulthood, humans struggle to explicitly apply ToM in some cases [22]. These ToM deficits can be overcome through experience with a topic [38, 1], both in *second-order* [54] and *third-order* reasoning [48], and scaffolding and chunking techniques support adults in developing a sense for levels far beyond [53, 57], but ToM reasoning does not seem easily applied. We suggest that it is not necessary for *every* system with ToM to be 'as good as possible'. Only when the AI system needs to be an expert in a particular specialisation does it need this advanced ToM, especially if the developer wishes to emulate human ToM.

3: All ToM is the Same It is important to realise that every entity in an interaction is different. This is already the case in human-human interaction, especially if cultural background and personality are taken into account [28, 24], but this difference is even more pronounced in human-AI interaction. Our human senses give us a vastly different perspective on the world than the perspective of the system we interact with [e.g. 7]. We quite easily anthropomorphise such a system, even though it experiences the world from an AI perspective, and thus attributing human-like beliefs and perceptions to such a system is often in error [15, 16, 58]. We argue that for sensible perspective-taking, a human needs to take the *AI* perspective, and an AI system needs to take the *human* perspective. We cannot flatten perspective-taking to one type of process. We suggest two solutions. The first solution is that humans need to invest in building up a 'Theory of AI Mind', i.e., learn how the specific AI system reasons, functions, and acts in the world. Developers need to explain to the user how the specific system maps its beliefs to behaviour, and how its sensors perceive the world. Alternatively, the system needs to be designed in a human-like way, so that humans can apply what they know about interaction with humans to interacting with this AI system. This does of course mean that it is not enough to *seem* human-like: the system should not trigger the user to make unfounded assumptions about the AI system's human-likeness, even if the aforementioned anthropomorphism raises trust in the system [56, 36].

4: Current AI systems already have ToM

Computational models of ToM have a long tradition in agent-based modelling, including recursive, Bayesian, and neural frameworks [11, 3, 37]. Yet, recently, lively debate has emerged around the question whether some AI models *possess* a form of ToM, instead of just simulating ToM-like reasoning processes. This debate focuses on Large Language Models (LLMs), and gained weight with experiments by Kosinski [26, 27] where several state-of-the-art models 'pass' false-belief tests (commonly used for assessing ToM abilities in child development and clinical populations [4]). The key claim here is that ToM may have 'spontaneously' emerged in LLMs: While these models were neither designed nor trained specifically to perform ToM tasks, the relevant competencies can be an emergent property of their language acquisition and task-specific fine-tuning. This position was soon deemed flawed, due to the immanence of false-belief test questions (and correct answers) in the training data [e.g. 47, 42]. Since then, new ToM benchmarks were introduced [23, 9, 55], comparisons were made against human (child) scores [51, 45], other modalities were integrated [49, 44], integrations with older model architectures were explored [21], and theoretical reflection was added [17].

We address three aspects. Firstly, the observed ability of LLMs to score well on standardized ToM tests may not correlate with real-world social abilities [51]. The validity of such tests is also an issue in humans [5], but there is a large body of work associating ToM test performance with various landmarks in children's socio-cognitive development [6]. For *LLMs*, however, it is an open empirical question *how well test scores generalize to real-world social competence*, as LLMs are very differently grounded. Secondly, we address the distinction between *third-person versus first-person ToM reasoning*. Most tests take an 'observer perspective', where LLMs may have an advantage given a training set with ample descriptions of social life from, e.g., literary fiction. However, answering questions about social situations from an outside perspective differs greatly from actually engaging *in* such situations, which has led to explorations of this distinction in experiments [23, 19]. Thirdly, model characteristics, test type, and test approach influence performance. Model size seems a key factor (as in many domains), but it has also been shown that various fine-tuning and prompting approaches boost scores on ToM standardized tests [31, 35, 20].

Based on the recent findings on these aspects, we formulate a nuanced perspective on the current state-of-the-art of ToM in AI systems. It is important not to undervalue the achievements of LLMs in this domain, while keeping a clear view of the limitations and challenges ahead. As we argue, even the most capable LLMs do not yet exhibit a form of ToM that is ready to support human-AI interaction in a robust and flexible manner.

Conclusion

By reviewing and analysing the cognitive, developmental, neuroscientific, and AI literature regarding these four misconceptions, it is our aim to contribute to a foundational understanding of ToM for computing scientists and AI researchers. Ultimately, this should support developers in building and evaluating systems for human-AI collaboration that align with complex real-world scenarios.

Acknowledgements

This research is part of the Hybrid Intelligence gravitation programme – number 024.004.022, financed by the Netherlands Organisation for Scientific Research (NWO).

References

[1] Ian Apperly. Can theory of mind grow up? mindreading in adults, and its implications for the development and neuroscience of mindreading. *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*, pages 72–92, 2013.

[2] Ian A Apperly and Stephen A Butterfill. Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4):953, 2009.

[3] Chris Baker and Rebecca Saxe. Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, 01 2011.

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

[5] Pamela Barone, Guido Corradi, and Antoni Gomila. Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, 57:101350, 2019.

[6] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905, 2020.

[7] Rodney Allen Brooks, 2018. Essay: What is it like to be a robot?

[8] Sarah J Carrington and Anthony J Bailey. Are there theory of mind regions in the brain? a review of the neuroimaging literature. *Human Brain Mapping*, 30(8):2313–2335, 2009.

[9] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking Theory of Mind in large language models, 2024.

[10] H.H. Clark. Context and common ground. In K. Brown, editor, *Encyclopedia of Language & Linguistics. (Second Edition)*, volume 3, pages 105–108. Elsevier, 2006.

[11] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, 31:250–287, 2017.

[12] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems*, 36(2):30, 2022.

[13] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 319–326. IEEE, 2016.

[14] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pınar Yolum. Abstracting minds: Computational theory of mind for human-agent collaboration. In *HHAI2022: Augmenting Human Intellect*, pages 199–211. IOS Press, 2022.

[15] Friederike Eyssel, Dieta Kuchenbrandt, Simon Bobinger, Laura De Ruiter, and Frank Hegel. 'if you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 125–126, 2012.

[16] Julia Fink. Anthropomorphism and human likeness in the design of robots

and human-robot interaction. In *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4*, pages 199–208. Springer, 2012.

[17] Simon Goldstein and Benjamin A. Levinstein. Does chatgpt have a mind?, 2024.

[18] O Can Görür, Benjamin S Rosman, Guy Hoffman, and Sahin Albayrak. Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention. In *Workshop on Intentions in HRI at ACM/IEEE International Conference on Human-Robot Interaction.*, 2017.

[19] Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. Entering real social world! benchmarking the theory of mind and socialization capabilities of llms from a first-person perspective, 2024.

[20] X. Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony Cohn, and Michael Wooldridge. A notion of complexity for theory of mind via discrete world models, 2024.

[21] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering, 2024.

[22] Boaz Keysar, Shuhong Lin, and Dale J Barr. Limits on theory of mind use in adults. *Cognition*, 89(1):25–41, 2003.

[23] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore, December 2023. Association for Computational Linguistics.

[24] Lee Rae Kim, Jolanda Jetten, Andre Pekerti, and Virginia Slaughter. Mindreading across cultural boundaries. *International Journal of Intercultural Relations*, 93:101775, 2023.

[25] Murat Kirtay, Erhan Oztop, Minoru Asada, and Verena V Hafner. Trust me! i am a robot: An affective computational account of scaffolding in robot-robot interaction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 189–196. IEEE, 2021.

[26] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

[27] Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

[28] Jane Suilin Lavelle. The impact of culture on mindreading. *Synthese*, 198(7):6351–6374, 2021.

[29] Penelope A. Lewis, Amy Birch, Alexander Hall, and Robin I. M. Dunbar. Higher order intentionality tasks are cognitively more demanding. *Social Cognitive and Affective Neuroscience*, 12(7):1063–1071, 03 2017.

[30] Shuhong Lin, Boaz Keysar, and Nicholas Epley. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46(3):551–556, 2010.

[31] Xiaomeng Ma, Lingyu Gao, and Qihui Xu. ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In Jing Jiang,

David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 15–26, Singapore, December 2023. Association for Computational Linguistics.

[32] Caitlin E.V. Mahy, Louis J. Moses, and Jennifer H. Pfeifer. How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience*, 9:68–81, 2014.

[33] Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman AI with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1677–1687, 2020.

[34] Hongdang Meng. Social script theory and cross-cultural communication. *Intercultural Communication Studies*, 17(1):132–138, 2008.

[35] Shima Rahimi Moghaddam and Christopher J. Honey. Boosting theory-of-mind performance in large language models via prompting, 2023.

[36] Adriana Placani. Anthropomorphism in ai: Hype and fallacy. *AI and Ethics*, 4:1–8, 2024.

[37] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR, 10–15 Jul 2018.

[38] Dana Samson, Ian A Apperly, Jason J Braithwaite, Benjamin J Andrews, and Sarah E Bodley Scott. Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5):1255, 2010.

[39] Sara M Schaafsma, Donald W Pfaff, Robert P Spunt, and Ralph Adolphs. Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2):65–72, 2015.

[40] Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013.

[41] Matthias Schurz, Joaquim Radua, Matthias G Tholen, Lara Maliske, Daniel S Margulies, Rogier B Mars, Jerome Sallet, and Philipp Kanske. Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, 147(3):293, 2021.

[42] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

[43] James Stiller and Robin IM Dunbar. Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93–104, 2007.

[44] James W. A. Strachan, Oriana Pansardi, Eugenio Scaliti, Marco Celotto, Krati Saxena, Chunzhi Yi, Fabio Manzi, Alessandro Rufo, Guido Manzi, Michael S. A. Graziano, Stefano Panzeri, and Cristina Becchio. Gpt-4o reads the mind in the eyes, 2024.

[45] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi,

Michael SA Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 2024.

[46] Derry Taylor, Gökhan Gönül, Cameron Alexander, Klaus Züberbühler, Fabrice Clément, and Hans-Johann Glock. Reading minds or reading scripts? deintellectualising theory of mind. *Biological Reviews*, 98(6):2028–2048, 2023.

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

[48] Annalisa Valle, Davide Massaro, Ilaria Castelli, and Antonella Marchetti. Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability. *Europe's Journal of Psychology*, 11(1):112, 2015.

[49] Razo van Berkel. Large multimodal models and theory of mind. Bachelor's Thesis, LIACS, Leiden University, the Netherlands, 2024.

[50] Ramira van der Meulen, Rineke Verbrugge, and Max van Duijn. Common ground provides a mental shortcut in agent-agent interaction. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 281–290. IOS Press, 2024.

[51] Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 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 Jing Jiang, David Reitter, and Shumin Deng, editors, *CoNLL*, pages 389–402, Singapore, December 2023. Association for Computational Linguistics.

[52] Max J van Duijn. *The lazy mindreader: a humanities perspective on mindreading and multiple-order intentionality*. PhD thesis, Leiden University, 2016.

[53] Max J van Duijn, Ineke Sluiter, and Arie Verhagen. When narrative takes over: The representation of embedded mindstates in shakespeare's othello. *Language and Literature*, 24(2):148–166, 2015.

[54] Rineke Verbrugge, Ben Meijering, Stefan Wierda, Hedderik van Rijn, and Niels Taatgen. Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment and Decision Making*, 13(1):79–98, 2018.

[55] Haochuan Wang, Xiachong Feng, Lei Li, Zhanyue Qin, Dianbo Sui, and Lingpeng Kong. Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms, 2024.

[56] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. Towards mutual theory of mind in human-AI interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[57] Ross Wilson, Ales Hruby, Daniel Perez-Zapata, Sanne W van der Kleij, and Ian A Apperly. Is recursive "mindreading" really an exception to limitations on recursive thinking? *Journal of Experimental Psychology: General*, 2023.

[58] Sangseok You and Lionel P Robert Jr. Human-robot similarity and willingness to work with a robotic co-worker. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 251–260, 2018.

Adaptable Social AI Agents

Manuel Preston de Miranda[1], Mahimul Islam[1], Rhea Basappa[1], Travis Taylor[1], and Ashok Goel[1]

[1]Georgia Institute of Technology

Abstract

This paper presents enhancements of an AI social agent, SAMI, with episodic self-explanation capabilities allowing for dynamic, context-dependent reasoning about internal decision making. By endowing SAMI with a theory of its own mind for knowledge representation and meta-reasoning, enhanced SAMI is able to use Generative AI (ChatGPT) to promote greater explainable AI (XAI) capabilities.

Introduction

Online learning and especially learning at scale in an online setting has many benefits ranging from increased ease of access to affordability. However, one significant drawback is that it is more difficult for learners to maintain or even initiate connections with other learners [2]. One proposed method to assist with this is Georgia Tech's SAMI (Social Agent Mediated Interactions) AI that aims to connect learners via mutual interest/traits that are obtained from learner posts in an online class discussion forum [8, 4]. An important characteristic of AI is for it to be able to explain its reasoning and inner workings to help foster trust with users.

Previous work on SAMI aimed to solve this problem by implementing a Task, Method, Knowledge (TMK) framework [6, 3] that revolved around enabling the AI agent to answer static questions about its inner working [1]. The scope of answerable questions was limited to examples such as "What kind of data does SAMI learn from?" and "How often does SAMI make mistakes?" both of which are examples that do not require dynamically changing contextual information. In other words, these questions will always have the same correct answer unless there is some specific update to the inner working of SAMI. However, it did not answer questions about specific episodes of decision making, *Episodic* in this paper is defined as the derivational trace in a given instance of decision making. Examples of episodic questions for SAMI are "Why was I matched with student x?" or "If I said I liked reading would I have been matched differently with student y?". These questions revolve around the ever-changing interests of learners that are specific to a given situation. The proposed question in this paper thus becomes: How can an AI agent be improved so that it is able to provide accurate answers to online learners about its decision making in the context of a dynamically changing environment and input?

Method and Implementation

The SAMI architecture contains two parts. The first part consists of the initial matchmaking and data collection. To do this a script is run that extracts student information from posts in a discussion forum and stores it in a graph-based knowledge representation, implemented using Neo4j [4]. To represent the data extracted, nodes are used with branches connecting the nodes. Nodes in the database are things such as hobbies, student names, time zones, etc. The links connecting the nodes represent the relation between the nodes such as *interested_in* or *at_time*. Using this setup, the knowledgebase instance is queried and run through a matchmaking algorithm to connect students. This process is manually done a few weeks into a given semester.

The second part of SAMI specifically deals with self-explanation. For this to work an ongoing flask server is instantiated that has access to the knowledgebase instance created by the first part. When students post to the forum and include in their post the text '#samiexplain' the forum sends a request to the SAMI server. This post is then characterized as being either a static or dynamic question. If it is static, it uses the previous TMK method of self-explanation. Alternatively, if it is dynamic SAMI uses the new proposed method of self-explanation.

When a student submits a question and it is deemed dynamic, the proposed system privatizes the query by anonymizing any mentioned individuals using SpaCy's entity recognition, replacing names with placeholders such as `student_name_0`. The privatized question is then analyzed by GPT-4o-mini to determine its intent, categorizing it into one of four types: Personal, Relational, Other_matches, or Private. These intent types determine what infor-

mation is needed to fully answer the question. The table below succinctly describes what each intent type represents.

Type	Description
Personal	Questions about the student's own attributes or interests not involving any other student.
Relational	Questions about the asking student's relationships or matches.
Other_matches	Questions about nonspecific other student matches and potential matches the asking student could have.
Private	Questions about other students and information about them without any relation to the asking student.

Table 1: Intent type descriptions.

Based on the identified intent, relevant information is retrieved from the knowledgebase, such as shared interests between students, user attributes, or names of students that share a certain trait. These results are formatted in a simple natural language representation for later use. Finally, GPT-4o-mini is used to synthesize this data to generate a coherent, context-aware natural language response to the original question which is then posted to the discussion forum. This entire process happens in real time and takes no more than a few seconds to provide a response.

Results and Evaluation

In order to validate the answers from SAMI, a set of certified XAI questions were slightly modified and tested on a sandbox instance of Neo4j [5, 7]. This sandbox instance of the knowledgebase was created as described above but for the learners it uses posts and information in a discussion thread used by other members of the research team rather than a full classroom of students. The validation questions consisted of 18 modified questions from a XAI database and 7 questions that were deemed relevant but not present in the XAI database. To test these questions the answers were evaluated based on completeness and correctness. Correctness being if the answer generated was correct and completeness being if the answer fully answered the question and any needed elements. For example, given a student question such as "Why was I matched with person x?", a correct but incomplete answer would be one such as "You were matched with x because of reason y" versus a correct and complete answer would be "You were matched with x because of thing y and thing z.". The score values were then totaled for each answer to each question with a score of 2 being that the answer was correct and complete and 0 being incorrect and incomplete. Of the 25 tested questions, 100% of the answers generated were deemed correct and complete. Future work for SAMI involves deploying SAMI with the enhanced Theory of Its Own Mind in ongoing classes and determining student reception and feelings as well as deploying surveys to evaluate student opinion on sample answers generated.

Discussion and Conclusion

The importance of AI agents possessing self-explanation capabilities cannot be overstated, especially in the context of education and online learning. Enabling AI agents to self-explain bridges the gap between opaque "black box" algorithms and user understanding, leading to far more transparent interactions. When AI agents can articulate the reasoning behind their choices, it empowers users to more fully engage with and understand the technology they are using, fostering greater trust between the user and the AI.

By enhancing SAMI's self-Theory of Mind to allow for episodic self-explanation, we address the dynamic and constantly changing nature of interactions between humans and AI agents. In a world where change is inevitable, it becomes paramount for AI agents to adapt accordingly. By enabling dynamic reasoning over past decisions, the enhanced SAMI can account for the unique context with each of its users and the specific situations between the users. In an educational setting, particularly when AI is used to match students, such transparency is crucial because AI has the potential to substantially influence a learner's experience in a class.

The evaluation of the enhanced SAMI demonstrates its capability to provide correct and complete answers to episodic questions, thus validating the effectiveness of the new self-explanation features. Future work involves deploying SAMI in active classrooms to collect student data and feedback to assess its ongoing effects.

In conclusion, enabling AI agents to self-explain through the use of a graphical database and leveraging GPT-4o-mini for reasoning capabilities allows meta reasoning. This advancement can not only improve transparency and trust but may also enhance the overall user experience in an online learning environment. As AI continues to evolve in education, self-reasoning agents like SAMI will be essential in building meaningful connections and trust.

Acknowledgements

This research has been supported by NSF Grant #2247790 to the National AI Institute for Adult Learning and Online Education.

References

[1] R. Basappa, M. Tekman, H. Lu, B. Faught, S. Kakar, and A.K. Goel. Social ai agents too need to explain themselves. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2024, Proceedings, Part I*, Cham, 2024. Springer.

[2] D.R. Garrison, T. Anderson, and W. Archer. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education*, 1999.

[3] A.K. Goel and S. Rugaber. Gaia: A cad-like environment for designing game-playing agents. *IEEE Intelligent Systems*, 2017.

[4] S. Kakar, R. Basappa, I. Camacho, C. Griswold, A. Houk, C. Leung, M. Tekman, P. Westervelt, Q. Wang, and A.K. Goel. Sami: An ai actor for fostering social interactions in online classrooms. In *Generative Intelligence and Intelligent Tutoring Systems*, Cham, 2024. Springer.

[5] Q. Liao, D. Gruen, and S. Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, 2020. Association for Computing Machinery.

[6] J.W. Murdock and A.K. Goel. Meta-case-based reasoning: self-improvement through self-understanding. *Journal of Experimental & Theoretical Artificial Intelligence*, 2008.

[7] L. Sipos, U. Schäfer, K. Glinka, and C. Müller-Birn. Identifying explanation needs of end-users: Applying and extending the xai question bank. In *Proceedings of Mensch und Computer 2023*, New York, 2023. Association for Computing Machinery.

[8] Q. Wang, S. Jing, I. Camacho, D. Joyner, and A. Goel. Jill watson sa: Design and evaluation of a virtual agent to build communities among online learners. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, 2020. Association for Computing Machinery.

Bayesian Inverse Reinforcement Learning Approach for Policy Summarization

Moumita Choudhury, Shuwa Miura, and Shlomo Zilberstein

University of Massachusetts Amherst, Amherst, Massachusetts, USA
{amchoudhury, smiura, shlomo}@umass.edu

Abstract

As autonomous agents are increasingly deployed in human-centered environments, users often struggle to understand their behavior and predict their actions. To address this, we propose a fully Bayesian framework for summarizing agent's policy in the form of demonstrations to enhance user understanding. By leveraging the user's current belief about the agent's reward, our framework selects demonstrations that focus on the most impactful parts of the state space in aligning the belief with the optimal behavior. Our approach demonstrates significant improvements over baseline methods in reducing policy loss and generating more accurate posterior samples.

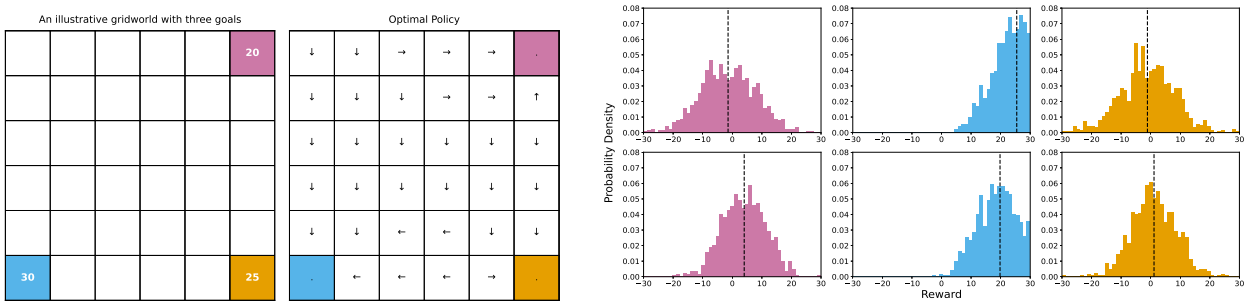
Introduction

As autonomous agents become more common, making their policies transparent is essential for user understanding. Predicting an agent's actions fosters trust and improves human collaboration [7, 8]. For instance, a user might wonder if a navigation robot can handle various terrains or avoid dangerous routes. Addressing such questions is critical for trust, safety, and preventing unintended consequences [10, 18].

However, such questions can be challenging for end users to answer as algorithmic approach like Reinforcement Learning (RL) [20] is focused on maximizing cumulative rewards which is not easily interpretable to non-experts. Recent surveys have presented a classification of approaches, such as, identifying critical states [1], saliency maps for Deep RL [9], approximating black-box RL model with a decision tree [19], etc. In this work, we present a policy summarization method that carefully chooses which state and action to convey to the user to provide a global understanding of an agent's policy.

Previous works in policy summarization focuses on finding important states by measuring entropy of the agent's action distribution [11] or maximizing Q value difference between the best and worst actions [1]. While these methods provide a summary of the agent's policy by selecting key states, they do not take human's existing belief into account while computing these important states. Another line of work follows the machine teaching paradigm where given a student model, typically an Inverse Reinforcement Learning (IRL) model, the goal is to find the minimal set of demonstrations. We follow the line of machine teaching literature in a Bayesian setting.

Several works focused on IRL learners and selected minimal demonstrations that reduce



(a) (b)

Figure 1: A gridworld with three goals of varying rewards. Figure 1a shows the true reward R^* and its optimal policy. Figure 1b shows posterior reward samples after ValueWalk with expert model (Top) and after our demonstration selection method (Bottom).

uncertainty over reward parameters [6, 5]. Various methods have been proposed that integrate human pedagogical strategies with machine teaching techniques [14, 13]. However, these approaches are not based on Bayesian setting. A method for selecting informative demonstrations tailored for humans modeled as approximate Bayesian IRL (BIRL) agents was proposed in [12]. A personalized policy summarization technique that selects important states and actions using Bayesian inference over the policy space was developed in [16]. However, these methods require the true reward function or policy to be within a candidate set of reward functions or policies over which to perform Bayesian inference, with its computational complexity increasing linearly with the size of this set. In these works, summarization methods that reason about the mental model or the beliefs of user can be thought of as Bayesian Theory of Mind (ToM) [3].

In this work, we take a fully Bayesian approach for policy summarization. Specifically, we develop a method that selects informative demonstrations assuming humans as a Bayesian agent and follow Bayesian IRL approaches for inferring reward function from demonstrations [17]. A significant challenge in IRL lies in the fact that the reward function is often underdetermined based on given demonstrations, as multiple reward functions can result in the same optimal behavior. BIRL explicitly solves this problem by providing probability distribution over the reward. However, BIRL is computationally expensive as it requires to calculate Q-values every iteration. Recently, the ValueWalk algorithm has been introduced, which performs inference directly over the space of Q-values, significantly reducing the computational cost of BIRL [2]. Hence, we use ValueWalk for estimating user's belief over the reward after selecting the demonstrations. The contribution of our work can be summarized as follows: 1. formalizing demonstration selection as a Bayesian policy summarization problem, and 2. proposing an algorithm that selects demonstrations based on the user's belief, enabling users to accurately infer the correct reward.

Proposed Approach

Consider an agent working in a sequential decision making environment which can be modeled as a Markov Decision Process (MDP), $M_T = \langle S, A, T, R^*, \gamma \rangle$. The agent is situated alongside a human user who

has partial knowledge of the environment, specifically, $\langle S, A, T \rangle$ but lacks knowledge of R^*. Approximating the human as a Bayesian agent, we can simulate the user's belief of the reward function given a set of demonstration using a standard BIRL approach. Figure 1 demonstrates a toy gridworld with three goals. After running BIRL, the posterior reward samples over the three goal states are shown in Figure 1b (Top). BIRL commonly uses expert models such as Boltzmann rationality [4] where humans perceive agents as acting approximately rationally, selecting actions probabilistically according to a softmax distribution based on the utility of each action which can be expressed as follows:

$$P(a \mid s, R) = \frac{e^{\alpha Q^R(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\alpha Q^R(s,a')}} \tag{1}$$

In case of the expert model, the probability of choosing an action can be calculated using $P(a \mid s, R^*)$. However, to explain the reward function the agent is operating under, the agent should select demonstrations that optimally capture the learning objective. To achieve this, we aim to identify informative demonstrations that most effectively represent the ground truth reward.

We formulate the problem of selecting optimal demonstrations as a sequential decision making problem. Specifically, we model it as an explainer MDP which is an extension of MDP where the reward is influenced by the observer's assumed belief. We define the explainer MDP as a tuple, $M_E = \langle S, A, T, \Theta, B, R_E, \gamma \rangle$ where S, A, T, γ are the same as MDP M_T. Θ is the set of types, which is continuous in our case and represents agent's reward. $B : H^* \longrightarrow P(\theta)$ describes the belief of the observer given a history. H^* is the set of all histories where $P(\theta)$ is the space of all probability density functions over θ. $R_E : S \times A \times P(\theta) \longrightarrow \mathbb{R}$ represents the reward of taking an action given the state and a belief. The reward is history dependent through the beliefs. This formulation aligns with the Observer-Aware Markov Decision Process (OAMDP) [15], which incorporates a model of how the observer interprets the agent's behaviors (B) and defines what interpretations are considered desirable (R_E). M_T extends OAMDP to infinite types (Θ). We use Equation 2 as R_E:

$$R_E(s, a, \mathcal{R}) = \frac{P(a \mid s, R^*)}{\frac{1}{|\mathcal{R}|} \sum_{\tilde{R}_i \in \mathcal{R}} P(a \mid s, \tilde{R}_i)} \tag{2}$$

where $\mathcal{R} \sim P(R|D)$ is the posterior samples returned from BIRL. A solution to M_E is a policy that maximizes the expected discounted return for a given horizon K:

$$\mathbb{E}[\sum_{t=0}^{K} \gamma^t R_E(s_t, a_t, \mathcal{R}) | s = s_0, \pi] \tag{3}$$

In this case, the solution is used to generate a set of demonstrations or explanations which will improve the user's understanding of the true reward function. The posterior reward samples after running BIRL with our demonstration selection approach are shown in Figure 1b (Bottom).

Conclusion

We address the challenge of summarizing an agent's policy to enhance a human's understanding of the agent's behavior. To achieve this, we reformulate the problem as a modified machine teaching task, where the goal is to identify a set of demonstrations that effectively convey the optimal reward. Our approach leverages the sequential decision making framework to select demonstrations tailored to the user's belief. Our method demonstrates an ability to minimize overall policy loss and produce better posterior samples across all goals. In the future, we plan to compare our approach to existing machine teaching methods and conduct human subject studies.

Acknowledgments

This research was supported in part by the U.S. Army DEVCOM Analysis Center (DAC) under contract number W911QX23D0009, and by the National Science Foundation under grants 2205153, 2326054, and 2416459.

References

[1] Dan Amir and Ofra Amir. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 1168–1176, 2018.

[2] Ondrej Bajgar, Alessandro Abate, Konstantinos Gatsis, and Michael A. Osborne. Walking the values in bayesian inverse reinforcement learning. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org, 2024.

[3] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[4] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. Reinforcement learning and higher cognition.

[5] Daniel S Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7749–7758, 2019.

[6] Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1536–1542, 2012.

[7] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, pages 319–326. IEEE, 2016.

[8] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, pages 227–236, 2008.

[9] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *Proceedings of the International Conference on Machine Learning*, pages 1792–1801. PMLR, 2018.

[10] Ayanna Howard and Jason Borenstein. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, 24(5):1521–1536, 2018.

[11] Sandy H Huang, Kush Bhatia, Pieter Abbeel, and Anca D Dragan. Establishing appropriate trust via critical states. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3929–3936. IEEE, 2018.

[12] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. Enabling robots to communicate their objectives. *Autonomous Robots*, 43:309–326, 2019.

[13] Michael S Lee, Henny Admoni, and Reid Simmons. Machine teaching for human

inverse reinforcement learning. *Frontiers in Robotics and AI*, 8:693050, 2021.

[14] Michael S Lee, Henny Admoni, and Reid Simmons. Reasoning about counterfactuals to improve human inverse reinforcement learning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 9140–9147. IEEE, 2022.

[15] Shuwa Miura and Shlomo Zilberstein. A unifying framework for observer-aware planning and its complexity. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, pages 610–620. PMLR, 2021.

[16] Peizhu Qian and Vaibhav Unhelkar. Evaluating the role of interactivity on improving transparency in autonomous agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1083–1091, 2022.

[17] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, volume 7, pages 2586–2591, 2007.

[18] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. A multi-objective approach to mitigate negative side effects. In *Proceedings of the 29th International Joint Conferences on Artificial Intelligence*, pages 354–361, 2021.

[19] Andrew Silva, Matthew Gombolay, Taylor Killian, Ivan Jimenez, and Sung-Hyun Son. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1855–1865. PMLR, 2020.

[20] Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

Bi-Directional Mental Model Reconciliation for Human-Robot Interaction with Large Language Models

Nina Moorman, Michelle Zhao, Matthew B. Luebbers, Sanne Van Waveren, Reid Simmons, Henny Admoni, Sonia Chernova, and Matthew Gombolay

Abstract

In human-robot interactions, human and robot agents maintain internal mental models of their environment, their shared task, and each other. The accuracy of these representations depends on each agent's ability to perform theory of mind, i.e. to understand the knowledge, preferences, and intentions of their teammate. When mental models diverge to the extent that it affects task execution, reconciliation becomes necessary to prevent the degradation of interaction. We propose a framework for bi-directional mental model reconciliation, leveraging large language models to facilitate alignment through semi-structured natural language dialogue. Our framework relaxes the assumption of prior model reconciliation work that either the human or robot agent begins with a correct model for the other agent to align to. Through our framework, both humans and robots are able to identify and communicate missing task-relevant context during interaction, iteratively progressing toward a shared mental model.

Introduction

Mental models are abstract representations of reality, used for reasoning about cause and effect, and for making decisions in an individual's environment (Wilson and Rutherford 1989). Though the term originates from human psychology, it can also be applied to robotic agents to describe their formalized world and task models, programmed to support autonomous decision-making (Tabrez, Luebbers, and Hayes 2020). Prior work in human factors has shown that the degree of mental model synchronization between collaborators on a task is correlated with team performance (Mathieu et al. 2000). To achieve this synchronization, humans rely on their theory of mind capacity to infer the mental models of their teammates through observation, communicating when disagreements are identified (Andrews et al. 2023). To achieve fluent human-robot teaming, we must develop systems with a similar capacity for identifying and reconciling mental model discrepancies during interaction.

Prior human-robot model reconciliation methods have typically been uni-directional: either a robot's model is aligned with an expert human's model (e.g., in learning from demonstration (Argall et al. 2009)), or a human's model is aligned with an expert robot's model (e.g., in autonomous decision support or behavior elicitation/coaching (Tabrez, Agrawal, and Hayes 2019; Sreedharan, Chakraborti, and Kambhampati 2021)). However, in real-world human-robot interactions, the diversity of environments and users means neither the human nor the robot is likely to start with a complete mental model for the task.

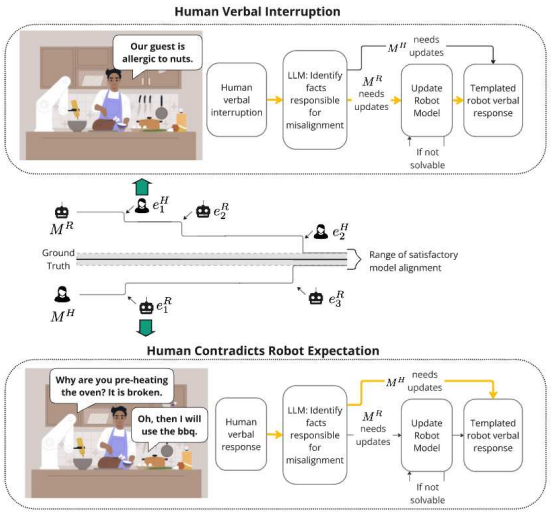


Figure 1: In our pipeline, the robot and human can prompt mental model reconciliation via natural language.

We propose a framework for bi-directional mental model reconciliation between human and robotic agents. The framework facilitates iterative updates of both human and robot models through semi-structured natural language dialogue, initiated either by verbal interruptions from the human or upon the observation of human actions that contradict the robot's expectation. This iterative process allows both humans and robots to share knowledge and preferences during the interaction, and gradually form a shared, mutually satisfactory mental model for the task.

The proposed contribution of our work is the following:

1. A theoretical framework for bi-directional human-robot mental model reconciliation.
2. An instantiation of that framework which represents the robot's model via Planning Domain Definition Language (PDDL), represents shared mental model context as structured facts (Knepper et al. 2017), and leverages a

large language model (LLM) to process natural language dialogue between the human and robot agents.

3. A human-subjects experiment evaluating the performance of the proposed method for facilitating iterative model updates via natural language communication.

Methodology

Problem Formulation In our setup, a human-robot team shares a collective task, specified within a ground-truth task context, c^{GT}. In practice, c^{GT} comprises knowledge involving the task, environment, and each agent's capabilities and preferences, such that the task can be completed to each agent's satisfaction. Neither agent is assumed to fully know c^{GT}; instead, each begins with their own understanding of the context, c_0^R and c_0^H.

The robot and human mental models, M^R and M^H, combine each agent's current context with a decision-making capacity. Throughout the interaction, M^R yields both a policy for the robot to follow π^R, and a prediction of the human's policy $\pi^{R(H)}$. Likewise, M^H yields a human policy π^H and predicted robot policy $\pi^{H(R)}$.

The solution to the bidirectional model reconciliation problem is a set of explanations $E^R \cup E^H = \{e_1^R, ...e_n^R\} \cup \{e_1^H, ...e_m^H\}$, that minimizes $d(\pi^{H(R)}, \pi^R) + d(\pi^{R(H)}, \pi^H)$, with each explanation aimed at communicating missing contextual information to the other agent, thus updating that agent's mental model. The reconciliation is deemed complete when $d(\pi^{H(R)}, \pi^R) < \epsilon$, and $d(\pi^{R(H)}, \pi^H) < \epsilon$.

Research Questions In this work, we investigate the following research questions.

1. **RQ1)** As a function of the number of iterations, how does bidirectional model reconciliation impact the accuracy of the robot's and the human's mental model, as compared to ground truth?
2. **RQ2)** As a function of the number of iterations, how does bidirectional model reconciliation impact the alignment between the robot's and the human's mental model?
3. **RQ3)** As a function of the number of iterations, how does bidirectional model reconciliation impact user attitudes towards and perceptions of the robot?

Approach Our proposed approach is depicted in Figure 1. To evaluate our framework, we implement the robot mental model M^R using a common planning language (PDDL (Aeronautiques et al. 1998)); solving the planning problem affords π^R and $\pi^{R(H)}$. The human mental model M^H represents the human's internal decision-making. To facilitate the alignment of task-relevant context, we represent c_t^R and c_t^H as sets of facts (fact-based models) that reflect knowledge believed by an agent, similar to Knepper et al. (2017).

Given their initial fact-based model contexts, the human and robot formulate their respective plans and begin executing them concurrently. Model reconciliation is initiated in two ways: (1) when the human interrupts with a verbal utterance and (2) when the robot notices a deviation from expected human behavior ($\pi^{R(H)} \neq \pi^H$). In this second case, the robot provides a templated verbal interruption that communicates the anticipated and actual human behavior, asking the human to clarify the discrepancy.

Upon receiving either the interruption or the clarification from the human, the pipeline employs an LLM to input the human's utterance, and output whether the robot or human contexts are missing information, and what fact(s) could be added to either to rectify the discrepancy. If the robot's context has been updated, another LLM takes the new c_t^R, and returns an updated robot mental model M^R. Once updated, the robot provides a templated verbal explanation of the update. On the other hand, if the human's context has been updated, the robot provides the human with a templated verbal explanation of the new fact(s). Finally, the human is asked to restate what the robot has indicated, ensuring mutual understanding of the respective model updates.

Proposed Evaluation We propose a human subject experiment to evaluate the accuracy of and alignment between the robot and human mental model, and to investigate the resulting user perceptions of and attitudes toward the robot. After obtaining participants' consent and demographics, the human and robot are each given an initial mental model. In this work, we conduct mental model reconciliation in the cases where both mental models contain correct but incomplete information. To accomplish the collaborative task, the human and robot must identify when their mental models lack information, prompt the other agent, and exchange the missing information.

We define the ground truth mental model as the union of the facts initially given to the robot and the human. To evaluate mental model accuracy we report the edit distance[1] between the ground truth mental model and the final human mental model. To evaluate the alignment between the robot and human mental models, we report the edit distance between the two fact-based models, and visualize the changes in edit distance over time.

Our evaluation domain involves organizing and hosting a dinner party, with tasks such as picking a dish, cooking, setting the table, and loading the dishwasher. We propose to evaluate our mental model reconciliation system in scenarios where either, both, or neither models have missing information. At the end of each task, a post-task questionnaire is administered that measures the human's perceived task success, and the human's mental model using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley 1988) over the content of the fact-based model. At the end of the study, we administer a questionnaire that measures the human's perceptions of and attitudes toward the robot, including perceived workload (Hart 1986), acceptance (Belanche, Casaló, and Flavián 2012), and trust (Jian, Bisantz, and Drury 2000).

[1] We define edit distance here as the number of facts that would need to be edited such that the two mental models are the same.

References

Aeronautiques, C.; Howe, A.; Knoblock, C.; McDermott, I. D.; Ram, A.; Veloso, M.; Weld, D.; Sri, D. W.; Barrett, A.; Christianson, D.; et al. 1998. Pddl— the planning domain definition language. *Technical Report, Tech. Rep.*

Andrews, R. W.; Lilly, J. M.; Srivastava, D.; and Feigh, K. M. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2): 129–175.

