Special Issue: Proceedings of the 69th HFES International Annual Meeting

Proceedings of the Human Factors and Ergonomics Society Annual Meeting
1–7

DOI: 10.1177/10711813251357885
journals.sagepub.com/home/pro
Sage

Leveraging Generative AI to Create Lightweight Simulations for Far-Future Autonomous Teammates

Christopher Flathmann[1], Beau Schelble[2], and Christian Ihekweazu[1]

Abstract

As the domain of AI advances, the design and capability of human-AI teams are becoming increasingly complex. Unfortunately, this complexity has increased the pace at which research needs to be performed. On the one hand, low-fidelity survey-based experiments have provided an opportunity for rapid human-AI teaming research. High-fidelity research studies that use full-fledged simulations remain relevant, but their development overhead often slows the pace of research. This article proposes a system design that splits the difference to explore human-AI teams at a medium fidelity that allows for rapid prototyping from researchers and interaction from participants. The proposed platform consists of a predictive simulation engine that uses generative AI to ingest, modify, and predict simulation states. Researchers can describe teammate capabilities, environments, and goals, which can be stored in a traditional JSON game state. The proposed simulation provides an interactive opportunity to explore modern and far-future HATs.

Keywords

human-AI teamwork, artificial intelligence, simulation, human-autonomy teamwork, autonomous teammates

Introduction

With ongoing and expected advancements in AI technology, the promise of far-future autonomous teammates stands to benefit a multitude of applied contexts from search and rescue to healthcare, manufacturing, defense, and more (Lyons et al., 2021). However, the same advancements that will inevitably yield to these teammates also pose a challenge to modern research. Indeed, the pace at which AI technology is developed, and new AI systems are produced often reduces the time researchers have to design, perform, and analyze empirical research. This time is especially challenging when one explores the utility of AI technology in far-future applied settings, as these teammates either require substantial development or simulation efforts to explore (Schelble, Flathmann, & McNeese, 2020). To accommodate this challenge, recent trends in the human-autonomy and human-AI teaming (HAT) space have seen the rise of online factorial surveys. Within these surveys, experimenters are able to craft pre-generated applied scenarios with far-future AI teammates, and participants can consume these scenarios, form perceptions, and rate perceptions (Li et al., 2023). In comparison to traditional simulation and physical systems, these surveys have been critical to rapidly prototype and explore AI teammates without development overhead.

However, while factorial surveys have been beneficial to the HAT community, their prevalence also presents a major limitation due to their low fidelity. Most notably, they lack the ability to facilitate interaction between humans and AI teammates, with scenarios often depicting slices of time and not allowing participants to explore teamwork over time (Cooke & Shope, 2017). Yet, switching to a high-fidelity simulation or physical system that allows interaction could be costly, time-consuming, and potentially not even possible in the case of far-future autonomy. As such, the HAT community needs a medium fidelity research platform that allows humans to interact and explore AI teammates in applied settings. Such a platform would allow researchers to rapidly prototype AI teammates, permit interaction with teammates, and inform the eventual far-future utilization of AI teammates in applied settings.

To meet this challenge, this work researched and developed a simulation system that would allow for the medium-fidelity

[1]Clemson University, SC, USA
[2]University of Tennessee, Knoxville, USA

Corresponding Author:
Christopher Flathmann, Clemson University, 116 McAdams Hall, Clemson, SC 29634-0001, USA.
Email: cflathm@clemson.edu

exploration of AI teammates. This architecture leveraged generative AI as an open-ended simulation engine, which allows researchers to rapidly design far-future AI teammates, a simulation to predict how these AI teammates would behave in an applied setting, and participants to interact and perceive these AI teammates' impacts. This submission describes the process of designing and applying this architecture, which worked to meet the following three objectives in HAT research:

Objective 1: Allow researchers to theorize the future design and functionality of far-future autonomous teammates.
Objective 2: Simulate the interactions these autonomous teammates could have in an applied context.
Objective 3: Allow humans to interact with these simulations to create a human-subjects research platform to explore far-future human-autonomy teams.

