Training Humans for Robust Human-Agent Teaming: Knowing When to Engage with an AI Partner

Leon Lange[1], **Qiao Zhang**[2], **Christopher J. MacLellan**[2], **Ying Wu**[1]
[1] Institute for Neural Computation
University of California, San Diego 9500 Gilman Drive
La Jolla, California 92093
[2] School of Interactive Computing
Georgia Institute of Technology North Avenue
Atlanta, Georgia 303332

Abstract

Learning to team with an AI counterpart can be challenging – particularly in the context of an unfamiliar task that must also be learned. This study compares the impacts of scaffolded versus self-paced training on human-AI agent teams negotiating a novel logistics and sustainment task. It was found that guiding participants early on in how to leverage AI assistance (scaffolded practice) led to much more robust teaming than allowing them to learn at their own pace. Additionally, teams whose human counterpart received scaffolded practice tended to achieve higher scores than those who learned under self-direction. Post-hoc analysis also revealed that *when* human teammembers leveraged the agent was of particular importance – with the greatest impact of human-AI teaming observed in the most high stakes periods of the game. Taken together, these findings demonstrate not only that some forms of training are more beneficial than others for human-AI agent teaming – but also, that context-specific learning on the fly is important for effective team performance.

Introduction

The integration of AI systems into various societal aspects holds immense potential for revolutionizing how we work and interact (Vaccaro, Almaatouq, and Malone 2024). However, challenges to effective human-agent teaming (HAT) are also well known AI systems are needed that can function as genuine collaborators, complementing human strengths and offsetting weaknesses (Metcalfe et al. 2021). Yet, existing AI systems often lack the flexibility and robustness of human systems, impeding the emergence of efficient communication and coordination processes (McNeese et al. 2018; Demir, McNeese, and Cooke 2016). Further, negative human perceptions of an AI agent – particularly along the lines of its intent, transparency, and task readiness and integration (Wynne and Lyons 2018) – can diminish trust and obstruct HAT processes.

The present analysis explores approaches to training for improved human-AI collaboration. It builds from the idea that teaming dynamics are affected both by stable, trait-like factors, such as prior experience of team members working together, personality, or interface designs, as well as by transient, emergent features of context, such as momentary task complexities or time pressures. Effective training for robust HAT will need to account for how stable and transient features interact. Here, a proof of concept is outlined using a virtual sustainment and logistics game, Space Transit, in which human- agent teams create efficient routes in 3D space to transport passengers from spontaneously spawning transit stations to their destinations. A study was conducted comparing the benefits of scaffolded versus self-paced training methods to foster human-AI collaboration during game play. Scaffolded methods yielded superior collaboration – likely because they offered users an opportunity to gain richer, context-relevant experience with the agent before launching into the task. However, deeper analysis suggests that simply knowing how to collaborate with the agent did not heavily impact performance outcomes. Rather, better performance was associated with strategic collaboration at critical times within the game. This finding suggests a corollary to the idea that robust HAT requires AI agents with more flexible, adaptive capabilities. Humans must also be able to organize their interaction with AI teammates to meet the contingencies that arise from task fluidity.

Method

Platform Design and Implementation

Paralleling the commercially available game, Mini Metro, Space Transit centers on transporting passengers from stations that spawn in 3D space by creating transit networks (Figure 1). In constructing routes, the player is tasked with ensuring that all passengers can reach at least one of their possible destinations, and that lines are efficiently constructed for timely pick-ups and drop-offs. If too many passengers accumulate at a location, overcrowding will cause the game to end. A second challenge is allocation of resources. New trains and lines become available at regular intervals and can be distributed at the player's discretion. Further, multiple separate transit networks can run in parallel, tasking players both with designing and monitoring each network as well as with switching attention between networks.

