Advancing Cognitive Science and AI with Cognitive-AI Benchmarking

Felix Binder[1], Logan Cross[2], Yoni Friedman[3], Robert Hawkins[4], Daniel Yamins[2], and Judith E. Fan[1,2]

[1]University of California, San Diego, [2]Stanford University, [3]Massachusetts Institute of Technology, [4]Princeton University

Abstract

What are the current limits of AI models in explaining human cognition and behavior? How might approaches from the cognitive sciences drive the development of more robust and reliable AI systems? The goal of this workshop is bring together researchers across cognitive science and artificial intelligence (AI) to engage with these questions and identify opportunities to work together to advance progress in both fields. In particular, we propose Cognitive-AI Benchmarking as a particularly promising strategy — that is, the community-coordinated establishment of common benchmarks, tools, and best practices for model-human comparisons across diverse and ecologically relevant domains and tasks. We will host a combination of talks, panel discussion, and breakout activities to: highlight past successes in Cognitive-AI Benchmarking and limitations of current approaches, share tools and best practices, and outline future challenges and goals for the field.

Overview & Motivation

In recent years, the field of artificial intelligence (AI) has seen an explosion of increasingly performant models in domains of long-standing interest to cognitive science, including vision (Krizhevsky, Sutskever, & Hinton, 2017; Ramesh et al., 2021; Adamkiewicz et al., 2022), decision-making (Mnih et al., 2015; Bakhtin et al., 2022; Chen et al., 2021), and language (Chowdhery et al., 2022; Ouyang et al., 2022). These models are distinguished by the capacity to act on naturalistic, high-dimensional input (e.g. raw pixels or text) and produce high-dimensional behavior (e.g. executing motor plans in realistic 3D physical environments or producing natural language). Such progress raises the possibility that these algorithms might embody dramatically improved cognitive models that succeed in describing human behavior in a wide range of complex, real-world settings. Reciprocally, systematic methods of evaluation pioneered in behavioral science will be of great value in identifying where AI models fail to match human performance levels or response patterns, clarifying the limits of existing technology and providing a guide for further AI advances.

We believe that this potential for a new wave of collaborations between cognitive science and AI researchers will be enabled by Cognitive-AI Benchmarking (CAB) — the systematic design and deployment of large-scale experiments to probe behavior in rich real-world domains and tasks that are of common interest in cognitive science and AI. This will involve the community-coordinated establishment of common benchmarks, tools, and best practices for model-human comparison across a set of diverse and ecologically relevant task domains. Any CAB effort will generally consist of: (1) large sets of high-quality stimuli providing dense coverage of each domain; (2) a spectrum of rich behavioral readouts; (3) high-throughput measurements of behavioral outputs to enable detailed comparison against models. Rather than using human behavior only as the "gold standard," CAB projects ask: What exactly are the current limits of AI models in explaining the full pattern of human behavior? What gaps exist in both our current best models and our ability to explain human behavior? How might insights from cognitive modeling drive the development of more robust and reliable AI systems?

Goals & Approach

Although Cognitive-AI Benchmarking is inspired by some of the most successful motifs in multidisciplinary collaboration, it is not business as usual in AI, psychology, or neuroscience. A CAB project is only successful if the proposed benchmark requires advancing both our mechanistic understanding of human behavior and generating meaningful algorithmic innovation. We expect that such projects will strongly benefit from coordination between larger teams of researchers spanning traditional scientific and engineering disciplines. By carefully measuring a comprehensive suite of cognitive benchmarks, we can simultaneously address fundamental questions about diverse human behaviors and abilities, and facilitate an empirically grounded, data-driven approach to developing more robust and human-like AI.

The goal of the workshop is to highlight recent efforts to benchmark state-of-the-art AI systems across several cognitive domains, aiming to both push the frontier of state-of-the-art methods in AI and to develop more unified mechanistic theories of human cognition. Brief presentations from invited speakers will help participants gain an understanding of the range of applications of this approach. Additionally, we will host a panel discussion to bring together researchers from a variety of communities within the cognitive sciences and AI to share their perspective on various high-level questions regarding the opportunities and challenges associated with CAB efforts. Finally, participants will have opportunities to learn about the tools used to perform this research.

Schedule

The workshop will open with brief remarks by the organizers introducing CAB and how it relates to other approaches in the field. We will then transition to a series of short invited talks showcasing examples of current CAB efforts spanning several cognitive domains, including perception, action, social cognition, and language.

Invited talks

Brief research talks will be delivered by the following invited speakers:

- **Kristen Grauman** is a Professor in Computer Science at the University of Texas at Austin, where she leads the

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Computer Vision Group. She will present the Ego4D dataset, a large dataset of egocentric video and benchmark suite collected across a wide range of locations (Grauman et al., 2022).