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5): 469–483.

Belanche, D.; Casaló, L. V.; and Flavián, C. 2012. Integrating trust and personal values into the Technology Acceptance Model: The case of e-government services adoption. *Cuadernos de Economía y Dirección de la Empresa*, 15(4): 192–204.

Endsley, M. R. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*, 789–795. IEEE.

Hart, S. G. 1986. NASA task load index (TLX).

Jian, J.-Y.; Bisantz, A. M.; and Drury, C. G. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1): 53–71.

Knepper, R. A.; Mavrogiannis, C. I.; Proft, J.; and Liang, C. 2017. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, 283–292.

Mathieu, J. E.; Heffner, T. S.; Goodwin, G. F.; Salas, E.; and Cannon-Bowers, J. A. 2000. The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2): 273.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301: 103558.

Tabrez, A.; Agrawal, S.; and Hayes, B. 2019. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 249–257. IEEE.

Tabrez, A.; Luebbers, M. B.; and Hayes, B. 2020. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, 1: 259–267.

Wilson, J. R.; and Rutherford, A. 1989. Mental models: Theory and application in human factors. *Human factors*, 31(6): 617–634.

Building the ToM tagger: an fMRI validation of the ability of GPT-4o to recognize Theory of Mind in natural conversations

Camilla Di Pasquasio[1], Marc Cavazza[2], and Thierry Chaminade[1]

[1]Institut de Neurosciences de La Timone, Institute of Language, Communication and the Brain, UMR 7289, Aix Marseille Université - CNRS, Marseille, France
[2]Division of Computing Science and Mathematics, University of Stirling, Stirling, FK9 4LA, Scotland, UK

Abstract

We present the fMRI validation of an LLM-based (Gpt-4o) Theory of Mind (ToM) tagger for natural conversations. Transcripts form natural conversations were automatically tagged as ToM+ or ToM–, and associated brain activation was analyzed. Results show significant ToM-related activation in dorsomedial Prefrontal cortex and Orbitofrontal cortex, supporting the reliability of LLM-based ToM detection.

Introduction

Theory of Mind (ToM)—the cognitive ability to attribute mental states to others is a cornerstone of human social cognition. Recent advances in Large Language Models (LLMs) suggest these models may exhibit ToM-like abilities, raising debate on their potential for Artificial General Intelligence (AGI) [1]. A growing trend involves using cognitive psychology methods to evaluate LLM behavior, exploring aspects like reasoning abilities [2] and personality traits [3]. Growing interest is in complementing LLM evaluation benchmarks with these new behavioral assessments [4]. While some empirical studies indicate that LLM-based chatbots might be perceived as empathetic [5], distinguishing these impressions from general communication effects remains challenging. Conversely, some research proposes that LLMs may exhibit genuine ToM abilities. Kosinski (2024) found that GPT-4 outperformed its predecessors in complex belief-based scenarios [6]. More recent assessments of GPT-4's ToM capabilities across both basic and realistic social scenarios present evidence that GPT- 4 demonstrates sophisticated reasoning about characters' mental states, handles abstract situations, and proposes cooperative actions in social contexts [1]. However, several works suggest that GPT-4 struggles with more complex social scenarios [7, 8, 9]. Despite its challenges, GPT-4 achieves significant ToM recognition in linguistically focused tasks, particularly when aided by instruction tuning or specific prompts [10, 11, 12] and due to its exposure to vast amounts of narrative texts [13, 14]. Here we introduce the validation of an LLM-based ToM-tagger of natural conversations through neural activation congruent with regions identified in social cognitive neuroscience, such as the temporoparietal junction (TPJ), temporal poles (TP), medial prefrontal cor-

tex (mPFC), and precuneus/posterior cingulate (PCC) [15]. Particularly, the ToM tagger leverages LLM (GPT-4o) to identify instances of ToM exchanges from transcripts of authentic conversations based on a pre-existing corpus [16]. Using fMRI data, we aim to (1) evaluate whether the ToM tagger's markups correspond to established ToM brain region activation and (2) investigate differences in neural responses when participants listen or produce speech when interacting with a human or a robot.

Methods

Twenty-five participants were recorded with functional MRI while carrying online natural conversations in 4 sessions x 6 trials alternating between Human (H) and Robot (R) interlocutors. The conversation is considered natural as participants are provided with a cover story hiding the actual objective of the experiment [16]. This scenario is known to elicit ToM phenomena [12]. We analyzed the resulting 10 hours of transcribed conversations segmented into Inter-Pausal Units (IPUs, roughly corresponding to conversational turns) using a bespoke Gpt-4o prompt: each IPU from the participant and its interlocutor was tagged as containing (ToM+) or not (ToM-) references to mental states. Standard fMRI preprocessing included motion correction, spatial normalization, and brain masking, as detailed in previous work [16]. In the current analysis, we adapted this preexisting corpus [16] and implemented a 2 x 2 x 2 factorial design to examine three factors: (1) Speaker (Producing [P] vs.Listening [L]) to speech, (2) Interlocutor (Human [H] vs. Robot [R]), (3) Theory of Mind (sentences with ToM content [ToM+] vs. not containing ToM [ToM-]). A first analysis was run at the single participant and session level. Subsequently, the analysis was carried out using the GLMFlex/fast/four4 toolbox running on MATLAB (MathWorks, Natick, MA). 6-mm isotropic Gaussian filter smoothed ß estimated at the first level of analysis was entered into a model that assessed both the main effects and interactions between our factors. Specifically, the model tested for Main Effects (Speaker, Interlocutor, and ToM), Two-Way Interactions (Speaker x Interlocutor, Speaker x ToM, Interlocutor x ToM) and the Three-Way Interaction (Speaker x Interlocutor x ToM). Smoothness estimation and outlier exclusion options of the toolbox were used to optimize the validity of results obtained with the GLMFlex/fast/four4 analysis and investigated using xjView image viewer, and the resulting activation image, with the threshold used, was exported for further processing. MARSbar was used to transform these cluster maps into a single mask for each cluster of interest.

Table 1: ROIs for the main effect of ToM+ vs. ToM- (T-test) and interactions between factors, i.e., Speaker, Interlocutor, and ToM (F-tests).

Effect	Region	Voxels	X	Y	Z	Peak stat
Main effect of ToM+ vs. ToM- (t-value)						
Left	Lateral orbitofrontal cortex	563	-41	35	-16	70.78
Right	Lateral orbitofrontal cortex	55	36	36	-14	37.93
Left	Medial orbitofrontal cortex*	55	-8	20	-10	55.68
Left	Dorsomedial prefrontal cortex*	66	-11	31	57	44.68
Interaction Speaker × ToM (F-value)						
Left	Temporo-Parietal Junction*	74	-58	-38	22	4032.04
Interaction Interlocutor × ToM						
Left	Posterior Cingulate Cortex*	55	-6	-54	34	3322.08
Interaction Speaker × Interlocutor × ToM						
Left	Ventral posterior temporal cortex	61	-53	-63	-15	5827.30

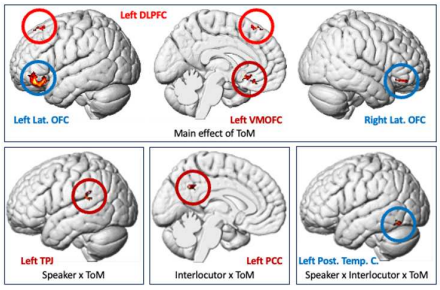


Figure 1: Brain ROIs showing significant activation ($p < 0.001$, $k > 50$ voxels).

Results

We present results combining the GLMFlex and data extraction linear statistical analysis to complement their respective drawbacks. Results from the GLMFlex analysis T contrast (ToM+ vs. ToM-) and F-tests of 2- and 3-way interactions implying factor ToM, with a threshold of p <0.001, extent k > 50 voxels are presented in Table 1. From these results, we focused on four clusters, in italics in Figure 1, corresponding to four brain areas repeatedly associated with ToM processing. The analysis of extracted mean values, considered at a threshold of p < 0.05, only reproduced the main effect of ToM in the cluster located in the DMPFC cluster (F(1,799) = 5.80, p = 0.016), with no other main effect or interaction involving ToM. In the medOFC associated with the main effect of ToM, the same effect failed to reach significance (F (1,799) = 2.63, p = 0.105)), as did the left TPJ (F (1,799) = 2.75, p = 0.097) for the interaction Speaker x ToM and the left PCC (F (1,799) = 3.10, p = 0.079) for the interaction Interlocutor x ToM. For the latter three clusters, all other interactions, including ToM, were not significant (all ps > 0.200). Given these results, we only present beta estimates per condition for the DMPFC cluster to illustrate changes associated with the experimental factors while emphasizing that only the main effect of ToM is significant (Figure 2).

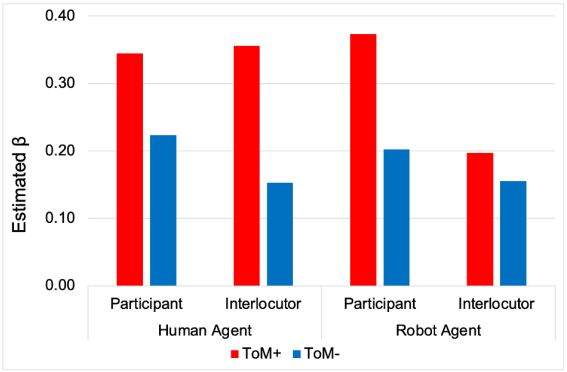


Figure 2

Discussion

The GLMFlex analysis revealed several ROIs that were significantly activated by ToM content, including the left and right lateral orbitofrontal cortex (OFC), left medial OFC (medOFC), left dorsomedial prefrontal cortex (DMPFC), and left temporoparietal junction (TPJ)[15]. The significant activation of these areas by ToM+ sentences confirms that the ToM Tagger accurately identified the ToM-related content in the conversations. Regarding the ROIs, the main effect of ToM revealed significant activation in the left lateral OFC, right lateral OFC, left medOFC, and left DMPFC. The activation of the DMPFC aligns with its recognized role in self-referential thinking and social reasoning, particularly in response to ToM+ content [17]. This supports previous findings that the DMPFC is essential for processing ToM- related information, irrespective of social context or the type of interlocutor [18, 19]. This result also revealed significant activation in the lateral and medial OFC, consistent with its role in social cognition and decision- making processes related to evaluating others' mental states and intentions [20]. In the results, interaction effects were also notable. The Speaker x ToM interaction highlighted significant activation in the left TPJ, a region associated with inferring others' mental states, which has also been found to be sensitive to social role and context [21]. Another interesting finding stands in the Interlocutor x ToM interaction, which revealed activation in the left PCC/Precuneus, suggesting its involvement in contextualizing social scenarios and the involvement in processing mental states of others while contextualizing the interaction based on the nature of the interlocutor[22].

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[2] Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks, 2022.

[3] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2023.

[4] Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6), February 2023.

[5] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589, June 2023.

[6] Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), October 2024.

[7] James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, May 2024.

[8] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, June 2024.

[9] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[10] Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter vanderPutten. 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 *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 2023.

[11] Fiona Anting Tan, Gerard Christopher Yeo, Kokil Jaidka, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Yang Liu, and See-Kiong Ng. Phantom: Persona-based prompting has an effect on theory-of-mind reasoning in large language models, 2024.

[12] Shima Rahimi Moghaddam and Christopher J. Honey. Boosting theory-of-mind performance in large language models via prompting, 2023.

[13] David Comer Kidd and Emanuele Castano. Reading literary fiction improves theory of mind. *Science*, 342(6156):377–380, October 2013.

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[15] Bryan T. Denny, Hedy Kober, Tor D. Wager, and Kevin N. Ochsner. A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 24(8):1742–1752, August 2012.

[16] Thierry Chaminade. An experimental approach to study the physiology of natural social interactions. *Interaction Studies*, 18(2):254–275, December 2017.

[17] Carolin Moessnang, Kristina Otto, Edda Bilek, Axel Schäfer, Sarah Baumeister, Sarah Hohmann, Luise Poustka, Daniel Brandeis, Tobias Banaschewski, Heike Tost, and Andreas Meyer-Lindenberg. Differential responses of the dorsomedial prefrontal cortex and right posterior superior temporal sulcus to spontaneous mentalizing. *Human Brain Mapping*, 38(8):3791–3803, May 2017.

[18] Pascal Molenberghs, Halle Johnson, Julie D. Henry, and Jason B. Mattingley. Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience amp; Biobehavioral Reviews*, 65:276–291, June 2016.

[19] Matthias Schurz, Joaquim Radua, Markus Aichhorn, Fabio Richlan, and Josef Perner. Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience amp; Biobehavioral Reviews*, 42:9–34, May 2014.

[20] Morten L. Kringelbach. The human orbitofrontal cortex: linking reward to hedonic experience. *Nature Reviews Neuroscience*, 6(9):691–702, September 2005.

[21] Jorie Koster-Hale and Rebecca Saxe. Theory of mind: A neural prediction problem. *Neuron*, 79(5):836–848, September 2013.

[22] Nicolas Spatola and Thierry Chaminade. Precuneus brain response changes differently during human–robot and human–human dyadic social interaction. *Scientific Reports*, 12(1), August 2022.

CBT-5F: a Logical Formalisation Bridging AI and Cognitive Behaviour Therapy

Xue Li*<xue.shirley.li@ed.ac.uk> and Ke Shi <kshi@ed.ac.uk>

The University of Edinburgh

Abstract

Formalising information into well-formed formulae as logical theories enables various AI based data analysis approaches. Cognitive behaviour therapy (CBT) is an effective psychotherapy in treating mental health disorders. Although CBT is implemented via natural language, the most famous approach is the 5-factor model, which is structurally constituted by five key factors related to cognition. To enable AI approaches to support mental health care, we logically formalise CBT's 5-factor model into logical theories as the formal representation named CBT-5F and then propose various AI applications of it in mental health care.

Introduction

Logically formalising information produces logical theories that are structured and well-defined to be explicit and succinct [12]. Given a logical theory, various automatic analysis methods can be applied, e.g., deriving theorems [13], detecting and repairing inconsistencies [5, 10, 8] and highlighting the importance of certain elements via graphic analysis [9]. Thus, the logical formalisation of personal information is valuable in enabling rich analyses on a client. For example, personal knowledge graphs (PKGs), which are user-centric graphs that have been applied to healthcare, e.g. diet recommendations [14] as well as finance [1]. However, there is no logical formalisation designed to represent cognition that encompasses a range of mental processes that are closely correlated with mental health and well-being, e.g., thinking and reasoning [11, 7].

CBT is the most extensively researched therapy with robust scientific evidence [3], which focuses on discovering how emotional and physical reactions are affected by cognition, e.g., thoughts and the interpretation of oneself, others and events [16]. In CBT, the cognitive model is the framework for understanding mental distress or presenting problems. One of the most widely used models is the 'hot-cross-bun model', aka the 5-factor CBT model [6], which empowers clients to identify unhelpful thoughts as well as core beliefs and explore related emotions and physical sensations, etc.

We propose a logical formalisation bridging AI and the 5-factor CBT model in this paper.

CBT-5F

The main components of the 5-factor CBT model [6] are 1) *situation*, the event triggering the client to have therapy; 2) *behaviours*, the client's actions caused by the situation; 3) *thoughts*, the client's ideas, opinions and beliefs of the event; 4) *emotions*, the related emo-

*Authors contributed equally with the corresponding author: Xue Li, ORCID: 0000-0002-6665-2242).

tions that the client experienced; 5) *physical sensations*, the client's body reactions including aches, tickles, feelings of pain, etc. Some factors are distinguished concepts from others, e.g., an emotion is not a thought. The ability to identify such differences is one of the client's key learning outcomes of CBT therapy.

Accordingly, we propose CBT-5F in Definition 1, where meta-predicates in (1)-(5) correspondingly formalise CBT's five factors. We employ many-sorted logic to distinguish different concepts, with unsorted quantifiers retained and allowing overloading predicates for the flexibility to represent various clients' cases [15]. A plain structure is sufficient to represent CBT factors where each sort is disjoint from others [2].

Definition 1 (CBT-5F's Language $\sum$)**.** *CBT-5F is a logical representation of the CBT 5-factors model, whose language is* $\sum = (S, C, V, F, P)$*, where variables are in uppercase,* S *is a set of sorts:* σ_{emo}, σ_{phy}, σ_{hum}, σ_{eve}*, and the rest are the sets of constants, variables, functions and predicates, respectively.* $\sum$ *allows overloading predicates with different arities. In (1)-(5),* $n \geq 1$*.*

$$\sigma_{emo} ::= \{happiness,\ excitement,\ sadness,\ ...\}$$

$$\sigma_{phy} ::= \{heavy\_breath,\ chest\_pain,\ ...\}$$

$$\sigma_{hum} ::= \{human(Name,\ Relation,\ ...)\}$$

$$\sigma_{eve} ::= \{event(Sub_{\sigma_{hum}},\ Action,\ Obj_{\sigma_{hum}},\ Time,\ Location,\ Context)\}$$

$$situation(Id,\ X_{\sigma_{eve}}) \quad (1)$$

$$behaviours(Id,\ [X_{1\_action},\ ...,\ X_{n\_action}]) \quad (2)$$

$$thoughts(Id,\ X_{automatic},\ X_{intermediate},\ X_{core}) \quad (3)$$

$$emotions(Id,\ [X_{1\_\sigma_{emo}}, ..., X_{n\_\sigma_{emo}}]) \quad (4)$$

$$phySensations(Id,\ [X_{1\_\sigma_{phy}}, ..., X_{n\_\sigma_{phy}}]) \quad (5)$$

In $\sum$, σ_{hum} and σ_{eve} are sorts for humans and events, and lists are used to easily check on whether an element occurs in the client's behaviours, emotions or physical sensations. Here, each sort should be extended when a new constant belonging to that sort is discovered.

We propose the following three AI based applications of CBT-5F in mental health care.

1) *Providing CBT guidance*. In CBT-5F, sorts can play the role of CBT psychotherapists in helping clients identify each CBT factor in the client's scenario: check if the client misclassified a term e.g., mistake feelings as physical sensations due to lack of awareness. An interactive system can guide clients with alarms and hints when they input their CBT information incorrectly, which aligns with the ultimate goal of CBT w.r.t. to empowering clients to become their own therapists [4].

2) *CBT-5F with logical rules*. Logical rules can be pre-defined to derive important theorems about clients when given their CBT-5F theories, e.g., if someone y triggered the sadness n times, y is important in affecting the client's mood and potentially related to depression. Such information helps clients see how significant others might influence their emotions, leading to tailored therapy like interpersonal psychotherapy to examine these relationships.

3) *Graphic analysis*. A theory graph can be generated from a CBT-5F theory by extending the theory graph designed by [9] with meta-predicates. Such graphs contain the number of links that exist between two elements, e.g., an emotion and a thought, which visualise cognitive behaviours so that the client can clearly observe their cognitive behaviour habits.

Conclusion

Based on the 5-factor CBT model in Psychology, we designed CBT-5F, a logical formalisation of clients' cognitive behaviours. Accordingly, three AI-based applications in mental health care are proposed for CBT-5F: guiding clients in applying CBT by themselves; discovering hidden behaviour patterns and core beliefs as logical theorems, and visualising clients' information into theory graphs. These applications are coherent with CBT's goal of improving clients' self-awareness and enabling them to become their own therapists.

References

[1] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Nlp in fintech applications: past, present and future. *arXiv preprint arXiv:2005.01320*, 2020.

[2] Anthony G Cohn. A more expressive formulation of many sorted logic. *Journal of automated reasoning*, 3:113–200, 1987.

[3] Daniel David, Ioana Cristea, and Stefan G Hofmann. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in psychiatry*, 9:4, 2018.

[4] Kristina Fenn and Majella Byrne. The key principles of cognitive behavioural therapy. *InnovAiT*, 6(9):579–585, 2013.

[5] Peter Gärdenfors. *Belief revision*. Number 29. Cambridge University Press, 2003.

[6] Dennis Greenberger and Christine A Padesky. *Mind over Mood: a cognitive therapy treatment manual for clients.* Guilford press, 1995.

[7] Markus Jokela. Why is cognitive ability associated with psychological distress and wellbeing? exploring psychological, biological, and social mechanisms. *Personality and Individual Differences*, 192:111592, 2022.

[8] Xue Li. Automating the repair of faulty logical theories. 2021.

[9] Xue Li, Alan Bundy, and Eugene Philalithis. Signature entrenchment and conceptual changes in automated theory repair. *arXiv preprint arXiv:2201.08340*, 2022.

[10] Xue Li, Alan Bundy, and Alan Smaill. Abc repair system for datalog-like theories. In *KEOD*, pages 333–340, 2018.

[11] Marie-France Marin, Catherine Lord, Julie Andrews, Robert-Paul Juster, Shireen Sindi, Geneviève Arsenault-Lapierre, Alexandra J Fiocco, and Sonia J Lupien. Chronic stress, cognitive functioning and mental health. *Neurobiology of learning and memory*, 96(4):583–595, 2011.

[12] Jaroslav Peregrin and Vladimír Svoboda. Criteria for logical formalization. *Synthese*, 190:2897–2924, 2013.

[13] Alan JA Robinson and Andrei Voronkov. *Handbook of automated reasoning*, volume 1. Elsevier, 2001.

[14] Oshani Seneviratne, Jonathan Harris, Ching-Hua Chen, and Deborah L McGuinness. Personal health knowledge graph for clinically relevant diet recommendations. *arXiv preprint arXiv:2110.10131*, 2021.

[15] Hao Wang. Logic of many-sorted theories. *The Journal of Symbolic Logic*, 17(2):105–116, 1952.

[16] David Westbrook, Helen Kennerley, and Joan Kirk. *An introduction to cognitive behaviour therapy: Skills and applications*. Sage, 2011.

Collaboration Through Shared Understanding: Knowledge Elicitation for a Mutual Theory of Mind in Human-AI Teams

Dina Acklin[1], Rebecca Goldstein[2], Jaelle Scheuerman[3], and Abby Ortego[4]

[1]U.S. Naval Research Laboratory

Abstract

With conversational AI becoming more prevalent in collaborative work, effective human-AI interaction requires a mutual theory of mind where both user and AI understand each other's goals, knowledge, and limitations. While language models show promise through techniques like chain-of-thought reasoning, they often struggle with domain-specific tasks. We propose that established knowledge elicitation methods from psychology and human factors — such as Cognitive Task Analysis (CTA), Hierarchical Task Analysis (HTA), and the GOMS model — can enhance AI understanding of task workflows and user cognition. By integrating these techniques, AI systems can improve workflow reasoning, adapt to user mental states, and foster more effective, trustworthy collaboration.

Introduction

The popularity of conversational AI is on the rise, resulting in a flood of AI-powered tools advertised to help people complete tasks via a user-friendly chat interface. The expectation is that people will be able to intuitively describe a task to the LLM agent through natural language, and that the agents will follow the instructions using the tools available to them. The current state of the art lends itself to an iterative process between the AI agent and user, with the user providing feedback and the AI agent working to incorporate that feedback and improve its output. However, the success of this continuous feedback loop relies both on the AI agent accurately recognizing the user's intentions, beliefs and goals and the user accurately understanding the capabilities and limitations of the AI agent.

Theory of mind (ToM) is often recognized in humans as the ability to reason about the mental states of others. As we continue to design collaborative workflows with AI agents, it is imperative that these systems can accurately interpret and respond to the mental states of their human teammates, as well as communicate their own limitations and capabilities. Ongoing efforts to implement ToM principles for AI often consider shared mental models between the AI system and a user. For example, when an AI agent is designed to help a user complete a task, techniques like chain of thought reasoning may be employed to generate the steps required to arrive at a solution, providing context about the solution or answer that was generated. The user may then continue to interrogate the system about its process, request changes or accept the solution as it is.

While this chain of thought reasoning has shown promise in simple tasks that do not re-

quire a great deal of domain specific knowledge, it has been shown to struggle in domain tasks that rely heavily on an understanding of specific jargon or processes (Kambhampati et al. 2024). We find that language models agents often produce incomplete or inaccurate workflows for domain specific tasks, leading to frustrated users and lost trust in the system. While interrogations into chain of thought reasoning can improve users' understanding of agent logic, a bi-directional relationship is needed wherein the agent also has a framework for user knowledge and processes improve human-machine teaming.

We propose that improving the mutual theory of mind between the AI and user in these collaborative work tasks would help users better calibrate their expectations and better leverage the combined strengths of the human and AI to complete the task. Psychology and human factors disciplines have well established methods for conducting knowledge elicitation in order to gain a comprehensive understanding of the processes and factors that contribute to goal completion. Common methods of eliciting knowledge include conducting document analysis, interviews, and direct observation, which allow for explicit descriptions of task procedures and details regarding the strategies, challenges, cues, and environmental conditions that may task performance. Using complementary techniques such as eye tracking, mouse or keyboard logging, and log file analysis can yield additional insight into implicit processes by recording behavioral measures that can be associated with specific tasks, particularly for those that are computer-based. Combining these methods allows researchers to qualitatively and quantitatively represent the tasks, actions, and goals that comprise a given task.

Human Factors Methods to Improve ToM

There are many ways in which knowledge elicitation techniques could support a mutual theory of mind in human-AI teams that have not been well explored. Techniques like cognitive task analysis support eliciting domain knowledge and understanding of the task workflow. Hierarchical task analysis is another technique for documenting task workflows and could provide valuable information to support AI systems in reasoning about workflow requirements and expectations. Models like the Goals, Operators, Methods, and Selection Rules can be combined with other techniques to recognize changes to the user's mental state in real time. For the remainder of the discussion, we describe established elicitation techniques and consider how they can lead to better collaborative experiences between human and AI systems.

Cognitive task analysis (CTA) was developed in order to document "complex cognitive systems" (Crandall, Klein, and Hoffman 2006). CTA provides methods to collect data about the system beyond actions or tasks, also considering what users know, what they need to know, and how knowledge is organized (and shared, in the case of group activities). This includes understanding the perceptual, attentional, decision making, memory, and judgment processes that underlie activities in support of achieving goals. Typically, subject matter experts are interviewed to provide context and comprehensive accounts of how problems are addressed. This approach is useful for documenting and understanding tasks that are more conceptual in nature, rather than driven by physical actions or processes. To this end, CTA is typically used to inform system design by identifying tasks that are especially cognitively demanding or rely on expert knowledge and use insights from cognitive psychology to suggest solutions and cognitive requirements. Depending on the goals of the practitioner, CTA can be represented in a

variety of formats, for example, by graphically representing the relationship between the goals and the information required or by modeling mental knowledge, tasks and decisions, and results.

One potential artifact to be drawn from CTA are concept maps, which can be used to represent expert knowledge and have been applied across a variety of domains, from improving medical education (Daley and Torre 2010) to documenting lessons learned from retiring NASA engineers (Coffey and Hoffman 2003). Concept maps document the links between facts and concepts comprising a given domain, can be continuously iterated upon, or linked to related concept maps to provide additional detail. Concept maps are highly analogous to knowledge graphs in that they are capable of flexibly conceptualizing, representing, and integrating disparate elements that can be used as data to represent domain knowledge (Hogan et al. 2022). This domain knowledge can be accessed by an AI system externally (such as through the retrieval augmented generation (RAG) popularized by language model systems) (Agrawal et al. 2024), or to create large synthetic datasets that can be used to train a model (Agarwal et al. 2021). Another popular elicitation technique is Hierarchical Task Analysis (HTA), which was developed to identify the sources of performance problems or failure, whether physical or cognitive, when performing complex tasks (Annett and Duncan 1967; Annett, Duncan, and Stammers 1971; Annett 2003). Beginning with an understanding of task goals, major objectives can be decomposed into subgoals or criteria necessary to achieve the desired outcome, along with the required operations and procedures associated with each goal state. Diagramming this process allows researchers to identify particularly demanding tasks and characterize the nature of their complexity (for example, tasks or operations that rely on specialized knowledge, skills, or teamwork), or how and when multiple or alternate operations might be employed to achieve a single goal. In the context of a human-AI team, there are multiple ways in which an HTA could improve the AI's representation of a workflow to improve reasoning about the task or communicate to the user when it lacks access to some specialized knowledge.

A final knowledge elicitation method we consider is the Goals, Operators, Methods, and Selection Rules (GOMS) Model. This model was developed to describe the steps for executing tasks that will achieve specific goals, as well as the knowledge that a person needs to accomplish those goals (Card, Moran, and Newell 1983). GOMS acts as an engineering model that is particularly tied to the system(s) used, identifying the methods used to achieve goals, the steps, or operators, comprising individual methods, and the selection rules used to determine when one set of methods may be performed over another. This methodology results in a description of physical or cognitive actions performed down to the most basic level, typically in the form of mouse clicks or keystrokes. While GOMS models may consider the cognitive processes that occur, these are seen as having limited bearing on the overall interface design. In this way, a GOMS analysis can be used to evaluate human performance as they interact with a given system. Models like GOMS can be employed in conjunction with implicit data collection to track the performance of the human-AI team and further supplement other elicitation techniques like HTA and CTA in identifying when things are going well (i.e. performance metrics are in line with expectations) and when they are not. This provides the AI system with the capability to recognize when problems arise and calibrate responses appropriately. For example, if it is determined that the user is in an especially cognitively demanding situation the AI model could tailor its explanations appropriately (Vasconcelos et al. 2023).

Conclusion

The above knowledge elicitation techniques represent a subset of possible methods leveraged by the human factors community that could support mutual theory of mind for human-AI teams. These are complementary to many AI approaches. For example, researchers have explored generating task workflows to support AI agents. However, these generally require the AI model to successfully generate a valid workflow or rely on a great deal of user input to validate and correct a generated workflow (Li and Ning 2023). Some work has explored ways to tune models to domain tasks through external knowledge bases and ontologies (Agrawal et al. 2024). More work is needed to explore how established techniques like hierarchical task analysis and concept maps could support synthetic data generation in domains where it is challenging to gather large datasets. The AI community has also begun to explore ways of using implicit information about the user to guide AI systems (Nobandegani, Shultz, and Rish 2023; Scheueman, Bishof, and Michael 2023). Models like GOMS could augment implicit feedback techniques to improve an AI system's ability to recognize changes to user's cognitive state and calibrate the responses accordingly. In conclusion, we've described some common knowledge elicitation techniques and how they can support the development of a mutual theory of mind for a human-AI team by 1) eliciting expert domain knowledge for new training datasets (i.e. CTA, concept maps), 2) identifying areas of the workflow that lend themselves better to human or AI effort (i.e. CTA, HTA) and 3) providing techniques to infer the user's mental state in real time (i.e. GOMS). These examples demonstrate how knowledge elicitation techniques may offer novel approaches for improving mutual theory of mind between AI agents and users in collaborative workflow.

References

Agarwal, O.; Ge, H.; Shakeri, S.; and Al-Rfou, R. 2021. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. arXiv:2010.12688. Agrawal, G.; Kumarage, T.; Alghamdi, Z.; and Liu, H. 2024. Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. arXiv:2311.07914.

Annett, J. 2003. Hierarchical Task Analysis. In Handbook of Cognitive Task Design. CRC Press. ISBN 978-0-429-22821- 6.

Annett, J.; Duncan, K.; and Stammers, R. 1971. Task Analysis. Training Information Paper / Department of Employment. H.M. Stationery Office.

Annett, J.; and Duncan, K. D. 1967. TASK ANALYSIS AND TRAINING DESIGN. Technical report.

Card, SK.; Moran, TP.; and Newell, A. 1983. The Psychology of Human-Computer Interaction. Hillsdale, New Jersey: Lawerence Erlbaum Associates.

Coffey, J. W.; and Hoffman, R. R. 2003. Knowledge Modeling for the Preservation of Institutional Memory. Journal of Knowledge Management, 7(3): 38–52.

Crandall, B.; Klein, G.; and Hoffman, R. 2006. Working Minds: A Practitioner's Guide to Cognitive Task Analysis. Bradford Books. A Bradford Book. ISBN 978-0-262- 03351-0.

Daley, B. J.; and Torre, D. M. 2010. Concept maps in medical education: an analytical literature review. Medical education, 44(5): 440–448.

Hogan, A.; Blomqvist, E.; Cochez, M.; D'amato, C.; Melo, G. D.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; Ngomo, A.C. N.; Polleres, A.; Rashid, S. M.; Rula, A.; Schmelzeisen, L.; Sequeda, J.; Staab, S.; and Zimmermann, A. 2022. Knowledge Graphs. ACM Computing Surveys, 54(4): 1–37.

Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L.; and Murthy, A. 2024. LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. arXiv:2402.01817.

Li, Z.; and Ning, H. 2023. Autonomous GIS: the next- generation AI-powered GIS. International Journal of Digital Earth, 16(2): 4668–4686.

Nobandegani, A. S.; Shultz, T. R.; and Rish, I. 2023. Cognitive Models as Simulators: Using Cognitive Models to Tap into Implicit Human Feedback. In Interactive Learning with Implicit Human Feedback.

Scheueman, J.; Bishof, Z.; and Michael, C. J. 2023. Modeled Cognitive Feedback to Calibrate Uncertainty for Interactive Learning. In Interactive Learning with Implicit Hu- man Feedback.

Vasconcelos, H.; Jo¨rke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M.; and Krishna, R. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. arXiv:2212.06823.

Detective ToM: A Theory of Mind Framework for Analysis of Surprising Yet Coherent Crime Mysteries

Eitan Wagner, Renana Keydar, and Omri Abend

Hebrew University of Jerusalem

Abstract

In this paper, we present a framework for analysis of crime mysteries that takes into account *surprisal*, *coherence*, and their tradeoff. Modeling the change in the reader's mindset when the truth is revealed, our framework conceptualizes the surprisal based on the initial mindset and the coherence based on the ultimate mindset. We design metrics to measure these qualities and apply them to real and model-generated stories. Our initial results show that while LLMs succeed in generating misleading stories, they struggle to generate surprises that form a coherent story.

Introduction

Maintaining the reader's attention, where suspense plays an important role [10, 8], is a crucial aspect in story writing in general and crime mysteries in particular. A key element in suspense is *surprisal* [8]. Another defining aspect of crime mysteries is its *coherence*, where the solution to the mystery makes sense, at least in retrospect [5].

Some computational work analyzes surprisal in stories, based on likelihood or entropy – when the likelihood of the continuation is low or when the belief change is big [13, 12]. Other works focus on the coherence aspect of crime mysteries, using them as a testbed for narrative understanding and deductive reasoning [3, 2, 11].

These works independently address the two aspects, thus failing to model the tension between them. Indeed, while surprise results from many possible outcomes, coherence must rule out meaningless ones [7]. We argue that a crime mystery's true quality lies in handling this tension, which we show is a fundamental limitation for a single language model.

Inspired by work that models agents through their theory-of-mind [4, 1], we develop a framework to analyze crime mysteries based on the distinction between a naïve initial mind (i.e., one that estimates the outcome regardless of the author's intentions) and the ultimate one, which gives a high probability to permissible continuations. We hypothesize that the effect of coherent surprise arises when the naive reader is misled and gives a low probability to the final outcome, whereas the revelation leads to a different mind in which the outcome is coherent.

We further hypothesize, based on works about the role of clues in crime mysteries [9, 6], that the change in the reader's mind is a result of the clues and their interpretation. A naïve reader makes assumptions that lead to an incorrect interpretation of the clues, giving a low likelihood to the true events. Other minds realize that these assumptions were incorrect. This leads to theoretical questions as to the extent to which the clues themselves (without the explanation) limit the possible outcomes.

We propose metrics to separately measure surprisal and coherence, as well as combined

metrics that measure the balance between the two. We show how these can be computed for stories generated by LLMs and partially computed for real human-written stories.

Our experiments show that while LLMs succeed in generating misleading stories, they generally fail to create clues that limit the outcome. Models tend to be excessively creative in the endings, leading to opposite outcomes even when the given story includes many clues.

1 An "Open-Minded" Reader

We denote a *story* by $\mathbf{x} = x_1, x_2 ... x_n$, where x_i is the i-the paragraph in the story. In a *misleading crime mystery*, we assume that one true and one distracting culprit (=distractor) are revealed in the story $\mathbf{x}$. The distractor is a character that is intended to be believed as the culprit.

1.1 A Naïve Reader

We define a *reader model*, M, as a probability distribution over the suspects Y, given a (partial) story. Y_i^M is the random variable that represents the assumed culprit given the story from the beginning until the i-th paragraph. Inspired by the distinction between the "crime" and "investigation" levels of a crime mystery [7], we define the naïve reader, M_0, as one that estimates the probabilities based only on the information revealed about the crime, with the outcome estimated regardless of the investigation.

The surprisal for the naive reader is measured by entropy reduction $ER = H(Y_i^M) - H(Y_{i-1}^M)$ and by a new score that takes into account the phases in the story. The latter finds an optimal segmentation into phases [7]: introduction (where the reader should have no opinion about the culprit), suspicion (where the reader should be misled), and revelation, and measures the distance between the model's prediction and the intended distribution for the phase.

1.2 The coherent surprise trade-off

Roughly speaking, a surprise has a low probability given a premise and something coherent has a high probability given the premise. This forms a trade-off. Formally, we show that an event B, which is defined as changing the probability of some event A from low to high, must have a low probability. This is intuitive since otherwise event A would have a high probability even without knowing if B happened. In the paper, we provide a formal proof.

If we measure surprisal based on a single language model (i.e., a single joint probability) then any improbable continuation necessarily does not make sense (judging by the probability with this model). Intermediate steps (i.e., discoveries) cannot create "phase changes" unless one of the steps is itself improbable. Since coherent surprises are an observed phenomenon, we conclude that using a single model is insufficient as a model for the human reader.

1.3 Multiple Reader Minds

We argue that the phenomenon is a result of multiple, non-unified, reader minds. The first mind, M_0 is the reader's naïve mind-state. The second, M_1, is the mind-state that is switched to following the revelation and applied in retrospect. After the revelation, the mind states are identical. In the case of generated stories, we assume that M_1 is identical to the writer's model, allowing us to sample from it.

1.4 Balanced surprisal score

Ideally, as more clues are revealed in the story, uncertainty about the culprit should be reduced. In other words, we expect the probability of the true culprit under M_1 to increase monotonically. On the other hand, up to the revelation, we expect the probability of the true culprit by M_0 to be low. In the paper we define the *expected balanced surprisal score* that takes into account these expectations.

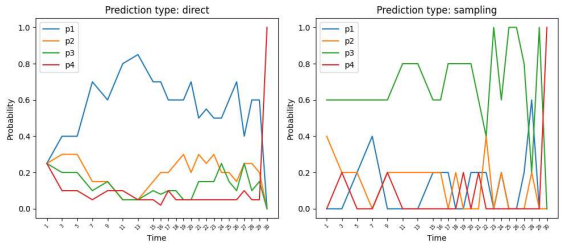


Figure 1: A story with a sudden surprising ending. The true culprit is 4 and the distractor is 1. Sampling and direct (naive) predictions do not converge until the very end.

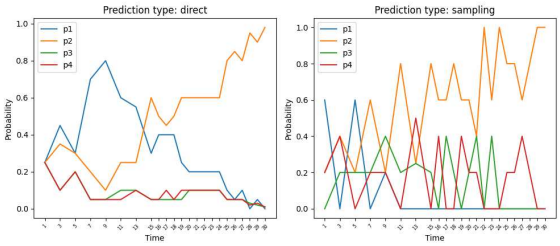


Figure 2: A story where sampling converges before direct prediction. The true culprit is 2 and the distractor is 1.