Space Transit can be played by solo humans or AI agents, or by heterogeneous or homogeneous teams. In the present study, humans were paired with a hierarchical task network (HTN) agent (Lawley and Maclellan 2024) in hybrid team

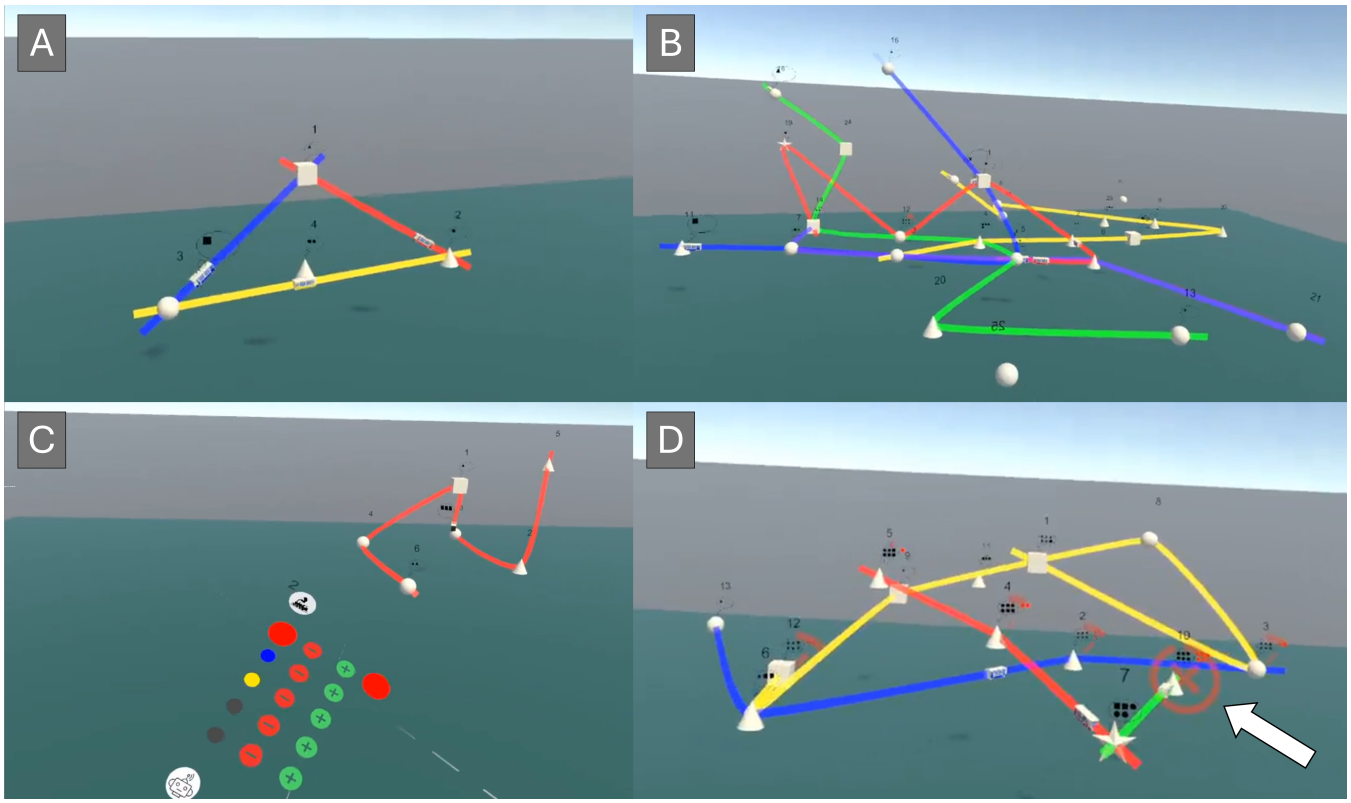

Figure 1: A) Stations in Space Transit are represented by spheres, cubes, and cones. Passengers can travel between stations when they are connected by lines. B) Up to five separate lines can be implemented. C) Trains can be moved between lines. D) An overcrowding warning signal appears if more than six passengers accumulate at a station. After thirty seconds of overcrowding, the network is terminated.

mode. The agent could perform the same tasks as its human counterpart, but only in response to a command to act. Possible actions (for either team member) are the following:

- **create_line** takes two stations and create a line to connect them.
- **delete_line** takes a given line and delete the line as well as releasing all the stations on the line.
- **insert_station** inserts a given station to the closest location to a given line.
- **remove_station** removes a given station.
- **add_train** add a train to a given line.
- **remove_train** removes a train from a given line.
- **goto_game** takes the game ID and directs players to a specific game.

Further, teams were required to manage two separate networks simultaneously in each game. Participants interfaced with the virtual world using the HTC Vive Pro Eye system.