- **Martin Hebart** is Professor of Computational Cognitive Neuroscience at Justus Liebig University Gießen and research group leader at the Max Planck Institute of Human Cognitive and Brain Science. He will present on THINGS, a collection of datasets using a shared set of images of objects and a variety of neuroimaging, behavioral and AI observations (Hebart et al., 2019).
- **Tal Linzen** is an Assistant Professor of Linguistics and Data Science at New York University, and a Research Scientist at Google. He will be presenting on the use of large-scale benchmarks as a means of accelerating progress in natural language processing.
- **Tianmin Shu** is a Research Scientist at MIT and incoming Assistant Professor at Johns Hopkins University. He will present his work on AGENT: a dataset and benchmark for human-like intuitive psychology for AI systems (Shu et al., 2021).
- **Jenn Hu** recently defended her Ph.D in Brain and Cognitive Sciences at MIT, and will be joining Harvard's Kempner Institute as a Research Fellow this Fall. She will present on benchmarks for measuring pragmatic knowledge in language models (Hu, Floyd, Jouravlev, Fedorenko, & Gibson, 2022; Hu, Levy, Degen, & Schuster, 2023).

Panel Discussion

Panelists from several different communities will be asked to share their perspective on several high-level questions. Some of these questions may include:

- How does comparing human cognitive behavior with that of AI models fit into the future of cognitive science?
- How can comparing AI models against humans advance the development of artificial intelligence?
- What makes for a good benchmark, and what are the important distinctions when benchmarking human versus AI abilities?
- What are the main bottlenecks for making progress? (Models? Measurements? Something else?)

Our panelists will include:

- **Fei-Fei Li** is Sequoia Capital Professor at Stanford University, where she works on cognitively inspired AI.
- **Moira Dillon** is Assistant Professor of Psychology at New York University. She works on infants' and children's knowledge and learning, with an eye towards improving commonsense AI.
- **Joshua Peterson** is a Postdoctoral Research Associate at Princeton University. His work explores how machine learning can be used as a tool in computational cognitive science.

Tools & best practices

Following the invited talks, the organizers will lead a session on best practices for conducting large CAB projects, including how to coordinate large teams, scale up human data collection for more complex tasks, and unify disparate model APIs. Additionally, the organizers will demonstrate a variety of software tools and systems that help facilitate CAB projects.

Breakout groups

Finally, the participants in the workshop will break out into small groups organized around different cognitive domains. Participants will be invited to self-assign to the domain that best matches their research interests. During the breakout groups, participants are encouraged to discuss potential tasks and models they would like to see unified within a CAB project. At the end of the workshop, ideas contributed by participants will be collected and will be published on the workshop website.

Organizers

- **Felix Binder** is a PhD student in the Cognitive Science department at UCSD. He works on planning and intuitive physics in human and AI systems.
- **Logan Cross** is a Postdoctoral Fellow in the Department of Computer Science in Stanford, working on curiosity-driven learning in AI agents.
- **Yoni Friedman** is a Research Engineer in the Computational Cognitive Science lab at MIT, and an incoming PhD student in MIT's EECS program, working on developing better benchmarks and models for visual scene understanding.
- **Robert Hawkins** is Postdoctoral Scholar at the Princeton Neuroscience Institute. His work investigates the cognitive mechanisms that allow people to flexibly communicate, collaborate, and coordinate on social conventions in groups.
- **Daniel Yamins** is Assistant Professor in the Departments of Psychology and Computer Science at Stanford university. His work sits at the intersection of artificial intelligence and neuroscience.
- **Judith Fan** is Assistant Professor in the Department of Psychology at UC San Diego and Stanford University. Research in her lab focuses on the use of physical representations of thought, including sketches and other objects, during learning, communication, and problem solving.

References

Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., & Schwager, M. (2022). Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, *7*(2), 4606-4613. doi: 10.1109/LRA.2022.3150497

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., ... Zijlstra, M. (2022). Human-level play in the game of ¡i¿diplomacy¡/i¿ by combining language models with strategic reasoning. *Science*, *378*(6624), 1067-1074. doi: 10.1126/science.ade9097

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., ... Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... others (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., ... Malik, J. (2022). Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF computer vision and pattern recognition (CVPR).*

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019, October). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), e0223792. doi: 10.1371/journal.pone.0223792

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). *A fine-grained comparison of pragmatic language understanding in humans and language models.*

Hu, J., Levy, R., Degen, J., & Schuster, S. (2023). *Expectations over unspoken alternatives predict pragmatic inferences.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, may). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, *60*(6), 84–90. Retrieved from https://doi.org/10.1145/3065386 doi: 10.1145/3065386

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015, February). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. Retrieved from https://doi.org/10.1038/nature14236 doi: 10.1038/nature14236

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). *Training language models to follow instructions with human feedback.* arXiv. Retrieved from https://arxiv.org/abs/2203.02155 doi: 10.48550/ARXIV.2203.02155

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021, 18–24 Jul). Zero-shot text-to-image generation. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 8821–8831). PMLR. Retrieved from https://proceedings.mlr.press/v139/ramesh21a.html

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., ... Ullman, T. D. (2021). AGENT: A Benchmark for Core Psychological Reasoning.