Building Shared Mental Models between Humans and AI for Effective Collaboration

Harmanpreet Kaur
harmank@umich.edu
Computer Science & Eng.
School of Information
University of Michigan

Alex C. Williams
alex.williams@uwaterloo.ca
Computer Science & Eng.
University of Waterloo

Walter S. Lasecki
wlasecki@umich.edu
Computer Science & Eng.
University of Michigan

ABSTRACT

Intelligent systems have become increasingly common in settings ranging from performing everyday tasks more easily to decision-making for complex domains (e.g., healthcare, autonomous driving, criminal justice). Given this rising ubiquity of artificial intelligence (AI), both researchers and industry practitioners are exploring ways to better integrate AI agents in tasks that people do at home or work. However, these systems are currently limited because of gaps in the understanding between humans and their AI counterparts. In this paper, we propose methods for building shared mental models between humans and AI to enable human-AI collaboration at a level where both can be equal partners working on a shared task. We ground our approach in existing literature from CSCW and UX design.

KEYWORDS

mental models, explainability, team cognition, human-AI collaboration

CHI'19, May 2019, Glasgow, Scotland

INTRODUCTION

As with human-human collaboration, the effectiveness of human-AI collaboration is rooted in a shared understanding of each collaborators' capabilities [9, 10]. On one hand, humans cannot always understand or anticipate what the AI agent will do, leading to a lack of trust and reliance on the agent [2]. User experience researchers call this the *gulf of execution* (i.e., the degree to which a person is able to accurately perceive the functionality related to a task [22]). On the other hand, without knowing how the human reasons about the task or what they do to accomplish it, the AI agent is not able to anticipate the human's intent. Thus, it is unable to reflect a change of state or output in a way that makes sense to the human, resulting in a *gulf of evaluation* (i.e., the degree to which the system makes its current state clear to a user [22]). Broadly, this lack of a shared understanding forms the core of the socio-technical gap: the fluid intents and interactions of humans remain unmatched by the discrete and brittle features of AI agents [1].

Applying strategies from effective human-human teamwork, we can build shared mental models between humans and AI agents for more effective human-AI collaboration [9, 10, 30]. Building these shared mental models could help both humans and AI agents better understand each other and anticipate which parts of a shared task they can each accomplish, leading to more effective collaboration. Further, this shared understanding helps bridge the socio-technical gap by providing a shared scaffolding that both humans and agents can use to align their outputs for a task.

In this paper, we ground the challenges of human-AI collaboration within existing HCI work on the socio-technical gap [1], and the gulfs of execution and evaluation [22]. We propose building shared mental models as a way of overcoming this gap and resulting in more effective human-AI collaboration. To that end, we describe a general approach for forming these shared mental models, and define evaluation metrics for our approach. We hope that future work on human-AI collaboration can use and extend this general approach for bridging the socio-technical gap in several domains.

BACKGROUND

Human-AI collaboration has become increasingly common in many domains, ranging from simple automation to complex decision-making tasks. What started with attempts to automate repetitive parts of people's workflows (e.g., using calendars to schedule recurring meetings rather than individual ones, using templates for writing common documents, etc.) has now become a suite of personalized assistants (e.g., [15, 25, 33]) and task collaborators (e.g., [2, 3]). Notably, we now have AI-assisted decision making for complex tasks in healthcare [5, 32], data science [4], education [35], criminal justice [28], and several other domains. When AI fails or cannot predict with high probability, we now have human-in-the-loop methods and algorithms to perform the task (e.g., [8] uses these methods for quantifying mental illness severity online, [6, 12–14] for visual question answering for people with

(1) Transactional Memory System (TMS):
"A set of individual memory systems in combination with the communication that takes place between in dividuals to access each other's memory systems" [31] . The spectrum ranges from an integrated memory system (where everyone on the team has the same notion about a task in their memory, e.g., sales pitch for a product), to a distributed memory system (where everyone has unique expertise but they are working towards a shared goal, e.g., more creative, distributed tasks) [7, 31].
(2) Group Learning:
Group learning pertains to how members within a team interact with each other, including communication, influence, and autonomy patterns [21].
(3) Cognitive Consensus:
A belief system of key definitions that is shared among all members of the team. This caters to the interpretation of the outcome, rather than the method of reaching the outcome. E.g., team members agreeing that classifying as "Yes" means "Not defective," but how they classify something as "Yes" is not included within this definition [21].