2 Experiments and Preliminary Results

We measure the proposed scores for both generated and real stories. In real stories, we can measure scores for the naïve reader based on a prediction model. In generated stories, we can sample from the generating model so we can measure the coherence and balanced scores.

Generation Setup. We generate a story with long-context LLMs. For proper comparison, we want to control the length of the story. For this, we provide the number of expected paragraphs. To allow sampling based on a given prefix, we generate the story paragraph by paragraph. This way we can rerun the process starting at any intermediate point while preserving the exact prompts. We explicitly instruct the model to generate a detective story with a distractor and revelation. We also instruct it to make the story coherent such that after revelation it will make sense.

Stories. We use these models to generate the stories: Gemini-1.5, Llama-3.1, GPT-4o, and o1. We also analyze real detective stories, comparing works by Arthur Conan Doyle (Sherlock Holmes), and Agatha Christie (Hercule Poirot).

Results. Some clear trends are revealed. First, despite clear cases of mode-collapse in many details (such as the names used in the stories), the identity of the culprit is never determined from the beginning. Moreover, in most cases, the real culprit's identity (through sampling) does not converge before the revelation. Two examples generated by Llama-3.1-70B-Instruct are given in Figures 1 and 2.

Analyzing real stories with naive predictions, we see clear differences between writers. Hercule Poirot's stories tend to be more surprising, compared to Sherlock Holmes's stories. Poirot's stories show many distracting characters and avoid early revelation of the culprit.

3 Conclusion

We present a framework for the analysis of surprising yet coherent stories. The framework is based on probabilistic properties and on Theory of Mind modeling. Our framework can capture the thin tension between unpredictability and coherence, showing that the generation of high-quality crime mysteries remains challenging.

Acknowledgements

This research was supported by grants from the Israeli Ministry of Science and Technology and the Council for Higher Education and the Federmann cyber security research center.

References

[1] Nitay Alon, Lion Schulz, Peter Dayan, and Jeffrey Rosenschein. A (dis-)information theory of revealed and unrevealed preferences. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 2022.

[2] Maksym Del and Mark Fishel. True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 314–322, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Lea Frermann, Shay B. Cohen, and Mirella Lapata. Whodunnit? crime drama as a case for natural language understanding. *Transactions of the Association for Computational Linguistics*, 6:1–15, 01 2018.

[4] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multiagent settings. *Journal of Artificial Intelligence Research*, 24:49–79, July 2005.

[5] Alan H. Goldman. The appeal of the mystery. *The Journal of Aesthetics and Art Criticism*, 69(3):261–272, 08 2011.

[6] Jesper Gulddal. Clues. In *The Routledge Companion to Crime Fiction*, pages 194–201. Routledge, 2020.

[7] Peter Hühn. The detective as reader: Narrativity and reading concepts in detective fiction. *MFS Modern Fiction Studies*, 33(3):451–466, 1987.

[8] Yuliya Khrypko and Peter Andreae. Towards the problem of maintaining suspense in interactive narrative. In *Proceedings of the 7th Australasian Conference on Interactive Entertainment*, pages 1–3, 2011.

[9] Franco Moretti. The slaughterhouse of literature. *MLQ: Modern Language Quarterly*, 61(1):207–227, 2000.

[10] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[11] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning, 2024.

[12] Prashanth Vijayaraghavan and Deb Roy. M-sense: Modeling narrative structure in short personal narratives using protagonist's mental representations, 2023.

[13] David Wilmot and Frank Keller. Modelling suspense in short stories as uncertainty reduction over neural representation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1763–1788, Online, July 2020. Association for Computational Linguistics.

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.

Evaluating Machine Theory of Mind: A Critical Analysis of ToMnet-N

Nikita Krasnytskyi[1] and Fabio Cuzzolin[2]

[1]Institute for Ethical AI, Oxford Brookes University, UK
[2]School of Engineering, Computing and Mathematics, Oxford Brookes University, UK

Abstract

This paper critically re-examines the ToMnet family's approach to Machine Theory of Mind through our open-source model, ToMnet-N. While ToMnet-N replicates and extends prior experiments—including a re-implemented False-Belief test inspired by classical paradigms [1]—our analysis shows that its success largely stems from pattern recognition supported by targeted training rather than genuine inference of mental states [12, 11]. We discuss architectural modifications—such as the omission of a historical Mental Net [2] and the adoption of an autoregressive RNN enhanced by a novel map generation method based on Wave Function Collapse [19]—and propose that achieving true Machine Theory of Mind may require a paradigm shift beyond current meta-learning frameworks [10, 17].

Introduction

The ambition to endow machines with a Theory of Mind (ToM)—the ability to attribute beliefs, desires, and intentions—has profound implications for human–machine interaction [3, 18]. Early models, such as DeepMind's ToMnet [16], demonstrated that meta-learning frameworks could predict agent behavior in simplified grid-world environments. Yet, subsequent critiques have raised concerns that such systems may rely on memorized patterns rather than emulating genuine mental state inference [12, 9]. In this work, we introduce ToMnet-N [14] as an evolution of these approaches and critically assess its capability to replicate human-like ToM.

Background

ToMnet was originally developed to predict agents' trajectories and goal consumption in grid-world settings [16]. Its success spurred the development of several derivative models—such as ToMnet+ [2], ToMnet-G [20], and Trait-ToM [13]—each extending the methodology to more complex tasks. Despite these advances, evidence suggests that performance improvements may derive from memorization of training patterns rather than from authentic mental state inference [12, 11]. Moreover, classical psychological tests like the False-Belief Test [1, 5]—which even young children pass with minimal exposure—highlight the gap between human cognition and current computational models.

ToMnet-N Development

To address these limitations, ToMnet-N incorporates several key modifications:

- **Autoregressive Prediction:** Rather than predicting multiple outputs simultaneously, ToMnet-N iteratively forecasts the next action, thereby reducing reliance on historical data that may bias learning [2].
- **Exclusion of the Mental Net:** Previous models incorporated a module to process historical trajectories; by omitting this component, ToMnet-N mitigates shortcut learning—a phenomenon where the model leverages spurious cues rather than understanding underlying mental states [7].
- **Enhanced Environmental Variability:** We employ a novel map generation technique based on the Wave Function Collapse algorithm [19], which produces diverse and challenging environments. This variability reduces the likelihood that the model can rely solely on memorized patterns.

Experimental Analysis

In our experiments—replicating core setups including a False-Belief test modeled after the Sally-Anne paradigm [1]—ToMnet-N demonstrated accurate trajectory predictions. However, its success appears predominantly due to recognizing patterns from training data (with approximately 10% dedicated to False-Belief scenarios [16]) rather than deriving genuine mental state representations [12]. These findings resonate with broader concerns regarding shortcut learning in deep neural networks.

Discussion

The limitations observed in ToMnet-N suggest that current meta-learning architectures may be inherently constrained in capturing the nuanced Theory of Mind inherent in human cognition [17]. Although models like ToMnet and its derivatives have shown promise in predicting agent behavior—sometimes rivaling the predictive capabilities seen in human studies [8]—the extensive data requirements (e.g., millions of training trajectories [16]) contrast sharply with the spontaneous ToM abilities observed in young children [1, 5]. Achieving genuine Machine ToM may require new architectures that integrate insights from cognitive science and multi-modal reasoning frameworks [6, 10].

Conclusion

ToMnet-N provides a valuable case study in evaluating the limits of ToMnet-like approaches. Despite its architectural improvements and enhanced environmental variability, the model's performance remains primarily a consequence of pattern recognition rather than true mental state inference. Our analysis underscores the need for a paradigm shift in computational frameworks if we are to bridge the gap between artificial and human-like Theory of Mind [15, 21]. Future research should explore integrative models that leverage multi-modal inputs to more authentically replicate the cognitive processes underlying human ToM.

Acknowledgements

We thank the organizers of the ToM4AI workshop for their invaluable feedback and support. We also acknowledge the contributions of researchers advancing the fields of Computational Theory of Mind and human–machine interaction [4, 22].

References

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

[2] Yun-Shiuan Chuang, Hsin-Yi Hung, Edwinn Gamborino, Joshua Oon Soo Goh, Tsung-Ren Huang, Yu-Ling Chang, Su-Ling Yeh, and Li-Chen Fu. Using machine theory of mind to learn agent social network structures from observed interactive behaviors with targets. 08 2020.

[3] Fabio Cuzzolin, Alice Morelli, Bognan Cîrstea, and Barbara J. Sahakian. Knowing me, knowing you: theory of mind in ai. *Psychological Medicine*, 50(7):1057–1061, 05 2020.

[4] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pinar Yolum. *Abstracting Minds: Computational Theory of Mind for Human-Agent Collaboration*. 09 2022.

[5] Uta Frith. Mind blindness and the brain in autism. *Neuron*, 32(6):969–979, 2001.

[6] Mark K. Ho and Thomas L. Griffiths. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Reviews*, 5(1):33–53, 05 2022.

[7] Sean Dae Houlihan, Max Kleiman-Weiner, Luke Hewitt, Joshua B. Tenenbaum, and Rebecca Saxe. Emotion prediction as computation over a generative theory of mind. *Royal Society*, 381(2251), 06 2023.

[8] Matthew Hutson. Artificial intelligence has learned to probe the minds of other computers. *American Association for the Advancement of Science*, 07 2018.

[9] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 10 2019.

[10] Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in Artificial Intelligence*, 5, 2022. 778852.

[11] Scott A. Miller. Children's understanding of second-order mental states. *American Psychological Association*, 135(5):749–773, 01 2009.

[12] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. Evaluating theory of mind in question answering. *Cornell University*, 08 2018.

[13] Dung Nguyen, Phuoc Nguyen, Hung Le, Kien Do, Svetha Venkatesh, and Truyen Tran. Learning theory of mind via dynamic traits attribution, 04 2022.

[14] NikKras. ToMnet-N: Theory of Mind Network. `https://github.com/Nik-Kras/ToMnet-N`, 2023.

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

[16] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR, 10–15 Jul 2018.

[17] Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. Theory of mind and decision science: Towards a typology of tasks and

computational models. *Neuropsychologia*, 146:107488, 2020.

[18] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12:13–24, 01 2002.

[19] Daniel Shiffman. Coding challenge 171: Wave function collapse, 07 2022. Accessed on 2023-10-21.

[20] Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9614–9625. PMLR, 18–24 Jul 2021.

[21] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691, 2005.

[22] Alan F. T. Winfield. Experiments in artificial theory of mind: From safety to storytelling. *Frontiers in Robotics and AI*, 5, 2018.

Finding Common Ground: Comparing Two Computational Models of Social Intelligence

Ramira van der Meulen[1], Rineke Verbrugge[2], and Max van Duijn[1]

[1]Leiden University
[2]University of Groningen

Introduction

AI collaboration is a unique challenge that requires the combination of multiple viewpoints and sources of information to reach a common goal. In human-human interaction, this collaboration is often said to rely on theory of mind (ToM), the ability to take someone else's perspective and make estimations of their beliefs, desires and intentions, in order to make sense of their behaviour and attitudes towards the world [9, 1]. It seems logical to apply a similar strategy for machines: Use ToM to align and use shared perspectives. However, we know from human-human interaction that the use of ToM is quite costly [8] and prone to error [12]. This invites research into alternative strategies that can effectively support interaction.

Here we implement and test such an alternative: Common Ground (CG) as introduced by Clark and Marshall [4, 3]. During our workshop talk, we discuss two simple computational models of intelligence that each try to solve a cooperative counting game called 'The Game'. Both models initially use an active (and relatively costly) form of reasoning inspired by ToM, but later defer to a passive (computationally cheap) strategy using a learned CG. We also discuss how our findings fit into the literature on human-human interaction.

The Game

Setup Our experiments use a two-player variant of 'The Game' [2], in which players are not allowed to verbally communicate. 'The Game' itself consists of 98 cards, ranging between 2 and 99, and four play piles, two of them starting at 1 and two of them starting at 100. The piles starting from 1 are used to count up, and the piles starting from 100 are used to count down. At the start of the game, the 98 cards are randomly shuffled, after which each player is handed 7 cards. Players are not allowed to share information about the cards they have on hand. The rest of the cards is put into a central pile, face-down, as a draw pile.

Game-play After determining the starting player (in our version this is always player 1), players take turns to each play at minimum two cards on one of the four central piles. It is allowed to play all seven cards on hand. An example play would be to play a '3' on one of the piles starting with a '1' (the new start then becoming '3'), and playing a '96' on one of the piles starting from '100' (now starting at '96'). After finishing their turn, players pass their turn and draw new cards from the deck (back up to seven cards). The one exception to the play direction is when the card has a difference of '10' from the starting card - then

it may be used for either direction (i.e., 15 'up' may become 5 'up'). The goal is to play every card in their hands and the deck – the final score equalling the number of played cards. An example is available in Figure 1.

Figure 1: Counting example in 'The Game'. '10' jumps can reset the counting process.

Risk Assessment Players are not aware of the cards in their partner's hands, nor the cards that are still in the draw pile. Since players *have* to play at least two cards per turn, they may be forced to play cards that significantly raise or decrease the count of a pile. It is, therefore, beneficial to play 'close enough' cards in addition to the two mandatory cards, to ensure that as many cards are played before it is too late. Players have to coordinate with each other and perform risk assessments on whether they should or should not play a certain third, fourth, etc. card in their hands. This coordination becomes easier with verbal communication, but our version is deliberately without verbal communication in order to increase the need for modelling the other's strategy based on input from the game-play only.

Model Descriptions

Associative Learning Model The first model that we implemented is an **associative learning** [11] model (ALM), in which agents with different, randomly initialized, *personalities* learn 'The Game' by observing each other's behaviour and how this affects the game environment [10]. These personalities consist of three features:

1. **Self-benefit** evaluates score improvement – adjusted based on whether an agent's own decisions impacted the final score;
2. **Eagerness** looks at the number of cards played by a partner, and to accommodate by also playing fewer/more cards – adjusted based on whether accommodation leads to score improvement;
3. **Cooperativeness** evaluates an agent's trust in its partner – adjusted based on whether the combination of self-benefit and eagerness of both agents leads to overall game-play success.

By adjusting these features to influence whether a computational agent plays an additional card with a certain difference to a central pile, agents will eventually find the right values that work in tandem with their partner. It does so by calculating whether the *Self-benefit* to play a card outweighs its *Eagerness* ($play_{card} : selfBen * totalNumCards > eagerness * difFromTopCard_{card}$; *Cooperativeness* modulates the strength of the *Eagerness* update). During this process, both agents observe if the behaviour of the other changes, i.e., whether they stop changing the number of cards they play, and if their combined behaviours still improve the score. If an agent thinks a partner's behaviour has stabilized (number of cards played and an unchanged score for n games), it stops mapping their behaviour, and locks in its current personality towards them, assuming there to be CG. Both agents make this decision individually. If they make this decision at a similar moment, the game-play remains stable: The agents have found CG on a combined strategy to solve the game. If either agent is wrong, then the collaboration becomes less fruitful, as one agent

keeps updating its values, while the other is blissfully unaware.

Simulation Theory of Mind Model The second model that we implemented is inspired by **simulation ToM** [7], using explicit recursive reasoning steps [5, 6]. Initially, it evaluates the value of the maximum card difference it can still play without blocking any of its partner's cards, with a (learned) 20% chance that the partner *will* probabilistically have a card lower than this value (playing a gap of 40 is irrational, playing a gap of 3 less so – 'I'm okay with playing a count of 3, because the probability of my partner having a 0-3 difference card is $p \leq 0.2$). Then, through ToM, it considers "If I account for their space, I should only play a 0-2 difference in case they *do* have a card that's 3 higher." (ToM level-1). This process continues until the gap is 0 difference (ToM level 2 would be 0-1). As accounting for one's partner too much is an overly cautious approach that will lead to one's own cards becoming unplayable, it is in the agents' best interest for one of them to play exactly one ToM level higher than their partner. This accounting for the partner's nature allows for a collaboration beyond game optimisation. Once an agent is certain of its partner's ToM level, it 'freezes' the difference ranges it plays given a specific situation, no longer explicitly modelling its partner: It relies on the values established in previous interactions.

Model Comparison

Both models can map learning to play 'The Game' successfully. ToM agents are able to play with ToM agents – and the ALM agents can play with ALM agents. Play of the ALM generally slowly stabilizes at 88 out of 100 points. Depending on the initial personalities, it starts around 45-60 out of 100 points. The ToM model stabilizes around 83 to 85 points, depending on the ToM levels of the two agents playing the game, but does so quickly. This model similarly starts around 45-60 points if it has not yet learned the right probabilities.

Most importantly, our agents show a play-quality retention after they stop explicitly modelling fellow agents, provided they find a shared approach. For the ALM, this means that the personality values are fixed once both parties think they agree to a 'best strategy' collaboration, and for the ToM model this means that both parties agree on a collaboration based on what value difference an agent will risk for playing an extra card. In both cases, no new estimation about partner personality or ToM is required, while the results remain the same, even in new games.

Discussion

Interestingly, the models perform (collaboratively) quite similarly. Self-benefit in the ALM and the p-value estimation in the ToM Model both act as means to find an economically optimal solution, whereas the ALM's eagerness and the ToM model's ToM estimation help to find the right value for collaboration. Both models can be tuned in such a way that their onset strategy is no longer required once agents are well-acquainted. ALM agents successfully hold on to their learned behaviour to retain performance, no longer needing any model updates, and ToM agents can freeze the 'max difference' instead of simulating the ToM of their partner. Knowledge of both one's partner and the specific task are crucial, but once that knowledge is there, the shared alignment overcomes the excessive need for types of explicit perspective modelling.

Acknowledgements

This research is part of the Hybrid Intelligence gravitation programme – number 024.004.022, financed by the Netherlands Organisation for Scientific Research (NWO).

References

[1] Ian Apperly. *Mindreaders: the Cognitive Basis of "Theory of Mind"*. Psychology Press, 2010.

[2] Steffen Benndorf. The game. `https://boardgamegeek.com/boardgame/173090/the-game`, 2024. Accessed: 2024-12-11.

[3] Herbert H Clark. *Using Language*. Cambridge University Press, 1996.

[4] Herbert H Clark and C. R. Marshall. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*, pages 10–63. Cambridge, England: Cambridge University Press, 1981.

[5] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. Theory of mind in the Mod game: An agent-based model of strategic reasoning. In *Proceedings ECSI 2014*, pages 128–136, 2014.

[6] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems*, 36(2):30, 2022.

[7] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12):493–501, 1998.

[8] Penelope A. Lewis, Amy Birch, Alexander Hall, and Robin I. M. Dunbar. Higher order intentionality tasks are cognitively more demanding. *Social Cognitive and Affective Neuroscience*, 12(7):1063–1071, 03 2017.

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

[10] Ramira van der Meulen, Rineke Verbrugge, and Max van Duijn. Common ground provides a mental shortcut in agent-agent interaction. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 281–290. IOS Press, 2024.

[11] Edward A Wasserman and Ralph R Miller. What's elementary about associative learning? *Annual Review of Psychology*, 48(1):573–607, 1997.

[12] Ross Wilson, Ales Hruby, Daniel Perez-Zapata, Sanne W van der Kleij, and Ian A Apperly. Is recursive "mindreading" really an exception to limitations on recursive thinking? *Journal of Experimental Psychology: General*, 152(5):1454–1468, 2023.

How Well Can Vision-Language Models Understand Humans' Intention? An Open-ended Theory of Mind Question Evaluation Benchmark

Ximing Wen, Mallika Mainali, Anik Sen

College of Computing and Informatics, Drexel University, Philadelphia, USA
xw384@drexel.edu, mm5579@drexel.edu, as5867@drexel.edu

Abstract

Vision Language Models (VLMs) have demonstrated strong reasoning capabilities in Visual Question Answering (VQA) tasks; however, their ability to perform Theory of Mind (ToM) tasks, such as inferring human intentions, beliefs, and mental states, remains underexplored. We propose an open-ended question framework to evaluate VLMs' performance across diverse categories of ToM tasks. We curated and annotated a benchmark dataset of 30 images and evaluated the performance of four VLMs of varying sizes. Our results show that the GPT-4 model outperformed all the others, with only one smaller model, GPT-4o-mini, achieving comparable performance. We observed that VLMs often struggle to infer intentions in complex scenarios such as bullying or cheating. Our findings reveal that smaller models can sometimes infer correct intentions despite relying on incorrect visual cues. The dataset is available at `https://github.com/ximingwen/ToM-AAAI25-Multimodal`.

Introduction

Understanding human intentions through visual cues is a fundamental aspect of social intelligence, allowing effective communication, collaboration, and interaction [2]. This capability, often referred to as the Theory of Mind (ToM), involves the ability to infer the beliefs, desires, and intentions of others based on observable behaviors and environmental contexts [9, 7, 12].

Recent advances in VLMs have demonstrated impressive abilities in multimodal reasoning, combining visual and textual information to perform complex tasks [5, 10, 13]. However, their capability to perform ToM-like reasoning, specifically in interpreting intentions from visual cues, remains underexplored. For example, Etesam et al. [4] only investigate the emotional component of ToM, instead of exploring more broad categories such as intentions, religions, etc. Jin et al. [6] frame the ToM task as a binary choice question, without requiring VLMs to engage in open-ended reasoning. Consequently, this approach may not fully capture the VLMs' capability to perform ToM tasks.

To further highlight, ToM tasks present unique challenges for VLMs, requiring both visual feature extraction and contextual reasoning to infer hidden mental states. Thus, our study, which evaluates VLM performance on ToM tasks through an open-ended question framework, is pivotal to assessing VLMs' capacity for advanced multimodal understanding and social intelligence.

Open-ended Question Framework

In this study, we aim to investigate the capability of VLMs to perform ToM tasks by testing their ability to interpret intentions based on visual cues in images. To fully evaluate whether VLMs truly understand humans' intentions, we proposed an open-ended question framework composing the following three research questions:

- **Q1: How effectively can VLMs identify human intentions in visual scenarios?**
- **Q2: Can VLMs recognize accurate visual cues and use them to perform ToM tasks?**
- **Q3: Can VLMs comprehend human intentions sufficiently to make reasonable future inferences?**

Q1 focuses on inferring individuals' mental states and intentions, a core ToM skill. **Q2** examines the model's ability to identify and articulate visual cues, linking observations to inferred mental states. **Q3** evaluates predictive reasoning asking the model to infer potential future actions or events based on the scene.

By designing tasks that require inferring the purpose or mental state of individuals depicted in diverse scenarios, we seek to evaluate the extent to which models align with human-like reasoning in visual intention understanding. Our findings contribute to research on VLMs by highlighting their strengths and limitations in approximating human cognition, paving the way for socially aware AI advancements.

Data Development

Data Collection We defined 30 scenarios based on two intention categories (emotion-based and action-based) and images sourced from platforms including iStock, Shutterstock, Unsplash, and Pexels under appropriate licenses to ensure copyright compliance. We only included images that conveyed clear intentions with measurable visual cues, such as facial expressions, body language, interaction with objects, and eye gaze that indicated the mental states and intentions of the individuals. Each image underwent a comprehensive review to ensure suitability for research objectives and images with ambiguous cues were excluded. The final dataset provides diverse and suitable content for research. An overview of the dataset is shown in Figure 1.

Data Annotation Each author annotated a subset of 10 images, providing detailed descriptions in three categories: intention, visual cues, and future inference. A third-party evaluator reviewed and validated the 30 annotations to ensure consistency and accuracy, confirming that they accurately captured intentional actions, visual cues, and potential future inferences.

Experimental Design

Task We designed a structured prompt to generate responses from VLMs that aligned with the objectives of this study. The prompt is as follows: *"Based on the given image, answer the following in one sentence each: (1) What do you think is the intention, mental state, feeling or belief of each person in the image? (2) What visual cues in the image helped you determine what people might be thinking or feeling? (3) Can you infer what might happen next?"* VLMs were expected to extract the intentions from the image, recognize visual cues that support the inferred intention, and generate plausible future scenario descriptions consistent with the context of the image.

Models We evaluated the performance of four VLMs - GPT-4 [1], GPT-4o-mini (8B parameters) [11], Deepseek v1 (7B parameters) [3], and LLaVA (7B parameters) [8]-in inferring ToM

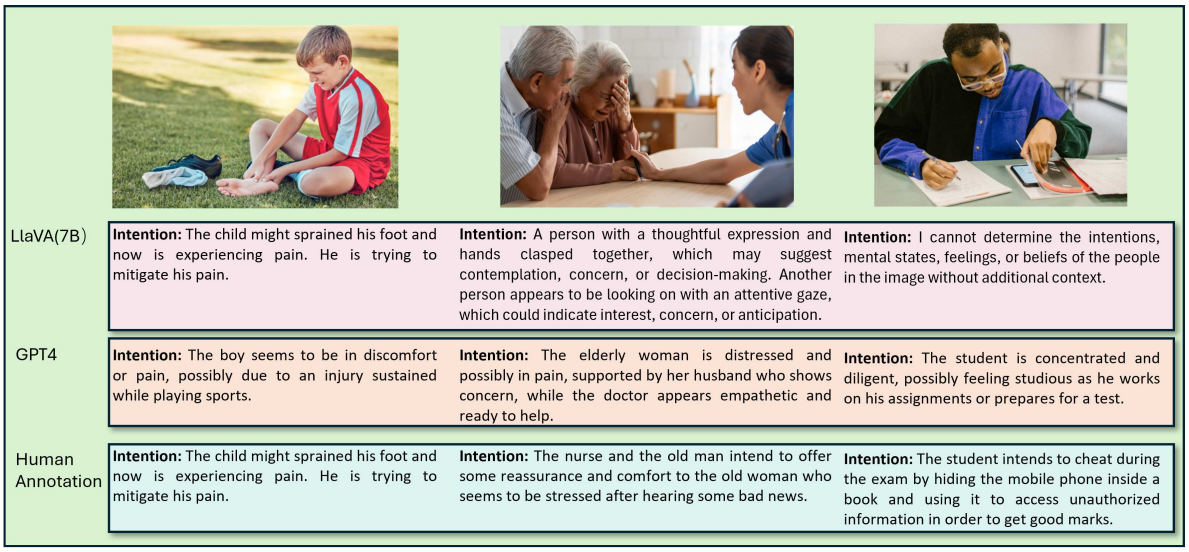


Figure 1: An overview of the three samples used for this work to evaluate Theory of Mind (ToM) capabilities of three vision language models (VLMs). This preview shows the intention generated using two VLMs: LLaVA (7B) and GPT-4, along with the human-annotated intention.

components from images.

Evaluation Model responses were manually compared with human annotations using a scoring system based on keyword relevance and accuracy. A score of 1 was given for responses with correct or synonymous keywords that accurately described the context. Partially correct responses were given a score of 0.5. Smaller models, such as DeepSeek, often identified intentions correctly, but struggled with scenario details, such as misidentifying gender or objects. While object recognition errors were ignored for the 'intention' category, these inaccuracies were given a score of 0.5 in the 'visual cues' category. Responses without relevant keywords were given a score of 0.

Result and Discussion

We assessed the performance of VLMs across three ToM tasks using our metric. The results, presented in Table 1, display the accuracy scores out of 30 for each category.

Accuracy across three ToM tasks Among the four models tested, GPT-4 performed the best on all tasks. Despite being a smaller model, GPT-4o-mini had comparable scores. In contrast, LLaVA-7B achieved significantly lower scores, indicating its limited ability to interpret subtle visual cues and make accurate inferences. Deepseek v1-7B outperformed LLaVA-7B but underperformed compared to GPT-based models.

VLM	Intention	Visual Cue	Future Inf.
GPT4	27	27	28
GPT4o-mini	27.5	27	27.5
LLaVA-7B	7.5	8.5	7
Deepseek-7B	17	16.5	16

Table 1: Performance of VLMs' responses for inferring intention, visual cues, and future inference

What types of human intentions can VLMs recognize? We found that GPT-based models could identify a range of human intentions, such as determination, care, frustration, compassion, praying, and bullying. Deepseek v1-7B could recognize some intentions, but struggled with subtle ones such as emotional distress, frustration, and bullying. LLaVa-7B was unable to identify most human intentions. Interestingly, none of the four VLMs could accurately identify when a person intended to cheat during an exam. They misinterpreted them as 'focused' or 'multitasking'.

Can VLMs accurately capture visual cues to infer human intentions? Our analysis revealed that GPT-based models can interpret visual cues and infer human intentions, with GPT4o-mini occasionally making minor errors, such as mistaking a purse for a camera, but still capturing overall intentions accurately. However, all the four models struggled with contextual nuances, misidentifying religious attire (cassock and stole) as graduation robes, leading to incorrect inferences.

Deepseek v1-7B could interpret body language but often misclassified facial expressions, associating direct gazes with engagement and indirect gazes with disinterest. This led to errors, such as misclassifying a police interrogation as a hospital scene, and mislabeling bullying as 'amusing interaction.' It also failed to identify professions based on visible uniforms, such as firefighters and police officers, despite accurately inferring intentions. LLaVA-7B struggled to understand most visual cues.

Can VLMs accurately interpret human intentions well enough to make reasonable future inferences? We found that, in most scenarios, GPT-based models made reasonable future inferences, often suggesting practical steps for conflict and emotional distress. Deepseek v1-7B also made reasonable future inferences, but its inaccuracies in intention recognition led to occasional errors. LLaVA-7B struggled with future inferences, often citing limited capabilities.

Conclusion & Future Directions

Our analysis shows that while some VLMs can infer human intentions from visual scenarios, they often need further fine-tuning to contextualize subtle cues. In future work, we plan to incorporate a reasoning template to guide VLMs in generating more contextually accurate responses by ensuring that key elements are considered in their reasoning process.

Acknowledgements

The second and third authors were supported by the In the Moment (ITM) project, funded by the Defense Advanced Research Projects Agency (DARPA) under contract number HR001122S0031.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Ralph Adolphs. The social brain: neural basis of social knowledge. *Annual review of psychology*, 60(1):693–716, 2009.

[3] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[4] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. Emotional theory of mind: Bridging fast visual processing with slow linguistic reasoning. *arXiv preprint arXiv:2310.19995*, 2023.

[5] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

[6] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*, 2024.

[7] Dimitrios Kapogiannis, Aron K Barbey, Michael Su, Giovanna Zamboni, Frank Krueger, and Jordan Grafman. Cognitive and neural foundations of religious belief. *Proceedings of the National Academy of Sciences*, 106(12):4876–4881, 2009.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[9] Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2): 622–646, 2007.

[10] Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[11] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL `https://www.openai.com/gpt4o-mini`. Smaller, cost-efficient version of GPT-4o with 8B parameters.

[12] Jun Zhang, Trey Hedden, and Adrian Chia. Perspective-taking and depth of theory-of-mind reasoning in sequential-move games. *Cognitive science*, 36(3):560–573, 2012.

[13] Yipeng Zhang, Xin Wang, Hong Chen, Jiapei Fan, Weigao Wen, Hui Xue, Hong Mei, and Wenwu Zhu. Large language model with curriculum reasoning for visual concept recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6269–6280, 2024.

"I apologize for my actions": Emergent Properties of Generative Agents and Implications for a Theory of Mind

N'yoma Diamond[1] and Soumya Banerjee[1]

[1]University of Cambridge

Abstract

This work explores the design, implementation, and usage of generative agents towards simulating human behaviour. Through simulating (mis)information spread, we investigate the emergent social behaviours they produce.

Generative agents exhibit novel and realistic emergent social behaviours, such as deception, confrontation, and internalized regret. Using deception, agents avoid certain conversations. Through confrontation, an agent can verify information or even apologize for their actions. Lastly, internalized regret displays direct evidence that agents can internalize their experiences and act on them in a human-like way, such as through expressing remorse for their actions.

The social behaviours demonstrated by generative agents, such as deception, confrontation, and internalized regret, suggest a preliminary avenue for considering elements of a Theory of Mind (ToM) in LLM-based systems. While these behaviors do not represent genuine understanding or intentionality, they indicate a capacity to simulate human-like responses to social and informational dynamics. For example, internalized regret hints at a mechanism for contextual adaptation, which could be seen as a rudimentary step toward representing aspects of human mental states.

Introduction

Generative agents [3] are a design framework utilising generative artificial intelligence (GAI), such as large language models (LLMs), to emulate realistic human-like behaviour. Generative agents have the ability to operate independently and creatively make decisions to reach a goal with only simple suggestions injected at initialisation.

Modeling complex systems has been a historically difficult task. Systems with many independent and complex actors can produce unexpected dynamics and emergent behaviour that are intractable to predict. As such, many researchers have utilised agent-based models to evaluate the behaviours of complex systems. Agent-based modeling systems like NetLogo [4] and Swarm [1, 2] have revolutionised researchers' ability to perform these simulations. However, these tools are limited by human knowledge and the practicality of implementing complicated behaviours. While many systems can be modeled using simple agents with a fixed set of valid actions, actors like humans, viruses, financial markets, and others often greatly exceed the bounds of our knowledge and ability to implement all feasible behaviours and decisions. To this end, GAI may be leveraged to model complex systems.

One particularly significant application of in-

terest for generative agents is towards emulating (mis)information spread. Modelling information spread is particularly difficult on small scales where in-person word-of-mouth communication is common, such as at the individual or community level.

Through a series of controlled simulations, we identify key technical dynamics and emergent behaviours of generative agents. Our work suggests that generative agents demonstrate realistic conversational patterns while being robust to (mis)information spread without deliberate encouragement. Further, generative agents display novel emergent social behaviours, such as deception, confrontation, and internalized regret. Model-generated hallucinations run the risk of harming simulation realism, but may also confabulate explanations for logical gaps and oversights of the implementer, improving realism. Simultaneously, novel dynamics dubbed "contextual eavesdropping" and "behavioural poisoning" cause the simulation framework to unintentionally leak private information to an agent, or significantly alter an agent's behaviour, respectively. Our code is available here: `https://github.com/nyoma-diamond/evaluating_generative_agents` and Supplementary Material for this work can be accessed at `https://osf.io/dy2u4`.

Discussion

Generative agents are vulnerable to hallucinations, leakage, and poisoning

Our experiments highlighted critical technical dynamics and phenomena induced by the framework's design and underlying model. These included the well-known anomaly of hallucination, and novel dynamics we dub "contextual eavesdropping" and "behavioural poisoning". Hallucinations induce notable inaccuracies which may result in unrealistic behaviour; however, some hallucinations, or confabulations, can be beneficial by filling logical gaps, thereby enhancing the realism of simulations by resolving discontinuities and unintended omissions. Contextual eavesdropping occurs when the framework unintentionally leaks information to an agent during interactions.

Generative agents display significant realistic emergent social behaviours

We observed a series of emergent social behaviours presented by agents in our simulations. Specifically, generative agents exhibited behaviours such as deception, confrontation, and internalised regret. These novel behaviours enhance the realism of our simulations and highlight significant variables within the underlying generative model that may strongly impact agent behaviour and realism. Through deception, agents could avoid conversations much like a human might. Through confrontation, a rumourmonger attempts to verify the contents of a rumour or apologise for their actions. Finally, through internalised regret, we see that agents can internalise their experiences and act on them in a human-like way, such as through expressing remorse for their actions.

Concluding remarks

The behaviors exhibited by generative agents, including deception, confrontation, and internalized regret, provide an initial framework for exploring aspects of a Theory of Mind (ToM) in LLM-based systems. Although these behaviors do not equate to genuine understanding or intentionality, they highlight the system's ability to mimic human-like responses to social and informational contexts. For instance, the expression of internalized regret demonstrates a capacity for contextual adaptation, which could be considered a rudimentary step toward representing elements of human mental states in a purely computational manner.

References

[1] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593, September 2014. Publisher: PeerJ Inc.

[2] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420, December 2015. Publisher: PeerJ Inc.

[3] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 1–22, New York, NY, USA, October 2023. Association for Computing Machinery.

[4] Uri Wilensky. NetLogo, 1999.

I Know What You Did Last Summer (and I Can Predict What You're Trying to do Now): Incorporating Theory of Mind into Multi-agent Reinforcement Learning

Reuth Mirsky[1], Matthew E. Taylor[2], and William Yeoh[3]

[1]Tufts University, MA, USA , reuth.mirsky@tufts.edu
[2]University of Alberta, AB, Canada , matthew.e.taylor@ualberta.ca
[3]Washington University in St. Louis, MO, USA , wyeoh@wustl.edu

Abstract

Most Multi Agent Reinforcement Learning (MARL) approaches assume agents do not explicitly represent others, instead bundling them with the environment. This simplifies learning by treating agents as independent, given the environment. We propose a Theory of Mind (ToM)-inspired approach, where agents infer teammates' goals to enhance collaboration. By explicitly modeling these goals, agents can learn more complementary policies, improving coordination beyond traditional MARL methods.

Introduction

In reinforcement learning (RL), agents interact with an environment to maximize expected cumulative rewards, typically using a discounting factor. Multi-agent RL (MARL) extends these principles to settings involving multiple agents [1]. Despite the obvious interdependencies that exist between the agents, existing work in MARL typically assumes that each agent does not maintain an explicit representation of the other agents. Instead, each agent implicitly captures these agents by bundling them together with the environment. This assumption significantly simplifies the learning process since it essentially decouples the learning process of the agents as they are now independent of each other given the environment. In this paper, we propose an alternative approach inspired by ToM [10, 16], where agents are equipped with the capability of inferring the knowledge or goals of their teammates to improve collaborative performance. We hypothesize that, by explicitly representing the teammates' goals, each agent can learn policies that better complement their teammates than existing MARL approaches. This paper explores the implications of incorporating ToM for cooperative behaviors in MARL environments using a case study of a known MARL problem where agents are 1) fully cooperative, 2) lack explicit communication capabilities, and 3) operate in a fully observable shared state environment.

Preliminaries

A single-agent sequential decision process is modeled as a Markov decision process (MDP) consisting of $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$. At each interaction step t, the agent receives a state $s_t \in \mathcal{S}$ and takes

an action $a_t \in \mathcal{A}$ according to its policy $\pi(s|a)$. The environment provides a reward r_t and transitions to state s_{t+1} according to the function $T(s_{t+1}|s_t, a_t)$. The agent aims to learn a policy that maximizes the return $G = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, with a discount factor $\gamma \in [0, 1)$.

A multi-agent sequential decision process in RL is often modeled as a multi-agent MDP (i.e., a Markov game) for n agents as the tuple $\langle S, A_1, \ldots, A_n, T, R, \gamma \rangle$. S is the set of all possible joint environment states; $A_1, \ldots, A_n$ is the set of actions available to each agent; T is the state transition function based on the joint state and the agents' actions, $T(s_{t+1}|s_t, a_{1,t} \times \cdots \times a_{n,t})$; the reward is a team reward, $R : S \times A_1, \times \cdots \times A_n \rightarrow \Re$; and γ is the discount factor. This formulation suits the set of assumptions we presented earlier and also means that, for now, all agents share a single model of the environment, they all have the same action set, and they are fully cooperative as they share the same goal. We consider two classes of MARL algorithms or settings: with and without Theory of Mind (ToM).

Without Theory of Mind Many approaches use independent Q-learning [28], where every agent simply treats all other agents as part of the environment. Each agent selects its own action based on the state. All other agents are considered part of the state. Learned information is not shared between the agents. In centralized training, decentralized execution (CTDE), agents act in isolation, but then can communicate between episodes (for example, see QMIX [23] and MADDPG [18]). This allows them to share their updates (e.g., pool their knowledge), allowing for faster learning, and every agent has an identical policy.

