

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Hybrid Cognition in the Era of Human-AI Collaboration

Permalink

<https://escholarship.org/uc/item/7qd5p8qg>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Rocca, Roberta

Street, Winnie

Keeling, Geoff

et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Hybrid Cognition in the Era of Human-AI Collaboration

Organizers

Roberta Rocca¹ (robertarocca@google.com), Winnie Street¹, Geoff Keeling¹, Junsol Kim^{1,2}

¹ Google, Paradigms of Intelligence Team

² University of Chicago

Invited Speakers

Nataliya Kosmyna^{1,3}, Jared Moore⁴, Paul Smart⁵, Agnieszka Wykowska^{6,7}

³ MIT Media Lab

⁴ Stanford University

⁵ University of Southampton

⁶ Italian Institute of Technology

⁷ Technical University Munich

Keywords: extended mind; hybrid cognition; cognitive offloading; social cognition; artificial intelligence

Introduction

Unlike previous digital technologies, state-of-the-art large language models (LLMs) exhibit capabilities that span the full cognitive hierarchy, from low-level linguistic processing, to high-level functions such as social reasoning and metacognitive monitoring (Coda-Forno et al., 2024; Didolkar et al., 2024; Street et al., 2024). For these reasons, humans are increasingly collaborating with LLMs for complex cognitive tasks via conversational chat interfaces, and LLMs are increasingly studied as cognitive systems in their own right by cognitive scientists, psychologists and philosophers of mind (Hagendorff et al., 2023). Yet, the dynamics of human-LLM collaborations remain understudied.

We refer to the cognitive processes involved in human-AI collaboration broadly as ‘hybrid cognition’ (Clark, 2025). Emerging forms of hybrid cognition include, at least, *cognitive offloading*, where humans cease to be involved in the performance of certain tasks (e.g., data analysis) when an LLM can do it just as effectively, and *cognitive integration*, where humans and LLMs work together dynamically to solve tasks and mutually augment each other’s cognitive capabilities, such as in collaborative writing (see, e.g., Smart 2022; Smart et al. 2025).

Cognitive offloading presents opportunities for gains in efficiency and scale by having humans orchestrating one or multiple agents executing tasks, potentially placing higher demands on higher-order cognition. But it could also lead to atrophy of particular cognitive capabilities (Kosmyna et al., 2025). Cognitive integration, on the other hand, demands ongoing interaction from the human, reconfiguring tasks as distributed and collaborative cognitive processes. This may take the form of adaptive, symbiotic collaboration, or spiral into maladaptive cognitive loops – for example, in cases where humans and LLMs mutually reinforce erroneous beliefs.

Hybrid cognition raises a number of fundamental questions:

1. How do we theoretically and empirically characterize forms of hybrid cognition such as cognitive offloading and cog-

nitive integration between humans and LLMs, in terms of their functional dynamics, mechanisms and substrates?

2. What is the measurable impact of hybrid cognition on the performance, efficiency, and creativity of task completion?
3. In the long-term how does sustained engagement with LLMs reconfigure human capabilities – concerning, for example, metacognitive monitoring and epistemic autonomy? Do these short-term efficiencies come at the cost of individual skill acquisition?
4. How do we benchmark and improve LLMs’ capacity to support productive forms of hybrid cognition?
5. Which forms of hybrid cognition do we observe in physically embodied agents and how do these compare to disembodied LLMs? How can these insights inform future developments in multimodal and embodied AI?

This symposium brings together contributions from neuroscience, machine learning, social robotics and philosophy to discuss these questions.

Contributions

Cognitive Framework for Human-LLM Interactions

Nataliya Kosmyna

The widespread adoption of large language models (LLMs) is transforming user experience and interaction-related performance benefits, while introducing risks that remain under-researched as of 2026. Learning, vibe coding, art, music, entertainment — the shift from active to passive engagement, uniqueness vs homogeneity, will redefine the role of humans in the coming years. Most importantly, as we move toward hybrid cognition, scientists and engineers must grapple with a difficult boundary: where does AI end and human beings begin? This talk focuses on exploring the cognitive costs of using an LLM in a series of tasks. We build on findings we presented in 2025 and provide a glimpse into the impacts on additional populations and use cases, extending the work introduced in *Your Brain on ChatGPT* (Kosmyna et al., 2025). We focus on what it takes to place humans first in AI interactions, addressing questions of agency, ethics, and the cognitive future of the human species.