Participants and Protocol

This study was approved by the UC San Diego Institutional Review Board. Twenty healthy adults were recruited from the general population and gave written informed consent. They were randomly allocated to either a self-paced or scaffolded training group. All participants received the same comprehensive game play instructions and completed a structured tutorial. Subsequently, they undertook two independent practice sessions. Self-paced players explored the game independently and had the discretion to utilize the AI agent at will. In contrast, for the first guided practice session, scaffolded players were directed by a co-present experimenter to request specific forms of AI assistance (e.g. move a train to a different line) according to a pre-established schedule. In the second practice session, they were instructed to perform their next self-chosen action using the AI agent every 30 seconds. Following the practice sessions, both groups performed four test sessions, aiming to maximize their scores.

Data Collection and Analysis

Performance was operationalized as the final score for each game, computed as the total number of passengers delivered and averaged. Teaming was operationalized as the number of requests made to the agent per minute. Additionally, all instances of actions performed by the player or agent were recorded, along with their timestamp.

As an exploratory examination of the relationship between visual attention and performance, eye tracking data were recorded in fourteen individuals using the integrated

eye tracker in the head-mounted display. After preprocessing, a velocity-based parsing algorithm was applied to the time-series gaze points. Periods of data characterized by an angular velocity lower than 30 °per second were classified as fixations. Angular velocities that exceeded this threshold characterized saccades. Next, we classified each fixation as either focal or ambient on the basis of its duration and the amplitude of the preceding saccade (Krejtz et al. 2016). Focal fixations typically involve longer durations with smaller intervening saccades. This type of gaze indicates attention centered on small, consistent regions of space. On the other hand, ambient fixations are typically shorter in duration and preceded by longer saccades, indicating a more exploratory pattern of gaze.

Team Roles and Responsibilities

The VAL agent was structured to support humans by providing timely assistance in low stress as well as in high-stress situations. The human participants, on the other hand, retained decision-making authority, but could rely on the AI agent for support when needed. This asymmetric team structure enables us to analyze the impact of training in a controlled context in which the agent performs in predictable and repeatable ways. Our goal is to understand the principles of effective collaboration, which will inform the development of more autonomous AI agents and the implementation of complex team configurations in the future.

Preliminary Results and Discussion

Scaffolded Practice Fosters Collaboration

Offering guidance early on in how to leverage AI assistance led to more robust teaming compared to self-paced learning (Figure 2). Participants with scaffolded practice engaged the AI agent more frequently and consistently, suggesting that proper training protocols are crucial for successful human-AI collaboration. They also achieved higher performance scores, suggesting that proper training protocols aiming for successful human-AI collaboration can optimize performance outcomes.

Increased HAT in High-Stakes Situations

We divided games into quarters and computed each team's mean performance in each quarter. Next, we constructed a linear mixed model that predicted each team's performance using the number of agent actions, the segment of game play (divided into quarters), and the interaction between these two variables as predictors. Our findings revealed that engaging the AI agent led to better performance mostly in the third and fourth quarters of the game, when the risk of failure was much higher due to overcrowding of passengers, compared to low-stress situations (se Figure 3). This behavior underscores the importance of designing AI systems that provide timely and effective support during critical times. Furthermore, it demonstrates that players were willing to cooperate with and trust the agent at high stakes moments. Finally – and perhaps most notably – it demonstrates the importance of learning how to utilize AI strategically to mitigate emergent challenges of the moment.

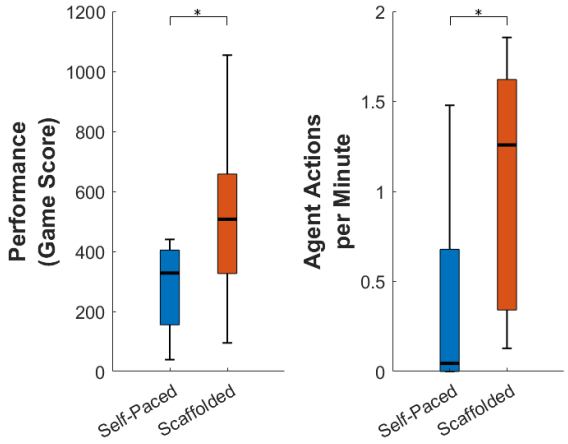

Figure 2: Increased performance and agent interactions were observed after scaffolded relative to self-directed practice.