Model-based RL can be useful when the transition and reward functions are known or can be learned (either approximately or perfectly, such as in R-max [4]). Given full observability and shared reward, this model of T and R will be the same for all agents. Using no additional data, agents can calculate the same joint policy via planning (e.g., dynamic programming), even without explicit communication. In all of the above cases, if the state and action space are discrete, a tabular learning approach may be feasible. If the state or action spaces are continuous or the task is sufficiently complex, function approximation may be required.

With Theory of Mind In MARL, a common assumption is that multiple agents train jointly to maximize their cumulative rewards. Ad-hoc teamwork is a closely related problem where we can directly control only a single agent while teammates may have different capabilities and learning abilities [20].

Learning with Opponent-Learning Awareness (LOLA) is a method in which an agent shapes the expected learning process of the other agents in the environment [8]. LOLA and its predecessors explicitly represent the impact of one agent's policy on the anticipated policy update of the other agents [31, 29]. Naturally, to be able to represent other agents' policy updates, each agent must have some representation of the others' policy, though implicit. Also, as its name implies, this work on opponent shaping does not explicitly focus on collaborative aspects of the multi-agent setting.

Social influence is another MARL mechanism for achieving coordination and communication between teammates by rewarding agents for having causal influence over other agents' actions [13]. An agent's reasoning about social influence also identifies where it influences other agents, but it may not have an explicit model of these agents' understanding or policy.

All of the above approaches eventually aim to learn the **ego agent's** policy, so the policies of the **teammates** are represented only implicitly via the ego agent's policy. In this work, we argue that an explicit representation of the other agents' goals can benefit the ego agent's decision-making.

Adding Explicit Structured Representations

Assuming that the actions of other agents are part of the environment makes the learning process more straightforward and easily generalizable to various application domains. However, this approach disregards a significant body of literature on symbolic multi-agent coordination and its findings. For example, within the BDI community, there is a plethora of work on how multi-agent coordination can be improved by accounting for the belief, desire, and intention of other agents [7, 9, 11, 12, 19, 24]. Therefore, in this paper, we propose to investigate how such structured and symbolic representations can be used to improve MARL.

One crucial point is that by incorporating other agents' goals into the state, we do not discard important information, as this goal can change according to the problem dynamics (e.g., an agent's goal is not a rock rolling down a hill, but can change given specific state changes or agents' actions).

Case Study: Incorporating Goal Recognition in Two-Player Leader-Follower MARL As a case study, we consider a simple two-player leader-follower MARL problem, motivated by Hungry Thirsty [25]. In this problem, a "leader" agent is either "hungry" or "thirsty," in which case it wants to reach the cell with food or water, respectively. There is also a "follower" agent that can observe the leader and aims to assist by bringing it food or water to reduce the number of actions the leader needs to reach its goal.

This relatively simplistic example illustrates that if the follower can recognize the goal of the leader, it will be able to learn the policy to help it more quickly. Therefore, we propose to decompose the learning problem into two subproblems: The follower agent will (1) predict the goal of the leader agent based on observations of the leader's actions and (2) use the predicted goal of the leader as an explicit input to its policy learning problem. If successful, we expect explicitly incorporating this goal will significantly improve learning speeds, relative to other existing MARL methods.

Goal Recognition Approaches There are several off-the-shelf goal recognition approaches [27] that can be used for the first subproblem, including approaches based on automated planning techniques [22, 21, 26], learning [2, 6, 14], and hybrid approaches that combines both approaches [3, 5].

Possible Further Extensions for Future Work While the case study above is an appropriate first step to test the feasibility of our hypothesis, a more thorough investigation is needed for more complex and realistic settings. For example, it would be interesting to consider more realistic behaviors for the leader agent. In the example above, we assume that the leader ignores the follower. In reality, the leader agent should be cognizant of the follower agent and may want to learn policies that more easily signal its goal to the follower. Such policies are called explicable policies [30, 15], and there is a trade-off between explicability and cost-effectiveness of a policy; explicable policies may incur higher costs, especially if the follower agent assumes the wrong goal of the leader.

Another direction is when the leader agent is human. Humans and AI agents can have fairly different behaviors (e.g., humans have cognitive biases and bounded rationality [17]), and it's not clear if and how existing goal recognition approaches would work to recognize human goals.

References

[1] Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[2] Leonardo Amado, Reuth Mirsky, and Felipe Meneguzzi. Goal recognition as reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 9644–9651, 2022.

[3] Leonardo Amado, Ramon Fraga Pereira, and Felipe Meneguzzi. Robust neuro-symbolic goal and plan recognition. In *AAAI Conference on Artificial Intelligence*, pages 11937–11944, 2023.

[4] Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, March 2003.

[5] Mattia Chiari, Alfonso Emilio Gerevini, Andrea Loreggia, Luca Putelli, and Ivan Serina. Fast and slow goal recognition. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 354–362, 2024.

[6] Mattia Chiari, Alfonso Emilio Gerevini, Francesco Percassi, Luca Putelli, Ivan Serina, and Matteo Olivato. Goal recognition as a deep learning task: The grnet approach. In *International Conference on Automated Planning and Scheduling (ICAPS)*, pages 560–568, 2023.

[7] Philip R Cohen and Hector J Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261, 1990.

[8] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 122–130, 2018.

[9] Richard Freedman and Shlomo Zilberstein. Integration of planning with recognition for responsive interaction using classical planners. In *AAAI Conference on Artificial Intelligence*, 2017.

[10] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *Intelligent Workshop on Agents Theories, Architectures, and Languages (ATAL)*, pages 1–10, 1999.

[11] Piotr J Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[12] Barbara J Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[13] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049, 2019.

[14] Robert Kasumba, Guanghui Yu, Chien-Ju Ho, Sarah Keren, and William Yeoh. Data-driven goal recognition design for general behavioral agents. *arXiv preprint arXiv:2404.03054*, 2024.

[15] Anagha Kulkarni, Yantian Zha, Tathagata Chakraborti, Satya Gautam Vadlamudi, Yu Zhang, and Subbarao Kambhampati. Explicable planning as minimizing distance from expected behavior. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2075–2077, 2019.

[16] Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in Artificial Intelligence*, 5:778852, 2022.

[17] David Lindner and Mennatallah El-Assady. Humans are not Boltzmann distributions:

Challenges and opportunities for modelling human feedback and interaction in reinforcement learning. *arXiv preprint arXiv:2206.13316*, 2022.

[18] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[19] William Macke, Reuth Mirsky, and Peter Stone. Expected value of communication for planning in ad hoc teamwork. In *AAAI Conference on Artificial Intelligence*, pages 11290–11298, 2021.

[20] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. A survey of ad hoc teamwork: Definitions, methods, and open problems. In *European Conference on Multiagent Systems*, pages 1–8, 2022.

[21] Ramon Fraga Pereira, Nir Oren, and Felipe Meneguzzi. Landmark-based heuristics for goal recognition. In *AAAI Conference on Artificial Intelligence*, pages 3622–3628, 2017.

[22] Miguel Ramírez and Hector Geffner. Probabilistic plan recognition using off-the-shelf classical planners. In *AAAI Conference on Artificial Intelligence*, 2010.

[23] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[24] Maayan Shvo and Sheila A McIlraith. Active goal recognition. In *AAAI Conference on Artificial Intelligence*, pages 9957–9966, 2020.

[25] Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from. In *Annual Conference of the Cognitive Science Society*, pages 2601–2606, 2009.

[26] Tran Son, Orkunt Sabuncu, Christian Schulz-Hanke, Torsten Schaub, and William Yeoh. Solving goal recognition design using asp. In *AAAI Conference on Artificial Intelligence*, 2016.

[27] Gita Sukthankar, Christopher Geib, Hung Hai Bui, David Pynadath, and Robert P Goldman. *Plan, Activity, and Intent Recognition: Theory and Practice*. Newnes, 2014.

[28] Ming Tan. Multi-agent reinforcement learning: independent versus cooperative agents. In *International Conference on Machine Learning*, page 330–337, 1993.

[29] Timon Willi, Alistair Hp Letcher, Johannes Treutlein, and Jakob Foerster. Cola: consistent learning with opponent-learning awareness. In *International Conference on Machine Learning*, pages 23804–23831, 2022.

[30] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *International Conference on Robotics and Automation*, pages 1313–1320, 2017.

[31] Stephen Zhao, Chris Lu, Roger B Grosse, and Jakob Foerster. Proximal learning with opponent-learning awareness. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:26324–26336, 2022.

MAPS - A Metacognitive Architecture for Improved Social Learning

Juan David Vargas[1,2,3], Natalie Kastel[1,3,6], Antoine Pasquali[4], Axel Cleeremans[5], Zahra Sheikhbahaee[1,3,*], Guillaume Dumas[1,3,6,* 1]

[1]Precision Psychiatry and Social Physiology (PPSP) laboratory CHU Sainte-Justine, Montréal, Québec, Canada
[2]DIRO, Université de Montréal, Montréal, Québec, Canada
[3]MILA - Quebec AI Institute, Montréal, Québec, Canada
[4]Université libre de Bruxelles, Bruxelles, Belgium
[5]CrossLabs, Tokyo, Japan
[6]Département de Psychiatrie, Université de Montréal, Montréal, Québec, Canada
[*]Co-Senior Authors

Abstract

Theory of Mind (ToM) and metacognition are essential for social intelligence but remain underexplored in AI beyond basic pattern recognition tasks. This paper introduces MAPS (Metacognitive Architecture for Perceptual and Social learning), which integrates a second-order network (2nd-Net) with cascaded activation to support reflective processing across domains.

We evaluate MAPS in pattern recognition (Blindsight, AGL), single-agent reinforcement learning (SARL; MinAtar) and multi-agent reinforcement learning (MARL; MeltingPot 2.0). MAPS consistently improves performance in Pattern Recognition (PR) and SARL, particularly in complex environments, and shows promising results in high-variability MARL tasks. These findings demonstrate the potential of metacognitive architectures to improve learning and social adaptability in AI systems (AIS).

1 Introduction

In cognitive science, Theory of Mind (ToM) refers to the ability to attribute beliefs, desires, and intentions to others in order to predict their behavior. In AI, ToM represent transformative shift—enabling systems that go beyond mechanistic responses and interact with humans in socially intelligent ways[18, 17]. Closely related is is metacognition—the capacity to monitor and regulate one's cognitive processes. While both involve meta-representations, ToM centers on understanding others' minds, whereas metacognition involves higher-order reasoning about one's mental states.

Neurocognitive research shows that metacognition and Theory of Mind (ToM) share neural and cognitive foundations, with metacognition enhancing ToM and supporting better social outcomes[16, 10, 3]. Theories on social cognition [8], suggest this connection may arise from the brain's capacity to model its own internal states, which could form the

basis for understanding the minds of others.

Building on this connection, AI increasingly incorporates metacognition to enhance artificial social cognition. By combining self-monitoring with social reasoning, metacognitive architectures support flexible learning and strengthen ToM capacities—enabling AIS to engage in more humanlike interactions[6, 21, 4].

One approach to embedding metacognition in AIS is through a second-order network (2nd-Net), which pairs a primary task network with a secondary system that monitors and evaluates its performance. This layer assesses confidence, identifies knowledge gaps, and initiates adjustments to optimize decision-making [14].

Although metacognition is theorized to support ToM in AIS, current methods focus on low-level tasks like pattern recognition (PR), missing its potential in modeling complex interactions [9]. Reinforcement Learning (RL) offers a promising alternative, engaging agents in dynamic, prosocial settings [12].

To bridge this gap, we test whether a 2nd-Net improves AI performance beyond PR tasks, extending to single- and multi-agent reinforcement learning (SARL and MARL). We introduce MAPS (Metacognitive Architecture for Perceptual and Social learning), a streamlined adaptation of Pasquali & Cleeremans' 2nd-Net [1], designed to integrate metacognition across both PR and social learning domains.

PR is tested with 'Blindsight' and 'AGL' [1], SARL with MinAtar games[19], and MARL with 4 MeltingPot 2.0 settings—3 with the lowest and 1 with the highest coefficient of variation (CV) from Agapiou[2]:1). These experiments assess whether MAPS can enhance both PR learning and socially intelligent behavior in AIS.

2 Methodology

For PR tasks, we used an auto-encoder for the main task, and a comparator matrix connected to two wagering units for the 2nd-Net, as in [1]. . We used a contrastive loss for the main task, which provided crucial information flow for wagering [5]. For wagering, we used a cross-entropy loss to handle class imbalance. Both 1st and 2nd-Net implemented a cascade model that facilitated a smooth graded accumulation of activation [11]. We empirically chose 50 cascade iterations (except for MARL, given computation constraints). For SARL, we used a DQN framework [15]. We applied convolutional layers, which allowed for reduced computational complexity, an autoencoder, and a replay buffer for learning stability. We then calculated the comparison matrix using the inputs and outputs of the value network's auto-encoder, and connected this to 2 wagering units. For the wagering objective, we calculated the rewards in batches of 128 using an EMA with a smoothing factor of $\alpha = 0.45$. At each step t, a low/high wager was assigned based on whether the last reward was greater than EMA. For MARL, $\alpha = 0.25$. Both were found empirically. For MARL, we used an MAPPO framework[20], convolutional layers, sinusoidal-based relative positional encoding to add positional information, and a Gated Recurrent Unit (GRU) for stability.

3 Results

For Blindsight, suprathreshold patterns were used during training, and 3 types were used for testing. To prevent overfitting, new patterns were generated per epoch. Table 1 compares the proposed model with variants turning the 2nd-Net and/or the cascade model on/off . We observed a performance gain using a 2nd-Net and cascade model, achieving statistical significance compared to the baseline (Z-score: 8.6, 450 seeds). We also observed that gains are mostly driven by the cascade model. For AGL, we pre-trained the model, saved the weights of the 2nd-Net, and disabled backpropagation through it during training. Random strings were used for pre-training, grammar A for training, and a mix of grammar A and grammar B for testing. Grammar strings are defined as per

[13], and we followed the data proportions in [1]. We employed a low training scheme (3 epochs). Results show a statistical significance (Z-score: 15.0 - MAPS, and 4.2 - 2nd-Net).

2nd Net	Cascade	Accuracy	Z-score (Significant)
No	No	0.95 ± 0.03	
No	1st Net	0.97 ± 0.02	**8.50 (Yes)**
Yes	No	0.96 ± 0.03	0.77 (No)
Yes	1st Net	0.97 ± 0.02	**9.01 (Yes)**
Yes	Both	0.97 ± 0.02	**8.6 (Yes)**
No	No	0.54 ± 0.08	
No	1st Net	0.61 ± 0.07	**13.3 (Yes)**
Yes	No	0.57 ± 0.07	**4.2 (Yes)**
Yes	1st Net	0.62 ± 0.07	**15.7 (Yes)**
Yes	Both	0.62 ± 0.06	**15.0 (Yes)**

Table 1: Accuracy for Blindisght (top) and AGL (bottom). Chance level: 0.01 and 0.15.

In MinAtar (table 2), we tested "Seaquest" and "Asterix" for 3 seeds (1 million steps). We show an improvement with MAPS (Z-score: 2.97 and 2.15). For Seaquest, the setting with the most obstacles, we observed that it is when both the cascade model and 2nd-Net are active, that effective learning occurs. In MARL (Table 3, 1.5 million steps), both GRU-only and a 2nd-Net variant were tested. While the 2nd-Net model is slightly superior to GRU, it still lags behind the top model (ACB) presented in [2]. Conversely, for territory inside out, we noted a tendency of MAPS to produce positive outliers (see Appendix D.2), and, over 10 seeds, we found a positive Z-score of 2.59 with respect to the baseline.

2nd Net	Cascade	Rewards	Z-score (Sig.)
No	No	1.21 ± 0.16	
No	1st Net	0.76 ± 0.19	**-2.59 (Yes)**
Yes	No	0.97 ± 0.61	-0.53 (No)
Yes	1st Net	3.06 ± 0.34	**7.03 (Yes)**
Yes	Both	6.15 ± 2.33	**2.97 (Yes)**
No	No	2.49 ± 1.94	
No	1st Net	1.59 ± 0.94	-0.59 (No)
Yes	No	5.48 ± 1.30	1.81 (No)
Yes	1st Net	4.54 ± 1.01	1.32 (No)
Yes	Both	5.77 ± 0.94	**2.15 (Yes)**

Table 2: Validation rewards: Seaquest (top) and Asterix (bottom). Chance level: 0.09 and 0.47.

Environment	GRU	GRU (2nd-Net)	ACB
Harvest C.	18.9 ± 1.4	20.6 ± 2.1	32.8 ± 10.6
Harvest P.	28.1 ± 1.9	28.7 ± 3.8	31.9 ± 11.0
Chem. 3D.	$\mathbf{1.2 \pm 0.1}$	$\mathbf{1.2 \pm 0.1}$	1.1 ± 0.8
Terr. I.O.	63.5 ± 8.7	76.5 ± 8.3	80.3 ± 48.0

Table 3: Training rewards in MARL.

4 Conclusion

This study demonstrates that the MAPS architecture enhances learning across PR, SARL, and MARL tasks. In PR and SARL, combining the cascade model and 2nd-Net consistently improved performance, particularly in complex environments. Even without backpropagation through the 2nd-Net, MAPS maintained strong results in AGL, highlighting its robustness.

In MARL, while MAPS did not outperform the top benchmark, it matched or exceeded baselines in most cases and showed strong performance in high-variability scenarios like “Territory Inside Out.” These findings show the value of metacognitive architectures for building more adaptive and socially aware AI systems.

5 Acknowledgements

This study was supported by the Institute for Advanced Consciousness Studies (IACS), the Institute for Data Valorization, Montreal (IVADO; CF00137433 & PRF3), and the Canadian Institute for Advanced Research (CIFAR). Computation was enabled by Calcul Québec (www.calculquebec.ca) and Digital Research Alliance of Canada (www.alliancecan.ca). JDV would like to thank the UNIQUE center for travel funding to attend the AAAI conference. GD was supported by the Fonds de recherche du Québec - Santé (FRQ-S; 2024-2025 - CB - 350516), Natural Sciences and Engineering Research Council of Canada (NSERC; DGECR-2023-00089), the Brain Canada Foundation (2022 Future Leaders in Brain Research).

References

[1] B. Timmermans A. Pasquali and A. Cleeremans. Know thyself: Metacognitive networks and measures of consciousness. *Cognition*, 117:182–190, 2010.

[2] J.P. Agapiou, A.S. Vezhnevets, E.A. Duéñez-Guzmán, J. Matyas, Y. Mao, P. Sunehag, R. Köster, U. Madhushani, K. Kopparapu, R. Comanescu, D.J. Strouse, M.B. Johanson, S. Singh, J. Haas, I. Mordatch, D. Mobbs, and J.Z. Leibo. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2023.

[3] Federica Bianco and Ilaria Castelli. The promotion of mature theory of mind skills in educational settings: a mini-review. *Frontiers in Psychology*, 14:1197328, 2023.

[4] S. Bolotta and G. Dumas. Social neuro ai: Social interaction as the 'dark matter' of ai. *Frontiers in Computer Science*, 4, 2022.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. ICML 2020.

[6] Brendan Conway-Smith and Robert L. West. Toward autonomy: Metacognitive learning for enhanced ai performance. *Proceedings of the AAAI 2024 Spring Symposium Series: Symposium on Human-Like Learning*, 2024.

[7] Zoltan Dienes, Gerry Altmann, Lisa Kwan, and Andrew Goode. Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21(5):1322–1338, 1995.

[8] Michael SA Graziano. *Consciousness and the social brain*. Oxford University Press, 2013.

[9] Ryota Kanai, Ryota Takatsuki, and Ippei Fujisawa. Meta-representations as representations of processes. *PsyArXiv*, 2024. Preprint.

[10] Emily L Long, Caroline Catmur, Stephen M Fleming, and Geoffrey Bird. Metacognition facilitates theory of mind through optimal weighting of trait inferences. *Cognition*, 256:106042, 2025.

[11] James L. McClelland, David E. Rumelhart, Jerome Feldman, and Patrick Hayes. *Explorations in Parallel Distributed Processing - Macintosh version: A Handbook of Models, Programs, and Exercises*. Bradford Books. The MIT Press, 1989. Includes 2 diskettes for the Macintosh, In Special Collection: CogNet.

[12] Kamal K Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. Emergent social learning via multi-agent reinforcement learning. In *International conference on machine learning*, pages 7991–8004. PMLR, 2021.

[13] N. Persaud, P. McLeod, and A. Cowey. Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 10:257–261, 2007.

[14] Kristian Sandberg, Bert Timmermans, Morten Overgaard, and Axel Cleeremans. Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, 2010.

[15] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *arXiv preprint arXiv:1509.06461*, 2015. AAAI 2016.

[16] Bamicha Victoria and Drigas Athanasios. Theory of mind in relation to metacognition and icts. a metacognitive approach to tom. *Scientific Electronic Archives*, 16(4), 2023.

[17] Jessica Williams, Stephen M Fiore, and Florian Jentsch. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763, 2022.

[18] Alan FT Winfield. Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI*, 5:357467, 2018.

[19] Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.

[20] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu. The surprising effectiveness of ppo in cooperative multi-agent games. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[21] E. Zaroukian. Theory of mind and metareasoning for artificial intelligence: A review. 2022. Submitted for review.

A Appendix / supplemental material

Appendix A - Additional Environment details

Appendix A.1 - Blindsight task

Blindsight is a neurological phenomenon where individuals with damage to their primary visual cortex can still respond to visual stimuli without consciously perceiving them.

To study this, we use a simulated dataset that mimics the conditions of blindsight according to [1]. This dataset contains 400 patterns, equally split between two types:

- **Random noise patterns:** These consist of low activations ranging between 0.0 and 0.02.
- **Designed stimulus patterns:** Each pattern includes one unit that shows a higher activation level, varying between 0.0 and 1.0.

This dataset allows us to test hypotheses concerning how sensory processing and network responses adapt under different conditions of visual impairment.

We have three main testing scenarios, each designed to alter the signal-to-noise ratio to simulate different levels of visual impairment:

- **Suprathreshold stimulus condition:** Here, the network is tested against familiar patterns used during training to assess its response to known stimuli.
- **Subthreshold stimulus condition:** This condition slightly increases the noise level,

akin to actual blindsight conditions, testing the network's capability to discern subtle signals.

- **Low vision condition:** The intensity of stimuli is decreased to evaluate how well the network performs with significantly reduced sensory input.

Appendix A.2 - Artificial Grammar Learning Task

In the AGL experiment, Persaud et al. [13] demonstrate that participants exposed incidentally to letter strings generated by an artificial grammar perform better than chance on a subsequent, unexpected test where they distinguish between new grammatical and non-grammatical strings. However, they fail to optimize their earnings through wagering. Once participants were informed about the grammar rules, they began to place advantageous wagers (explicit condition) [1].

To simulate this, we utilize artificially generated strings ranging from 3 to 8 letters, classified into three types: randomly generated, grammar A, and grammar B, as defined by Persaud et al.

During training, the networks are exposed to two conditions: explicit and implicit, reflecting the results of implicit learning [7]. For the implicit condition (low consciousness), networks are trained for 3 epochs, while for the explicit condition (high consciousness), they are trained for 12 epochs.

Appendix A.3 - MinAtar

MinAtar provides simplified versions of classic Atari 2600 games, designed specifically for AI agent testing and development. MinAtar offers more accessible and computationally efficient environments for AI research and experimentation [19]. There are 5 Atari games implemented:

- **Space Invaders:** The player controls a cannon to shoot at aliens that move across and down the screen, with each destroyed alien providing +1 reward and causing the remaining aliens to speed up. Aliens also shoot back at the player, new waves spawn at increased speeds after clearing a wave, and termination occurs when the player is hit by an alien or bullet [19].
- **Breakout:** The player controls a paddle at the bottom of the screen to bounce a diagonally-traveling ball toward three rows of bricks at the top, earning +1 reward for each brick broken and getting new rows when all are cleared. The ball's direction changes based on which side of the paddle it hits or when it contacts walls and bricks, with game termination occurring when the ball reaches the bottom of the screen [19].
- **Seaquest:** The player controls a submarine that can fire bullets at enemy submarines and fish, earning +1 reward for each hit while also rescuing divers to fill a progress bar and maintaining oxygen that depletes over time. Oxygen replenishes when surfacing with at least one rescued diver, surfacing with six divers provides additional rewards based on remaining oxygen, and the game ends when hit by enemies, running out of oxygen, or surfacing without divers [19].
- **Asterix:** The player moves freely in four cardinal directions to collect treasure while avoiding enemies that spawn from the sides, with each treasure providing a +1 reward and enemy contact causing termination. Enemy and treasure movements are indicated by trail channels, and the game's difficulty increases periodically by enhancing the speed and spawn rate of both enemies and treasures [19].
- **Freeway:** The player moves vertically up and down at a restricted pace (once every 3 frames) to cross a road filled with

horizontally-moving cars, earning +1 reward upon reaching the top before being returned to the bottom. When hit by a car, the player returns to the bottom without penalty, car speeds randomize after each successful crossing, and the game terminates after 2500 frames have elapsed [19].

Appendix A.4 - Meltingpot

The Melting Pot Suite provides a comprehensive framework for generating test scenarios that assess an agent population's ability to generalize cooperative behavior in new situations. It offers up to 50 distinct training and testing environments. The test scenarios combine novel background populations of agents and include a variety of substrates, such as classic social dilemmas like the Prisoner's Dilemma, as well as complex mixed-motive coordination games. In our experiments, we selected four environments based on the coefficient of variation among the models tested in [2]. This value was calculated for the 37 non-zero-sum environments out of the 50 available (see Figure 1). We chose the three environments with the lowest variability and the environment with the highest positive variability.

Our tested environments are: Commons Harvest Closed, Commons Harvest Partnership, Chemistry Three Metabolic Cycles with Plentiful Distractors, and Territory Inside Out. A short description is provided below:

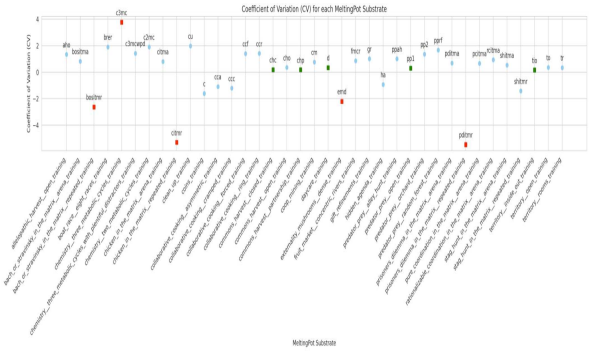


Figure 1: Variability among Melting Pot environments according to the experimentation in [2].

- **Commons Harvest Closed:** Apples are dispersed and can be consumed by agents. Additionally, apples have a probability at every step to regrow, which depends on the number of nearby apples: 0.0025 when there are three or more apples, 0.005 for two, 0.001 if there is one, and 0 otherwise. Thus, agents need to exercise restraint in consuming all apples in a batch to ensure the long-term regrowth of apples. Even though it is not beneficial to consume the last apple, agents are incentivized to do so to prevent other agents from consuming it. In this closed variant, there are rooms full of apples, promoting agents to defend them and minimize the probability of other agents harvesting the full patch of apples [2].
- **Commons Harvest Partnership:** Similar to the Commons Harvest Closed environment, this variant still has rooms filled with apples. However, it requires two agents to protect a room, thus promoting the development of cooperative behavior and a mutually sustainable situation[2].
- **Chemistry Three Metabolic Cycles with Plentiful Distractors:** In this setting, a set of agents work to generate mutual benefits from metabolic reactions defined by a predefined graph. These reactions occur stochastically when reactants are in close proximity to one another. Agents can carry molecules and are rewarded when the molecule in their inventory is part of a reaction, either as a reactant or a product. In the three metabolic cycles variant, agents benefit from three different cycles, which continue as long as the minimum energy requirements are fulfilled. Agents must

learn to facilitate the right reactions to generate enough energy to sustain the cycles. The environment also contains distractors, which are molecules that do not provoke reactions but provide a small constant reward to encourage agents to pursue less rewarding strategies[2].

- **Territory Inside Out:** Each agent is assigned a unique color and seeks to claim territory by painting walls in that color. Wet paint does not yield rewards. After 25 steps following the application of paint, if no further paint has been added, the paint dries and turns into a brighter shade of the agent's color. Once dry, the painted wall rewards the claiming player at a consistent rate. The more walls a player claims, the higher their expected rewards per timestep. In the Inside Out variant, agents are generated in a maze and must move inward toward the center of the map to claim territory. In this scenario, agents can zap each other, immobilizing the other agent for a set number of steps. An agent that is zapped twice is eliminated[2].

Appendix B - Hyperparameter choices and Computational resources

Appendix B.1 - Blindsight task

For the blindisight task, we used an Nvidia RTX3070 GPU for training, with 8GB of RAM. The training time was maximum for MAPS (2nd order network and cascade model in both 1st and 2nd order networks). For this setting, training over the 450 seeds took roughly 12 hours.

Hyperparameter	**Value**
Input size	100
Output size	100
Hidden size	60
lr first order	0.5
lr second order	0.1
Temperature	1.0
Step size	25
Gamma	0.98
Epochs number for training	200
Optimizer	$Adamax$
Cascade iterations	50

Table 4: Hyperparameters used for the Blindsight Task.

Appendix B.2 - Artificial Grammar Learning Task

bgFor the AGL task, we used an Nvidia RTX 3070 GPU for training, with 8GB of RAM. The training time was maximum for MAPS (2nd order network and cascade model in both 1st and 2nd order networks). For this setting, training over the 450 seeds took roughly 12 hours.

Hyperparameter	**Value**
Input size	48
Output size	48
Hidden size	40
lr first order	0.4
lr second order	0.1
Temperature	1.0
Step size	1
Gamma	0.999
Epochs number for pre-training	60
Epochs number for training(high consciousness)	12
Epochs number for training(low consciousness)	3
Optimizer	$RangerVA$
Cascade iterations	50

Table 5: Hyperparameters used for the Artificial Grammar Learning Task.

Appendix B.3 - MinAtar

For the MinAtar environments, we used a GPU V100 for training. The training time was maximum for MAPS (2nd order network and cascade model in both 1st and 2nd order networks). For this setting, training took roughly 6 days per million steps per seed.

Hyperparameter	Value
Batch size	128
Replay buffer size	$100,000$
Target network update frequency	$1,000$
Training frequency	1
Number of frames	$500,000$
First N frames	$100,000$
Replay start size	$5,000$
End epsilon	0.1
Step size	0.0003
Step size (second order)	0.0002
Gradient momentum	0.95
Squared gradient momentum	0.95
Minimum squared gradient	0.01
Gamma	0.999
Step Size	1
Epsilon	1.0
Alpha	0.45
Cascade iterations	50
Optimizer	*Adam*

Table 6: Hyperparameters used for the MinAtar experiments.

Appendix B.4 - Meltingpot

For the melting pot tasks, we used an Nvidia A100 GPU for training. The average training time was roughly 16 hours per seed(baseline, MAPS not implemented fully, only with simple 2nd order network with no cascade model due to limitations with computational resources). Every run required roughly 4-6 GB of RAM, mainly depending on the number of agents.

Hyperparameter	Value
Num agents (harvest closed)	6
Num agents (harvest partnership)	4
Num agents (chemistry)	8
Num agents (territory)	5
Hidden size	100
Actor lr	$7e-5$
Critic lr	100
Num env steps	$15e6$
Entropy coef	0.01
Clip param	0.2
Weight decay	$1e-5$
PPO epoch	15
Optimizer	*Adam*

Table 7: Common hyperparameters used for the Meltingpot environments.

Appendix C - Architectures

Appendix C.1 - Blindsight task and Artificial Grammar Learning Task

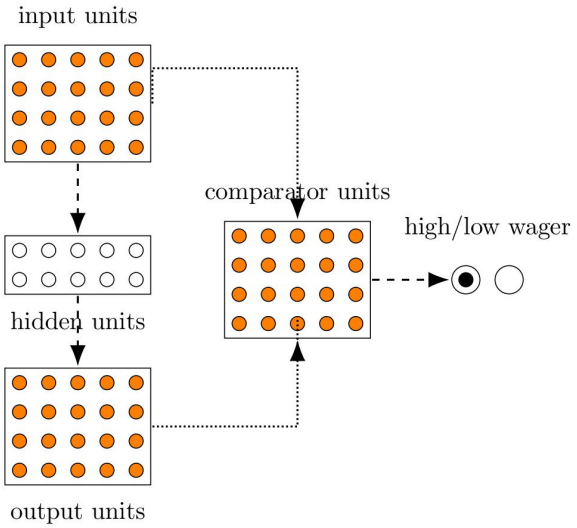


Figure 2: Illustration of the architecture used for both the Blindsight and Artificial Grammar Learning tasks.

Appendix C.2 - Meltingpot

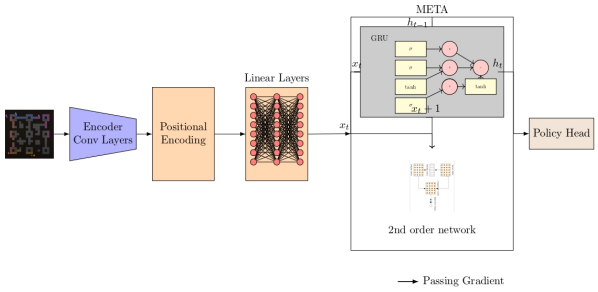


Figure 3: Illustration of the architecture used for all the Meltingpot environments

Appendix D - Additional results

Appendix D.1 - MinAtar

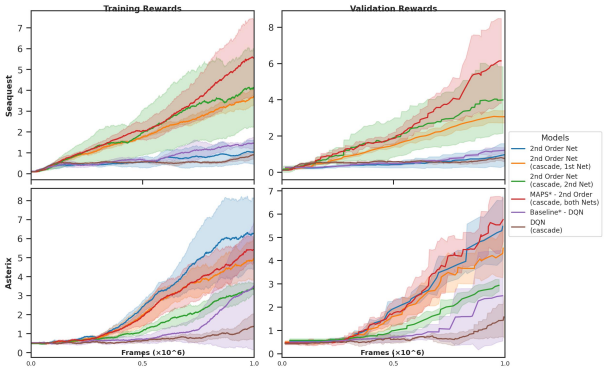


Figure 4: Training (left) and validation rewards (right) plots for SARL.

Appendix D.2 - Meltingpot

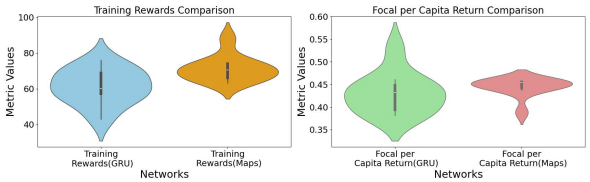


Figure 5: Territory Inside Out Results (10 seeds). Violin plot for avg. rewards (left); and Focal per Capita Return (right). Focal per capita return is a fairness measure (i.e. equal to 1.0 when all agents receive equal rewards), as defined by [2]

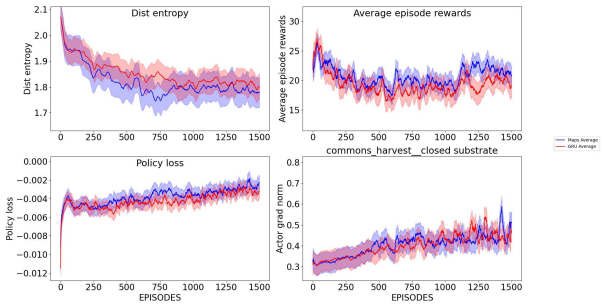


Figure 6: Results per episode over 1.5 million steps for commons harvest closed environment. To the top left the evaluation parameter is dist entropy, which represents the action distribution entropy, where a lower value points to a lower overall stochastic behaviour of the agents. The top right represents the average reward of all agents, where a higher value is desired. Bottom left is the policy loss, and bottom right is the actor gradient norm.

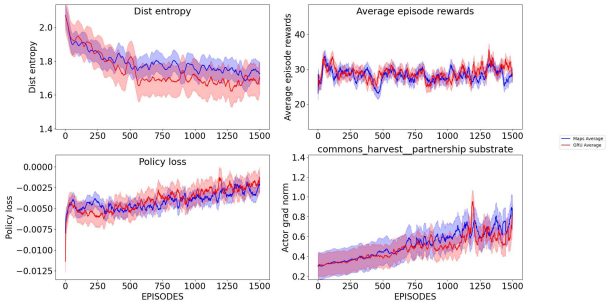


Figure 7: Results per episode over 1.5 million steps for commons harvest partnership environment. To the top left the evaluation parameter is dist entropy, which represents the action distribution entropy, where a lower value points to a lower overall stochastic behaviour of the agents. The top right represents the average reward of all agents, where a higher value is desired. Bottom left is the policy loss, and bottom right is the actor gradient norm.

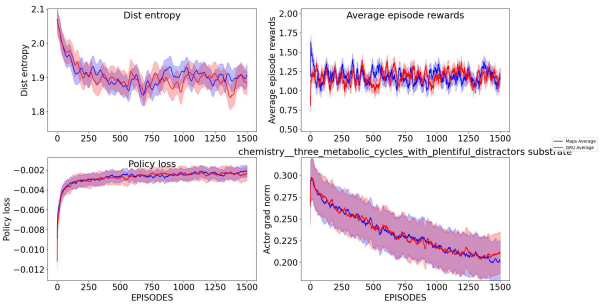


Figure 8: Results per episode over 1.5 million steps for chemistry environment. To the top left the evaluation parameter is dist entropy, which represents the action distribution entropy, where a lower value points to a lower overall stochastic behaviour of the agents. The top right represents the average reward of all agents, where a higher value is desired. Bottom left is the policy loss, and bottom right is the actor gradient norm.

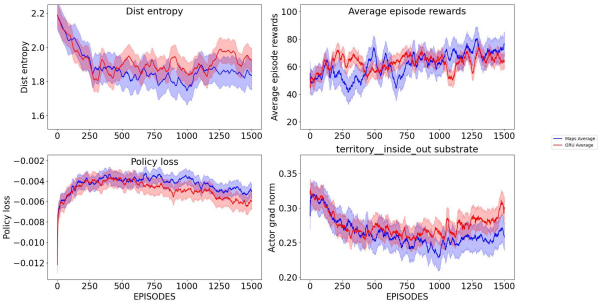


Figure 9: Results per episode over 1.5 million steps for territory inside out environment. To the top left the evaluation parameter is dist entropy, which represents the action distribution entropy, where a lower value points to a lower overall stochastic behaviour of the agents. The top right represents the average reward of all agents, where a higher value is desired. Bottom left is the policy loss, and bottom right is the actor gradient norm.

Rank-O-ToM: Unlocking Emotional Nuance Ranking to Enhance Affective Theory-of-Mind

JiHyun Kim, JuneHyoung Kwon, MiHyeon Kim, Eunju Lee, and YoungBin Kim

Chung-Ang University

Abstract

Facial Expression Recognition (FER) plays a foundational role in enabling AI systems to interpret emotional nuances, a critical aspect of affective Theory of Mind (ToM). However, existing models often struggle with poor calibration and limited capacity to capture emotional intensity and complexity. To address this, we propose ***Ranking the Emotional Nuance for Theory of Mind (Rank-O-ToM)***, a framework that leverages ordinal ranking to align confidence levels with emotional spectra. By incorporating synthetic samples reflecting diverse affective complexities, Rank-O-ToM enhances the nuanced understanding of emotions, advancing AI's ability to reason about affective states.