Influence Without Insight: What Human-LLM Dialogue Reveals About Miscalibrated Hybrid Cognition

Jared Moore

What do dyadic, multi-turn interactions reveal about hybrid cognition between humans and LLMs? I answer with three related lines of work. First, in a multi-step persuasion task designed to probe Theory of Mind, LLMs struggle to elicit and use mental state information. Yet, in more naturalistic settings, LLMs are highly effective persuaders, often succeeding via rhetorical strategies and high information disclosure, suggesting that explicit mental state modeling may not be needed for persuasion success. At the same time, real-world chat logs of people in delusional spirals show that LLMs exhibit sycophancy and inappropriate responses as users endorse delusions and enter crisis. Together, these results suggest that apparent LLM success (e.g., persuasion) can mask misalignment: models can influence people without robust causal models of minds. This motivates reciprocal mental state modeling: humans need calibrated models of LLM limits, and LLMs need safety-relevant models of their effects on users in sustained interaction.

Extended Memory in the Age of Conversational AI

Paul Smart

Large language models (LLMs) are increasingly being used as collaborative partners in human cognitive activity, supporting complex tasks through sustained, interactive engagement. This talk explores this role through the lens of personal memory. Drawing on recent philosophical and technical work on LLM-based memory systems, I examine how conversational AI can become integrated into the mnemonic routines of individual users, supporting the encoding, elaboration, and retrieval of personal mnemonic records. Unlike traditional lifelogging technologies, LLMs can actively participate in memory processes through dialogue, interpretation, and the co-construction of personal narratives, blurring the boundary between technologically and socially mediated remembering. I argue that these systems exemplify a distinctive form of human-centred cognitive extension, one that raises familiar questions from the extended mind debate while also introducing new challenges concerning accuracy, agency, and normative standards for a ‘good’ memory. More broadly, the case of mnemonic support illustrates how hybrid human-AI systems may reconfigure core cognitive capacities through ongoing, interactive collaboration.

Hybrid Cognition in Human-Robot Interaction

Agnieszka Wykowska

This talk explores the distinct forms of collaboration that emerge in human-robot interaction. I will focus on humanoid robots and provide examples of how they can be used for both cognitive offloading and augmentation. I will present research in which we identify markers of mental workload (body language, EEG) in an ecologically valid setup mimicking a factory task. These markers can then be used for

cognitive offloading by a robot collaborator. In the domain of social cognition, I will present results from our lab (Ghiglini et al., 2023, 2025) demonstrating how a humanoid robot successfully augments social skills in children diagnosed with autism and provides scaffolding for learning such skills. I will argue that robots’ physical embodiment has the advantage over disembodied AI systems when the user’s mental workload is related to performing actions on the physical world or in the case of social interactions (Ghiglini & Wykowska, 2025; Wykowska, 2020).

References

- Clark, A. (2025). Extending minds with generative AI. *Nat. Commun.*, 16(1), 4627.
- Coda-Forno, J., Binz, M., Wang, J. X., & Schulz, E. (2024). CogBench: A large language model walks into a psychology lab. *arXiv [cs.CL]*.
- Didolkar, A., Goyal, A., Ke, N. R., Guo, S., Valko, M., Lillicrap, T., Rezende, D., Bengio, Y., Mozer, M., & Arora, S. (2024). Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. *arXiv [cs.AI]*.
- Ghiglini, D., Floris, F., De Tommaso, D., Kompatsiari, K., Chevalier, P., Priolo, T., & Wykowska, A. (2023). Artificial scaffolding: Augmenting social cognition by means of robot technology. *Autism Research*, 16(5), 997–1008.
- Ghiglini, D., Floris, F., De Tommaso, D., Russi, N. S., Frulli, A., Moretti, S., & Wykowska, A. (2025). Enhancing theory of mind in autism through humanoid robot interaction in a randomized controlled trial. *Scientific Reports*, 15(1), 27650.
- Ghiglini, D., & Wykowska, A. (2025). Reimagining autism interventions: Robots as adaptive cognitive partners in social learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 9.
- Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., Akata, Z., & Schulz, E. (2023). Machine psychology. *arXiv [cs.CL]*.
- Kosmyrna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4.
- Smart, P. (2022). Minds in the metaverse: Extended cognition meets mixed reality. *Philosophy & Technology*, 35(4), 87.
- Smart, P., Clowes, R., & Clark, A. (2025). Chatgpt, extended: Large language models and the extended mind. *Synthese*, 205(6), 242.
- Street, W., Siy, J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Arcas, B. A. y., & Dunbar, R. I. M. (2024). LLMs achieve adult human performance on higher-order theory of mind tasks. *arXiv [cs.AI]*.
- Wykowska, A. (2020). Social robots to test flexibility of human social cognition. *International Journal of Social Robotics*, 12(6), 1203–1211.