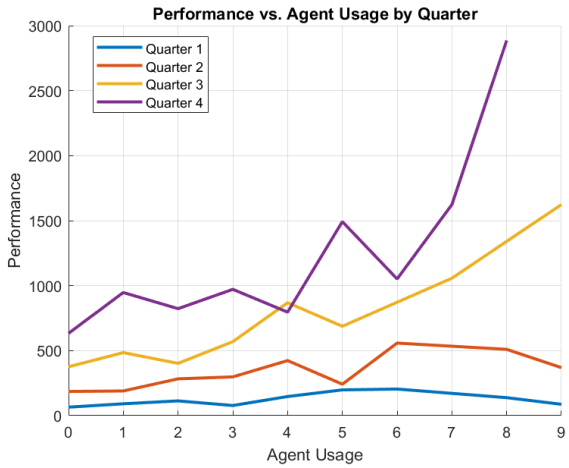

Figure 3: Agent interactions during critical, high-stress game phases were linked to better performance.

Enhancing HAT with Psychophysiological Data

We investigated the link between psychophysiological data and HAT, focusing on eye gaze metrics. Our analysis revealed that shorter fixation durations and longer saccade amplitudes, characteristic of ambient fixation modes, lead to improved performance (Figure 4). Building on these findings, our future research will explore approaches to training and agent assistance that stimulate ambient gaze patterns. We will also utilize diverse psychophysiological metrics (e.g. eye-gaze, pupillometry, cardiological indices, and EEG) to gain deeper insights into the mental states of individuals and understand how these states influence human-AI collaboration. We will also use psychophysiological metrics to guide and inform AI agents in their reasoning and decision-making by providing them with information about human mental states.

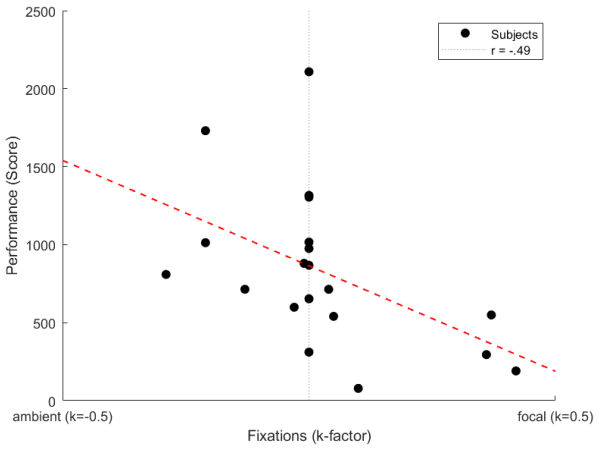

Figure 4: Ambient viewing behavior was linked to better performance.

Conclusion

The present analysis explores a framework for studying factors that can modulate the robustness of human- agent collaboration. Our findings highlight the importance of structured training in fostering effective human-AI collaboration, as scaffolded practice was shown to promote engagement with AI agents. By understanding and incorporating contextual characteristics and psychophysiological metrics, such as eye gaze patterns, collaborative AI systems can be designed that are flexible, robust, and capable of genuine collaboration with humans.

Acknowledgments

This study was made possible by a grant from the Army Research Laboratory (W911NF2120126).

References

Demir, M.; McNeese, N. J.; and Cooke, N. J. 2016. Team communication behaviors of the human-automation teaming. In *2016 IEEE international multi-disciplinary conference on cognitive methods in situation awareness and decision support (CogSIMA)*, 28–34. IEEE.

Krejtz, K.; Duchowski, A.; Krejtz, I.; Szarkowska, A.; and Kopacz, A. 2016. Discerning ambient/focal attention with coefficient K. *ACM Transactions on Applied Perception (TAP)*, 13(3): 1–20.

Lawley, L.; and Maclellan, C. 2024. VAL: Interactive Task Learning with GPT Dialog Parsing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18.

McNeese, N. J.; Demir, M.; Cooke, N. J.; and Myers, C. 2018. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors*, 60(2): 262–273.

Metcalfe, J. S.; Perelman, B. S.; Boothe, D. L.; and Mcdowell, K. 2021. Systemic oversimplification limits the potential for human-AI partnership. *IEEE Access*, 9: 70242–70260.

Vaccaro, M.; Almaatouq, A.; and Malone, T. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1–11.

Wynne, K. T.; and Lyons, J. B. 2018. An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, 19(3): 353–374.