Table 1: Mohammed and Dumville's [21] Components of Shared Mental Models.

visual impairments, [27] for 3D training data for autonomous vehicles). These human-in-the-loop methods are further empowered by approaches that enable real-time aggregation of feedback from multiple people [17, 19, 26], and tools that improve the ease of data collection from crowd workers [18].

Real-world applications of human-AI collaboration in complex domains are contingent on two key factors: (1) trust, transparency, and accountability of the AI partner involved [2], and (2) users' ability to understand and predict agent behavior (i.e., explainability and intelligibility [20]). These factors will continue to serve as a bottleneck regardless of how complex, nuanced, and robust the AI systems get, causing gulf of evaluation and execution, and consequently socio-technical gap, to exist.

Forming accurate mental models of the systems we use is key to their usability, and for bridging both the gulfs described above. Per Norman's definition of mental models [23], they serve three functional purposes: (1) representing a person's beliefs about the system, acquired via observation, instruction, or inference; (2) a mapping between the observable features of a system as intended by the designer, and the features and functionality perceived by the user; and (3) a predictive power for anticipating the system's output in a given situation.

Going from simply interaction to effective human-AI teamwork requires the existence of a shared mental model among team members, i.e., both the human and the AI agent need to form mental models of each other [16]. Team cognition literature defines these shared mental models for human-human teams: "an emergent state that refers to the manner in which knowledge important to team functioning is mentally organized, represented, and distributed within the team and allows team members to anticipate and execute actions" [10]. Mohammed and Dumville [21] describe three key components for a shared mental model to exist between team members (Table 1).

A GENERAL APPROACH FOR BUILDING SHARED MENTAL MODELS

As a concrete example of our approach, we use a binary classification task presented to a human-AI team: a person and an AI agent working together to classify objects as defective or not.

Building Mental Models of AI Agents

To build an accurate mental model of the AI agent, people must understand the agent's output, and the process underlying its decision. Further, they must be able to anticipate this – TMS requires existing knowledge of the expertise and the process of each team member [21]. Prior work defines related concepts, such as explainability and trust in AI (e.g., [2, 3, 20, 34]), and provides guidelines for generating explanations. For example, Fourney et al. [11] use Query-Feature Graphs to create a mapping between people's task goals and the interface features that would help them complete these, and then design tutorials based on these mappings. Further, there now exist automated approaches for providing explanations of black box models, making it easier to share these with the human [24].

Before we can extend explainability towards building mental models, two questions need to be answered for the target domain: (1) how do people explain the mental model of the AI agent they are collaborating with; and (2) what is the effect of task and team complexity on people's mental model of their AI counterpart? The former can provide better definitions for explainability that are based on the subjectivity of the user, and the latter helps us characterize the difficulty of forming mental models.

For our example task, we could conduct these studies in controlled setups — proxies of the domain of interest — and varying task and team complexity to learn the unknown boundary of error. Task complexity could be varied by adding more features defining an object (e.g., include color, size, pattern, border, etc.), and team complexity by adding more AI agents and human members in the team. Doing this in a controlled setup allows us to understand the performance drop. In addition to quantitative measures, we can use open-ended questions at the end of the task to ask people about their understanding of the agent. Further, we can code the responses to these open-ended questions and match them to the actual definition of the AI agent that we use in our controlled setup. Using inter-rater reliability metrics, such as Cohen's Kappa, can: (1) give us a sense of how often mental models were accurately formed, and (2) provide categories of errors for when people were not able to explain the AI agent's underlying model, for improving explainability mechanisms. Integrating this to improve mental model accuracy would be a step towards bridging the gulf of execution, and consequently more effective collaboration.