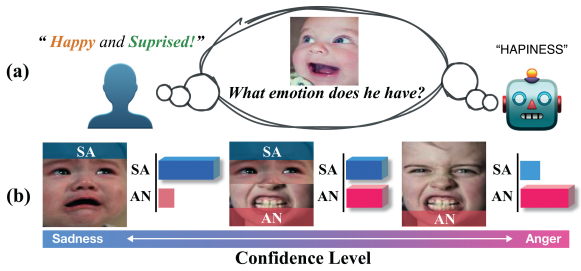


Figure 1: (a) Affective ToM challenge: interpreting nuanced emotional states, such as blended emotions. (b) Rank-O-ToM blends basic expressions into synthetic samples with ranked confidence scores to capture the emotional spectrum.

Introduction

Bridging the gap between AI and human understanding requires accurate emotional recognition. To foster trust and empathetic communication, AI models must recognize basic emotions and interpret nuances—a capability rooted in the affective Theory of Mind (ToM) [16, 15]. Facial Expression Recognition (FER) is crucial, as facial expressions universally convey emotions and intentions [1]. By interpreting these expressions, AI models can respond effectively to emotional cues in applications like compassionate healthcare [17] and adaptive education [12].

At the heart of affective ToM is the ability to understand emotions along with their intensity and complexity [6], as shown in Figure 1 (a). Humans often convey subtle and blended emotions, such as simultaneous happiness and surprise, with varying intensity [5], which helps infer deeper mental states. However, existing FER frameworks rely heavily on datasets with single basic emotion labels (e.g., happiness, anger) [10], limiting their ability to generalize and interpret emotional intensity and complexity. An advanced framework that can recognize nuanced emotional states and produce appropriate confidence scores (i.e., the model's certainty about its predictions for ba-

sic emotion categories) while reflecting varying intensity and complexity is urgently needed in the field of affective ToM.

Inspired by human cognition to interpret emotional nuances, we propose a novel FER framework, ***Ranking the Emotional Nuance for Theory of Mind (Rank-O-ToM)***, addressing affective granularity. Our method synthesizes diverse training samples by blending basic emotions to capture real-world affective complexity. To interpret these variations, the model employs a ranking mechanism aligning confidence levels with emotional intensity and clarity, enabling it to distinguish subtle and blended affective cues. Grounded in these principles, Rank-O-ToM enhances AI's capacity for nuanced affective reasoning, advancing affective ToM.

Method

AI models struggle to interpret complex affective states due to the limited diversity of basic emotions in FER datasets. To address this, we propose Rank-O-ToM, combining synthetic emotion blending and a ranking-based loss function, as shown in Figure 1 (b), to enhance emotional granularity.

Synthetic Sample Generation. Humans express emotions through subtle variations in facial regions (e.g., upper face for surprise, lower face for happiness)[11]. To capture this, we synthesize samples by horizontally blending images annotated with basic emotions, enriching the training data to reflect real-world affective diversity better. Additionally, we incorporate samples from face recognition (FR) datasets to further enhance diversity (details in Appendix).

Ranking Loss for Ordinal Relationships. We propose a ranking-based loss function aligning the model's confidence levels with the hierarchical structure of human perception for ordinal relationships between affective states [2, 4]. This ensures the model assigns higher confidence scores to original samples (representing clearer or more intense emotions) and lower scores to synthetic samples (representing blended or less intense emotions). The loss function is as follows:

$$\begin{aligned}\mathcal{L}_{rank} = max(0, \max_{c_1} p_{c_1}^{\text{syn}} - \max_{c_1} p_{c_1}^{\text{fer}} + \delta) \\ + max(0, \max_{c_2} p_{c_2}^{\text{syn}} - \max_{c_2} p_{c_2}^{\text{fr}} + \delta)\end{aligned} \quad (1)$$

Here, p_{c1}^{syn} and p_{c1}^{fer} are the confidence scores for synthetic and original samples in the basic emotion category c_1, respectively, and p_{c2}^{syn} and p_{c2}^{fer} are their counterparts for category c_2. By enforcing a meaningful separation with the margin δ, the loss ensures the model assigns higher confidence to clear, intense emotions in original samples while producing lower confidence for blended or less intense emotions in synthetic samples. This mechanism enables the model to capture subtle variations in affective states and maintain consistent confidence across diverse emotional expressions.

Experiment

To evaluate whether AI models can replicate humans' natural ability to assess emotions by accurately categorizing emotional states and gauging their intensity, we assess classification accuracy and confidence calibration metrics, including Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Adaptive ECE (AECE). These metrics align predicted probabilities with emotional intensity, reflecting the system's ability to interpret nuanced affective states. We compare a Rank-O-ToM with existing FER approaches on benchmarks, RAF-DB [10], FERPlus [3], and AffectNet [14] (details in Appendix).

Table 1 shows that Rank-O-ToM outperforms state-of-the-art FER methods on RAF-DB and FERPlus, with comparable results on AffectNet, demonstrating its ability to capture emotional categories and intensities. Additionally, the superior calibration performance underscores the effectiveness of our ranking-based loss in interpreting emotional granularity.

Method	RAF-DB				FERPlus				AffectNet			
	Acc ↑	*AECE* ↓	*MCE* ↓	*ECE* ↓	*Acc* ↑	*AECE* ↓	*MCE* ↓	*ECE* ↓	*Acc* ↑	*AECE* ↓	*MCE* ↓	*ECE* ↓
SCN [19]	88.72	5.51	18.18	5.46	79.32	15.79	34.94	15.84	48.31	21.20	13.56	21.21
EAC [23]	89.96	6.23	34.48	4.87	83.99	12.64	36.16	12.71	52.26	25.91	**0.98**	25.92
RAC [22]	92.11	25.4	27.4	25.39	85.92	19.16	24.50	19.73	54.14	**9.03**	11.39	9.20
Ours	**95.00**	**3.06**	**3.19**	**2.74**	**86.44**	**9.57**	**18.08**	**9.59**	**54.84**	10.28	10.47	**5.93**

Table 1: Comparison of Top-1 accuracy (Acc, %), AECE (%), MCE (%), and ECE (%) for FER methods on RAF-DB, FERPlus, and AffectNet. Top performances are bolded, and second-best results are underlined.

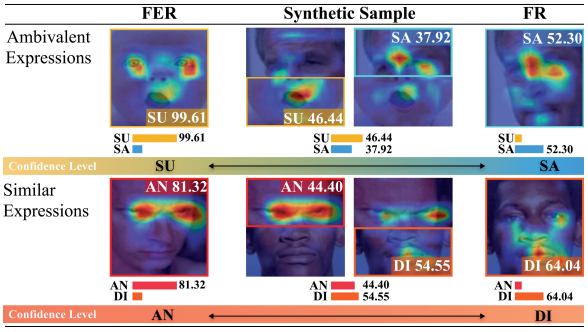


Figure 2: CAMs and confidence scores for FR, FER, and synthetic samples, showing activations for ambivalent (top) and similar (bottom) expressions with predicted labels.

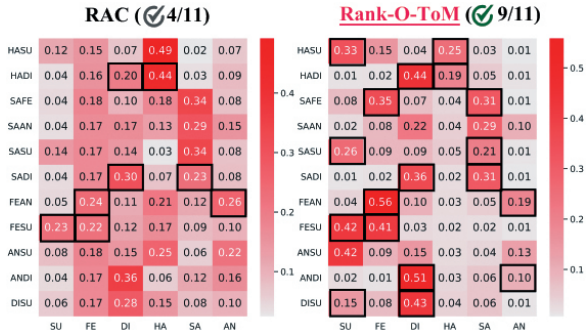


Figure 3: Confidence heatmaps for RAF-DB compound set: Basic expressions (x-axis) and compound expressions (y-axis) with bold squares marking correct Top-2 matches.

Figure 2 demonstrates how our model interprets emotional expressions using class activation maps (CAMs) [24] and confidence scores for original and synthetic samples. The CAMs reveal that synthetic samples activate a broader range of facial regions, showcasing a more extensive focus on diverse affective features in Rank-O-ToM. Moreover, the confidence scores indicate that our model is susceptible to complex affective cues, capturing blended and subtle emotions. Figure 3 illustrates the evaluation of compound emotions (e.g., happily surprised) [10], assessing whether Top-2 confidence scores match their components (e.g., happiness and surprised). Unlike the existing method, our model effectively aligns confidence scores with constituent emotions, reflecting a nuanced understanding of emotional complexity (details in Appendix).

Conclusion

Rank-O-ToM advances FER by capturing emotional granularity through synthetic blending and ordinal ranking, aligning AI predictions with human-like reasoning about affective states. Demonstrating superior accuracy and calibration across datasets, it bridges the gap between AI and human cognitive capabilities, enabling nuanced emotional understanding critical for affective ToM.

Acknowledgements

This research was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01341, Artificial Intelligence Graduate School Program, Chung-Ang University).

References

[1] Simon Baron-Cohen, Therese Jolliffe, Catherine Mortimore, and Mary Robertson. Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. *Journal of Child psychology and Psychiatry*, 38(7):813–822, 1997.

[2] Lisa Feldman Barrett and Eliza Bliss-Moreau. Affect as a psychological primitive. *Advances in experimental social psychology*, 41:167–218, 2009.

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016.

[4] Roberto Cittadini, Christian Tamantini, Francesco Scotto di Luzio, Clemente Lauretti, Loredana Zollo, and Francesca Cordella. Affective state estimation based on russell's model and physiological measurements. *Scientific Reports*, 13(1):9786, 2023.

[5] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *In Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014.

[6] Edith Theresa Gabriel, Raphaela Oberger, Michaela Schmoeger, Matthias Deckert, Stefanie Vockh, Eduard Auff, and Ulrike Willinger. Cognitive and affective theory of mind in adolescence: developmental aspects and associated neuropsychological variables. *Psychological research*, 85:533–553, 2021.

[7] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *In Proceedings of the International Conference on Learning Representations*, 2015.

[9] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4166–4175, 2022.

[10] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017.

[11] James Jenn-Jier Lien, Takeo Kanade, Jeffrey F Cohn, and Ching-Chung Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000.

[12] Haozhuo Lin and Qiu Chen. Artificial intelligence (ai)-integrated educational applications and college students' creativity and academic emotions: students and teachers' perceptions and attitudes. *BMC psychology*, 12(1):487, 2024.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017.

[14] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[15] Michele Poletti, Ivan Enrici, and Mauro Adenzato. Cognitive and affective theory of mind in neurodegenerative diseases: neuropsychological, neuroanatomical and neurochemical levels. *Neuroscience & Biobehavioral Reviews*, 36(9):2147–2164, 2012.

[16] Elizabeth Stewart, Cathy Catroppa, and Suncica Lah. Theory of mind in patients with epilepsy: a systematic review and meta-analysis. *Neuropsychology Review*, 26:3–24, 2016.

[17] Banafsheh Tehranineshat, Mahnaz Rakhshan, Camellia Torabizadeh, and Mohammad Fararouei. Compassionate care in healthcare systems: a systematic review. *Journal of the National Medical Association*, 111(5):546–554, 2019.

[18] Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023.

[19] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020.

[20] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6023–6032, 2019.

[21] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[22] Yuhang Zhang, Yaqi Li, Xuannan Liu, Weihong Deng, et al. Leave no stone unturned: mine extra knowledge for imbalanced facial expression recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision*, pages 418–434, 2022.

[24] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Appendix

This appendix provides additional explanations, analytical experiments, visualizations, and implementation details for the proposed Rank-O-ToM framework. Section 1 details the process of integrating the FR dataset and generating synthetic samples. Section 2 includes implementation details for all experiments and an additional experiment using the compound subset of RAF-DB. Finally, Section 3 presents further analytical experiments and visualizations to deepen the understanding of Rank-O-ToM's performance and capabilities.

1 Details on Synthetic Sample Generation

To enhance the model's ability to interpret complex affective states—a core aspect of affective ToM—we detail the integration of an unlabeled FR dataset, which lacks emotion category labels, and the generation of synthetic samples with varied emotional expressions. These methods address the limitations of relying solely on FER datasets, which may fail to capture the full spectrum of affectivity, leading to overly simplistic comparisons and suboptimal calibration [18].

Integrating FR Dataset with Pseudo-labeling. Human affective understanding involves interpreting subtle cues and varying emotional intensities across contexts. To emulate this, we integrate an unlabeled FR dataset into training using a confidence-based dynamic thresholding mechanism to assign pseudo-labels. This approach adapts thresholds for each class based on the model's confidence during training, addressing FER data imbalance and enhancing diversity. By incorporating pseudo-labeled FR data, the model learns to represent nuanced and realistic emotional expressions, aligning its learning process with human-like affective reasoning.

First, consider the labeled FER dataset $\mathcal{D}_{\text{fer}} = \{(x_i^{\text{fer}}, y_i^{\text{fer}})\}_{i=1}^{n}$. For each sample, we obtain the predicted label $\hat{y}_i^{\text{fer}}$ for the i-th sample at epoch t. To determine class-specific confidence thresholds, we define the set of correctly predicted samples for each class c as $\mathcal{D}_{\text{fer}}^{cs} = \{(x^{\text{fer}}, y^{\text{fer}}) | \hat{y}^{\text{fer}} = y^{\text{fer}} = c\}$. For each sample in this set, we calculate the confidence score $\tilde{p}^{\text{fer}} = \max_c p_c(y^{\text{fer}}|x^{\text{fer}})$. For $(x^{\text{fer}}, y^{\text{fer}}) \in \mathcal{D}_{\text{fer}}^{cs}$, we calculate the confidence $\tilde{p}^{\text{fer}} := \max_c p_c(y^{\text{fer}}|x^{\text{fer}})$. For each epoch t and class c, the confidence threshold $\mathcal{T}_c^t$ is computed as follows:

$$\begin{aligned}\mathcal{T}_c^t &= \frac{\beta}{1+e^{-t}} \cdot \frac{1}{|\mathcal{D}_{\text{fer}}^{cs}|} \sum_{i=1}^{|\mathcal{D}_{\text{fer}}^{cs}|} \tilde{p}_i^{fer} \\ &= \frac{\beta}{1+e^{-t}} \cdot \frac{1}{|\mathcal{D}_{\text{fer}}^{cs}|} \sum_{i=1}^{|\mathcal{D}_{\text{fer}}^{cs}|} \max_c p_c(y_i^{\text{fer}}|x_i^{\text{fer}}) \\ &\quad , (x_i^{\text{fer}}, y_i^{\text{fer}}) \in \mathcal{D}_{\text{fer}}^{cs} \end{aligned} \tag{2}$$

The hyperparameter $\beta \in (0, 1)$ moderates confidence levels. The term $\frac{1}{1+e^{-t}}$ ensures that the threshold adapts dynamically over training epochs, reflecting the model's evolving understanding—analogous to how human perception becomes more refined with experience.

During training, FER models exhibit varied patterns across classes due to data imbalance [9]. Initially, there's a narrow confidence gap between classes with abundant and scarce samples. As training advances, this distinction becomes more pronounced, encompassing both clear and ambiguous emotional expressions. The dynamic threshold $\mathcal{T}_c^t$ adjusts to these shifts, enabling the model to calibrate its confidence in a manner akin to human affective judgment.

Subsequently, for the unlabeled FR dataset $\mathcal{D}_{\text{fr}} = \{x_i^{\text{fr}}\}_{i=1}^{m}$, we use distinct weak augmentations to obtain two transformed samples, $x_{\mathtt{a}}^{\text{fr}} = \mathtt{Aug_a}(x_i^{fr})$ and $x_{\mathtt{b}}^{\text{fr}} = \mathtt{Aug_b}(x_i^{fr})$. Then, through the model $\mathcal{F}(\cdot)$, we determine the probability distributions $p_{\mathtt{a}}^{\text{fr}}$ and $p_{\mathtt{b}}^{\text{fr}}$ for facial expressions' classes for the two transformed samples $x_{\mathtt{a}}^{\text{fr}}$ and $x_{\mathtt{b}}^{\text{fr}}$. To emulate the human ability to integrate multiple subtle cues when interpreting emotions, we perform class-wise interpolation to obtain the aggregated probability distribution $\hat{p}$ for each sample:

$$\hat{p}_c = \lambda_c \cdot p_{\mathtt{a}}^{\text{fr}}(c|x_{\mathtt{a}}^{\text{fr}}) + (1-\lambda_c) \cdot p_{\mathtt{b}}^{\text{fr}}(c|x_{\mathtt{b}}^{\text{fr}}) \tag{3}$$

Here, c represents the class in FER, and λ_c denotes the interpolation ratio for class c. We assign $\hat{y}^{\text{fr}}$ based on the class with the highest confidence among those exhibiting confidence

higher than the threshold $\mathcal{T}_c^t$, as follows:

$$\hat{y}_i^{\text{fr}} = \text{argmax}_c(\mathbb{1}(\hat{p}_c > \mathcal{T}_c^t) \cdot \hat{p}_c) \quad (4)$$

where $\mathbb{1}(\cdot)$ indicates whether the probability score belonging to a specific class c is larger than the threshold $\mathcal{T}_c^t$. This method enables the integration of the unlabeled FR dataset during training, increasing the diversity of synthetic samples and enhancing the efficiency of ranking relationships. It allows the model to learn from a wider array of facial expressions, capturing subtle variations and complexities—mirroring the human capacity for nuanced affective understanding central to affective ToM.

Synthesizing Samples with Diverse Emotional Expressions. To enhance the model's ability to interpret complex affective states, we generate synthetic samples that capture a broader range of emotional expressions by blending facial images with different labels. This emulates the human ability to perceive blended or intermediate emotions, which are often absent in FER datasets.

For example, combining a sample labeled "sadness" with one labeled "anger" creates an image representing an intermediate emotion like "frustration." This aligns with the human capacity to interpret compound emotions, a key aspect of affective ToM. We achieve this using an adapted CutMix augmentation method [20], which preserves the semantic integrity of facial expressions. CutMix is an augmentation technique that generates synthetic samples by cutting and pasting regions from two training images. Given two samples, (x_i, y_i) and (x_j, y_j), CutMix creates a new sample $(\tilde{x}, \tilde{y})$ as follows:

$$\tilde{x} = M \odot x_i + (1 - M) \odot x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda) y_j, \quad (5)$$

where M is a binary mask indicating the region to be replaced, $\odot$ denotes element-wise multiplication, and $\lambda \sim Beta(1, 1)$ controls the combination ratio, ensuring a balanced mix of the two images. The mask M is determinded by a randomly generated bounding box $B = (b_x, b_y, b_w, b_h)$, where:

$$b_x \sim \text{Uniform}(0, W), \quad b_w = W\sqrt{1-\lambda}, \quad (6)$$

$$b_y \sim \text{Uniform}(0, H), \quad b_h = H\sqrt{1-\lambda}. \quad (7)$$

In standard CutMix, the position and size of the bounding box B are randomly determined, which may disrupt the semantic structure of facial expressions critical for emotion recognition.

To preserve the semantic regions associated with different emotional expressions—such as the eyes and mouth—we adapt the CutMix method by fixing the bounding box to horizontally split the image. Specifically, we set the top-left coordinates to $b_x = 0$ and $b_y = 0$, and define the width and height as $b_w = W$ and $b_h = 1/2H$, respectively. This effectively divides the face into upper and lower halves, each containing distinct emotional cues. Since both facial regions contribute equally to the combined image, we fix the combination ratio $\lambda = 0.5$. This ensures that the synthetic sample $\tilde{x}$ integrates the upper half from one image and the lower half from another, maintaining the balance of emotional expressions.

By horizontally bisecting and combining facial images in this manner, we preserve the critical semantic information necessary for affective understanding. This approach aligns with the human ability to integrate facial cues from different areas of the face—a key aspect of affective ToM.

2 Experimental Details

2.1 Datasets

We train Rank-O-ToM on the FER benchmark datasets, including RAF-DB [10], FERPlus [3], and AffectNet [14]. Additionally, we incorporate the LFW face recognition dataset [7], which does not include emotion class labels, into our training process. Overview of each datasets is shown

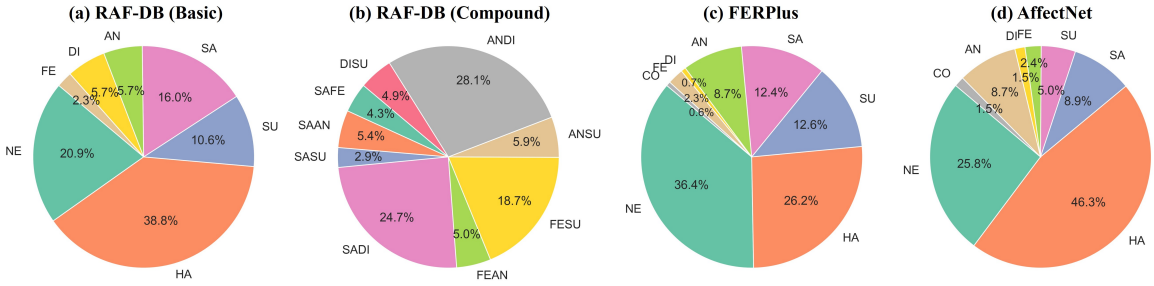


Figure 4: **Class proportions of FER datasets** The pie charts display results for (a) RAF-DB (Basic), (b) RAF-DB (Compound), (c) FERPlus, (d) AffectNet.

Datasets		# Annotated images (train / val / test)	# Classes	Emotional labels
RAF-DB [10]	Basic	15,339 (12,271 / 3,068 / -)	7	neutral, happiness, surprise, sadness, anger, disgust, fear
	Compound	3,954 (3,162 / 792 / -)	11	happily surprised, happily disgusted, sadly fearful, sadly angry, sadly surprised, sadly disgusted, fearfully angry, fearfully surprised, angrily surprised, angrily disgusted, disgustedly surprised
FERPlus [3]		35,887 (28,558 / 3,579 / 3,573)	8	neutral, happiness, surprise, sadness, anger, disgust, fear, contempt
AffectNet [14]		291,650 (287,651 / 3,999 / -)	8	neutral, happiness, surprise, sadness, anger, disgust, fear, contempt

Table 2: **Overview of FER datasets** This presents the datasets used to evaluate FER performance. It includes the number of annotated images, the number of classes, and the emotional labels used in experiments. The datasets covered are RAF-DB (both Basic and Compound), FERPlus, and AffectNet.

in Table 2, and the proportions of each dataset are shown in Figure 4.

RAF-DB [10] is a facial expression dataset comprising 29,672 individual facial images, including basic or compound expressions. In this work, we utilize facial images with 7 expressions (i.e., “surprise”, “fear”, “disgust”, “happiness”, “sadness”, “anger”, “neutral”), including 12,271 images as training data and 3,068 images as test data. In addition, we employ the *compound set* for validation in Figure 3 to evaluate our framework’s performance on more complex and nuanced emotional expressions. The compound dataset comprises 11 distinct classes, each representing a combination of basic emotions (i.e., “happily surprised”, “happily disgusted”, “sadly fearful”, “sadly angry”, “sadly surprised”, “sadly disgusted”, “fearfully angry”, “fearfully surprised”, “angrily surprised”, “angrily disgusted”, and “disgustedly surprised”). Images are aligned and cropped using three landmarks, then resized to 224×224 pixels.

FERPlus [3] is an extended version of the original FER dataset, designed to improve label accuracy and reliability through enhanced annotations provided by 10 annotators via crowdsourcing. It includes 35,887 facial images labeled with 8 basic expressions: “neutral”, “happiness”, “surprise”, “sadness”, “anger”, “disgust”, “fear”, and “contempt”. Additionally, FERPlus utilizes a multi-label annotation, allowing images to be associated with multiple emotions, reflecting the complexity of human facial expressions. Images are aligned and

cropped using key facial landmarks, then resized to 224×224 pixels for consistency with our processing pipeline.

AffectNet [14] is one of the most comprehensive and sizable facial expression datasets, containing 287,651 training images and 3,999 test images manually labeled into eight classes. While AffectNet includes a wide range of annotations, in this work, we utilize only the images labeled with eight basic emotion categories: "neutral", "happiness", "surprise", "sadness", "anger", "disgust", "fear", and "contempt". This selection excludes images with labels such as "none", "uncertain", and "non-face", which do not correspond to any specific emotion.

LFW [7] is a popular FR dataset that has not been labeled for emotion categories. LFW comprises 13,000 facial images, representing over 5,000 identities. In our study, we pseudo-label the LFW dataset into seven emotion classes, the same as the RAF-DB dataset. We utilize the MTCNN [21] alignment method to detect and align faces within the LFW dataset, resizing them to 224×224 pixels.

2.2 Evaluation Metrics

When evaluating AI agents in the context of affective ToM, it is crucial to assess not only their accuracy but also how well their confidence—derived from softmax scores—aligns with the true likelihood of their predictions. This alignment, quantified through calibration metrics, ensures that the model's confidence levels are appropriate for reliable human-AI interactions, particularly when interpreting nuanced emotional expressions.

To assess calibration, predictions are grouped into M bins of equal size intervals. For each bin B_m containing samples whose confidence scores fall within bin m, we define accuracy and confidence as follows:

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i) \tag{8}$$

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \tilde{p}_i \tag{9}$$

where $\tilde{p}_i$ represents the confidence of sample i.
Expected Calibration Error (ECE) measures the average discrepancy between a model's accuracy and confidence across bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\ \mathrm{acc(B_m)} - \mathrm{conf}(B_m)\ | \tag{10}$$

Adaptive Expected Calibration Error (AECE) extends ECE by dynamically adjusting the binning process to better match the distribution of predicted probabilities, offering a finer-grained evaluation of calibration.
Maximum Calibration Error (MCE) highlights the largest miscalibration by capturing the maximum discrepancy between accuracy and confidence across bins:

$$\mathrm{MCE} = \max_{m \in (1,\ldots M)} |\ \mathrm{acc(B_m)} - \mathrm{conf}(B_m)\ | \tag{11}$$

These metrics enable the evaluation of how well a model's confidence aligns with its true performance, ensuring reliable and interpretable predictions. Such alignment is crucial for interpreting complex human emotions, as it allows the model to emulate human-like reasoning and maintain appropriate confidence levels in affective ToM tasks.

2.3 Implementation Details

We utilize the Adam optimizer [8] with a learning rate of 5×10^{-4}. Training proceeds with a mini-batch size of 64 over 60 epochs. The initial threshold for class-wise dynamic pseudo-labeling starts at 0.95. Weak augmentations apply, including RandomCrop and RandomHorizontalFlip. Hyperparameters stand at $\lambda_c = \frac{1}{2}$ and $\beta = 0.97$. We evaluate the performances

on calibration metrics, setting all bins to 15. Ranking loss uses a margin of $\delta \in [0.1, 0.3]$ and a balancing scalar $\beta = 1$, which balances the focal loss, which is employed to classify emotion categories as the default values unless otherwise specified. The hyperparameter $\gamma = 2$ and $\alpha = 0.25$ are used in focal loss [13].

Implementation details on an experiment with a compound set of RAF-DB We elaborate on the details of the experiment regarding compound emotions in Rank-O-ToM. We utilize a compound RAF-DB dataset composed of compound expressions. The compound RAF-DB consists of data annotated with compound expressions where two basic expressions (e.g., "sad" and "angry") are combined to form a compound expression (e.g., "sadly angry"). A well-trained FER framework should demonstrate high confidence levels in the basic expressions that comprise the compound expression when presented with samples of compound expressions, provided it has learned a diverse spectrum of expressions. For instance, with a sadly angry" sample, a well-calibrated FER framework trained with basic emotions should output high probabilities for "sad" and "angry."

The heatmaps in Figure 3 and 5 depict the average confidence for samples containing compound expressions, as evaluated by both FER methods and Rank-O-ToM on the RAF-DB dataset. Additionally, we evaluate whether the Top-2 confidence scores inferred by the model match the basic emotions constituting the compound expression. Experimental results indicate that Rank-O-ToM activates the constituent expressions in a balanced manner compared to other approaches.

3 Additional qualitative examples on CAM

In Figure 6, we present additional qualitative results on the CAMs produced by our framework for two original samples—from the FER and FR datasets—and the synthetic sample (SYN) generated by combining them. These visualizations illustrate various combinations of emotions, demonstrating how our method captures nuanced affective states essential for human-like emotion recognition.

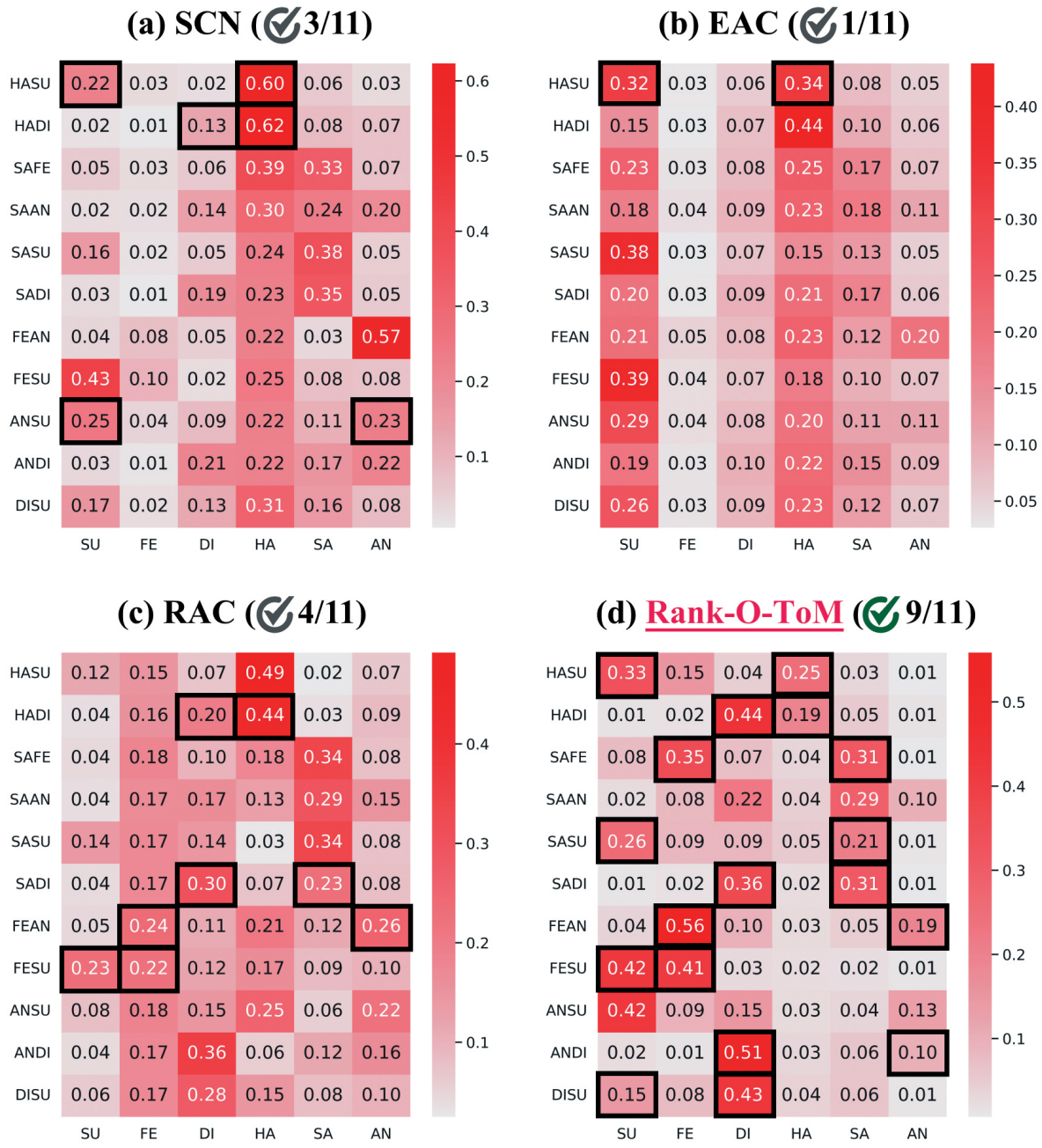


Figure 5: **Additional heatmaps for RAF-DB compound set**. Basic expressions (x-axis) and compound expressions (y-axis) are depicted, with bold squares marking correct Top-2 matches. SCN and EAC models are presented here as supplementary to the main text, which focuses on RAC and Rank-O-ToM.

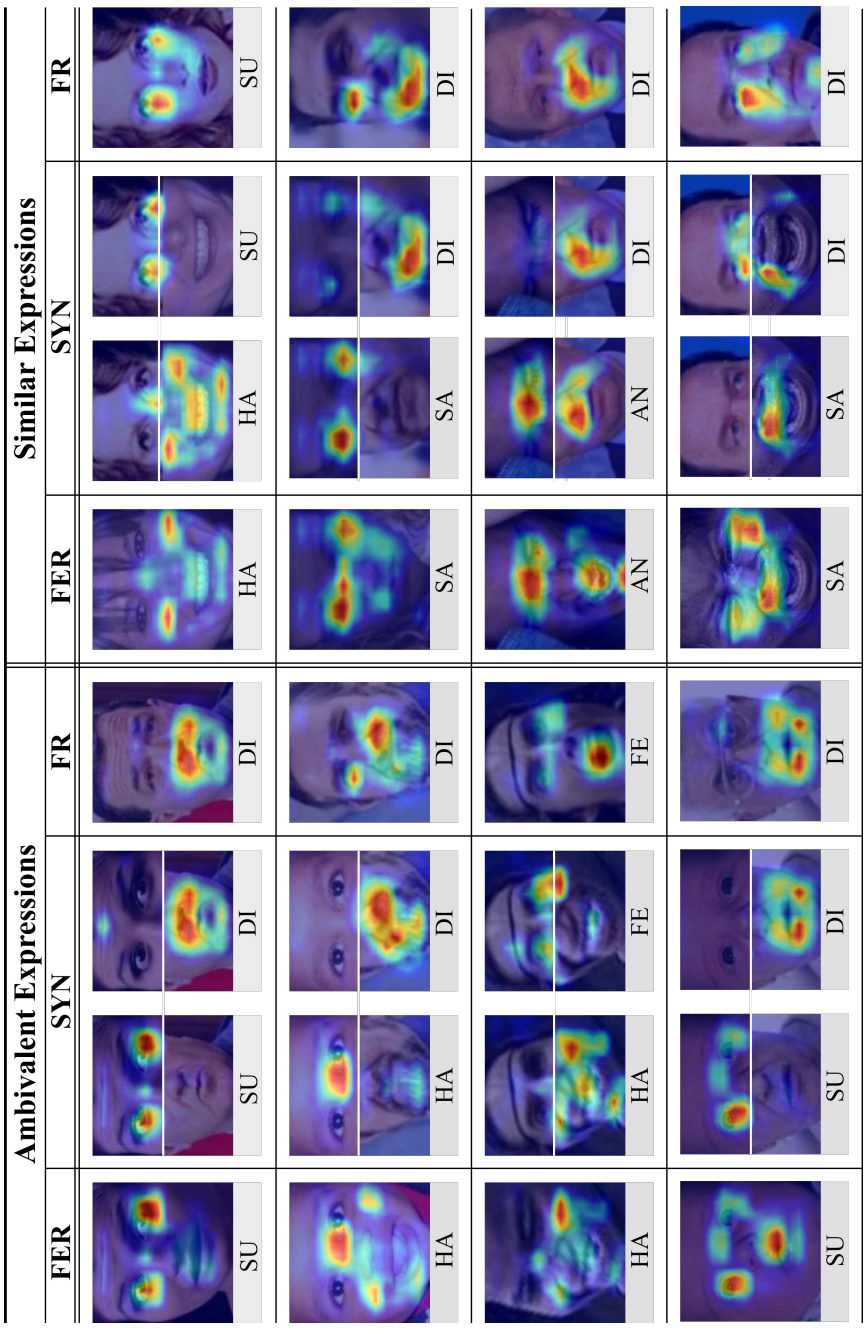


Figure 6: **Qualitative results of CAM on original samples and synthetic samples generated by Rank-O-ToM**

Relational Closure for Reasoning

Arun Kumar and Paul Schrater

Dept. of Computer Science, University of Minnesota, Minneapolis, MN, USA

Abstract

Meta-reasoning demands higher-order thinking over relational structures and their composition. While embedding-based graph learning approaches propagate and aggregate statistics in neighborhoods to learn embeddings, such approaches have two major drawbacks. First, when there is no evidential support to ground the claimed relations and types, embeddings make unsupported connections. Second, the embedding statistics lack semantic consistency in grounded relations across neighborhoods. Relational closure addresses them by providing closure for coverage - evidentiary support and closure for coherence - consistency support across neighborhoods. Relational algebra on graphs by placing relations at the pivot of reasoning offers a promising formalism with the ability to define natural systems.

Introduction

Meta-reasoning processes [3, 1] are higher-order processes that monitor and govern the underlying resources with the goal of problem solving. They evaluate the available information to determine the next action an agent can take and possible effects on the environment, particularly on objects, object relations, and intrinsic constraints. However, the majority of current systems such as graph networks prioritize object properties, where relations are merely used for learning object embeddings by aggregating neighborhood statistics leading to several drawbacks. An issue with statistical embeddings is that when there is insufficient evidence for a relation or the relation is not supported as in the claims, the relational connections made by the embeddings are unreliable, resulting in unwanted connections or failure to make certain connections. Furthermore, local neighborhood relations must be consistent with those in the larger semantic structure. As more relations expand from the local neighborhood, there is a need for consistency checks across neighborhoods.

To address challenges of grounding relations and coherence, we require a mechanism to determine whether sufficient support exists for grounding relations or establish support if more evidence is required, as well as ensuring consistency between relations for coherent relational structures. Relational algebra [6] treats relations between objects in the form of relational matrices. We conjecture a relational algebraic approach of relation support and coherence using relational closure and show with an example how relational closure allows us to answer a meta-level question by emulating the addition of a new relation to a graph and its utility. Support for relations and consistency does not have to be direct, so the relational matrices can be chained together until a support is found or established by adding a new relation, which is the basic idea behind relational closure. Closure support for relations is a local view, but this is where inconsistency might occur; so, a consis-

tency check, a global view, is imposed through coherence closure, thus defining a natural system.

Relational closure for reasoning

Graphs represent the environment as objects and the relations between objects. A graph $G = (V, E)$ is defined by a set of nodes V and a set of edges E that correspond to relations between the nodes. A relation r is an ordered pair (u, v) denoted by urv, where $u, v \in V$. This type of graph of relations is referred to as a directed graph or digraph. Different types of relations are allowed between node pairs. Given the prevalence of binary relations in the real world, it is reasonable to represent graphs using binary relational matrices [6, 4].

The graph has a finite number of nodes, $n = |V|$, and a relation matrix, $R^{d\times n\times n}$, which is a stack of binary relational matrices with d denoting the relational layer. The relation matrix R_k of the k-th relation layer is given as

$$R_k[i,j] = \begin{cases} 1 & \text{if } ir_kj, \\ 0 & \text{otherwise} \end{cases}$$

If u and v are two nodes in a graph, then a path between them goes from u to v along the graph's edges, i.e. a path is composed of relations between the two nodes. When there are multiple relation types, a path is a composition of relations that are compatible with each other. If urv and vrw, then urw, implying that the relation is transitive. The transitivity property of relations provides insight into how to expand a graph's relational structure by composing relations to create new relations, as well as how to analyze the effects of modifying relations in graphs.

Composition in binary relational algebra

If R_1 and R_2 are two relation matrices, the binary matrix multiplication R_1R_2 is the composition of two relations. The composition of a relation matrix R with itself R^2 can be computed similarly. Composition operation enables us to create new composite relations from existing relations. As the graph has a finite number of nodes, the transitive closure of its relation matrix is the union of the first n powers of R, $R^+ = R \cup R^2 \cup ... \cup R^n = \cup_{i=1}^{n} R^i$.

