

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Laying the foundations for foundation models of cognition

Permalink

<https://escholarship.org/uc/item/0bc9g9w3>

Journal

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

Authors

Binz, Marcel

Collins, Katherine M

Ying, Lance

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

Laying the foundations for foundation models of cognition

Marcel Binz (marcel.binz@helmholtz-munich.de)¹, Katherine M. Collins², Lance Ying^{2,3},
Joshua B. Tenenbaum² & Eric Schulz¹

¹Helmholtz Munich, Munich, Germany

²Massachusetts Institute of Technology, Cambridge, MA, USA

³Harvard University, Cambridge, MA, USA

Overview and motivation

Foundation models, large-scale artificial intelligence (AI) systems trained on vast amounts of data and able to engage with a wide range of tasks, have captivated public attention and promise transformative impact across many fields, from mathematics to education (Bommasani, 2021). The capacity for such models to fluidly engage in natural language carries huge potential implications for cognitive science: for the first time, we have computational models that bring us a significant step toward the generality of human cognition, thereby inviting us to reconceptualize how we may build models of the human mind (Binz et al., 2025; Wong et al., 2025). Yet, major questions remain as to how to build such models, what kind of data is needed for this, and what the implications are for cognitive science.

This workshop brings together researchers to examine how foundation models can inform and advance cognitive modeling, and what it may take to build towards a grand goal of general-purpose foundation models of human cognition. To this end, we have invited leading experts to present recent advances across a broad set of questions, including:

1. how to combine black-box models with existing theories about human cognition (Griffiths et al., 2025);
2. how to use foundation models as part of a broader toolkit for building cognitive models (Wong et al., 2025; Wüst et al., 2025);
3. how to integrate language into cognitive architectures (Wray et al., 2025);
4. what would it mean to have a true foundation model of human cognition that can capture both population-level regularities and individual differences;
5. how do we need to evaluate such models to obtain insights that human scientists care about?

In doing so, we hope to foster a dialogue on how to incorporate key characteristics of foundation models into cognitive modeling. This powerful synthesis – of advances in large-scale language modeling, alongside other computational approaches developed in computational cognitive science (e.g., for program synthesis) – could soon lead to computational models that capture human behavior not just within a single experiment or domain, but across many of them, taking steps towards a long-standing goal of cognitive science (Newell, 1994). This offers a promising route toward addressing one

of psychology’s biggest challenges – the generalization crisis – by enabling models to synthesize findings across traditionally isolated areas of cognition. Furthermore, we envision that such models will be stimulus-computable, able to operate on rich, naturalistic stimuli, such as raw text or images, thereby moving cognitive modeling toward more ecologically valid environments.

Goals and scope

The goal of this workshop is to lay the foundations for foundation models of cognition – models that are capable of reasoning about any kind of problem people can, and ideally do so in a way that advances our understanding of human cognition. Building such models raises a number of fundamental questions, including: (1) how can and should such models be built? (2) at what level of abstraction should they operate? (3) what are evaluation protocols that allow us to meaningfully compare competing models against each other and against humans? (4) what are the implications for cognitive science as a field if we can build such models? By tackling these questions, the workshop will serve as a catalyst for new research directions. We expect it to kickstart collaborations, inspire new modeling approaches, datasets, and evaluations for such models, and help chart a roadmap for years (even decades) to come.

Target audience

We aim to target an interdisciplinary group almost as wide as the main conference itself, including cognitive psychologists, neuroscientists, linguists, philosophers, and computer scientists. This wide scope is motivated by the workshop’s central goal – to lay the groundwork for foundation models of cognition – which, by design, requires integrating insights from cognitive science in its entirety.

Organizers and presenters

We have invited a diverse group of speakers, aiming for broad representation across disciplines, genders, institutions, and career stages:

Marcel Binz (Organizer) is a Staff Scientist at the Institute for Human-Centered AI at Helmholtz Munich. He will talk about *foundation models of cognition*.

Katherine M. Collins (Organizer) is a Postdoctoral Fellow in the Computational Cognitive Science Group at MIT and Visiting Postdoc at the Princeton AI Lab. She will give the introductory presentation and moderate the fireside chat.