Building Mental Models of Humans

For AI agents, building mental models of their human counterpart essentially translates to a personalized model of how the user performs the task. Having this mental model would not only enable the AI agent to anticipate the human's action for their shared task, but also update the representation of its output to best fit the user. The latter is key to potentially bridging the gulf of evaluation, wherein a human is unable to understand the states and representations of the agent.

One initial approach for doing this could apply reinforcement learning algorithms (e.g., Q-learning), where an AI agent follows the human as it does the task, and builds a cognitive network graph of how the human connects the different pieces of the task. For our binary classification example, this information could be elicited by having a simple interface where the AI agent predicts the human's answer, then have the human either agree with the AI's prediction or provide their actual answer. Starting with a random selection model, over time via Q-learning, the agent can learn the human's mental model for performing that task. Further feedback could be provided by asking the human for the features used to make a classification prediction for each object (e.g., "I made this decision based on the shape and the color of the object").

Evaluating Shared Mental Models

Once we have characterized and implemented potential ways of building mental models for the human and the AI agent collaborating on a task, we can re-evaluate their task performance by having them work together again. They key things we would be looking for are the existence of a transactive memory system (TMS), group learning, and cognitive consensus [21] (Table 1).

We can test whether each of these exist via different metrics, while keeping the controlled experiment setup we described above. An existing TMS makes a team work efficiently – the performance is consistent or better, with reduced time and effort spent on the task. We could compare the performance of a human-AI team before and after building shared mental models, to observe whether mental models for both the human and the AI agent served to improve performance costs. Additionally, we could apply Bayesian statistics to compare the optimal curve for learning a pattern that we would expect a human to follow (e.g., by using theories about people's pattern recognition capabilities, such as Cognitive Load Theory [29]) with the actual curve they follow. Group learning could be evaluated via an open-ended survey: asking people whether they understand the communication from the AI agent using open-ended questions. Finally, cognitive consensus could be evaluated by examining when the human and the agent have the same predicted output for each object in their shared classification task. Together, these three concepts give us an understanding of whether an effective shared mental model has been formed.

CONCLUSION

We have presented an approach to studying shared mental model formation between human and AI agents collaborating on a task. The bottleneck in human-AI collaboration is the socio-technical gap, which will persist until both humans and AI can better understand each other's capabilities. We propose ways to bridge this gap — via shared mental model formation — with the hope that this work will guide future research on human-AI teams.

ACKNOWLEDGEMENTS

We would like to thank Cliff Lampe and Stephanie O'Keefe for their feedback on these ideas. This work was supported in part by the University of Michigan and IBM.

REFERENCES

[1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *Proceedings of the 33rd AAAI conference on Artificial Intelligence*. AAAI.

[4] Gagan Bansal and Daniel S Weld. 2018. A Coverage-Based Utility Model for Identifying Unknown Unknowns. In *Proceedings of the 32nd AAAI conference on Artificial Intelligence*. AAAI.

[5] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M Mack, George Ruiz, Mark S Smith, and Eric Horvitz. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one* 9, 10 (2014), e109264.

[6] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 333–342.

[7] David P Brandon and Andrea B Hollingshead. 2004. Transactive memory systems in organizations: Matching tasks, expertise, and people. *Organization science* 15, 6 (2004), 633–644.

[8] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1171–1184.

[9] Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991 (1991), 127–149.

[10] Leslie A DeChurch and Jessica R Mesmer-Magnus. 2010. The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology* 95, 1 (2010), 32.

[11] Adam Fourney, Richard Mann, and Michael Terry. 2011. Query-feature graphs: bridging user vocabulary and system functionality. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 207–216.