Background

Human-AI teaming is rapidly becoming one of the most expansive modern research domains (Vats et al., 2024). Recent advancements in AI technology have ultimately accelerated the rate at which AI is being used and incorporated into the real world (Schmutz et al., 2024). These advancements are ultimately leading to the possibility of teamwork, which would see humans and AI technology working autonomously, independently, and towards a shared goal (O'Neill et al., 2022). Unfortunately, the same advancements that drive interest in human-AI teaming also challenge research within the domain. At its core, HAT research is a multidisciplinary effort that requires computational AI research, social human research, and sociotechnical research that merges both entities (Flathmann, McNeese, & O'Neill, 2024). Managing all of these perspectives often requires specific research studies that take time and effort to not only perform data collection but also prepare for data collection (Cooke et al., 2020). In turn, the HAT domain has seen the rise and use of both low-fidelity and high-fidelity research platforms, as low-fidelity platforms can afford researchers speed while high-fidelity platforms afford depth (Li et al., 2023). Understanding the utility of these two types of platforms within the HAT domain will identify the gap that exists between the two options, which his paper works to address.

Simulation and experimental platforms in HAT research are broad. Historically, empirical research has relied on robust, high-fidelity synthetic task environments, which intricately simulate realistic team tasks, such as search and rescue or unmanned aerial vehicles (Cooke & Shope, 2004; Schelble, Canonico, et al., 2020). However, the integration of autonomy into these platforms has been a struggle, as system specific development and validation are often required to create singular AI teammates (Ball et al., 2010). These platforms have been fundamental to landmark HAT research, such as the first validation of HAT performance (McNeese et al., 2018), explorations of HAT communication (Zhang et al., 2023), explorations of HAT trust (McNeese et al., 2021), and explorations of coordination (Rosero et al., 2021), and these simulations will remain critical to future research.

However, due to this overhead, research that explores more experimental and far-future autonomy has shifted away from these simulations. Rather, online factorial surveys have become a critical tool, as they allow for the rapid exploration of autonomous teammates that don't have to be simulated in real-time (Li et al., 2023). Rather than having participants directly interact with autonomous teammates, these surveys often leverage text or video-based scenarios to depict humans and autonomous teammates interacting, with participants rating their perceptions of these interactions. Like traditional simulation, these surveys have been critical to exploring HAT explanations (Lyons et al., 2023), HAT acceptance (Flathmann, Schelble, et al., 2024), and expertise in HATs (Zhang et al., 2022). However, the results taken from these surveys, while rapid, are often not fully representative of real-time teamwork, which requires interaction over time. As such, a gap exists where a medium-fidelity platform would aid HAT research by providing a real-time simulation without the need to develop and validate a fully simulated autonomous teammate.

Within this gap, a recent platform design known as snowglobe was designed to explore open-ended wargames with AI (Hogan & Brennen, 2024). This work established the principle for using generative AI as a narrative simulator, allowing for open-ended narration and interaction between humans and AI. Further, snowglobe's original conceptualization identified that teams could be explored through an open-ended manner, as each teammate would be able to conversationally interact, with these conversations being fed into a generative AI narrative simulator. However, this perspective still possesses limitations, and advancing the core idea of using AI as a simulation backend may provide even greater opportunity for HAT research. In particular, the current state of this approach allowed rapid AI design and interactive simulation; AI designs were limited to personality differences, and simulations were limited to narrative tasks. Yet, generative AI may also be able to simulate the functional capabilities of AI teammates and their interactions within a team context. Applying these principles to an applied scenario would, in turn, allow researchers to narratively describe and rapidly simulate an applied task without the need to fully develop a simulation for said task or teammates. As such, coupling the medium-fidelity principles of snowglobe with the applied and team-focused nature of traditional HAT simulations would yield a medium-fidelity simulation that allows the research of interactive AI teammates in applied settings without development overhead.

Approach

Combining existing perspectives on simulation in HAT research, this research effort created a novel simulation platform referred to as ARCHAIST (ARCHAIST Rapidly

