Toward a Generalized Model of Human–AI Team Effectiveness

Lixiao Huang

Arizona State University, Mesa, AZ 85212
lixiao.huang@asu.edu

Abstract

This extended abstract proposes a generalized model for evaluating human–AI team effectiveness, capturing its dynamic nature and the complex relationships among contributing variables. While future research may enrich or expand upon the set of variables, the underlying framework is intended to remain conceptually robust.

Position Description

Human–AI teaming has become a cornerstone of modern operations across numerous sectors, necessitating a robust framework to assess and enhance team effectiveness. AI's flexibility allows it to take on various roles in human–AI teams. It may serve as an advisor, overseeing and guiding individuals or teams during complex tasks (Freeman et al. 2021; Huang et al. 2022a), a facilitator coordinating human discussions, or a decision aide that helps humans make faster and more accurate decisions. Additionally, AI may act as a task performer, either completing tasks that are typically handled by humans (Wang et al. 2023), or working alongside one or more humans in dynamic, action-based environments (e.g., DARPA ADAPTII.2). The AI systems may or may not have a physical embodiment, which can influence users' perceptions of their affordances. However, all AI systems must provide an interface for human communication, whether through natural language, touchscreens, signal displays, or other sensory-based actuators (Baker et al. 2021). Depending on the requirements of the task domain and the AI's roles and capabilities, the environment, personnel, resources, adversaries, team goals, and task procedures may all vary. The interaction processes likewise vary depending on the context. These variables pose challenges for measuring team effectiveness across domains.

Traditional IPO (McGrath 1964) and IMO (Ilgen et al. 2005; Mathieu et al. 2008) models have limitations to capture the dynamic and complex relationships among the variables in evaluating human–AI team effectiveness. (Cooke et al. 2013) suggested team cognition is best measured through team activities. (Mathieu et al. 2008) also distinguished team states from team processes within the IMO model, though without specifying the explicit relationships between them. To address this gap in the literature, we propose a generalized model of human-AI team effectiveness (see Figure 1) that can be applied across diverse human–AI teaming scenarios using customizable components. This modified Input-Processes-States-Outcomes (IPSO) model of team effectiveness (Huang et al. 2020) provides a foundational framework for understanding and evaluating various human–AI teams. Unlike IPO/IMO models, it separates interaction processes from team states, emphasizes the dynamic nature of human–AI interactions, and requires non-intrusive and dynamic measures and modeling. Each component can be tailored to a specific domain and analyzed based on stakeholders' priorities. The evaluation criteria—whether the team performed well on selected measures for each variable—should be determined based on hypotheses from literature reviews, input from subject matter experts, and the distribution of the data. These criteria can also be cross validated using other variables and data modalities (e.g., whether communication data aligns with survey responses and objective team scores). By recognizing the iterative feedback processes and multiple interaction pathways, the model enhances our ability to apply the universal team effectiveness model in a targeted and effective manner.

Distributed Dynamic Team Cognition Perspective. Human–AI teaming is not limited to dyadic interactions but unfolds within broader team networks. The concept of Distributed Dynamic Team Trust (D2T2) (Huang et al. 2021) introduces an added layer of team complexity. It considers multiple stakeholders (e.g., engineers, pilots, trainers, managers, and AI components), and multimodal communication among members shapes how individuals perceive and respond to one another—and how their attitudes toward AI components are formed and transferred. Measures for evaluating these dynamics range from one-time surveys to dynamic action tracking and physiological monitoring, depending on team size and operational tempo (see Figure 2). To date, more research has focused on static than dynamic measures. However, well-designed experiments with time-stamped data can support advanced analytics techniques such as machine learning, even when working with relatively small datasets (Bustamante Orellana et al. 2025). The D2T2 framework also provides a foundation for studying broader distributed and dynamic team cognition concepts.

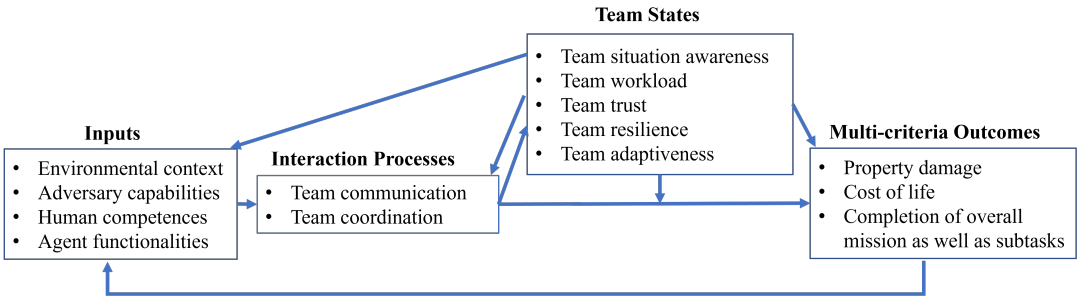

Figure 1: A Generalized Model of Human-AI Team Effectiveness (Huang et al. 2020)