[12] Danna Gurari, Qing Li, Chi Lin, Anhong Guo, Yinan Zhao, Abigale J Stangl, and Jeffrey P Bigham. 2019. Detecting private information and its purpose in pictures taken by blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[13] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.

[14] Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey P Bigham, Jaime Teevan, Ece Kamar, and Walter S Lasecki. 2017. Crowdmask: Using crowds to preserve privacy in crowd-powered systems via progressive filtering. In *Proceedings of the AAAI Conference on Human Computation (HCOMP 2017)., HCOMP*, Vol. 17.

[15] Harmanpreet Kaur, Alex C Williams, Anne Loomis Thompson, Walter S Lasecki, Shamsi T Iqbal, and Jaime Teevan. 2018. Creating Better Action Plans for Writing Tasks via Vocabulary-Based Planning. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 86.

[16] Walter S Lasecki. 2019. On Facilitating Human-Computer Interaction via Hybrid Intelligence Systems. In *Proceedings of the 7th annual ACM Conference on Collective Intelligence*. ACM.

[17] Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. 2011. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 23–32.

[18] Sang Won Lee, Rebecca Krosnick, Sun Young Park, Brandon Keelean, Sach Vaidya, Stephanie D O'Keefe, and Walter S Lasecki. 2018. Exploring real-time collaboration in crowd-powered systems through a ui design tool. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 104.

[19] Alan Lundgard, Yiwei Yang, Maya L Foster, and Walter S Lasecki. 2018. Bolt: Instantaneous crowdsourcing via just-in-time training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 467.
[20] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).
[21] Susan Mohammed and Brad C Dumville. 2001. Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 22, 2 (2001), 89–106.
[22] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Constellation.
[23] Donald A Norman. 2014. Some observations on mental models. In *Mental models*. Psychology Press, 15–22.
[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. [n. d.]. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.
[25] Greg Smith, Patrick Baudisch, George Robertson, Mary Czerwinski, Brian Meyers, Daniel Robbins, and Donna Andrews. 2003. Groupbar: The taskbar evolved. In *Proceedings of OZCHI*, Vol. 3. 10.
[26] Jean Y Song, Raymond Fok, Juho Kim, and Walter S Lasecki. 2019. FourEyes: Leveraging Tool Diversity as a Means to Improve Aggregate Accuracy in Crowdsourcing. In *Transactions on Interactive Intelligent Systems (TiiS)*. ACM.
[27] Jean Y Song, Stephan J Lemmer, Michael Xieyang Liu, Shiyan Yan, Juho Kim, Jason J Corso, and Walter S Lasecki. 2019. Popup: Reconstructing 3D Video Using Particle Filtering to Aggregate Crowd Responses. In *24th International Conference on Intelligent User Interfaces*. ACM.
[28] Nigel Stobbs, Mirko Bagaric, and Dan Hunter. 2017. Can sentencing be enhanced by the use of artificial intelligence? *Criminal Law Journal* 41, 5 (2017), 261–277.
[29] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4, 4 (1994).
[30] Warren Weaver. 1953. Recent contributions to the mathematical theory of communication. *ETC: a review of general semantics* (1953), 261–281.
[31] Daniel M Wegner. 1987. Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior*. Springer, 185–208.
[32] Jenna Wiens, John Guttag, and Eric Horvitz. 2016. Patient risk stratification with time-varying parameters: a multitask learning approach. *The Journal of Machine Learning Research* 17, 1 (2016), 2797–2819.
[33] Alex C Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T Iqbal, and Jaime Teevan. 2018. Supporting Workplace Detachment and Reattachment with Conversational Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 88.
[34] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.
[35] Haoyu Zhou, Haifeng Zhang, Yushan Zhou, Xinchao Wang, and Wenxin Li. 2018. Botzone: an online multi-agent competitive platform for AI education. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. ACM, 33–38.