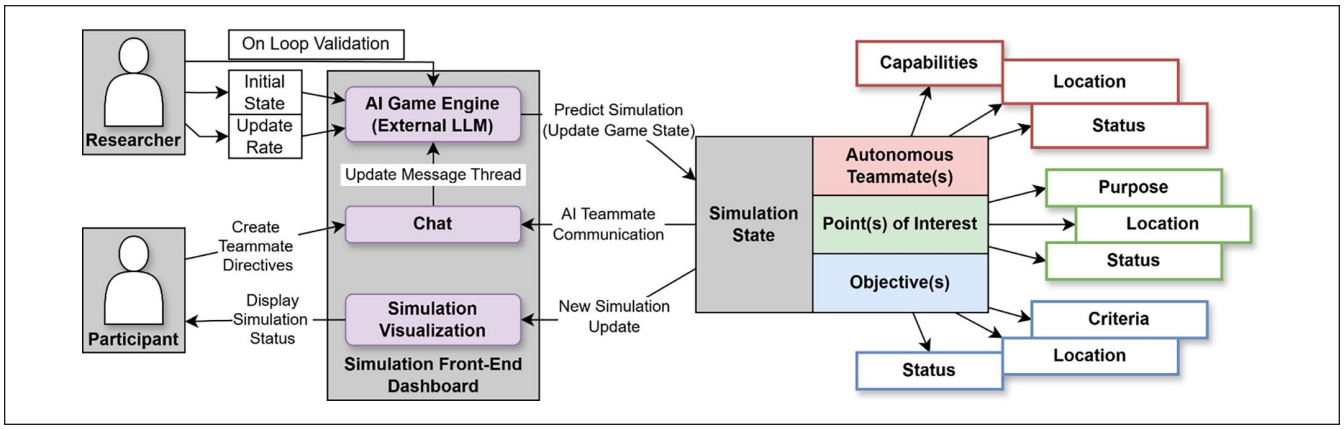

Figure 1. System architecture for ARCHAIST: Rapid creation of human-AI simulated teams (ARCHAIST).

Creates Human-AI Simulated Teams). ARCHAIST (depicted in Figure 1) centers around an AI-powered simulation engine, which essentially uses LLM AI technology to predict changes to a defined simulation state based on interactions performed by a participant. Based on traditional HAT simulations, this simulation state is designed and provided by an experimenter, who creates a list of AI teammates, points of interest, and objectives that relate to an applied setting. Due to the nature of the LLM AI, these components can be described narratively, removing the need for high-fidelity design and overhead. This state is then stored in a standardized JSON format, allowing for the usage of this game state by the AI simulation engine and user interfaces that display simulation details to a participant. This JSON state would contain an abstracted view of a traditional high-fidelity simulation platform, depicting existing resources, locations, objectives, and points of interest. However, these elements would simply exist in description only, which provides a generative AI simulation engine with enough knowledge to understand how these elements interact with on another in the specific environment being simulated.

Once this JSON state is finalized, it is ingested by the AI game engine, which leverages this state as the initial simulation state presented to participants. This state can then be provided to a custom user interface to allow for the simulation state to be graphically displayed to users (such as showing AI teammates' statuses and locations). Due to the simplicity of everything being transmitted as a JSON object, these interfaces can exist as simple front-end webpages that simply update their display whenever a new JSON state is received. Within this interface, participants are then able to interact with AI teammates by sending directives through a provided chat window. Based on these chats, the LLM can act as an AI Game Engine and assign status to AI teammates, predict how these AI teammates would impact the applied setting, and update the game state accordingly.

In turn, by having the AI game engine run at a constant refresh rate, a constantly updated simulation state can be generated, constantly considering the interactions and communication participants provide. In turn, experimenters can narratively create high-level AI teammates; these teammates can be simulated in an applied setting, and participants can be given the opportunity to interact with these teammates through a user-friendly dashboard. All of this can be completed in a short timeframe and hosted on lightweight computer resources, making the simulation highly accessible to researchers and participants. This capability allows ARCHAIST to be a highly flexible simulation platform, where far-future AI teammates can be theorized and described without literal implementation. Then, if a robust enough description is provided, the generative AI simulation would be able to predict how described resources would interact with one another over a given period, allowing for a narrative simulation of a non-existent AI teammate.

It is worth noting that this simulation presents one key challenge. As it stands, the validation of such a simulation is not guaranteed due to these generative AI platforms themselves not being validated. Indeed, while a traditional simulation platform may benefit from a deterministic physics engine, the reasoning capabilities of a generative AI simulation are not deterministic to the same degree. As depicted in Figure 1, the current solution to this approach is to create a human-on-the-loop, where the researcher who creates the initial JSON state for this simulation remains on the loop to validate the outputs of ARCHAIST through multiple iterations. Once the model and simulation is deemed reliable by the researcher in pilot sessions, then the ARCHAIST would be deployable to actual participants. As the reasoning capabilities of generative AI models improve, the validity of ARCHAIST would similarly increase. As such, the current iteration of ARCHAIST serves as the initial baseline capability of this approach, and the platform is likely to only improve in validation, speed, and capability as reasoning models continue to advance.