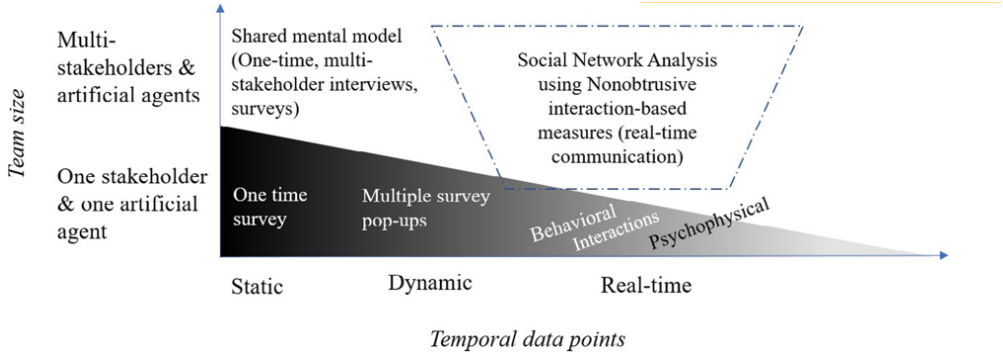

Figure 2: Measures as a Function of Team Size and Tempo (Huang et al. 2021)

Lifecycle Perspective. By accounting for product cycles, the model aligns with the lifecycle design and development of AI systems and their iterative interactions with human users (Miller et al. 2023). Traditional human–AI teaming has focused on the phase where human subjects are recruited to test the prototypes. However, systematic evaluation should span the entire AI system lifecycle—including initial prototyping, deployment, system upgrades, integration with complementary systems, and eventual retirement—with each stage presenting distinct priorities and challenges. Even with limited resources, conducting extensive evaluations across multiple updated versions of the AI system is more valuable than relying on a small number of prototype tests. This approach enables participants to build calibrated trust relationship with the evolving system.

Model Application. The generalized evaluation model applies to datasets from virtual environments—where timestamps and event logs are more readily available—as well as physical environments, where capturing and synchronizing time-stamped data may be more challenging. Researchers should maximize the utility of available data to assess team effectiveness. Applying the model to existing datasets requires analyzing the study's background, design rationale, and the availability and structure of collected data. The model can also support future studies by informing experimental design, defining team roles, allocating tasks, and developing data collection strategies (Cooke, Demir, and Huang 2020). This workshop presentation uses the DARPA ASIST and ADAPTII.2 datasets (see Table 1) as illustrative examples of how key components of team effectiveness can be examined. These datasets vary in team composition (e.g., team size, intact vs. ad hoc teams, fixed vs. exchangeable roles, member competencies, environmental risks, adversary capabilities, and task procedures), interaction processes (e.g., communication and coordination), team states (e.g., trust and workload), and outcomes (e.g., survival duration, damage taken, victims rescued, and task efficiency). Measuring these variables enables statistical and qualitative analyses, dynamic modeling, and novel illustrations of variable relationships.

Applying this model to existing and future research may enrich or expand the set of variables, while the underlying framework is designed to remain conceptually robust. As empirical findings accumulate, the model can evolve into a comprehensive library of effectiveness measures tailored to specific domain contexts. Integrating dynamic measures of human–AI team effectiveness will further enable researchers to explore how real-time feedback mechanisms enhance AI-adaptive behavior in team environments. By offering a flexible yet rigorous structure, the model has the potential to unify diverse lines of human–AI–team research under a common evaluative lens—bridging theoretical development with applied team performance assessment. The author welcomes inquiries on the model's use or refinement.