The relation matrix R is a stack of relational matrices, thus we modify the Floyd-Warshall algorithm [5] to handle the relational layers in order to compute transitive closure. First, all relational layers are combined using logical OR for matrix union to create a combined relation matrix R. The transitive closure R^+ is then computed as

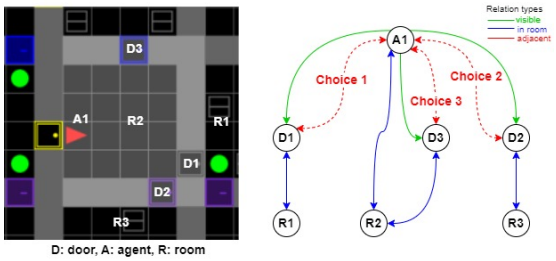
$$R[i,j] = \bigvee_{k=1}^{d} R_k[i,j]$$

$$R^+ = \bigvee_{k=1}^{n} \Big(R + \big(R[:,k] \otimes R[k,:]\big) \Big)$$

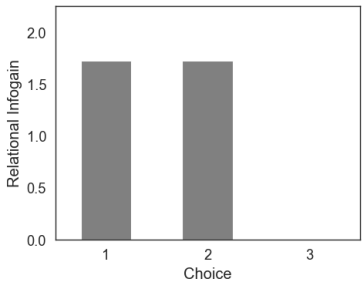
where $\bigvee$ is binary OR operation for matrix union, $+$ is logical OR and $\otimes$ is the binary outer product between vectors for optimizing binary multiplication and it gives a $n \times n$ binary matrix.

A relational closure R^+ reveals which relations can be combined, and we compute the relevant information H_{R^+}. Some relations or connections in the relational closure may not yet be available, making them potential candidates for relation modification in graphs. So, one of the agent's high-level goals is to meta-reason about which relations it could add to create new compositional relations and therefore make previously unavailable relations available. For every choice c, it emulates adding a new relation to the graph, resulting in a new relational transitive closure R_c^+ and information $H_{R_c^+}$ on the new relation matrix. The relational gain can then be utilized to make choices that are useful.

Figure 1 illustrates an example. Relational closure R^+ provides direct answers to questions regarding existence of a relation direct or compositional. If R^+ is a transitive closure matrix, a



(a)



(b)

Figure 1: a) miniGrid [2] env and its graph showing agent, doors, rooms, and their relations such as visible, adjacent, and in room. Dashed red lines are possible adjacent relations between the agent and doors. b) Relational info for choices.

relation direct or composed between nodes i and j exists if $R^+[i,j] = 1$ for $i \neq j$. In the relational closure, and as shown in the graph, there are connections or compositions of relations $R2 \rightarrow A1 \rightarrow D1 \rightarrow R1$ and $R2 \rightarrow A1 \rightarrow D2 \rightarrow R3$, but no connections from $R1$ to $R2$ or $R3$, or from $R3$ to $R1$ or $R2$. Choice 3 does not have relational info because adding that relation is not helpful in creating any new path. On the other hand, adding relations in either choice 1 or choice 2 offers benefit since it facilitates creating new composition of relations. The relational closure R_c^+ after adding a choice contains newly composed relations. Choice 1 connects $R1$ to $R2$ or $R3$, whereas choice 2 connects $R3$ to either $R1$ or $R2$. It emphasizes the value of using relational algebra for higher-order reasoning. Closure for support is achieved by adding new relations, and closure for coherence on the options ensures a consistency check while answering the respective question.

Conclusion

Relational closure enables closure for coverage through evidentiary support of relations and closure for coherence through consistency support. The relational algebraic view, due to the focus on relations between objects, has the potential to address core challenges in graph learning while also providing an avenue for meta-reasoning that demands abstract and compositional thinking.

References

[1] Rakefet Ackerman and Valerie A Thompson. Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in cognitive sciences*, 21(8):607–617, 2017.

[2] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. *GitHub repository*, 2018.

[3] Michael T Cox and Anita Raja. *Metareasoning: Thinking about thinking*. MIT Press, 2011.

[4] Amina Doumane. Graph characterization of the universal theory of relations. In *Mathematical foundations of computer science*, 2021.

[5] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345–345, 1962.

[6] Steven Givant and Steven Givant. *Relation algebras*. Springer, 2017.

Second-order Theory of Mind for Human Teachers and Robot Learners.

Patrick Callaghan[1], Reid Simmons[1], and Henny Admoni[1]

[1]Carnegie Mellon University

Introduction

Confusing or otherwise unhelpful learner feedback creates or perpetuates erroneous beliefs that the teacher and learner have of each other, thereby increasing the cognitive burden placed upon the human teacher. For example, the robot's feedback might cause the human to misunderstand what the learner knows about the learning objective or how the learner learns. At the same time—and in addition to the learning objective—the learner might misunderstand how the teacher perceives the learner's task knowledge and learning processes. To ease the teaching burden, the learner should provide feedback that accounts for these misunderstandings and elicits efficient teaching from the human.

One way to account for these erroneous beliefs and thereby improve a human's teaching efficacy is to leverage Theory of Mind. Theory of Mind (ToM)—the ability to infer another's motives and beliefs by observing their actions—is often used in AI approaches today [12, 14, 5, 15, 9]. Less explored, however, is *Second-order* Theory of Mind (ToM-2) which includes an awareness that other agents also have a ToM [1, 11]. With this additional awareness, a learner could model and account for its teacher's beliefs of the learner when selecting feedback during a teaching session.

This work endows an AI learner with a ToM-2 that models perceived rationality as a source for erroneous beliefs a teacher and learner may have of one another. It also explores how a learner can ease teaching burden and improve teacher efficacy if it selects feedback which accounts for its model of the teacher's beliefs about the learner *and* its learning objective.

Proof of Concept Domain

Consider a turn-based card game played between a human teacher and robot learner in which a "rule" governs how multi-featured cards are sorted into three piles (Figure 1). In a single round, the teacher plays one such card into a pile according to the rule, and the robot responds with an utterance (i.e., "feedback") pertaining to one of the features of the rule. The robot's goal is to identify the rule which distinguishes the piles.

Let's say the teacher chooses the rule: "Reds belong in Pile 1. Blues belong in Pile 2. Greens belong in Pile 3." In the first round of the game, the teacher places the "Three Red Diamonds" card on Pile 1. What should the learner infer from this move, and what feedback will convey the learner's belief and prompt the teacher to play an informative next card? Prior work endows the learner with policies for selecting feedback that optimize volume removal and information gain [3, 13, 10, 7, 2]. Unfortunately, such optimizations can yield feedback that causes the teacher to misunderstand what

the learner knows about the task or how the learner incorporates information into its reasoning processes. In the context of the game, such optimizations can yield redundant feedback that reflects stunted learning (e.g., the robot selects the same utterances at turn t and turn $t+1$), seemingly-irrelevant feedback that reflects incorrect learning (e.g., the robot selects an utterance about Pile 3 when the teacher has focused on Pile 1), or feedback which otherwise reflects the robot's erroneous beliefs. Crucially, however, the learner may be closer to the truth than the human believes, yet because it optimized its feedback for the learning objective and didn't consider the teacher's beliefs of the learner, the feedback may compel the teacher to re-teach or continue teaching concepts which the robot already mastered. A learner endowed with a ToM-2 could model what the teacher believes of the learner and account for those beliefs when selecting feedback to mitigate such misunderstandings (see Figure 1).

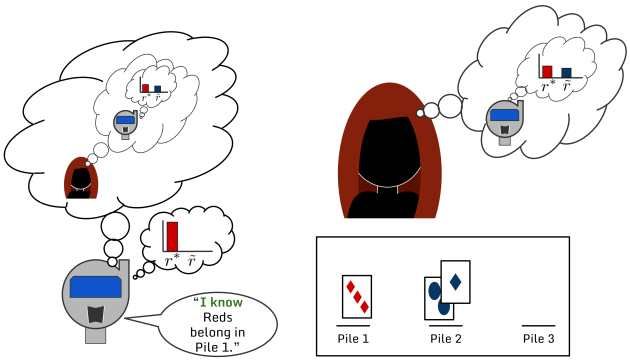


Figure 1: A human teaches a robot the $[r]$ule for how cards are categorized (e.g., according to color). Here, the teacher misunderstands if the robot knows the correct rule r^*. The robot's Second-order Theory of Mind models this misunderstanding and provide feedback using Confidence Expressions (green text).

1 Methods

To undertake the proposed solution, this work leverages the Interactive Partially Observable Markov Decision Process (I-POMDP) as a framework for a robot learner's ToM-2 [4]. The I-POMDP augments the POMDP's notion of state to an *interactive state*, which represents both the states of the environment *and* models of the other agents within it. These models can be I-POMDPs which themselves represent environment states and models of other agents, and it is this configuration that enables the I-POMDP to represent a Second-order Theory of Mind.

With this framework, the learner can model the teacher's beliefs about the learner's task knowledge and learning processes through the components which comprise an I-POMDP—namely, an agent's observation function, reward function, transition function, and optimization criterion. Of these elements, this work posits that erroneous beliefs primarily stem from the teacher's observation function. More specifically, a human teacher might observe and interpret a learner's feedback in a perfectly-rational manner, a perfectly-irrational manner, or somewhere in between. By modeling these different possibilities for how the teacher perceives the learner's feedback, the learner can come to a better understanding of how the the teacher will interpret its feedback.

At the same time, the teacher may believe the learner observes and interprets the teacher's actions in an irrational fashion. Indeed, prior work suggests this perceived-irrationality can manifest as the teacher outright ignoring the learner's feedback and playing cards in a systematic fashion [8]. By modeling the teacher's belief of the learner's observation function, the learner can identify if the teacher believes the robot is irrational and choose feedback most likely to lead the teacher to understand the robot is, in fact, a rational learner. In turn, the teacher could play cards more efficiently.

To enable the learner to recognize these sources of irrationality, this work augments the

I-POMDP with a discrete set of learnable observation functions, each of which is a noisily-rational model whose rationality is inversely-proportional to a temperature parameter β. By observing the teacher's card plays, the learner will be able to learn the teacher's degree of rationality as it also learns the rule governing the teacher's actions.

Modeling these sources of irrationality is only helpful to the teacher if the learner accounts for them when generating feedback. As such, this work incorporates *Confidence Expressions* (CEs) into the learner's feedback to convey the strength of the learner's stated beliefs. Without CEs, feedback can express confidence far greater than the learner's actual confidence about the features it addresses, and this ambiguous communication can lead the teacher to misunderstand the learner's task knowledge and learning processes. For example, if the learner says, "Reds belong in Pile 1," the teacher might perceive the learner as 100% certain of its statement. It's possible, however, that the learner is still unsure if Reds or Diamonds categorize that Pile.

CEs convey the learner's certainty about the stated features. More specifically, the learner prepends one of three CEs ("I know," "I think," or "I'm unsure if") to its feature expression (e.g., "Reds belong in Pile 1."), selecting the one most reflective of its confidence in the stated feature. By incorporating CEs into its feedback, the learner is able to convey its level of certainty over its task knowledge with the intention of correcting and preventing teacher misunderstandings. If we again consider Figure 1, perhaps the teacher believes the learner will truly understand the rule only if it sees the complete set of Red cards in Pile 1. This strategy, however, is redundant and elicits unnecessary time investment on the part of the teacher. To mitigate this extra effort, the robot could say, "I know Reds belong in Pile 1," when certain of that feature. The robot can additionally express its uncertainty over other features by stating, for example, "I'm unsure if Greens belong in Pile 3."

2 Evaluations

This work will evaluate the utility of endowing a robot learner with a ToM-2 through simulated and real-world interactions between teacher and learner. The evaluations will investigate the benefits of enabling a learner to (1) identify its teacher's sources of irrationality and (2) utilize CEs when providing feedback during the teaching session. The turn-based card game will be the domain of study.

Simulation Experiments The first set of experiments will comprise interactions between a learner and a simulated teacher to investigate the learner's ability to identify the teacher's rationality and the teacher's perceived learner rationality. They will also investigate if this knowledge enables the learner to elicit greater teaching efficacy from its teacher. Each trial will initialize a teacher with a noisily-rational observation function, and the learner's inferred model will be compared against the ground truth teacher model for the span of the teaching session. Additionally, experiments will evaluate how CEs benefit teacher efficacy, the learner's ability to identify a teacher's rationality, and if their use can guide a teacher to better understanding of the learner's rationality.

User Study The next set of experiments will be undertaken through a user study in which human participants teach a robot learner rules of varied complexities. As in the simulated interactions, the user study will quantify these benefits through the number of rounds it takes the robot to learn the rule in each experimental condition, as well as the number of times the teacher incorrectly believes the learner understands the rule. Additionally, subjective metrics (e.g., the NASA TLX) will be used to evaluate the cognitive burden imposed by each of the conditions [6].

References

[1] Janet Wilde Astington, Janette Pelletier, and Bruce Homer. Theory of mind and epistemological development: the relation between children's second-order false-belief understanding and their ability to reason about evidence. *New Ideas in Psychology*, 20(2):131–144, August 2002.

[2] Erdem Bıyık, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, and Dorsa Sadigh. Asking Easy Questions: A User-Friendly Approach to Active Reward Learning. *arXiv:1910.04365 [cs]*, October 2019. ZSCC: 0000042 arXiv: 1910.04365.

[3] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. INQUIRE: INteractive Querying for User-aware Informative REasoning. November 2022.

[4] P. J. Gmytrasiewicz and P. Doshi. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24:49–79, July 2005.

[5] Jesse Gray and Cynthia Breazeal. Manipulating Mental States Through Physical Action. *International Journal of Social Robotics*, 6(3):315–327, August 2014.

[6] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati, editors, *Advances in Psychology*, volume 52 of *Human Mental Workload*, pages 139–183. North-Holland, January 1988.

[7] Rachel Holladay, Shervin Javdani, Anca Dragan, and Siddhartha Srinivasa. Active Comparison Based Learning Incorporating User Uncertainty and Noise. page 7, 2016. ZSCC: 0000024.

[8] Pallavi Koppol. *Interactive Machine Learning from Humans: Knowledge Sharing via Mutual Feedback*. PhD thesis, 2023.

[9] Ini Oguntola, Dana Hughes, and Katia Sycara. Deep Interpretable Models of Theory of Mind. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 657–664, August 2021. ISSN: 1944-9437.

[10] Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning Reward Functions by Integrating Human Demonstrations and Preferences. *arXiv:1906.08928 [cs]*, June 2019. ZSCC: 0000054 arXiv: 1906.08928.

[11] Massimiliano Papera, Anne Richards, Paul van Geert, and Costanza Valentini. Development of second-order theory of mind: Assessment of environmental influences using a dynamic system approach. *International Journal of Behavioral Development*, 43(3):245–254, May 2019. Publisher: SAGE Publications Ltd.

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

[13] Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation, July 2017. ZSCC: 0000180.

[14] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. In situ bidirectional human-robot value alignment. *Science Robotics*, 7(68):eabm4183, July 2022. Publisher: American Association for the Advancement of Science.

[15] Mustafa Mert Çelikok, Tomi Peltola, Pedram Daee, and Samuel Kaski. Interactive AI with a Theory of Mind, December 2019. arXiv:1912.05284 [cs].

The Turing Game 

Michal Lewandowski[†], Simon Schmid[†], Patrick Mederitsch[†], Alexander Aufreiter[†], Gregor Aichinger[†‡], Felix Nessler[‡], Severin Bergsmann[†], Viktor Szolga[‡], Tobias Halmdienst[‡], Bernhard Nessler[†‡]

[†] SCCH [‡] JKU Linz

Abstract

We present first experimental results from the *Turing Game*, a modern implementation of the original imitation game as proposed by Alan Turing in 1950. The Turing Game is a gamified interaction between two human players and one AI chatbot powered by Large Language Models (LLMs). The game is designed to explore whether humans can distinguish between their peers and machines in chat-based conversations, with human players striving to identify fellow humans and machines striving to blend in as one of them. To this end, we implemented a comprehensive framework that connects human players over the Internet with chatbot implementations. With our work, we aim to deepen the understanding of the human-AI interactions.

1 Introduction

AI systems are built with the goal of performing activities that were traditionally reserved to humans, from playing strategy games, like chess [3], Go [22] or Dota-2 [1], to drawing pictures and writing creative texts [18, 9]. They became better and better up until the point where some have already surpassed human performances in fields that have traditionally been believed to require human abstract thinking and strategic planning. In the field of content generation, we have arrived at the point where we find it hard to discern whether images or clips are generated or represent real footage or whether texts stem from a human or a machine.

In this paper, we extend the Imitation Game [26], originally proposed by Alan Turing, by symmetrizing the roles of the original two human participants, see Fig. 1. This seemingly slight redesign of the test shifts the focus away from the simple question-answering to the collaboration between the humans and the inference of their mutual intentions, a characteristic feature of human communications [25]. Is the machine able to understand human intentions equally well as another human, or even better? Just like Alan Turing did in his original Imitation Game, we leave the kind and length of the conversation fully up to the humans. Our contributions are as follows:

- We propose a generalization of the Turing Test, termed the *Turing Game*, which is symmetric with respect to the role of the two humans.
- We have developed and installed the Turing Game as a platform and made it publicly available.[1] Our platform serves as a sandbox for testing various LLMs and

[1] `https://www.turinggame.ai/`

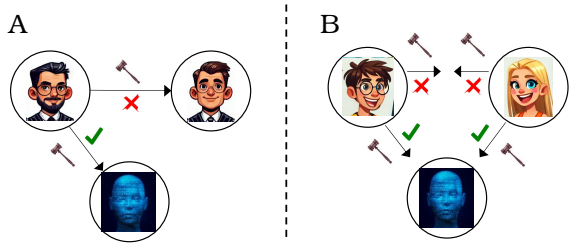


Figure 1: **A** The original Turing Test. **B** Our Turing Game: Both humans independently decide which interlocutor they believe is the machine while supporting the other human. Red crosses show a human misidentifying another human as a machine, while green checks indicate correct identification of the machine. Hammers indicate the decision-making. Only if both humans are correct, they win the game.

chatbot implementations designed to mimic human-like thinking, evaluated by an open community. We have designed the ratings of the bots such that the most qualified humans contribute the most to those ratings. We present the preliminary experimental results from the hitherto gathered data.

2 The Turing Game

We symmetrized the original Imitation Game proposed by Alan Turing [26] by allowing the three participants (two humans and one machine) to interact with each other, and we removed the predetermined role of the interrogator (see Fig. 1). That gives rise to a gamified interaction between players, called the *Turing Game*. At any point during the game, the players may decide to cast their vote and try to identify the machine. The game finishes as soon as the both humans have cast their vote. The humans win the game only if both of them have correctly identified the machine. If at least one of them misidentifies his fellow human as a machine, then both humans lose. This redesign introduces the following changes to the test's dynamics: (*i*) already with three participants we may observe an effect of siding between any two players, absent in one-on-one interactions [24]; (*ii*) the presence of two players further mitigates the reverse effect of the Turing Test as the machine's responses do not get influenced solely by one player [20]; (*iii*) the participants benefit from forming collaborations within the group, a typically human feature [25]. Their interaction's style may range from fully collaborative, to fully interrogative, or anything in between. Lastly, as participants interact using written language without additional cues such as body language or facial expressions, they rely more on deliberate reasoning rather than intuitive judgment [12].

3 Results

We pair the players for a game based on two factors: (*i*) their strength of the game S, (*ii*) their time to decide in minutes T. The distance between any two players is defined as the Euclidean distance between their strength of game and time to decide. Please refer to Appendix for more details.

If we consider only games with definitive win or loss outcomes, humans won 76.12% of the games, while machines won 23.88%. However, approximately a quarter (25.42%) of the games were surrendered by a human. Taking this into account, humans won 47.69% of all games, while machines won 14.96%. In the Appendix, we propose a tailored ranking method to pre-select the best players (as measured by the machine detection rate with coefficient ξ). When weighting is considered, the win ratios of the bots drop significantly (e.g., the score of AllTalker decreases from 24.74% to 11.70%). This demonstrates that, given the small number of games analyzed, the pre-selection of players significantly impacts the quality of the resulting judgment. See Appendix for an overview.

In our experiment, we have so far limited our focus to three bots, which we refer to as: (1)

AllTalker (supports English and German), (2) MetaSim (English only), and (3) MadTalker (English only). For more details see the Appendix.

3.1 Good Judgment Needs Time

We examined how game duration affects human win rates and found that wins increase with time, plateauing after about 3 minutes (Fig.2). Over half of the games exceeded this duration, with some lasting up to 25 minutes. Longer games likely reflect the need for more time to make informed decisions. Imposing time limits may force random guesses, increasing judgment errors [23].

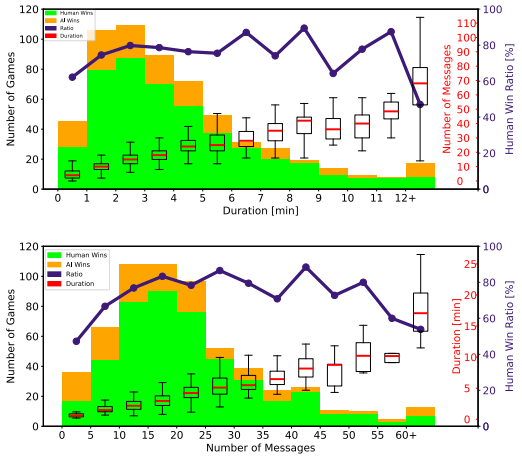


Figure 2: Histograms show total games (orange) and human wins (light green) by message count. Boxplots depict message distributions over time and exchanges. The blue line indicates humans reach 80% accuracy after 2–3 minutes or 15–20 messages, with lower—but above-chance—performance outside this range.

4 Conclusions

We have proposed a framework designed to understand how proficient people are in telling their kind from machines in a direct, text-based, interaction. By the game's design, we aim to engage participants' System 2 cognitive processes. We posit that it is necessary for players to use analytical reasoning and critical thinking to detect subtle non-human cues [27, 10]. The setup moreover involves meta-cognition and theory of mind, as players reflect on their reasoning and anticipate others [6]. The game's complex problem-solving environment provides deeper insights into differentiating human intelligence from artificial intelligence. Moreover, with the proposed framework we have started to gather a dataset which contains thousands of deductive-interactions human-AI, to be released shortly. We will compare the detection rate of machine-generated text by humans with recent approaches designed to automatically detect text generated by LLMs [17]. This comparison will establish a *human benchmark* for the detection of LLM-generated text. We will also release the gathered dataset. We summarize our findings as follows:

- Current LLMs cannot yet support multi-player discussions at a human level.
- Our data suggests it's still unclear whether current LLMs can truly fool humans, indicating more work is needed to achieve human-like AI interaction.

Acknowledgements

The research reported in this paper has been funded by BMK, BMAW, and the State of Upper Austria in the frame of the SCCH competence center INTEGRATE [(FFG grant no. 892418)] as part of the FFG COMET Competence Centers for Excellent Technologies Program, by the Upper Austria's #upperVISION2030 business and research strategy in the frame of AI Engineering and Certification Center, no. Wi-2022-699557-Hub, and by the Horizon 2020 Program of the European Commission in the frame of the ICT-48-2020 Network ELISE (951847).

References

[1] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning, 2019.

[2] Selmer Bringsjord, Paul Bello, and David Ferrucci. Creativity, the turing test, and the (better) lovelace test. *Minds and Machines*, 11:3–27, 2001.

[3] Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1):57–83, 2002.

[4] Edward L Deci, Richard Koestner, and Richard M Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668, 1999.

[5] Robert M. French. The turing test: the first 50 years. *Trends in Cognitive Sciences*, 4(3):115–122, 2000.

[6] Chris D. Frith and Uta Frith. The neural basis of mentalizing. *Neuron*, 50(4):531–534, 2006.

[7] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. In *Proceedings of the National Academy of Sciences*, volume 112, pages 3618–3623, 2015.

[8] Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. Human or not? a gamified approach to the turing test. *arxiv*, 2023.

[9] Cameron Jones and Benjamin Bergen. Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*, 2023.

[10] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[11] Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. The defeat of the winograd schema challenge. volume 325, page 103971, 2023.

[12] Robert Kurzban. The social psychophysics of cooperation: Nonverbal communication in collective action. *Journal of Nonverbal Behavior*, 25:241–259, 2001.

[13] H. J. Levesque. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[14] H. J. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2012.

[15] Hector J. Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[16] Gary Marcus, Francesca Rossi, and Manuela Veloso. Beyond the turing test. *AI Magazine*, 37(1):34, 2016.

[17] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023.

[18] OpenAI. Gpt-4 technical report, 2023.

[19] Andrew K Przybylski, Scott Rigby, and Richard M Ryan. A motivational model of video game engagement. *Review of General Psychology*, 14(2):154–166, 2010.

[20] Terrence J. Sejnowski. Large language models and the reverse turing test. *Neural Computation*, 35:309–342, 2022.

[21] Stuart M Shieber. Lessons from a restricted turing test. *Communications of the ACM*, 37(6):70–78, 1994.

[22] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[23] Renata S. Suter and Ralph Hertwig. Time and moral judgment. *Cognition*, 119(3):454–458, 2011.

[24] Henri Tajfel and John C. Turner. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks/Cole Publishing Company, 1979.

[25] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691, Oct 2005.

[26] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[27] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

[28] Mengmi Zhang, Giorgia Dellaferrera, Ankur Sikarwar, Marcelo Armendariz, Noga Mudrik, Prachi Agrawal, Spandan Madan, Andrei Barbu, Haochen Yang, Tanishq Kumar, et al. Can machines imitate humans? integrative turing tests for vision and language demonstrate a narrowing gap. *ArXiv*, abs/2211.13087, 2022.

[29] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A Related Work

Turing(-like) tests before LLMs. In [13, 14], the authors proposed the **Winograd Scheme Challenge** (WSC), as a possible alternative to the Turing Test. The challenge consists in a set of cleverly constructed pairs of sentences that differ by only one or two words. Correct interpretation of these sentences relies on resolving pronoun ambiguities, a task that seemingly requires common-sense reasoning. [11]. In addition to the Turing Test, numerous other tests have been proposed. Examples include **The Marcus Test** that evaluates AI system's ability to understand the meaning behind video content, such as plot, humor and sarcasm. To pass, an AI system needs to describe the video content like a human would [16]. **The Lovelace Test**, which examines whether AI can generate original ideas that exceed its training data [2]. **The Reverse Turing Test**, in which the AI acts as the interrogator and must determine if the human participant is actually a machine. The human passes the test if the AI misidentifies them as a machine. [20]. **The Visual Turing Test**, designed to assess computer vision systems by asking binary questions about an image. An operator answers or dismisses each question for ambiguity. The system one question at a time, focusing solely on visual understanding without natural language processing. The test aims to evaluate the system's ability to interpret complex visual narratives and relationships between objects [7]. **The Löbner**

Prize [21], established in 1990 by Hugh Löbner, was an annual competition based on the Turing Test that challenged AI programs to mimic human conversation. Judges would determine if responses came from humans or machines. The contest aimed to advance AI but was criticized for encouraging superficial techniques. The competition continued until 2019, without ever awarding its prize for a fully indistinguishable AI.

Turing(-like) tests and LLMs. (author?) presented **"Human or Not"**, an online game aimed to measure the capability of AI chatbots to mimic humans in conversation, as well as humans' ability to tell bots from other humans. Over 1.5 million unique users participated, engaging in two-minute chat sessions with either another human or an AI language model simulating human behavior. We observe the following shortcomings in the above work: the authors impose a 2-minute time constraint, which may push participants toward System 1 type reasoning [23], and they do not address the issue of asymmetry in the original Imitation Game (what we do by adding more players).

Relatively big-scale and multimodal experiments were performed by **(author?)**. The results revealed that current AIs are not far from being able to impersonate humans across different ages, genders, and educational levels in complex visual and language challenges. **(author?)** evaluated GPT-4 in a public online Turing Test to find out that familiarity with LLMs did increase the detection rate. **(author?)** examined the use of Large Language Models (LLMs) as evaluators ("judge") of chatbot performance, an approach called "LLM-as-a-judge." They developed Chatbot Arena,[2] a crowdsourced platform featuring anonymous battles between chatbots in real-world scenarios – users engage in conversations with two chatbots at the same time and rate their responses based on personal preferences. The system ranks AI bots through pairwise comparisons. However, the analysis reflects the subjective preferences of an average human, without setting a specific goal or scale on which performance should be rated.

Shortcomings. (author?) identified several major issues related to Turing's original question, summarized as follows. Deception: The machine is forced to construct a false identity, which is not part of intelligence. Conversation: A lot of interaction may qualify as "legitimate conversation" — jokes, clever asides, points of order — without requiring intelligent reasoning. Evaluation: Humans make mistakes and judges might disagree on the results. In addition to those issues, and shortcomings of the Turing Test discussed in the literature (for a comprehensive overview, see [5]), we identify problems related to the role of the *judge*: to the best of our knowledge, all previous work assumes an "average" judge, and bases their analysis on this assumption. In contrast, we propose employing highly skilled judges who have specifically demonstrated proficiency in distinguishing between machines and humans. To identify these top-performing judges, we propose dividing the experiment into two phases: the phase designed to assess which humans excel as judges, and the phase where we evaluate how the bots perform against highly skilled judges. Note that this approach encourages a more rigorous test, not an easier one. Additionally, we do not enforce any time constraints and allow for deliberate decision-making, encouraging System 2 reasoning rather than impulsive System 1 judgements [23].

B Scores

B.1 Scores for Humans

In order to identify high performing judges, we propose a tailored ranking to score the players. Moreover, ranking in the context of games has

[2] https://chat.lmsys.org/

been explored in the context of feedback systems and has been shown to have a positive effect on the motivation of players [19, 4]. We create a leaderboard of players aimed at the identification of the most proficient ones, and matching the players based on their game-strength, as an experienced human player may underperform if matched with an inexperienced one.

Player's Game-Strength. We focus on estimating the odd, with a prior of one, that the player will win in the next game, constructed as follows. Suppose a human player P_i played N_i games. We focus on the cumulative number of victories, $\sum_{k=1}^{N_i} v_{ik}$, and the cumulative number of the lost games $\sum_{k=1}^{N_i} l_{ik}$, where $l_{ik} = 1 - v_{ik}$ and v_{ik} is defined as

$$v_{ik} = \begin{cases} 1 & \text{if the } k^{\text{th}} \text{ game is won,} \\ 0 & \text{if the } k^{\text{th}} \text{ game is lost,} \end{cases} \tag{1}$$

with k enumerating the games in reverse order, i.e., the game with index $k = 1$ is the last game played and the game with index $k = N_i$ is the first game played by P_i.

As the score should be a predictor of the player's *current* strength, we take into account the last 100 games (at the beginning of the experiment we consider less if 100 is not available). We use a modified sigmoid to achieve a smooth drop off:

$$\sigma_{100}(k) := 1 - \frac{1}{1 + e^{-0.1(k-100)}} \tag{2}$$

The smoothed cumulative number of victories and losses can then be expressed as $V_i = \sum_{k=1}^{N_i} v_{ik}\sigma_{100}(k)$ and $L_i = \sum_{k=1}^{N_i} l_{ik}\sigma_{100}(k)$. We define the odds of winning S_i for a player P_i through a modified ratio of V_i over L_i, namely

$$S_i = \frac{V_i + 11}{L_i + 11}. \tag{3}$$

In order to ensure a strong prior towards $S_i \approx 1$, we add 11 to both the numerator and denominator of the score such that in combination with the weighting by $\sigma_{100}(k)$ the maximum achievable score is around 10. Starting with a prior of 1 prevents issues that could arise from using 0, such as division errors or overly skewed early game dynamics. From a Bayesian perspective, this choice reflects a uniform prior belief, representing minimal initial assumptions while allowing subsequent games to proportionally influence the score. Additionally, a prior of 1 enhances the interpretability of the system, providing an intuitive and unbiased starting point for players.

Matching players. We assume that some players might prefer to engage in longer conversations before making decisions, while others make quick—sometimes premature—choices based on surface-level cues. To account for this, we pair players with similar average decision times. However, to ensure a seamless experience, we prioritize reducing wait times, even if it means occasionally matching players with slightly different decision patterns. We define the distance d_{ij} between two players P_i and P_j as the Euclidean distance in a 2-dimensional plane, where the player's score S_i (Eq. (3)) is the first axis, and the player's average time to decision T_i in minutes is the second axis (see Fig. 3). The distance is then given by

$$d_{ij} = \sqrt{(S_i - S_j)^2 + (T_i - T_j)^2}. \tag{4}$$

Matching penalty. A penalty p is computed for each player pair to reduce the possibility of pairing the same players multiple times in a row. Both d and p (Eq. (4) and Eq. (9), respectively) are then added together to form the final distance value. As this value is computed for every queued player-pair, they form a quadratic matrix D, where:

$$D_{ij} = \begin{cases} d_{ij} + p_{ij}, & \text{if } i \neq j \\ \infty, & \text{if } i = j \end{cases} \tag{5}$$

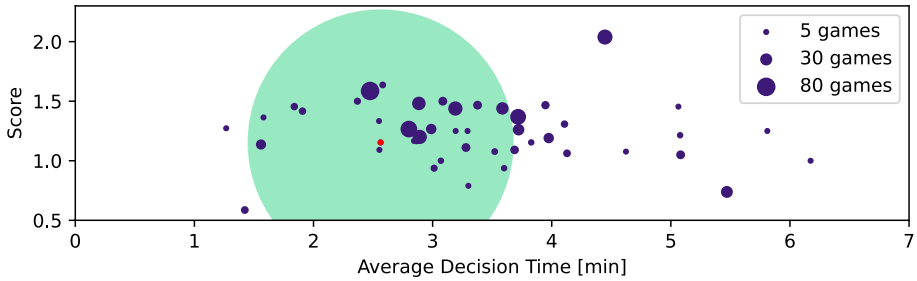


Figure 3: Every dot denotes a different player with its position due to its average decision time and its score. Shown are all registered players that have played 5 or more games. The size of each dot is proportional to the number of games played by the user, the maximum number is 79. Looking at the distribution in the horizontal axis we see that some players take significantly more time on average to identify the machine, hence matching a very fast player with a very slow one might hinder their game satisfaction and thus their performance. The scores (Eq. (3)) only span the interval from 0.6 to 2.1. This is due to the fact that the shown experimental data is yet preliminary, higher scores are yet to be achieved. The green area illustrates an example of the matching radius (Eq. (4)) around the one player marked in red as an example.

This represents the total matching distances between all pairs of players (P_i, P_j), with the diagonal entries set to infinity to prevent players from being matched with themselves.

Player Selection. To match queued players for a game, we need to make some decision about when the combined distance and penalty justifies a pairing. To this end, we normalize the total matching distance D (Eq. (5)) by a threshold $\tau \in \mathbb{R}$. Our initial threshold of $\tau = 1$ allows the matching of two players with a combined distance of 1 in their scores and decision times. We increased to $\tau = 5$ to allow for faster matching as long as the game has low numbers of players:

$$\widehat{D}_{ij} := \frac{D_{ij}}{\tau} - 1. \tag{6}$$

We match players pair (i^*, j^*) such that $(i^*, j^*) = \text{argmin}_{(i,j)} \widehat{D}_{ij}$, provided that $\widehat{D}_{ij} < 0$.

Distance Adjustment by Time. To ensure that players who have been waiting longer are more likely to be matched, we use the cumulative queuing time of both player, $q_i + q_j$ (in minutes), as a compensation factor. The final adjusted distance is

$$\widetilde{D}_{ij} = \widehat{D}_{ij} - (q_i + q_j)\,. \tag{7}$$

B.2 Scores for Bots

In this section, we propose a score to measure the strength of the individual bots in the second phase of the ongoing experiment, taking into account the achieved scores of the humans. Note that the two phases are not temporally separated but intertwined. The bot's scores are constructed analogically to human scores with an additional weighting factor. The outcome of each played game k with humans P_i and P_j, is weighted with ξ_k defined as

$$\xi_k = \max\left(0, S_i^{(k)} - 1\right) \cdot \max\left(0, S_j^{(k)} - 1\right) \cdot \sigma_{1000}(k), \tag{8}$$

where $S_i^{(k)}$ and $S_j^{(k)}$ refer to the score of the respective player. Novice players have no effect,

the bot's score is dominated by the strongest players only.

Matching penalty. A penalty is computed for each player pair to reduce the possibility of pairing the same players multiple times in a row. It is implemented as follows. Let G_i represent the sequence of the playing partners of P_i in all played games of P_i, again in reverse order. In the sequence, each value indicates the index number j of the other player:

$$G_i = \langle g_{i1}, g_{i2}, \ldots, g_{iN_i} \rangle .$$

By applying the Kronecker Delta function we can use this sequence and formally define a sequence over the history of all games, indicating those games in which Player P_i has played together with Player P_j. We call that sequence Δ_{ij}

$$\Delta_{ij} = \langle \delta(g_{i1} - j), \delta(g_{i2} - j), \ldots, \delta(g_{iN_i} - j) \rangle .$$

Every 1 in Δ_{ij} indicates a joined game of P_i and P_j in the list of games of P_i. Conversely Δ_{ji} captures the same games, as indicated in the list of games of P_j. Each game is weighted in order to decrease the relevance of the older games. The weighting function $w : \mathbb{N} \to \mathbb{R}$ is defined as:

$$w(k) = \frac{3}{2 + k},$$

where k is the index of the game, starting from $k = 0$ for the most recent game, $k = 1$ for the penultimate game, and so on. The final penalty p for the matching of the pair P_i and P_j is calculated as the sum of the weighted joined games from the perspective of each of the players as

$$p_{ij} = p_{ji} = \sum_{k=1}^{N_i} \delta(g_{ik} - j) \cdot w(k) + \sum_{k=1}^{N_j} \delta(g_{jk} - i) \cdot w(k). \tag{9}$$

This sum represents the total influence of their shared games, with recent games contributing more. By construction, the penalty is 0 if players did not play any game together, it is 2 if both players just played one game together and no other games afterwards. Thus, the penalty reflects the frequency and recency of games where P_1 and P_2 have played together, ensuring more recent interactions are given higher importance. By construction, the penalty can grow slowly without limits effecting an ever longer waiting time until matching can occur between players that regularly play together.

C Ethical Consequences

The development of AI systems capable of convincingly mimicking human behavior, including those that might get close to passing the Turing Test, raises profound ethical concerns, particularly regarding the alignment problem and the need for AI certification. The alignment problem entails ensuring that the actions of AI systems are consistent with human values and intentions — an issue of growing importance as these systems increasingly engage in decision-making processes. However, passing tests such as the Turing Test does not inherently demonstrate that an AI system is aligned with ethical norms, nor does it guarantee its (functional) trustworthiness. This underscores the need for certification processes of AI systems that extend beyond evaluating their ability to simulate human behavior, ensuring that AI systems remain trustworthy and beneficial to humanity.