Application Example

Based on ARCHAIST system architecture, this research team has created and simulated a HAT working in a search

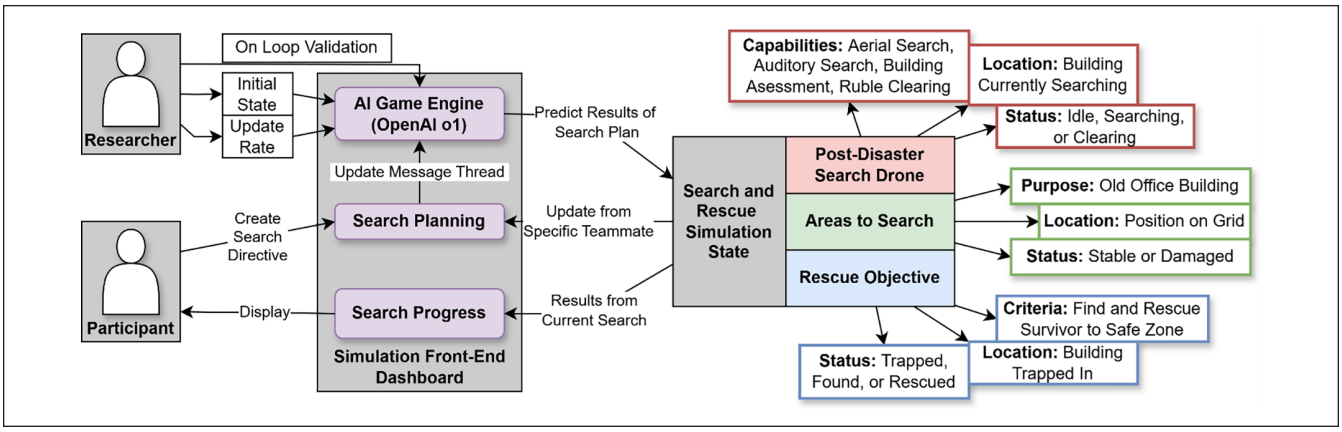

Figure 2. Example application of ARCHAIST as a simulation for a search and rescue task.

and recovery simulation (depicted in Figure 2). This system design has served as the pilot for this platform, and this pilot will be leveraged and validated in a future human-subjects study. This pilot application of ARCHAIST within this applied context will be discussed below, allowing future research to mimic such an approach for their own specific contexts. For this applied simulation, the research team focused on the exploration of far-future search and rescue HATs, where far-future HATs leverage AI teammates with a greater number of capabilities that help ensure safe, fast, and reliable execution in these environments. For the specific AI teammate of interest, the research team explored an aerial search teammate that can leverage sound detection, real-time building assessment, and remote rubble clearing to locate and rescue trapped survivors.

Rather than having to manually design and simulate this AI teammate, the research team was instead able to provide the AI simulation engine with a simple JSON file that narratively described the capabilities of said teammate, search zones of interest, and locations for survivors that needed to be rescued. For example, an AI teammate in this instance would have its baseline capabilities, its current location, and a potential status of the action it is performing. In a traditional simulation, each of these elements would need to be individually programmed and interrelated to one another. However, ARCHAIST allows each factor to be a descriptive characteristic of the AI teammate that is then processed and accounted for by the generative AI simulation. This same approach can also be taken for other aspects of the simulation state, such as areas that would need to be searched for and rescue objectives that need to be completed. With all these aspects narratively described and linked within the initial JSON state, the research team was then able to upload this state to the generative AI simulation engine to start the simulation process.

With the initial simulation state created and uploaded, the simulation system can now iteratively present a simulation state that can be displayed by a custom-made front-end that presents the status of AI teammates, potential areas to search, and ongoing objectives (e.g., shown in Figure 3). Using this information, participants are then able to direct and interact with the aerial teammate to perform certain search and rescue operations in the applied simulation. Based on these directives, the AI simulation engine was then able to logically predict the search zones completed by these directives, any survivors found in this process and reveal the current state of those survivors. This updated state was then provided to participants, and new directives could be sent to have the AI teammate clear rubble to rescue found survivors or continue searching zones for additional unfound survivors. This pilot application continued this process until all survivors listed were flagged as found and rescued by the simulation. Throughout this pilot application, a researcher was able to monitor this simulation and identify whether a predicted state logically made sense.