Lance Ying (Organizer) is a PhD student affiliated with MIT Brain and Cognitive Science, CSAIL and Harvard SEAS. He will provide logistical support to the workshop.

Joshua B. Tenenbaum (Organizer) is a Professor at the Department of Brain and Cognitive Sciences at MIT. He will take part in the fireside chat.

Eric Schulz (Organizer) is the Director of the Human-Centered AI at Helmholtz Munich. He will take part in the fireside chat.

Tyler Brooke-Wilson is an Assistant Professor of Cognitive Science and Philosophy at Yale. He will talk with Lio Wong about *rational foundations for human cognitive architecture*.

Molly Crockett is a Professor of Psychology at Princeton University. They will talk about *foundation models and illusions of generalizability in cognitive science*.

Judith E. Fan is an Assistant Professor of Psychology at Stanford University. She will talk about *generative behaviors as key targets for advancing cognitive theory-building*.

José Hernández-Orallo is the Director of Research at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. He will talk about *the intrinsic structure of individual AI capabilities*.

Lio Wong is a Stanford Human-Centered Artificial Intelligence Postdoctoral Fellow working with the Cognitive Tools Lab. They will talk with Tyler Brooke-Wilson about *rational foundations for human cognitive architecture*.

Robert E. Wray is a Principal Scientist and Co-Director of the Center for Integrated Cognition. He will talk about *the architectures of agentic systems: recurring patterns of cognitive function*.

Antonia Wüst is a PhD student at the AIML Lab CS Department at TU Darmstadt. She will talk about *from images to rules: neuro-symbolic program synthesis for visual inductive reasoning*.

Workshop structure

We propose a full-day workshop consisting of seven talks, a fireside chat focused on the design of foundation models of cognition, and a broader panel discussion on opportunities, pitfalls, and open challenges for the field. Each speaker session will involve two parts: an approximately 20 minute talk, followed by a 15 minute discussion period. We deliberately design sessions to have long discussion periods following talks to encourage conversation and constructive debate amongst the audience and speakers. While our workshop will likely focus on in-person activities, we are actively exploring possibilities for hybrid engagement to broaden access and discourse around these topics. The topics of the talks are as follows:

Presenter	Topic
Collins	Introduction
Binz	Foundation models of cognition
Wong & Brooke-Wilson	Rational foundations for human cognitive architecture
Coffee and discussion	
Wray	The architectures of agentic systems: recurring patterns of cognitive function
Wüst	From images to rules: neuro-symbolic program synthesis for visual inductive reasoning
Schulz & Tenenbaum	Fireside chat
Lunch and discussion	
Fan	Generative behaviors as key targets for advancing cognitive theory-building
Crockett	Foundation models and illusions of generalizability in cognitive science
Coffee and discussion	
Hernández-Orallo	The intrinsic structure of individual AI capabilities Panel discussion
Ying	Closing remarks

References

- Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. (2025). A foundation model to predict and capture human cognition. *Nature*, 1–8.
- Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Griffiths, T. L., Lake, B. M., McCoy, R. T., Pavlick, E., & Webb, T. W. (2025). Whither symbols in the era of advanced neural networks? *arXiv preprint arXiv:2508.05776*.
- Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.
- Wong, L., Collins, K. M., Ying, L., Zhang, C. E., Weller, A., Gerstenberg, T., O'Donnell, T., Lew, A. K., Andreas, J. D., Tenenbaum, J. B., et al. (2025). Modeling open-world cognition as on-demand synthesis of probabilistic models. *arXiv preprint arXiv:2507.12547*.
- Wray, R. E., Kirk, J. R., & Laird, J. E. (2025). Applying cognitive design patterns to general llm agents. *International Conference on Artificial General Intelligence*, 312–325.
- Wüst, A., Stammer, W., Shindo, H., Helff, L., Dhami, D. S., & Kersting, K. (2025). Synthesizing visual concepts as vision-language programs. *arXiv preprint arXiv:2511.18964*.