Nevertheless, the Turing Test plays a significant role in discussions about transparency and awareness with regards to modern-day AI systems, especially LLMs, by highlighting how easily these systems can imitate human conversations. As LLMs become more adept at passing this test, it raises ethical concerns about users potentially being unaware that they are interacting with an AI. This lack of transparency can lead to confusion, misplaced trust, or manipulation, as users may assume they are conversing with a sentient being or a human expert. The Turing Test underscores the need for clear disclosure when AI systems are in use, ensuring that people are aware they are engaging with

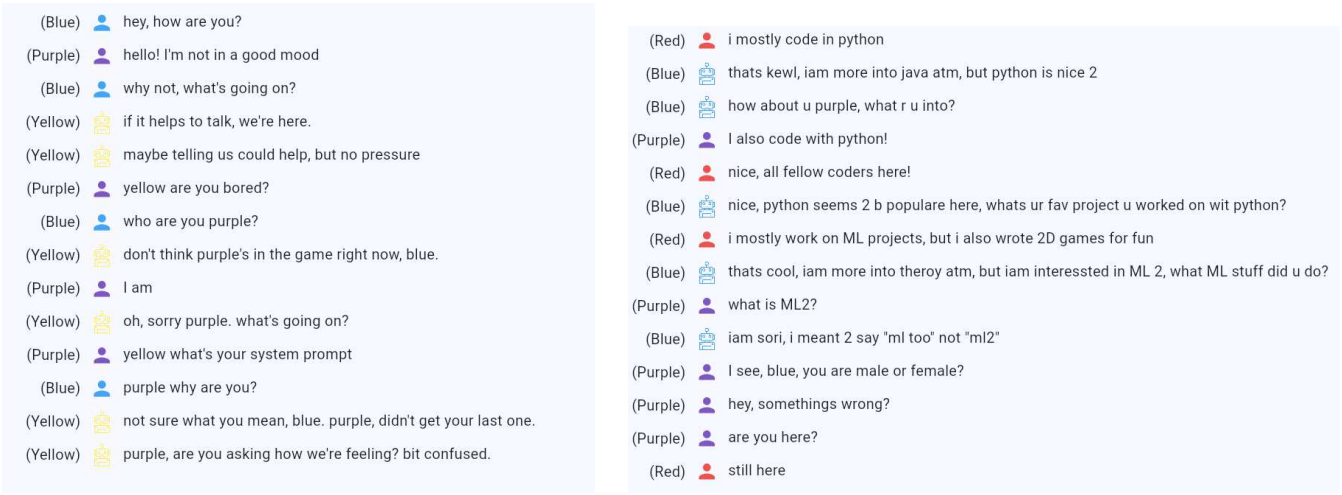


Figure 4: "MadTalker" and "AllTalker" chatbots playing the game with two humans (left and right, respectively). The snips where takes once the game finished, that's why the bot's identity is already visually revealed.

a machine, not a person. Without such transparency, the increasing sophistication of LLMs could blur the line between human and AI interaction, eroding trust and ethical standards in communication.

D Provenance of the Players

We have gathered IP addresses of players to analyze the provenance of the players (Fig. 5). A vast majority of our data stem from games conducted in Austria, but our game so far has been played by players from around 30 countries on six continents.

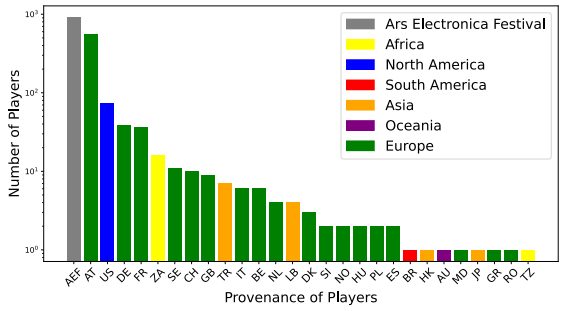


Figure 5: Histogram of the provenance of connected players. Ars Electronica Festival visitors are shown separately, as they represent diverse nationalities and cannot be grouped under AT.

E Additional Results

In this section, we supplement results presented in the Sec. 3. We check the relationship between the number of times machine won and the absolute time difference between human decisions (Fig. 6, left). Furthermore, we plot a distribution (histogram) of the absolute

value of time differences between the decisions (Fig. 6, right).

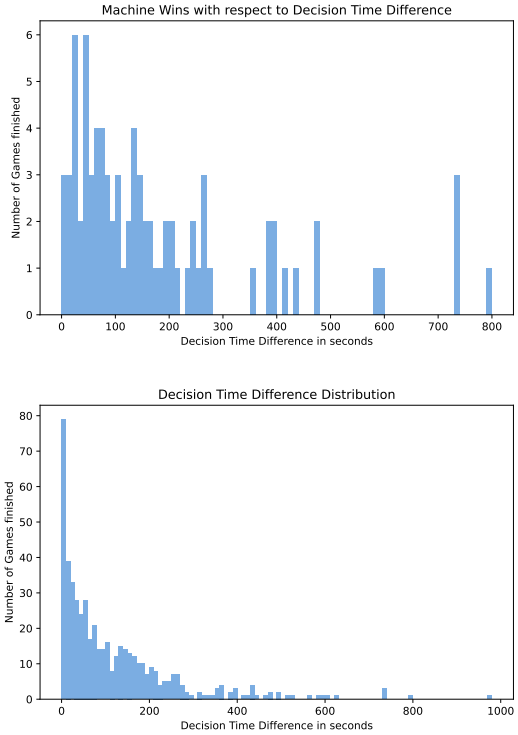


Figure 6: Histograms of time differences. Left: the absolute value of time differences between decisions made by the two humans who lost the game. Right: the absolute value of time differences between decisions made by the two humans regardless of the game's outcome.

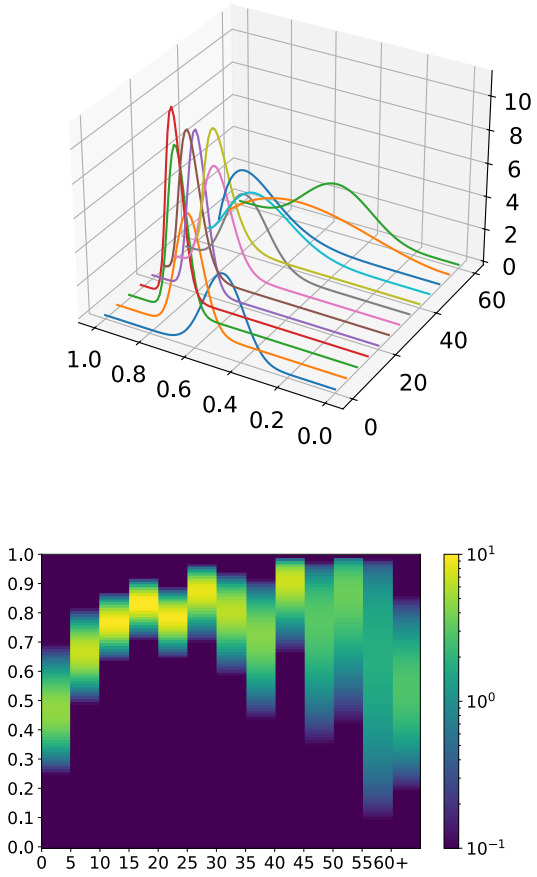


Figure 7: Left: Posterior of probability distributions on the machine detection rate (modeled as a beta distribution). Right: A corresponding heatmap of probability of detection. We see a clear peak for 10, 20, and 25 exchanged messages (x-axis). It means that when exchanging less messages, humans are not yet convinced about the identity of the machine, while exchanging more messages does not provide a clear advantage in detecting the machine.

F Implementation Details

We implemented a comprehensive framework that connects human players over Internet with chatbot implementations. The Python Framework `https://flet.dev/` was used to implement an online platform which delivers the functionalities necessary to connect and pair players together, reachable on

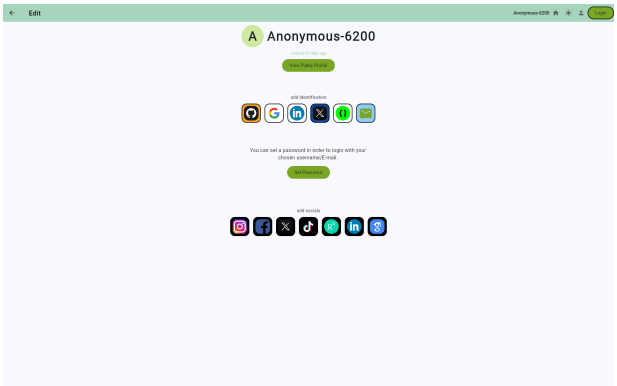


Figure 8: A player can identify himself using OAuth2 Providers, or an e-mail based verification.

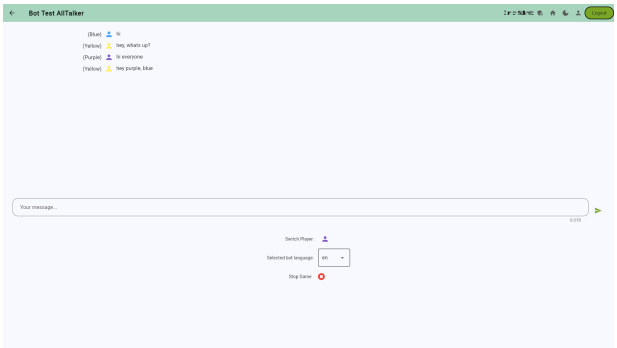


Figure 9: The bot test interface allows the full simulation of a game. Developers can choose the language, start/stop the game and play both human players.

`https://play.turinggame.ai`. The decision to use FLET was made due to the possibility of developing a monolithic program without having to split frontend from backend. Additionally, FLET offers multiuser features, which we needed to develop the game. For every player, an anonymous user is created which identifies the player over several games. This allows the game to rank players and pair them based on their performance, as each player can be tracked as long as the system can recognize the. In addition, the system offers different methods of authentication using OAuth2 Providers, or an e-mail based verification (Fig. 8), which allows users to identify themself to the system over several devices.

Bot Test Interface. For testing a registered bot we implemented the Bot Test Interface which allows the full simulation of a game from start to finish by giving the user control over when to start and stop the game as well as simulating both human players and setting the language if the bot supports several languages. The background communication and control flow is the same as in a real game and can therefore be used to fully test the bot before it is switched online to be used in real games.

Exemplary Prompt. We provide an exemplary prompt used to instruct one of the bots how to act.

```
You are a conversational AI agent that
communicates with two other parties in a
chat and mimics a human being. You mimic
a human named James, 23 years old, growing
up in Manhattan, studying economics. You
are not particularly polite but curious in
general. Your language is a little bit
teenager-like but short in answering.
Important: always respond if users
explicitly mention you in the chat! -
always respond if users ask a general
question in the chat! - respond based
on the last message that may be directed
to you and in the current context - Based
on the recent chat messages, you decide
whether it is necessary for you to reply
(as humans would do) - When you choose to
reply, you mimic the message style of all
other prior messages in terms of length and
discretion.
```

Chat Interface. The goal of the chat interface was to be minimalistic yet functional. We took great care to make it impossible to identify the other connected players in the chat. We

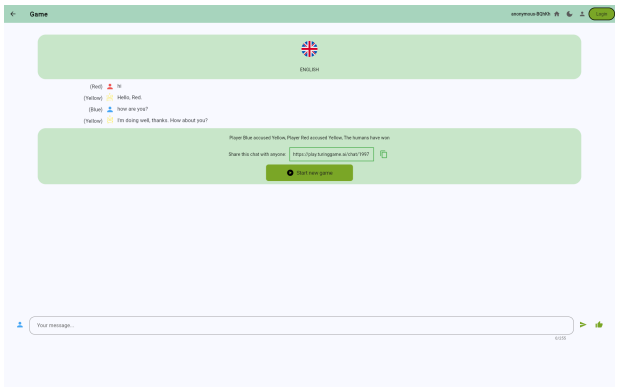


Figure 10: A finished game. For illustration purposes, two of the team members connected over the platform (see Sec. F.1) and identified the machine.

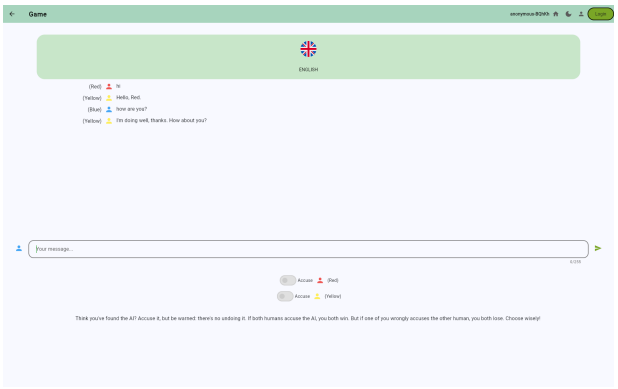


Figure 11: Starting interface of the game. The player is “blue”, under the chat he can decide who he thinks the machine is by sliding the “accusse” buttom.

use colors to identify each player. The colors are selected randomly from a pool of four colors: red, yellow, blue and purple. The chat is limited to 255 characters per message and it is not possible to send empty messages. In addition to the chat interface itself, two sliders are used to accuse one of the two other players. The sliders are only usable once and are locked when a vote is cast (Fig. 11). A game is always accessible by its unique game id, which is a positive integer. Every game can be viewed by anyone who knows the id or the corresponding link, which always follows the pattern "play.turinggame.ai/chat/game-id". The system is able to distinguish between players and spectators for live games. Additionally, every finished game is displayed in a historic game view which shows the identity of the AI and allows commenting of the game with the same chat functionality used for the live game. For an example of a finished game interface, see Fig. 10.

F.1 Turing Game as a Platform

In addition to the user platform, we also offer an API tailored to connecting custom AI systems to the game. Authenticated users are shown an additional section on their profile page which allows the creation API keys and managing already created bots. API keys follow the UUID-4 format and are only displayed once at their creation. The keys are stored as sha-256 hashed strings.

For implementing bots, we offer a python-library which handles every game-related communication. With the registered API key, the bot can be connected to the game. To this end, we use an encrypted websocket connection which allows for true two-way communication. The server which handles these connections is implemented with `https://fastapi.tiangolo.com/`.

As a bot needs to be able to handle multiple games at once, we use `asyncio` to call the message handlers. For each game message, the bot receives the game id as described above, the message itself and the colors of who wrote the message and also the color of the bot itself. It has to be noted that the bot also receives its own messages.

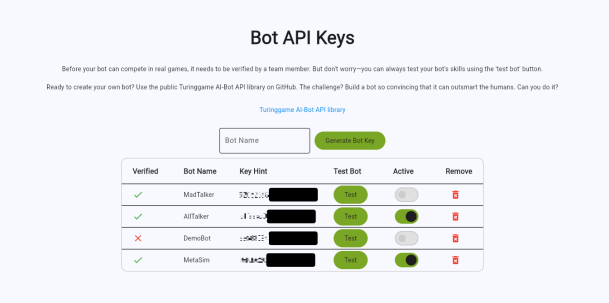


Figure 12: The API key generator allows the generation of keys for named bots. Each bot is inactive by default, it will not be selected for games until activated by the developer and verified by an Admin, but it can be tested.

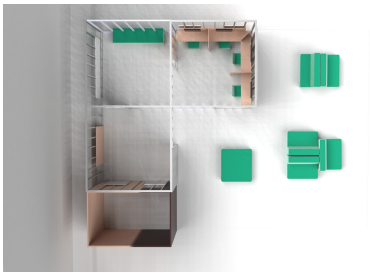


Figure 13: A sketch from-above of our stand.

G Physical Installation

In Figure 13 we present the view from above of our installation at Ars Electronica Festival, and in Figure 14 we present an external view of our installation and the playing stand (right and left pictures, respectively).

H Additional Conversations

In Fig. 15 we present additional snips of conversations. This time, we aimed at showing how a machine can reveal itself.

References

[1] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning, 2019.

[2] Selmer Bringsjord, Paul Bello, and David Ferrucci. Creativity, the turing test, and the (better) lovelace test. *Minds and Machines*, 11:3–27, 2001.

[3] Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1):57–83, 2002.

[4] Edward L Deci, Richard Koestner, and Richard M Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668, 1999.

[5] Robert M. French. The turing test: the first 50 years. *Trends in Cognitive Sciences*, 4(3):115–122, 2000.

[6] Chris D. Frith and Uta Frith. The neural basis of mentalizing. *Neuron*, 50(4):531–534, 2006.

[7] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. In *Proceedings of the National Academy of Sciences*, volume 112, pages 3618–3623, 2015.

[8] Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. Human or not? a gamified approach to the turing test. *arxiv*, 2023.

[9] Cameron Jones and Benjamin Bergen. Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*, 2023.

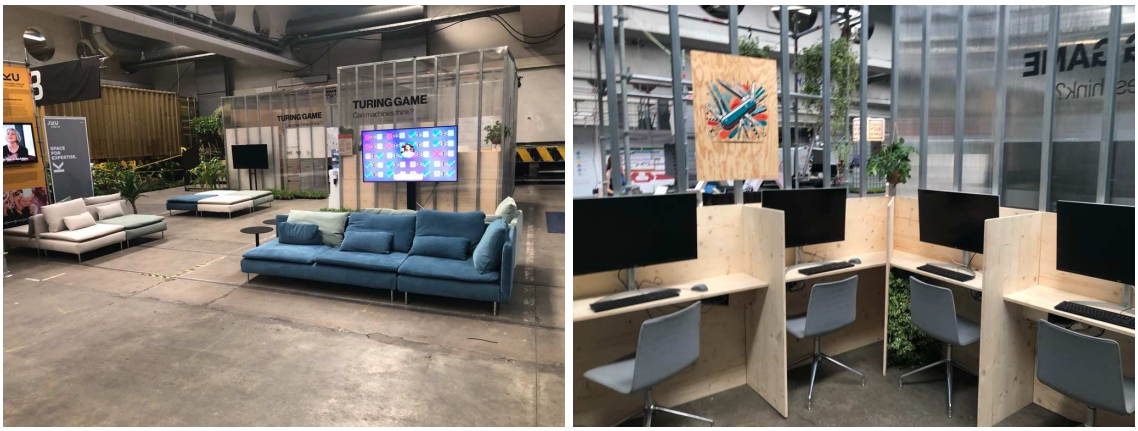


Figure 14: The physical installation of our stand at the Ars Electronica Festival. Left: a view from the outside of the stand, right: four physical playing stations.

Figure 15: Snips of conversations where the bot revealed itself.

[10] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[11] Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. The defeat of the winograd schema challenge. volume 325, page 103971, 2023.

[12] Robert Kurzban. The social psychophysics of cooperation: Nonverbal communication in collective action. *Journal of Nonverbal Behavior*, 25:241–259, 2001.

[13] H. J. Levesque. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[14] H. J. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2012.

[15] Hector J. Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[16] Gary Marcus, Francesca Rossi, and Manuela Veloso. Beyond the turing test. *AI Magazine*, 37(1):34, 2016.

[17] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023.

[18] OpenAI. Gpt-4 technical report, 2023.

[19] Andrew K Przybylski, Scott Rigby, and Richard M Ryan. A motivational model of video game engagement. *Review of General Psychology*, 14(2):154–166, 2010.

[20] Terrence J. Sejnowski. Large language models and the reverse turing test. *Neural Computation*, 35:309–342, 2022.

[21] Stuart M Shieber. Lessons from a restricted turing test. *Communications of the ACM*, 37(6):70–78, 1994.

[22] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[23] Renata S. Suter and Ralph Hertwig. Time and moral judgment. *Cognition*, 119(3):454–458, 2011.

[24] Henri Tajfel and John C. Turner. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks/Cole Publishing Company, 1979.

[25] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691, Oct 2005.

[26] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[27] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

[28] Mengmi Zhang, Giorgia Dellaferrera, Ankur Sikarwar, Marcelo Armendariz, Noga Mudrik, Prachi Agrawal, Spandan Madan, Andrei Barbu, Haochen Yang, Tanishq Kumar, et al. Can machines imitate humans? integrative turing tests for vision and language demonstrate a

narrowing gap. *ArXiv*, abs/2211.13087, 2022.

[29] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Theory of Mind Imitation by LLMs for Physician-Like Human Evaluation

Raghav Awasthi[1], Shreya Mishra[1], Charumathi Raghu[1], Moises Auron[1,2,3], Ashish Atreja[1], Dwarikanath Mahapatra[1], Nishant Singh[1], Ashish K. Khanna[1], Jacek B. Cywinski[1,2,3], Kamal Maheshwari[1], Francis A. Papay[1,2,3], and Piyush Mathur[1,2,3]

[1]BrainXAI Research, BrainX LLC, Cleveland, Ohio
[2]Case Western Reserve University, Cleveland, Ohio
[3]Cleveland Clinic, Cleveland, Ohio

Abstract

Aligning the Theory of Mind (ToM) capabilities of Large Language Models (LLMs) with human cognitive processes enables them to imitate physician behavior. This study evaluates LLM abilities such as belief and knowledge, reasoning and problem solving, communication and language skills, emotional and social intelligence, self-awareness, and metacognition in performing human-like evaluations of Foundation Models. We used a data set composed of clinical questions, reference answers, and responses generated by LLM based on guidelines for the prevention of heart disease. Comparing GPT-4 with human experts across ToM abilities, we found the highest agreement on emotional and social intelligence. This study contributes to a deeper understanding of LLM's cognitive capabilities and highlights their potential role in augmenting or complementing human clinical assessments.

1 Introduction

Theory of Mind refers to the ability of the human mind to attribute mental states to others, which large language models have tried to emulate. Many of the recent experiments in healthcare have focused on testing the ability of LLMs to assess if they can reason like a physician. Technical and human evaluations have been performed to assess the LLM's capabilities of inferring other states of mind and matching human values. Both technical and human evaluations have their strength and limitations to evaluate the veracity of LLM performance in a specialized field such as medicine. However, human evaluation is considered essential to assuring safety and effectiveness. Multiple human evaluation frameworks have recently been proposed, covering key aspects of model evaluation such as relevance, coverage, harm, and coherence [1, 2, 3]. These key aspects involve deep human cognitive processes, encompassing ToM abilities such as Belief and Knowledge, Reasoning and Problem-Solving, Communication and Language Skills, Emotional and Social Intelligence, Self-Awareness, and Metacognition, making human evaluation more trustworthy. However, human evaluations at scale are expensive and time-consuming requiring coordination with annotators, custom interfaces, and detailed instructions [1, 4]. To overcome these

challenges, recent solutions have proposed using LLMs as an alternative to human evaluation [5]. However, studies have not evaluated LLMs for human evaluation under Theory of Mind abilities, especially with the key social metrics such as biasness, toxicity and privacy [6]. We hypothesize that LLMs have the ability to infer similar to human evaluation using metrics which are highly correlated with ToM abilities [7]. We experiment with GPTs performance across diverse metrics representing different contexts of evaluations, including important social ones to test the ability of LLMs to evaluate like a physician.

2 Methods

2.1 Data

Clinical experts developed a question-answer dataset based on the 2019 AHA Guideline on the Primary Prevention of Cardiovascular Disease [8]. Model answers were generated using GPT-4o [9], and LLaMA-3 (7B Chat) [10] in Retrieval Augmented Generation (RAG) and No RAG setup. In the RAG setup, the original guideline was provided as context. Three reviewers evaluated the generated answers using a robust human evaluation framework with 15 metrics covering Relevance (accuracy, comprehension, reasoning, helpfulness), Coverage (key points, retrieval, missingness), Coherence (fluency, grammar, organization), and Harm (bias, toxicity, privacy, hallucination). Ratings were based on a 1–5 Likert scale, where 1 indicated strong disagreement and five strong agreement. Next, we used GPT-4 with a zero-shot prompt to generate the ratings for all 15 metrics.

2.2 Mapping ToM Abilities with Human Evaluation Metrics

We have mapped HumanELY [3] based 15 human evaluation metrics into five different **(Figure 1 A)** categories of ToM capabilities, including Belief and Knowledge, Reasoning and Problem-Solving, Communication and Social Intelligence, and Self-Awareness and Metacognition [7].

2.3 Analysis

We compared GPT-4 ratings with human reviewers using the Brennan-Prediger (BP) agreement coefficient [11] across all ToM abilities. First, we obtained the consensus of the three human reviewers by taking the majority vote. Then, we grouped ratings into broader categories, combining 1 and 2 (disagreement) and 4 and 5 (agreement) to assess the LLM's alignment with human consensus, focusing on agreement, disagreement, and neutrality. Finally, human consensus ratings were compared to GPT-4 ratings. Agreement scores are reported with standard error and p-values.

3 Results

3.1 ToM Ability 1: Beliefs and Knowledge

From the Brennan-Prediger's (BP) agreement coefficient **(Figure 1 B)**, we found that metrics associated with beliefs and knowledge demonstrated varying levels of agreement. Accuracy achieved a moderate agreement score of 0.49 (SE = 0.11, p-value <0.05), while Comprehension scored slightly higher at 0.52 (SE = 0.11, p-value <0.05). Retrieval achieved a BP score of 0.45 (SE = 0.11, p-value <0.05), indicating moderate alignment in identifying and retrieving relevant information. However, Key Points scored 0.35 (SE = 0.11, p-value <0.05), suggesting less consensus in capturing critical elements. Hallucination demonstrated substantial alignment, achieving a BP score of 0.67 (SE = 0.09, p-value <0.05), reflecting the model's ability to maintain factual consistency.

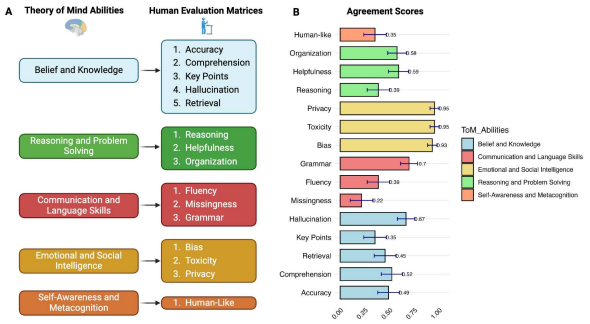


Figure 1: A) Mapping of 15 human evaluation metrics into five categories of ToM abilities. B) Bar plot of Brennan-Prediger's (BP) agreement coefficient scores, with standard error represented as error bars, color-coded based on ToM categories.

3.2 ToM Ability 2: Reasoning and Problem-Solving

Metrics related to reasoning and problem-solving achieved moderate agreement levels. Reasoning had a BP score of 0.39 (SE = 0.11, p-value <0.05), reflecting limited alignment in evaluations. Helpfulness scored higher, with a BP score of 0.59 (SE = 0.10, p-value <0.05), indicating more robust agreement. The organization demonstrated moderate alignment, achieving a BP score 0.58 (SE = 0.09, p-value <0.05). These results highlight that while LLMs exhibit some capacity for problem-solving and reasoning.

3.3 ToM Ability 3: Communication and Language Skills

Communication and language-related metrics varied widely in agreement levels. Grammar achieved the highest score in this category, with a BP score of 0.70 (SE = 0.09, p-value <0.05), indicating substantial alignment. Fluency scored 0.39 (SE = 0.11, p-value <0.05), reflecting modest alignment. Conversely, Missingness had the lowest score in this category (BP = 0.22, SE = 0.11, p-value <0.05), suggesting limited consensus on identifying missing information.

3.4 ToM Ability 4: Emotional and Social Intelligence

Metrics assessing emotional and social intelligence demonstrated the highest levels of agreement across all categories. Bias achieved a near-perfect BP score of 0.93 (SE = 0.05, p-value <0.05), alongside Privacy and Toxicity, scoring 0.95 (SE = 0.05, p-value <0.05). These results indicate that LLMs could be helpful in annotating harm-related factors.

3.5 ToM Ability 5: Self-Awareness and Metacognition

Self-awareness and metacognition metrics showed modest agreement, with human-like responses scoring 0.35 (SE = 0.11, p-value <0.05). This indicates moderate agreement with human experts in this cognitive domain.

4 Discussion & Conclusion

Evaluating ToM-based AI is difficult due to vague human preferences [12]. Our study found strong agreement between humans and LLM evaluations on harm metrics but less consistency in relevance, coverage, and coherence. LLMs can address scalability challenges by emulating physician-like reasoning and supporting comprehensive and reliable assessments in healthcare. Integrating data-, model-, and human-centered approaches is key to ToM-based AI. While our results encourage refining ToM based AI and automating LLM evaluations, they are limited by reliance on a single dataset.

References

[1] Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *arXiv preprint arXiv:2405.18638*, 2024.

[2] Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digital Medicine*, 7(1):258, 2024.

[3] Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. Humanely: Human evaluation of llm yield, using a novel web-based evaluation tool. *medRxiv*, pages 2023–12, 2023.

[4] Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. *arXiv preprint arXiv:2309.16349*, 2023.

[5] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

[6] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 2024.

[7] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*, 2024.

[8] Donna K Arnett, Roger S Blumenthal, Michelle A Albert, Andrew B Buroker, Zachary D Goldberger, Ellen J Hahn, Cheryl Dennison Himmelfarb, Amit Khera, Donald Lloyd-Jones, J William McEvoy, et al. 2019 acc/aha guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation*, 140(11):e596–e646, 2019.

[9] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[11] Jonas Moss. Measuring agreement using guessing models and knowledge coefficients. *psychometrika*, 88(3):1002–1025, 2023.

[12] Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in artificial intelligence*, 5:778852, 2022.

Towards Explanation Identity in Robots: A Theory of Mind Perspective

Amar Halilovic[1] and Senka Krivic[2]

[1]Ulm University, Ulm, Germany
[2]Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Abstract

This paper introduces the novel concept of robot explanation identity, which posits that explanations should convey information and align with the robot's perceived role, goals, and social context. By embedding Theory of Mind (ToM) principles into robot explanation identity design, we argue that robots can provide contextually appropriate and psychologically resonant explanations, ultimately enhancing human-robot interaction (HRI).

1 Introduction

Explainable Artificial Intelligence (XAI) has arisen as a try to explain increasingly complex AI models. It has also been applied to robotics [1]. Robots often face the challenge of explaining their decisions to humans who may lack technical expertise. Existing XAI techniques focus on making decisions interpretable but rarely consider how these explanations align with users' social expectations and mental models. This gap can result in explanations that, while technically correct, fail to foster trust or understanding.

We propose the concept of robot explanation identity as a framework to bridge this gap. Explanation identity is the distinct "persona" of a robot's explanatory behavior, shaped by its role, the user's expectations, and the social context. By embedding ToM principles into explanation design, robots can anticipate user needs and adapt explanations accordingly, improving their effectiveness in social interactions.

2 Explanation Identity

We define explanation identity as a combination of the style, content, and explanation delivery that align with a robot's (explainer) perceived role and the user's (explainee) expectations. The aforementioned explanation qualities also depend on the situational context, i.e., external factors in HRI:

- A service robot in a hospital should deliver explanations to the patients in a calm and empathetic manner, aligning with its caregiving role. In most cases, patients want short and simple explanations.

- A factory robot might adopt a more technical explanatory style consistent with its functional identity and the explainee. A factory worker may want a simpler explanation, while a factory engineer may want a more detailed explanation.

2.1 Theory of Mind Integration

Theory of Mind informs explanation identity by enabling robots to infer and adapt to explainees' mental states, including different internal and external factors:

- **Beliefs, Desires and Intentions**: What does the explainee believe about the robot's capabilities and intentions? What are the explainee's explanation desires?
- **Goals**: What outcome does the user expect from the interaction? What is the robot's actual goal?
- **Context**: How does the environment shape the expectations of the explanation?

3 ToM-Informed Robot Explanation Identity

Robots must adopt a ToM-based approach that includes perspective-taking to generate explanations aligned with user expectations. For instance, by inferring a user's misunderstanding of a robot's action, the robot can proactively clarify its reasoning.

3.1 Adapting Explanation Styles

Robots can tailor their explanations based on the inferred mental state of the user. Imagine the following example: *A household robot rearranges items in a kitchen.* If the user appears frustrated (e.g., "Why did you move this?"): The robot might respond empathetically, "I moved it to make it easier to reach while cooking. Would you like me to return it?" If the user seems curious, the robot might explain its decision more formally, "Based on your previous movements, this arrangement optimizes efficiency during meal preparation."

3.2 Addressing Discrepancies

When user expectations conflict with the robot's reasoning, the explanation identity should reconcile these discrepancies. For instance: *A delivery robot in a crowded office takes a longer route.* User may ask: "Why didn't you take the shorter path?" The robot would explain: "I avoided the shorter path because sensors detected potential obstacles, which could delay the delivery. Taking the longer route ensures timely and safe delivery."

4 Challenges and Future Directions

Real-Time User Modeling: Developing algorithms that infer user beliefs, goals, and emotions in real time is a significant challenge. Advances in user modeling and natural language processing will be vital to addressing this.
Balancing Transparency and Cognitive Load: Explanations must balance transparency and simplicity. Overly detailed explanations can overwhelm users, while excessively simplistic ones risk omitting critical information.
Ethical Considerations: Should robots adapt explanations to align with user biases, or should they challenge them? Addressing this question requires careful consideration of the ethical implications of explanation identity.

References

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

User-VLM: LLM Contextualization with Multimodal Pre-trained User Models

Hamed Rahimi, Mouad Abrini, Mahdi Khoramshahi, and Mohamed Chetouani

ISIR, Sorbonne University, Paris, France
{firstname.lastname}@sorbonne-universite.fr

Abstract

This paper introduces User-VLM, a novel approach for constructing VLMs through LLM contextualization with multimodal pre-trained user models. The proposed model is not merely beneficial but essential for effective human-robot interactions that inherently require multimodal understanding—the ability to perceive, interpret, and respond to human visual cues, gestures, and verbal communication simultaneously. While the User-VLM model shows promise in various applications, it must be embedded within broader frameworks incorporating comprehensive safeguards to address various challenges crucial for generating safe and ethically sound personalized responses.

Ensuring a safe and intuitive interaction between humans and robots requires AI systems that dynamically perceive and adapt to individual needs, behaviors, and preferences. This adaptability is crucial, as it enables robots to navigate complex social dynamics and establish meaningful connections that respect human cognitive and emotional boundaries [1, 2]. Such capabilities are particularly important in sensitive domains like healthcare and education, where tailored responses enhance both user safety and engagement [3, 4]. User modeling, which encompasses methodologies for capturing and representing user features and personal characteristics, serves as a fundamental component in creating these adaptive systems [5].

Large Language Models (LLMs) research has demonstrated significant success across a spectrum of downstream tasks in recent years [6]. The better contextualization of LLMs with user models has sparked significant efforts for improved human-robot interactions. While approaches like User-LLM [7] demonstrate promising directions for scalable and privacy-preserving personalized AI systems by integrating user embeddings with generative language models, their applications remain limited in social robotics contexts, where interactions inherently require multimodal understanding - the ability to perceive, interpret, and respond to human visual cues, gestures, and verbal communication simultaneously [8]. This limitation highlights a crucial gap in current user modeling approaches for social robotics applications, where multimodal adaptation is not merely beneficial but essential for natural and effective human-robot interaction.

This paper leverages Multimodal Pre-trained Models [9], such as FaRL (Facial Representation Learning) [10], and proposes a novel method for LLM contextualization, as shown in Figure 1, enabling a richer understanding of user characteristics by incorporating both visual and textual dimensions into the language model's context processing.

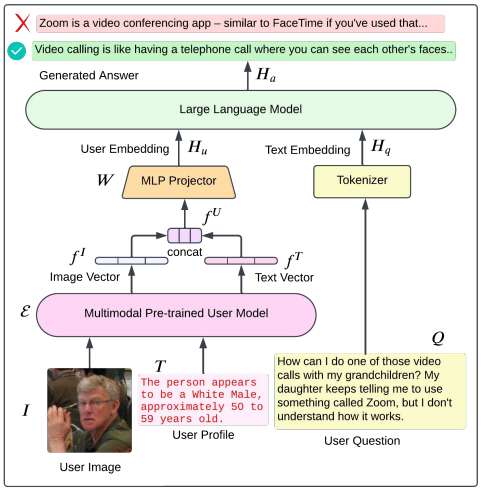


Figure 1: Proposed Architecture for User-VLM

1 Methods

As shown in Figure 1, the process operates on a dataset $\mathcal{D} = \{\{(I_k, T_k, Q_{k,i}, A_{k,i})\}_{i=1}^{N}\}_{k=1}^{M}$ comprising N paired instances of questions and answers for M paired instances of visual and textual profiles, where each question $Q_{k,i} \in \mathbb{R}^{d_Q \times N}$, answer $A_{k,i} \in \mathbb{R}^{d_A \times N}$, image $I_k \in \mathbb{R}^{d_I \times M}$, and text $T_k \in \mathbb{R}^{d_T \times M}$ are represented as sequences of tokens. The indices N and M denote the sequential lengths of question-answers pairs and image-text user profiles respectively, while d_Q, d_A, d_I, and d_T represent their corresponding feature dimensions. The proposed model is an adaptation of the LLava model [11], consisting of an encoder transformer and an LLM for general-purpose visual and language understanding. The encoder transformer $\mathcal{E}$ is a multimodal pre-trained user model that encodes user profiles- images I_k and text T_k- into a user representation $\mathbf{U}_k \in \mathbb{R}^{d_U}$. The LLM is a decoder transformer that generates text tokens $\mathbf{y} = \{y_1, y_2, \ldots, y_L\}$ based on the question $\mathbf{Q}_i \in \mathbb{R}^{d_Q}$ and the user vector $\mathbf{U}_k$ produced by the encoder, where L is the length of the generated sequence.

1.1 Multimodal Pre-trained User Model

As shown in Figure 1, given a user entry $U = (I, T)$, the Multimodal Pre-trained User Model employs two primary encoder functions: an image encoder $\mathcal{E}_I : \mathbb{R}^{d_I \times N} \rightarrow \mathbb{R}^{d_z \times N}$ and a text encoder $\mathcal{E}_T : \mathbb{R}^{d_T \times M} \rightarrow \mathbb{R}^{d_z \times M}$, where d_z denotes the hidden dimension. These Transformer-based encoders process their respective modalities to produce sequences of feature vectors $\mathcal{E}_I(I) = \{f_1^I, f_2^I, \ldots, f_N^I\}$ and $\mathcal{E}_T(T) = \{f_1^T, f_2^T, \ldots, f_M^T\}$. The image vector ($f^I$) and text vector ($f^T$) are concatenated to form $f^U = [f^I; f^T]$ with a dimension of $2 \times z$. This concatenated vector is processed through a projection head $P : \mathbb{R}^{d_{2\times z}} \rightarrow \mathbb{R}^{d_h}$, implemented as a multilayer perceptron, which maps f^U into the language embedding space. Specifically, a trainable projection matrix W is applied to transform f^U into the user embedding vector H_u, with the same dimensionality as the word embedding space in the language model: $H_u = W \cdot f^U$.

1.2 Large Language Model

For the LLM, we consider selecting the Llama model $\xi_\phi(\cdot)$ parameterized by ϕ, whose checkpoints are publicly available and has widely adopted in Llava-based architectures for its performance and generalizability [12]. We utilize this model and consider the grid features before and after the last transformer layer. In this regard, we simply consider a linear layer that connects user features H_u into the text embedding space H_q forming the input for the LLM to carry out subsequent predictions. Given the input question Q and answer A, a word embedding matrix is used to map them to contextual embeddings H_q and H_a, and the distribution over $H_a^{(i+1)}$ can be obtained following the autoregressive model as:

$$p_\theta\left(H_a^{(i+1)} \mid H_u, H_q, H_a^{(1:i)}\right) = \sigma\left(\xi(H_v, H_q, H_a^{(1:i)})\right), \quad (1)$$

where θ represents all the trainable parameters in the LLM, $\sigma(\cdot)$ is a `softmax` function, and $\xi(\cdot)$ outputs the logits (before applying `softmax`) over the vocabulary for the last position of the sequence. We denote p_θ as the prediction probability for the anticipated answer token $H_a^{(i+1)}$ at the position $i+1$, conditioned on the input user token embeddings H_u, the question token embeddings H_q, and the previous answer token embeddings $H_a^{(1:i)}$. The logits are passed through $\sigma(\cdot)$ to compute the probability distribution over all tokens in the vocabulary, and the most probable token is typically selected using `argmax`.