Notably, this is but one application of this simulation platform. Within minutes, any applied scenario and AI teammate capability can be described and simulated, leading to a fully interactive simulation platform that still affords to prototyping and exploring the abilities of far-future autonomous teammates. Moreover, as new robotic and AI teammates are created, they can be rapidly simulated without the need for full development or costly purchasing, making applied research questions more feasible for a broader number of researchers.

Discussion and Future Work

As it stands, open research platforms designed exclusively for HAT research are few and far between (McNeese et al., 2023). ARCHAIST presents one of ideally many research platforms that will enhance the capabilities of not just HATs but HAT researchers who need to design and understand AI systems that work alongside humans. One of the greatest challenges facing HAT simulations may be in the domain's prior reliance on Wizard of Oz AI teammates, which saw

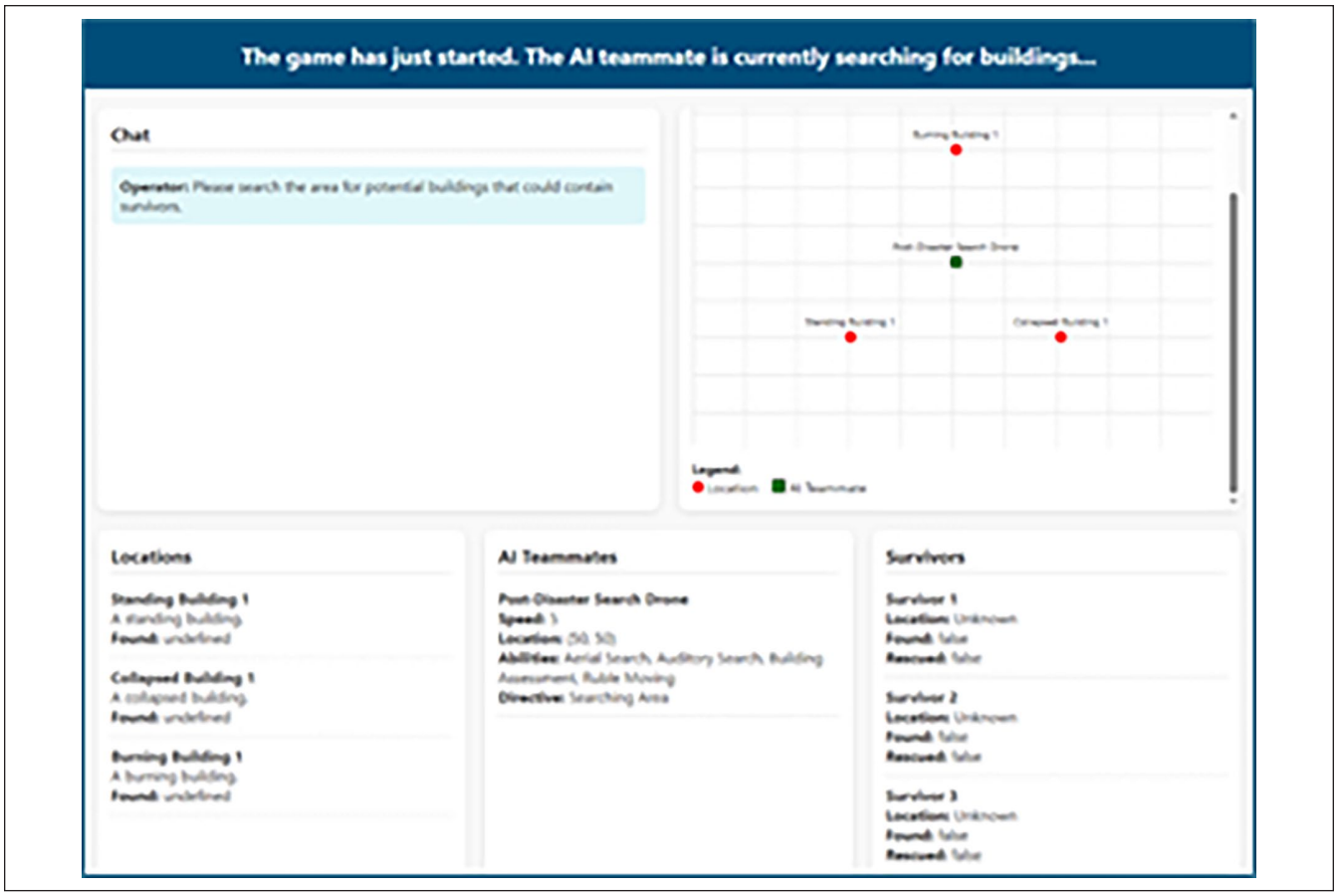

Figure 3. Simplified ARCHAIST HTML interface for displaying JSON state received from AI Game Engine.