Acknowledgments

This extended abstract builds upon previous work funded by ARL under the Contract No. W911-NF-182-0271 (NGCV)

	ASIST Study 2	ASIST Study 3	ASIST Study 4	ADAPTII.2
Task Scenario	USAR	USAR	Bomb disposal	Tower defense
AI Role	Passive monitor	Advisor	Advisor	Task performer
Data Type	Screen recordings, natural language, surveys, time-stamped testbed logs	Screen recordings, natural language, surveys, time-stamped testbed logs, physiological sensors	Natural language, surveys, time-stamped testbed logs	Screen recordings, natural language, surveys, and time-stamped testbed logs
Public Availability	(Huang et al. 2022c)	(Huang et al. 2022b; Pyarelal et al. 2023)	(Huang et al. 2024)	Anticipated in Fall 2025

Table 1: Summary of DARPA ASIST and ADAPTII.2 Datasets Characteristics

and the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0130 (ASIST) and 140D0423C0095 (ADAPTII.2). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

Baker, A.; Fitzhugh, S.; Huang, L.; Forster, D.; Scharine, A.; Neubauer, C.; Lematta, G.; Bhatti, S.; Johnson, C. J.; Krausman, A.; et al. 2021. Approaches for assessing communication in human-autonomy teams. *Human-Intelligent Systems Integration*, 3(2): 99–128.

Bustamante Orellana, C.; Rodriguez Rodriguez, L.; Huang, L.; Cooke, N.; and Kang, Y. 2025. Machine learning for automation usage prediction: identifying critical factors in driver decision-making. *Applied Intelligence*, 55(1): 12.

Cooke, N.; Demir, M.; and 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*, 134–146. Springer.

Cooke, N.; Gorman, J.; Myers, C.; and Duran, J. 2013. Interactive team cognition. *Cognitive science*, 37(2): 255–285.

Freeman, J.; Huang, L.; Wood, M.; and Cauffman, S. 2021. Evaluating artificial social intelligence in an urban search and rescue task environment. In *Aaai fall symposium*, 72–84. Springer.

Huang, L.; Cooke, N.; Johnson, C.; Lematta, G.; Bhatti, S.; Barnes, M.; and Holder, E. 2020. Human-autonomy teaming: Interaction metrics and models for next generation combat vehicle concepts. *ARIZONA STATE UNIV EAST MESA AZ, Tech. Rep*.

Huang, L.; Cooke, N. J.; Gutzwiller, R.; Berman, S.; Chiou, E.; Demir, M.; and Zhang, W. 2021. Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In *Trust in human-robot interaction*, 301–319. Elsevier.

Huang, L.; Fouse, A.; Cooke, N.; and Weiss, E. 2024. Artificial Social Intelligence for Successful Teams (ASIST) Study 4 Dragon Testbed Dataset. https://doi.org/10.48349/ASU/ZO6XVR.

Huang, L.; Freeman, J.; Cooke, N.; Colonna-Romano, J.; Wood, M.; Buchanan, V.; and Cauffman, S. 2022a. Exercises for Artificial Social Intelligence in Minecraft Search and Rescue for Teams. https://doi.org/10.17605/OSF.IO/JWYVF.

Huang, L.; Freeman, J.; Cooke, N.; Colonna-Romano, J. Wood, M.; Buchanan, V.; and Cauffman, S. 2022b. Artificial Social Intelligence for Successful Teams (ASIST) Study 3. https://doi.org/10.48349/ASU/QDQ4MH.

Huang, L.; Freeman, J.; Cooke, N.; Dubrow, S.; Colonna-Romano, J. Wood, M.; Buchanan, V.; Cauffman, S.; and Yin, X. 2022c. Artificial Social Intelligence for Successful Teams (ASIST) Study 2. https://doi.org/10.48349/ASU/BZUZDE.

Ilgen, D.; Hollenbeck, J.; Johnson, M.; and Jundt, D. 2005. Teams in Organizations: From Input-Process-Output Models to IMOI Models. *Annual review of psychology*, 56: 517–43.

Mathieu, J.; Maynard, M. T.; Rapp, T.; and Gilson, L. 2008. Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of management*, 34(3): 410–476.

McGrath, J. E. 1964. Toward a "theory of method" for research on organizations. *New perspectives in organization research*, 533: 533–547.

Miller, C.; Barber, D.; Holder, E.; Huang, L.; Lyons, J.; Roth, E.; and Wauck, H. 2023. LifeCycle Transparency: Why, and How, Transparency Information Exchange Should be Distributed throughout the Life of Technology Usage. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 67, 409–413. SAGE Publications Sage CA: Los Angeles, CA.

Pyarelal, A.; Duong, E.; Shibu, C. J.; Soares, P.; Boyd, S.; Khosla, P.; Pfeifer, V.; Zhang, D.; Andrews, E.; Champlin, R.; Raymond, V.; Krishnaswamy, M.; Morrison, C.; Butler, E.; and Barnard, K. 2023. The ToMCAT Dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.