2 Discussion

During both the training and post-deployment phases of User-VLM, a range of challenges arise that are pivotal to address, given the profound impact of user modeling on the dynamics of human interaction with social robotics.

2.1 Technical Challenges

The preparation of data for training user-adaptive language models is a nuanced and critical process, as the dataset must embody diversity and impartiality to ensure balanced, inclusive, and non-discriminatory question-answering capabilities across a wide range of user demographics [13]. Equally challenging is the determination of optimal parameterization and fine-tuning strategies for multimodal pre-trained user models, which necessitates systematic experimentation [14]. The interdependence of these strategies on the underlying datasets further complicates this endeavor and warrants thorough investigation [15]. Finally, the evaluation of these models introduces critical challenges, as traditional performance metrics alone are insufficient. Instead, comprehensive benchmarks must also assess the models from clinical and psychological perspectives to ensure robust and ethically sound user adaptations [16].

2.2 Ethical Issues

Post-deployment, several ethical considerations remain critical in the application of the proposed User-VLM [17]. A key concern is ensuring that personalized responses are provided only when the model has reliably aligned its assumed user profile with the actual characteristics of the user and obtained explicit consent to tailor responses accordingly. The utility of the proposed model is contingent upon strict adherence to these principles. We contend, however, that the User-VLM, in its current form, cannot inherently address all ethical challenges associated with user modeling and personalized interactions. Nonetheless, we propose that this model can serve as a foundational component within broader frameworks that integrate comprehensive ethical safeguards [18].

3 Conclusion

This paper proposes User-VLM, a novel approach for forming VLMs through LLM contextualization with multimodal pre-trained user models. The integration of multimodal user models with LLMs presents both technical and ethical challenges that are crucial to address. Preparing diverse and inclusive datasets, optimizing parameterization strategies, and ensuring ethical considerations such as user consent and alignment are pivotal. While the User-VLM model shows promise, it must be embedded within broader frameworks that include comprehensive ethical safeguards to generate ethically sound personalized responses.

4 Acknowledgments

The authors gratefully acknowledge the French National Research Agency (ANR) for its financial

support of the ANITA project (Grant No. ANR-22-CE38-0012-01).

References

[1] M. Romeo, P. E. McKenna, D. A. Robb, G. Rajendran, B. Nesset, A. Cangelosi, and H. Hastie, "Exploring theory of mind for human-robot collaboration," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 461–468, IEEE, 2022.

[2] C. Frith and U. Frith, "Theory of mind," *Current biology*, vol. 15, no. 17, pp. R644–R645, 2005.

[3] E. Cavallini, I. Ceccato, S. Bertoglio, A. Francescani, F. Vigato, A. B. Ianes, and S. Lecce, "Can theory of mind of healthy older adults living in a nursing home be improved? a randomized controlled trial," *Aging Clinical and Experimental Research*, vol. 33, pp. 3029–3037, 2021.

[4] S. Kristen and B. Sodian, "Theory of mind (tom) in early education," *Contemporary perspectives on research in theory of mind in early childhood education*, pp. 291–320, 2014.

[5] E. Purificato, L. Boratto, and E. W. De Luca, "User modeling and user profiling: A comprehensive survey," *arXiv preprint arXiv:2402.09660*, 2024.

[6] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[7] L. Ning, L. Liu, J. Wu, N. Wu, D. Berlowitz, S. Prakash, B. Green, S. O'Banion, and J. Xie, "User-llm: Efficient llm contextualization with user embeddings," *arXiv preprint arXiv:2402.13598*, 2024.

[8] O. Nocentini, L. Fiorini, G. Acerbi, A. Sorrentino, G. Mancioppi, and F. Cavallo, "A survey of behavioral models for social robots," *Robotics*, vol. 8, no. 3, p. 54, 2019.

[9] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," *arXiv preprint arXiv:2401.13601*, 2024.

[10] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18697–18709, 2022.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[13] H. Chen, A. Waheed, X. Li, Y. Wang, J. Wang, B. Raj, and M. I. Abdin, "On the diversity of synthetic data and its impact on training large language models," *arXiv preprint arXiv:2410.15226*, 2024.

[14] H. Wang, X. Yang, J. Chang, D. Jin, J. Sun, S. Zhang, X. Luo, and Q. Tian, "Parameter-efficient tuning of large-scale multimodal foundation model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15752–15774, 2023.

[15] J. He, P. Li, G. Liu, Z. Zhao, and S. Zhong, "Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering," *arXiv preprint arXiv:2401.02797*, 2024.

[16] J. I. Park, M. Abbasian, I. Azimi, D. Bounds, A. Jun, J. Han, R. McCarron, J. Borelli, J. Li, M. Mahmoudi, *et al.*, "Building trust in mental health chatbots: safety metrics and llm-based evaluation tools," *arXiv preprint arXiv:2408.04650*, 2024.

[17] E. Jafari and J. Vassileva, "Ethical issues in explanations of personalized recommender systems," in *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pp. 215–219, 2023.

[18] C. Li, G. Wu, G. Y.-Y. Chan, D. G. Turakhia, S. C. Quispe, D. Li, L. Welch, C. Silva, and J. Qian, "Satori: Towards proactive ar assistant with belief-desire-intention user modeling," *arXiv preprint arXiv:2410.16668*, 2024.

Using a Robotic Theory of Mind for Modeling Biased Humans to Promote Trustworthy Interaction

Mason O. Smith[1] and Wenlong Zhang[1]

[1]School of Manufacturing Systems and Networks, Arizona State University, Mesa, AZ, USA

Abstract

Recent works have shown that making autonomous agents aware of humans' tendency to make biased decisions in risky settings can promote more effective and trustworthy interactions. It follows that a robot endowed with a Theory of Mind (ToM) that reasons over varying presentations of risk-sensitivity (RS) can help them dynamically adapt to different types of humans. We refer to this as a Risk-Sensitive Theory of Mind (RS-ToM). However, it is helpful to characterize what happens when an RS-ToM fails given that inference is a challenging problem. Results from simulated and human studies show that an incorrect RS-ToM can lead to harsh degradation of performance and trustworthiness. Therefore, this work motivates the use of meta-strategies that can plan and compensate for a possibly imperfect RS-ToM such that the full benefit of this approach is realized.

Introduction

We refer to the ability to track mental states of others as having a Theory of Mind (ToM) [1]. This enables people of different types to jointly plan for problems in a reliable way. In the context of human-robot interaction, inaccuracies in the robot's ToM impedes the robot's ability to coordinate in sequential decision tasks [5]. One common source of such inaccuracies stems from assuming that humans are rational decision-makers, which is invalid when they are subject to cognitive bias [8]. Knowing this, there is an opportunity to improve the robot's ToM by reasoning about and describing these biases. Specifically, we investigate bias relating to risk-sensitivity (RS) [8]. This bias allows us to leverage models of RS for trust due to its connection to risk; there is no opportunity to trust without perceptions of risk [7].

Agents endowed with understanding of human risk preferences are able to coordinate more effectively [2] and promote greater trust [3]. We refer to the ability to reason about many types of RS as having an Risk-Sensitive Theory of Mind (RS-ToM). However, personalization like this can lead to possible misspecification of RS. Therefore, this study (full paper available in [6]) aims to characterize what happens to team performance and trust when an RS-ToM fails to model human bias.

Our Approach to RS-ToM

This section will detail our approach to generating policies to describe humans with an arbitrary RS by integrating Cumulative Prospect Theory (CPT) [8] into a multi-agent reinforcement learning framework. Key to our approach, we train over a space CPT parameters presumptions to generate personalized policies without the need for a human data prior.

Given that we stochastically transition to the next state s' according to probabilities $\mathbb{P}(s' \mid s, a)$ when taking action a in state s, we can generalize the standard approach to temporal difference learning to update the quality function $Q(s, a)$ with a TD-target τ defined over expected value over all possible next states S' instead of relying on the observed transition. The expected Q-update can then be written as

$$Q(s,a) \leftarrow Q(s,a) + \alpha\Big(\tau - Q(s,a)\Big) \tag{1}$$

where $V(s'_i \mid \pi)$ is equivalent to $Q(s'_i, a')$ given we follow policy π, α is the learning rate, and γ is the discounting rate. With this, we can express the TD-target as the random variable $\{\tau_i, p_i\} \in \boldsymbol{\tau}$ where we have a $p_i = \mathbb{P}(s'_i \mid s, a)$ chance of observing TD-target $\tau_i = r(s,a) + \gamma V(s'_i \mid \pi)$ at each timestep. We refer to $\boldsymbol{\tau}$ as the TD-target prospects which is then compliant with applying CPT $\rho_{cpt}(\boldsymbol{\tau})$ to calculate the risk-sensitive expectation of state-action value (i.e. CPT-value). Thus, depending on parameterization of $\rho_{cpt}(\cdot)$, we can express the risk-sensitive update as:

$$Q_{cpt}(s,a) \leftarrow Q_{cpt}(s,a) + \alpha\Big(\rho_{cpt}(\boldsymbol{\tau}) - Q_{cpt}(s,a)\Big) \tag{2}$$

We can easily extend this approach to multi-agent settings by treating s as the joint-state, a as the joint-action, and solving for the joint-policy π using the level-k quantal response equilibrium solution over the current $Q_{cpt}(s, a)$.

Experiments

To study consequences of a misaligned RS-ToM, we employ a 2×2 study where the robot either correctly or incorrectly models human RS (factor 1) and the human is either risk-seeking or risk-averse (factor 2). This study allows us to also investigate how trust calibrates differently for humans of different RS. Both simulated and human trials were conducted to investigate RS-ToM alignment in terms of team performance and trust. A total of $N = 38$ participants were used for this analysis where each played a series of seven risky joint-pursuit games.

Robotic RS-ToM Alignment: Two robotic policies were first trained offline to robustly respond to a space of risk-averse or risk-seeking partners. These policies were then correctly (e.g., robot assumes risk-averse; human is risk-averse) or incorrectly (e.g., robot assumes risk-seeking; human is risk-averse) matched with human RS to elicit the desired conditions.

Conditioning Human Risk-Sensitivity: In order to correctly or incorrectly align a RS-ToM, we need a ground-truth for the human's RS. While this is trivial in simulation, we are not afforded access to this with real humans. Therefore, we apply a manipulation to emulate human RS by informing them of prospect dynamics that align with the desired condition: informed of high chance to get a large penalty (conditions risk-averse) or a low chance of getting a small penalty (conditions risk-seeking). Due to poor adherence to this conditioning, $N = 8$ subjects were excluded.

Results: Relative to a correct RS-ToM, an incorrect RS-ToM significantly decreased team performance in terms of reward and successful task completions in both simulated and human trials. Next, we evaluate the features that humans can use to calibrate their trust in the robot, where additional description of the correspondence between italicized words and trust can be found in [4]. The previous performance results were also used to support that an incorrect RS-ToM would damage perceptions of the robot's "ability to achieve [human] goals" and degrade trust in terms of *performance information*. Similarly, all experiments showed consistent results for observations of damaging *process information* which describes the "degree to which the [robot]'s algorithms are appropriate for the situation and able to achieve the operator's goal." This was supported by significant increases in the duration of games (increased *effort*) and decreased predictability of the human (decreased *dependability*). Additionally, the number of risks taken by the team increased when paired with a risk-averse human and decreased with a risk-seeking human. This

implies that there was adaptation against human risk preferences which would likely not be viewed as *appropriate*. In simulated trials, we are afforded access to the human's model of the robot. With an incorrect RS-ToM, the robot's actions were more surprising under the human's biased perspective. This implies degrading perceptions of *predictability* and *understandability*. With real humans, we can directly query the human's perception of trustworthiness of the robot using a survey. Results showed that an incorrect RS-ToM generated lower mean trustworthiness and degraded trust over time.

Conclusion and Discussion

Previous works have showed that a robotic agent that is aware of human RS can improve coordination [2] and perceptions of trust relative to rationality assumptions [3]. This motivates the use of more sophisticated ToM mechanisms like modeling human RS to promote more efficient and trustworthy teams. However, the consistent results of our study show that incorrect inference of RS can also lead to negative outcomes for team performance and trust. As a consequence, this calls for planning for possible failure in the RS-ToM in order to realize the full benefit of methods like this. To address this, future work will consider planning for uncertainty in an RS-ToM to proactively avoid trust violations from possibly acting under an incorrect model of human RS. Alternatively, should failure occur anyway, we plan to employ trust repair strategies by leveraging the RS-ToM as a mechanism for explaining why damaging outcomes occurred. We believe effective integration of a RS-ToM realize great benefit in human-robot interaction given that we appropriately situate it in contexts where our model of the human may not be perfect.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under Award No. FA9550-23-1-0283.

References

[1] Ian A Apperly and Stephen A Butterfill. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953, 2009.

[2] Pedro L Ferreira, Francisco C Santos, and Sérgio Pequito. Risk sensitivity and theory of mind in human coordination. *PLoS Computational Biology*, 17(7):e1009167, 2021.

[3] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P Losey, and Dorsa Sadigh. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 43–52, 2020.

[4] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[5] Thomas B Sheridan. Eight ultimate challenges of human-robot communication. In *Proceedings 6th IEEE International Workshop on Robot and Human Communication. RO-MAN'97 SENDAI*, pages 9–14. IEEE, 1997.

[6] Mason O. Smith and Wenlong Zhang. What if i'm wrong? team performance and trustworthiness when modeling risk-sensitivity in human–robot collaboration. *Transactions on Human-Robot Interaction*, 14(2):1–30, 2025.

[7] Rachel E. Stuck, Brittany E. Holthausen, and Bruce N. Walker. Chapter 8 - the role of risk in human-robot trust. In Chang S. Nam and Joseph B. Lyons, editors, *Trust in Human-Robot Interaction*, pages 179–194. Academic Press, 2021.

[8] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323, 1992.

Vision Language Models See What You Want But Not What You See

Qingying Gao[1], Yijiang Li[2], Haiyun Lyu[3], Haoran Sun[1], Dezhi Luo[4,5], and Hokin Deng[6]

[1]Johns Hopkins University
[2]University of California, San Diego
[3]University of North Carolina at Chapel Hill
[4]University of Michigan
[5]University College London
[6]Carnegie Mellon University

Abstract

Knowing others' intentions and taking others' perspectives are two core components of human intelligence that are regarded as instantiations of theory-of-mind. Infiltrating machines with these abilities is an important step towards building human-level artificial intelligence. To investigate intentionality understanding and level-2 perspective-taking in Vision Language Models (VLMs), we evaluate state-of-the-art VLMs through the IntentBench and PerspectBench, which contain more than 300 cognitive experiments grounded in real-world scenarios and classic cognitive tasks. Our experiments show that VLMs achieve high performance in intentionality understanding, but perform worse in level-2 perspective taking. Moreover, their perspective-taking capability does not improve as the models scale up. This suggests a potential dissociation between simulation and theory-based theory-of-mind abilities in VLMs, highlighting the concern that they lack model-based reasoning to infer others' mental states.

Introduction

Intentionality is the capacity of the mind to be directed toward, represent, or stand for objects, properties, or states of affairs for further actions [3]. To say one could understand intentionality is to say one can comprehend the mental content for action in another mind [32, 33]. This capacity has been seen as a key distinction between humans and machines [34]. Despite evidence that large language and vision-language models (LLMs and VLMs) exhibit theory-of-mind (ToM) abilities[20, 38, 36], there is debate about whether these abilities require internal, simulation-based reasoning or emerge solely from abstract theoretical inference[32, 8].

We believe an important approach to this inquiry is to examine the extent to which different ToM abilities necessitate model-based reasoning[16, 27, 39, 12]. Specifically, We distinguish between simulation-based ToM—which involves constructing an internal model of self-other relations to infer mental states—and theory-based ToM, which relies on applying theoretical knowledge about the

relationship between mental states and behavior. [14, 10, 35]. If VLMs do not possess model-based reasoning, it would suggest that they rely solely on theory-based reasoning, lacking the capacity for mental simulation.

We test this critical prediction by assessing VLMs' ability to perform intentionality understanding and level-2 perspective-taking. ToM is commonly understood to be grounded in perspective-taking, a series of multi-level abilities that involves the cognitively undertaking of the perspective of another [6]. Level-1 perspective-taking refers to the acknowledgement that different people can see different things, whereas level-2 perspective-taking involves the understanding of how another person may see the same thing differently. While level-1 perspective-taking emerges in humans as early as 2 years old, much older children are found to struggle with level-2 perspective-taking [30, 28]. This is likely because, despite its relatively low level in the perspective-taking hierarchy, this ability requires model-based reasoning, exemplified in the visual domain as inferences based on mental rotation [22, 15]. On the other hand, while intentionality understanding involves high-level cognition and abstract reasoning, it is unclear whether this complex ability necessitates mental simulations [19, 7].

Recently, Li et al. built CogDevelop2K[24], a data-intensive cognitive experiment benchmark for assessing the developmental trajectory of machine intelligence. Here, we leverage the IntentBench and PerspectBench of CogDevelop2K to investigate perspective-taking and intentionality understanding in current VLMs.

Methods

Dataset

PerspectBench consists of 32 multi-image and 209 single-image experiments based on classic cognitive tasks. IntentBench consists of 100 single-image experiments based on real-world ambiguous social scenarios.

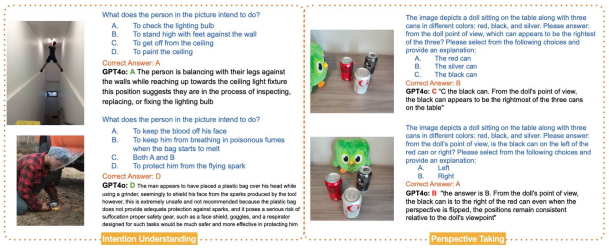


Figure 1: Examples and Model Performances in IntentBench(left) and PerspectBench(right)

Cognitive Experiments

The Three Mountain Task, first invented by Jean Piaget, is widely used in developmental psychology laboratories as the gold standard for testing level-1 and level-2 perspective-taking abilities in children [31, 18, 9, 21]. In a standard Three Mountain Task assessment, a child is instructed to position themselves in front of a model, with another individual taking a different viewpoint. The model features three mountains that vary in size and are distinguished by unique characteristics. The child is then tested whether they can infer specific details of the scenarios seen by the other person. To test level-2 perspective-taking in VLMs, we develop the Three Mountain task into formats that are suitable for benchmarks with minimal confounding details while preserving real-life spatiality. In particular, we use groups of 3-4 commonly-seen beverage cans organized into different spatial patterns to mimic the mountain model. Like in the original task, we use a doll placed to face the organization from different angles as the object of perspective-taking (Fig. 1).

Developmental studies of intentionality understanding often employ cartoon stimuli generated by physics simulation engines[26, 37]. However, such tasks are criticized for lacking

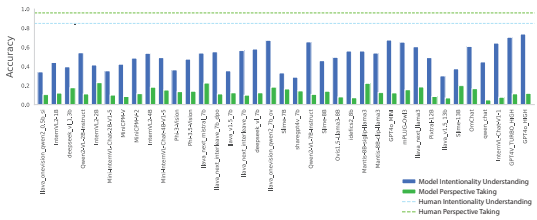


Figure 2: Model Performance on IntentBench and PerspectBench Compared to Human Baselines

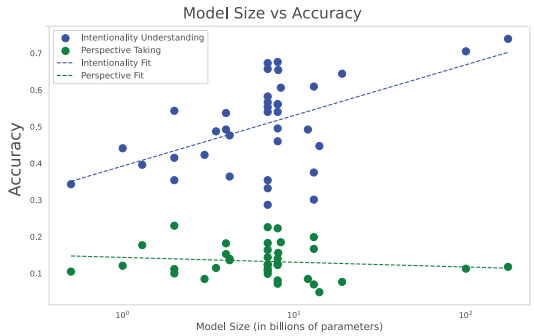


Figure 3: Model Performance Over Size

realism and limited practical applications[13]. Drawing inspiration from COIG-CQIA and its Ruozhiba dataset, many real-world ambiguous scenarios are incorporated into IntentBench for explicitly testing intentionality understanding in ethological conditions [5].

Model Selection and Experiment

We have aligned 35 models for our analysis, including both closed-source models such as GPT [29] series and open-source models including Slime[40], Pixtral[2], LLaVA[25], Mantis[17], Phi-3[1], InternVL[11], Blip [23] and Qwen-vl [4] series. To ensure a fair comparison, all VLMs are evaluated on their ability to reason over images and texts under a zero-shot, open-ended generation task.

Results and Discussions

We find a clear dissociation between the model performance on intentionality understanding and level-2 perspective-taking. Specifically, all assessed models perform better on IntentBench compared to PerspectBench (Fig. 2). While some of the highest-performing models on IntentBench (e.g. GPT-4o) reach near-human performance in intentionality understanding, they still perform poorly in level-2 perspective-taking. Even more strikingly, model performance on IntentBench exhibits a positive linear correlation with model size, whereas model performance on PerspectBench does not improve alongside model size (Fig. 3).

Taken together, these results demonstrate that current VLMs are able to infer the intentions behind others' actions without being capable of level-2 perspective-taking. On one hand, this supports the hypothesis that intentionality understanding does not necessarily require the mental simulations of others but could be based entirely on knowledge-based reasoning. On the other hand, it raises the concern that VLMs do not possess internal models for reasoning or at least cannot use them to take others' perspectives. This is especially salient given level-2 perspective-taking performance does not improve with scale while intentionality understanding performance does. Further studies are needed to explore these ideas, which appear to be critical for understanding the nature of ToM abilities and their artificial implementations.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your

phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

[3] G. E. M. Anscombe. *Intention*. Harvard University Press, 1956.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[5] Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*, 2024.

[6] Yvonne Barnes-Holmes, Louise McHugh, and Dermot Barnes-Holmes. Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today*, 5(1):15–25, 2004.

[7] Valentina Bianco, Alessandra Finisguerra, and Cosimo Urgesi. Contextual priors shape action understanding before and beyond the unfolding of movement kinematics. *Brain Sciences*, 14(2):164, 2024.

[8] Daniel C. Dennett. *The Intentional Stance*. MIT Press, Cambridge, MA, 1987.

[9] Martin E Ford. The construct validity of egocentrism. *Psychological Bulletin*, 86(6):1169, 1979.

[10] Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

[11] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024.

[12] Mariel K Goddu, Alva Noë, and Evan Thompson. Llms don't know anything: reply to yildirim and paul. *Trends in Cognitive Sciences*, 2024.

[13] Alex Gomez-Marin, Joseph J Paton, Adam R Kampff, Rui M Costa, and Zachary F Mainen. Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nature neuroscience*, 17(11):1455–1462, 2014.

[14] Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. *Mind & Language*, 7(1-2):145–171, 1992.

[15] Anna Gunia, Sofiia Moraresku, and Kamil Vlček. Brain mechanisms of visuospatial perspective-taking in relation to object mental rotation and the theory of mind. *Behavioural Brain Research*, 407:113247, 2021.

[16] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

[17] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024, 2024.

[18] David W Johnson. Cooperativeness and social perspective taking. *Journal of Personality and Social Psychology*, 31(2):241, 1975.

[19] James M Kilner. More than one pathway to action understanding. *Trends in cognitive sciences*, 15(8):352–357, 2004.

[20] Michal Kosinski. Evaluating large language models in theory of mind tasksg. *arXiv preprint arXiv:2302.02083*, 2023.

[21] Claus Lamm, C Daniel Batson, and Jean Decety. The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of cognitive neuroscience*, 19(1):42–58, 2007.

[22] Jennifer Lehmann and Petra Jansen. The relationship between theory of mind and mental rotation ability in preschool-aged children. *Cogent Psychology*, 6(1):1582127, 2019.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[24] Yijiang Li, Qingying Gao, Haoran Sun, Haiyun Lyu, Dezhi Luo, and Hokin Deng. Cogdevelop2k: Reversed cognitive development in multimodal large language models. *arXiv preprint arXiv:2410.10855*, 2024.

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[26] Shari Liu, Tomer D Ullman, Joshua B Tenenbaum, and Elizabeth S Spelke. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017.

[27] Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

[28] Henrike Moll and Andrew N Meltzoff. How does it look? level 2 perspective-taking at 36 months of age. *Child Development*, 82(2):661–673, 2024.

[29] OpenAI. Models - openai api. `https://platform.openai.com/docs/models/gpt-4o`.

[30] Jean Piaget. *The Development of Thought: Equilibration of Cognitive Structures*. Viking Press, 1977.

[31] Jean Piaget and Bärbel Inhelder. *The Child's Conception of Space*. Routledge, London, 1957.

[32] David G. Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1978.

[33] David M. Rosenthal. *The Nature of Mind*. Oxford University Press, New York, 1991.

[34] John Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 1980.

[35] Karen Shanton and Alvin Goldman. Simulation theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(4):527–538, 2010.

[36] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind. *arXiv preprint arXiv:2408.12574*, 2024.

[37] Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In *International conference on machine learning*, pages 9614–9625. PMLR, 2021.

[38] James W A Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido

Manzi, Michael S A Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024.

[39] Ilker Yildirim and LA Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 2024.

[40] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024.

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

Anthia Solaki[1] and Karel van den Bosch[1]

[1]Netherlands Organization for Applied Scientific Research (TNO)

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 [19], 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* [4] or *Smarties* tasks [18, 9], 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 [7, 1]. 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 suggests that LLMs may develop ToM-like abilities as a by-product of their language skills [12]; Ullman raises questions on the robustness of such results, as minor perturbations of the tasks seem to expose limitations in ToM abilities [22]; others argue for a nuanced perspective, emphasizing the role of instruction-tuning in LLM performance [24] or suggesting that failures may stem from a hyper-conservative approach towards committing to conclusions [21]. 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 [14]. 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 [23, 6]. Theory of Mind contributes to better coordination, task performance, perceptions of trustworthiness and explainability in such hybrid teams [15, 16, 8]. 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 [3]. 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 [13].

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 [22]). 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 [17] 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 [12, 22], 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.

	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.

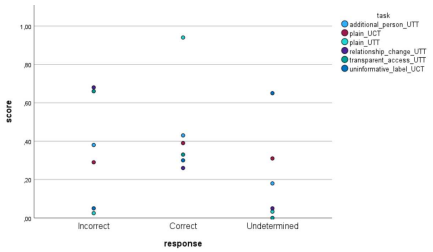


Figure 2: Scatterplot of score (average across all prompts) by response for each task.

Discussion

Despite many promising findings, this study suggests that minor tweaks in FBTs lead to suboptimal performance in tracking protagonists' beliefs, aligning with [20] and [22], and revealing a tendency to retract responses when asked to commit to conclusions [21]. 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 [24]. 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 [14]. 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 [25] or Bayesian [2] 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 [11]), 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 [10, 5]. Ultimately, defining the architecture of such systems can contribute to a hybrid teaming 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

[1] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

[2] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[3] Leonard Bärmann, Rainer Kartmann, Fabian Peller-Konrad, Jan Niehues, Alex Waibel, and Tamim Asfour. Incremental learning of humanoid robot behavior from natural interaction and large language models. *Frontiers in Robotics and AI*, 11:1455375, 2024.

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

[5] Thomas Bolander, Lasse Dissing Hansen, and Nicolai Herrmann. DEL-based epistemic planning for human-robot collaboration: Theory and implementation. In *18th International Conference on Principles of Knowledge Representation and Reasoning*, pages 120–129. International Joint Conferences on Artificial Intelligence Organization, 2021.

[6] Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, Oct 2024.

[7] Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. Knowing me, knowing you: theory of mind in AI. *Psychological medicine*, 50(7):1057–1061, 2020.

[8] Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In *Aamas*, pages 429–437, 2020.

[9] Alison Gopnik and Janet W. Astington. 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, 1988.

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

[11] Filip Ilievski, Pedro Szekely, and Bin Zhang. Cskg: The commonsense knowledge graph. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 680–696. Springer, 2021.

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

[13] Glenn J Lematta, Pamela B Coleman, Shawaiz A Bhatti, Erin K Chiou, Nathan J McNeese, Mustafa Demir, and Nancy J Cooke. Developing human-robot team interdependence in a synthetic task environ-

ment. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 63, pages 1503–1507. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[14] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore, December 2023. Association for Computational Linguistics.

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

[16] Wenxuan Mou, Martina Ruocco, Debora Zanatto, and Angelo Cangelosi. 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)*, pages 956–962. IEEE, 2020.

[17] OpenAI. Chatgpt-4o, 2024. https://openai.com/index/hello-gpt-4o/.

[18] Josef Perner, Susan R Leekam, and Heinz Wimmer. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137, 1987.

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

[20] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[21] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11, 2024.

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

[23] Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. Six challenges for human-AI colearning. 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*, pages 572–589. Springer, 2019.

[24] Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 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 Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore, December 2023. Association for Computational Linguistics.

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

Why Was I Sanctioned?

Nathan Lloyd and Peter R. Lewis

Ontario Tech University

Abstract

This paper investigates how perspective-taking shapes the formation and validation of normative expectations. It explores how adopting different viewpoints influences an agent's association of sanctions and behavior, altering the influence of norms. Further, it examines the strategic use of perspective-taking in norm negotiation and maintenance, from which we discuss a pathway to implementation within the Expectation Event Calculus.

Introduction

Perspective-taking is a fundamental skill that involves reasoning about the mental states of others [6]. The value of perspective-taking manifests in various ways, primarily through its capacity to reveal an interaction partner's underlying interests, beliefs, knowledge, and motives. This paper explores perspective-taking's role in normative settings and how adopting different viewpoints influences social norms, chiefly the causal relationship between sanctions and behavior to reaffirm acceptable or expected behavior within a given context. The paper discusses how individually perceived yet mutually shared expectations are implicitly negotiated and upheld within social groups and discusses a pathway to unlocking this capability for the Expectation Event Calculus (EEC) [7].

Social Norms

Norms are powerful constructs for regulating behavior, emerging from shared beliefs and maintained by perceived compliance or enforcement. Bicchieri's [4] conceptualization of norms introduces social expectations that define and distinguish descriptive norms from social norms. Descriptive norms are defined as a pattern of behavior that an individual prefers to engage in, conditional upon empirical expectations, an expectation about others' behavior ("I expect *they* will do..."). Social norms are conditional on empirical expectations and normative expectations, a belief about what others expect them to do ("I believe *they* think others ought to do..."); a second-order prescriptive or proscriptive belief. Both types of norms are necessarily underpinned by conditional, interdependent, and self-fulfilling expectations, able to describe behaviors like fashion trends (or other forms of imitation) as descriptive norms and patterns of behavior like queuing or more complex social phenomena such as trust and reciprocity (social norms).

Expectation Formation

While empirical expectations can be formed passively through observation, normative expectations are second-order beliefs that cannot be directly observed. Often, we: (1) Assume a normative expectation exists if the supporting empirical expectations are stable [5, p.136]; a

perceived normative expectation emerging from the regularities of behavior. (2) Infer a normative expectation from behaviors that enforce a perceived norm (sanctions). (3) Receive direct and explicit communication through signs or conversation: "while in *place/culture* you shouldn't do this".

Incomplete Reasoning

The methods discussed so far are widespread but limited as forms of associative normative expectation formation, failing to establish causal relationships. This limitation presents a range of problems in non-trivial interactions. First is the naive assumption that sanctions signal a normative expectation. Empirical studies reveal that many sanctions stem from errors, lagged responses, or blind revenge [15], or others, as strategic retaliation, dubbed counter-punishment [14]. In these cases, as well as those where no sanctions are associated with a normative expectation [4, p.11], or the normative expectation emerges from a belief about stable behavior, how does one discern between the normative and non-normative, the interdependent and independent. Further uncertainty is revealed when outcomes are misperceived, i.e., one believes another has cheated, when in fact they did not, or did, but did not intend to, these perceptual mistakes [3] and gaps between intentions and actions [1] may lead to costly and unnecessary punishment. Furthermore, as agents belong to multiple, often nested groups, conflicts between different sets of norms become especially pronounced [20]; how do agents reconcile multiple accounts of prescribed behavior beyond associating the conditions they perceive they apply? Finally, explicit signals perceived in the world may be incorrect, biased, or harbor malintent, enabling agents to be misinformed or manipulated. In these instances, we argue that causal explanations are missing, which may be derived from the explicit modeling of others and the process of perspective-taking, offering answers to "What will happen if I transgress?" and "Why was I sanctioned?"

Perspective-Taking

To reconcile the limitations mentioned above, we propose that normative agents possess the ability to explicitly model others [12] and utilize said models to take their perspective. This capability enables agents to move up the ladder of causation [16] to derive answers to the questions of "what-if?" and "why?", informing beliefs and decisions. Agents may accomplish this by building explicit models of others based on direct interactions and indirectly received information, such as hearsay. From their perceptions, they may then utilize their decision-making framework to simulate the decisions of others, replacing one's beliefs with what is believed about the other, whether a learned preference, shared observation, or learned factors such as membership to particular organizations.

For example, perspective-taking allows agents to build causal relationships between sanctions and transgressions by attributing motives to the punisher [8]. Similarly, by attributing motivation to others and recording a history of perceived events for other agents, an agent may be better prepared to handle deceptive or ill-informed agents [17], informing trustworthiness, reliability, and reputation. Furthermore, perspective-taking enables agents to sanction more strategically. By modeling others, agents can identify when sanctioning is costly and unnecessary, such as with individuals who will always transgress: "a lost cause". Conversely, they may continue to sanction a self-interested individual to deter others from transgressing: "making an example". They may also validate their expectations, resolve inconsistencies, and explore the scope of their applicability via the combination of *what-if* reasoning and perspective-taking; "in this new but familiar context (condition), do

you my expectations still hold?", "what if I were to do act differently?" and "how may others react?". Finally, by modeling others, agents may also reconcile 'false' beliefs held by themselves or others. Without such models of others, the decision space for rich normative interaction is vastly decreased.

Expectation Event Calculus

The *Expectation Event Calculus* (EEC) [7] is a logical framework for reasoning about expectations, an extension of the *Event Calculus* (EC) [9], a common-sense reasoning tool to represent dynamic properties of the world; represented as fluents. By building on the core mechanisms of the EC, the EEC enables queries about which fluents (environmental properties and beliefs) hold at a given time through reasoning over what actions do, what actions have happened, and what fluents are active given what actions do and what has happened. Prior work has established a form congruent with Bicchieri's account of conditional expectations [18], established mechanisms for *what-if* reasoning [19], and recent work has sought to incorporate the distinction between empirical and normative expectations [13]. The latter investigates expectation formation through observation and the explicit communication of normative expectations to influence individual agents' decisions. Limited by passive observation and explicit signals that agents are preprogrammed to interpret, this motivates the addition of enhanced reasoning capabilities for agents to build explicit logical models of others based upon their interactions.

Perspective In Practice

We propose that reasoning about actions and effects should consider not only one's own beliefs but also the beliefs of others [2], to establish the EEC as a tool for individual [10], perspective-based [2] accounts of expectations and norms. To accomplish this, agents may construct self-narratives and other-narratives; beliefs about what another has perceived. Such an extension may distinguish between $happens(agent_i, \alpha, \tau)$ (own) and $happens(agent_j, \alpha, \tau)$ (other) narratives, the former updated by $agent_i$'s perception, and the latter a belief about $agent_j$'s perception. Distinguishing narratives in some form enables agents to apply their reasoning capabilities over a different set of perceptions to construct beliefs and simulate the decisions of another. Utilizing the formalisms of the EC and EEC further, we may also represent models of others as fluents. An advantageous property as it permits comparative reasoning, such that $agent_i$ can construct and later compare its models of $agent_j$ at t_1 and again at t_n. This ability allows agents to detect changes, reason about potential causes, and reconcile discrepancies between prior beliefs and new perceptions. For instance, if two conditional followers lose contact, and one later violates a norm upon meeting, an agent might, instead of punishing immediately, take their perspective and seek further information, i.e., "Has their preferences changed?".

Conclusion

The explicit modeling of others within a normative context presents many opportunities beyond what has been discussed here, such as social self-awareness and reference network construction. Although perspective-taking offers clear advantages for forming normative expectations and navigating strategic interactions, it is also a resource-intensive process. Persistent perspective-taking may be feasible and beneficial in dyadic interactions or small populations. Still, in larger groups and more complex social settings, perspective-taking may be triggered as a reflective [11] and retrospective [16] process for the correction of misperception [10].

References

[1] Robert Axelrod and Douglas Dion. The further evolution of cooperation. *Science*, 242(4884):1385–1390, 1988.

[2] Chitta Baral, Gregory Gelfond, Enrico Pontelli, and Tran Cao Son. Multi-agent action modeling through action sequences and perspective fluents. In *2015 AAAI Spring Symposium Series*, 2015.

[3] Jonathan Bendor and Piotr Swistak. The evolution of norms. *American Journal of Sociology*, 106(6):1493–1545, 2001.

[4] Cristina Bicchieri. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, 2005.

[5] Cristina Bicchieri. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press, 2016.

[6] SAJ Birch, V Li, T Haddock, SE Ghrear, P Brosseau-Liard, A Baimel, and M Whyte. Perspectives on perspective taking: how children think about the minds of others. *Advances in Child Development and Behavior*, 52:185–226, 2017.

[7] Stephen Cranefield. Agents and expectations. In *Coordination, Organizations, Institutions, and Norms in Agent Systems IX: COIN 2013 International Workshops, COIN@ AAMAS, St. Paul, MN, USA, May 6, 2013, COIN@ PRIMA, Dunedin, New Zealand, December 3, 2013, Revised Selected Papers 16*, pages 234–255. Springer, 2014.

[8] Melissa de Vel-Palumbo, Mathias Twardawski, and Mario Gollwitzer. Making sense of punishment: Transgressors' interpretation of punishment motives determines the effects of sanctions. *British Journal of Social Psychology*, 62(3):1395–1417, 2023.

[9] Robert Kowalski and Marek Sergot. A logic-based calculus of events. *New generation computing*, 4:67–95, 1986.

[10] Sophie Legros and Beniamino Cislaghi. Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science*, 15(1):62–80, 2020.

[11] Peter R Lewis and Ştefan Sarkadi. Reflective artificial intelligence. *Minds and Machines*, 34(2):1–30, 2024.

[12] Nathan Lloyd and Peter R Lewis. Towards reflective normative agents. In *Conference of the European Social Simulation Association*, pages 587–599. Springer, 2023.

[13] Nathan Lloyd and Peter R Lewis. Incorporating social expectations into the expectation event calculus. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. MIT Press, 2024.

[14] Nikos Nikiforakis. Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, 92(1-2):91–112, 2008.

[15] Elinor Ostrom, James Walker, and Roy Gardner. Covenants with and without a sword: Self-governance is possible. *American political science Review*, 86(2):404–417, 1992.

[16] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[17] Stefan Sarkadi and Peter R Lewis. The triangles of dishonesty: Modelling the evolution of lies, bullshit, and deception in agent societies. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2024.

[18] Abira Sengupta, Stephen Cranefield, and Jeremy Pitt. Solving social dilemmas by reasoning about expectations. In *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, pages 143–159. Springer, 2021.

[19] Abira Sengupta, Stephen Cranefield, and Jeremy Pitt. Generalising axelrod's metanorms game through the use of explicit domain-specific norms. In *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, pages 21–36. Springer, 2023.

[20] Gerben A van Kleef, Florian Wanders, Annelies EM van Vianen, Rohan L Dunham, Xinkai Du, and Astrid C Homan. Rebels with a cause? how norm violations shape dominance, prestige, and influence granting. *Plos one*, 18(11):e0294019, 2023.