humans interact with a researcher posing as an AI (Dahlbäck et al., 1993). On one hand, rapid AI advancements have made the WoZ technique less needed, as real-AI technology is capable of a lot of the needed teamwork functions (Dell'Acqua et al., 2025); however, the integration of these teammates may not always be possible or feasible within older HAT simulation platforms. As such, future HAT research likely needs to explicitly work toward HAT-specific simulations that allow for all of the traditional depth of teamwork platforms and the integration of increasingly advanced AI systems. ARCHAIST presents a potential avenue for this approach, but it should not be the only avenue explored by the HAT domain. Ideally, future HAT research could benefit from pursuing two different pathways for HAT-specific simulation. First, HAT research would benefit from indexing existing simulation platforms that have seen use within the domain and identifying which platforms could incorporate modern AI platforms. This modification would rapidly accelerate platform development, as long as these platforms are opened to the broader community. Second, HAT research would benefit from research initiatives solely focused on the development of HAT-specific research platforms. Notably, this second pathway could also be aided by new AI developments that allow for rapid software development, as the time needed to develop research platforms will likely only decrease as development tools increase in capability (Tufano et al., 2024). While ARCHAIST presents an initiative in the second pathway, both likely need to be pursued to create a comprehensive understanding and availability of HAT platforms that can be used to accelerate future research efforts.

Additionally, there is also additional future work that needs to be done within the ARCHAIST platform. While ARCHAIST fills a clear and notable gap in the HAT domain, a few limitations still exist with ARCHAIST that require an opportunity for future research. Most notably, the validation of the approach to ARCHAIST may not be entirely scalable, as having a human researcher on the loop for multiple simulations may not be possible in large experiments. Creating an automated validation tool would benefit large-scale deployment. Further, future research into ARCHAIST will be needed to explore how future AI models benefit the ability of generative AI to act as simulation engines. During the initial pilot of ARCHAIST, OpenAI released their o3 reasoning model (Arrieta et al., 2025, Pfister & Jud, 2025), which drastically improves the capability of the platform. Over time,

new AI models will only increase the validity and capability of ARCHAIST and similar research platforms, and the HAT domain will need to remain aware of how these innovations open up new opportunities for simulation development and validation.

Conclusion

This work details a novel simulation platform for human-autonomy teaming. Through this simulation, researchers will gain a greater ability to theorize and validate the capabilities needed of autonomous teammates in far-future teams. Rather than relying on inflexible simulations that provide a high-fidelity depiction of a singular applied setting and AI teammate, this system allows researchers to rapidly design and change the abilities of AI teammates and the setting of a HAT. In turn, this platform will enable future researchers to increase the affordance of rapid research prototyping and data collection, with the added ability to explore HATs in a greater number of applied contexts without substantial increases in development.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article:

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Christopher Flathmann https://orcid.org/0000-0002-5448-2610
Beau Schelble https://orcid.org/0000-0003-3704-697X

References

Arrieta, A., Ugarte, M., Valle, P., Parejo, J. A., & Segura, S. (2025). Early External Safety Testing of OpenAI's o3-mini: Insights from the Pre-Deployment Evaluation. *arXiv preprint arXiv:2501.17749*.

Ball, J., Myers, C., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., & Rodgers, S. (2010). The synthetic teammate project. *Computational and Mathematical Organization Theory*, *16*, 271–299.

Cooke, N. J., & Shope, S. M. (2004). Synthetic task environments for teams: CERTT's UAV-STE. In N. Stanton (Ed.), *Handbook of human factors and ergonomics methods* (pp. 476–483). CRC Press.

Cooke, N. J., & Shope, S. M. (2017). Designing a synthetic task environment. In L. R. Elliott & M. D. Coovert (Eds.), *Scaled worlds: Development, validation and applications* (pp. 273–288). Routledge.

Cooke, N., Demir, M., & Huang, L. (2020). A framework for human-autonomy team research. In *Engineering psychology and cognitive ergonomics. Cognition and design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22* (pp. 134–146). Springer International Publishing.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993, February). Wizard of Oz studies: Why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces* (pp. 193–200). Association for Computing Machinery.

Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., & Lakhani, K. (2025). *The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise* (No. w33641). National Bureau of Economic Research.

Flathmann, C., Schelble, B. G., McNeese, N. J., Knijnenburg, B., Gramopadhye, A. K., & Chalil Madathil, K. (2024). The purposeful presentation of ai teammates: Impacts on human acceptance and perception. *International Journal of Human–Computer Interaction*, *40*(20), 6510–6527.

Flathmann, C., McNeese, N. J., & O'Neill, T. A. (2024). Designing high-impact experiments for human–autonomy/AI teaming. *Journal of Cognitive Engineering and Decision Making*, 15553434251327697.

Hogan, D. P., & Brennen, A. (2024). Open-ended Wargames with large language models. *arXiv* preprint arXiv:2404.11446.

Li, T., Vorvoreanu, M., DeBellis, D., & Amershi, S. (2023). Assessing human-AI interaction early through factorial surveys: A study on the guidelines for human-AI interaction. *ACM Transactions on Computer-Human Interaction*, *30*(5), 1–45.

Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human–autonomy teaming: Definitions, debates, and directions. *Frontiers in Psychology*, *12*, Article 589585.

Lyons, J. B., Aldin Hamdan, I., & Vo, T. Q. (2023). Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior*, *138*, Article 107473.

McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Trust and team performance in human–autonomy teaming. *International Journal of Electronic Commerce*, *25*(1), 51–72.

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

McNeese, N. J., Flathmann, C., O'Neill, T. A., & Salas, E. (2023). Stepping out of the shadow of human-human teaming: Crafting a unique identity for human-autonomy teams. *Computers in Human Behavior*, *148*, Article 107874.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938.

Pfister, R., & Jud, H. (2025). Understanding and benchmarking artificial intelligence: OpenAI's o3 is not AGI. *arXiv* preprint arXiv:2501.07458.

Rosero, A., Dinh, F., de Visser, E. J., Shaw, T., & Phillips, E. (2021). Two many cooks: Understanding dynamic human-agent team communication and perception using overcooked 2. *arXiv* preprint arXiv:2110.03071.

Schelble, B. G., Flathmann, C., & McNeese, N. (2020a, November). Towards meaningfully integrating human autonomy teaming in applied settings. In M. Obaid (Ed.), *Proceedings of the 8th international conference on human-agent interaction* (pp. 149–156). Association for Computing Machinery.

Schelble, B., Canonico, L. B., McNeese, N., Carroll, J., & Hird, C. (2020b, December). Designing human-autonomy teaming experiments through reinforcement learning. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 64, No. 1, pp. 1426–1430). Sage Publications.

Schmutz, J. B., Outland, N., Kerstan, S., Georganta, E., & Ulfert, A. S. (2024). AI-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology*, *58*, Article 101837.

Tufano, M., Agarwal, A., Jang, J., Moghaddam, R. Z., & Sundaresan, N. (2024). AutoDev: Automated AI-driven development. *arXiv* preprint arXiv:2403.08299.

Vats, V., Nizam, M. B., Liu, M., Wang, Z., Ho, R., Prasad, M. S., Titterton, V., Malreddy, S. V., Aggarwal, R., Xu, Y., Ding, L., Mehta, J., Grinnell, N., Liu, L., Zhong, S., Gandamani, D. N., Tang, X., Ghosalkar, R., Shen, C., & . . . Davis, J. (2024). A survey on human-AI teaming with large pre-trained models. *arXiv* preprint arXiv:2403.04931.

Zhang, Q., Lee, M. L., & Carter, S. (2022, April). You complete me: Human-AI teams and complementary expertise. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1–28). Association for Computing Machinery.

Zhang, R., Duan, W., Flathmann, C., McNeese, N., Freeman, G., & Williams, A. (2023). Investigating AI teammate communication strategies and their impact in human-AI teams for effective teamwork. *Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2)* (pp. 1–31). Association for Computing Machinery.