Framing the problem: Abstraction and task representations in human problem solving

David G. Nagy[1,2], Hanqi Zhou[1,2,3], Shuze Liu[4], Tracey Mills[5], Tony Chen[5], Zergham Ahmed[4] & Mark K. Ho[6]

[1]TU Darmstadt, [2]Max Planck Institute for Biological Cybernetics, [3]IMPRS-IS, [4]Harvard University, [5]Massachusetts Institute of Technology, [6]New York University

Overview and Motivation

Cognitive scientists have long sought to explain the human capability to solve complex and often ill-defined problems, ranging from puzzle games to the challenges of everyday life. Early work cast problem solving as search: given a fixed representation of states, actions, and transitions, an agent must find a sequence of actions that carries it from an initial state to a goal state, often guided by heuristic search (Newell, Simon, et al., 1972). This framework proved successful in explaining human behavior in lab-controlled puzzle tasks, where experimenters could hand-craft fixed problem representations, allowing models to focus analysis on the downstream search process (De Groot, 2008; Kotovsky et al., 1985). Yet, it left a more fundamental question comparatively unexplored: how do humans construct the problem representations that make search possible in the first place (Dunbar, 2017; Ohlsson, 2012; Simon, 1973)?

This "problem of problem representation" marks a deep conceptual obstacle for theories of problem solving. A core challenge is that of relevance — echoing classic discussions of the "frame problem" in AI and philosophy — namely, that agents must determine which facts, features, and contingencies matter for the task at hand, and which can be safely ignored, despite limited time and computation (Dennett, 1990; Icard III & Goodman, 2015). In reinforcement learning and robotics, the "problem of problem representation" is often framed as one of learning effective abstractions (e.g., state abstractions, options/skills, or symbolic predicates) for a given environment (Abel, 2022; Abel et al., 2018; Konidaris, 2019). As in cognitive modeling, these abstractions have typically been hand-crafted by researchers.

Recent advances in cognitive modeling and experimental methodology have shifted attention toward more naturalistic problem settings—where the need to form appropriate representations is especially acute. At the normative level, resource-rational frameworks have proposed that humans construct representations that balance the utility of resulting plans against representational cost (Ho et al., 2022; Icard III & Goodman, 2015), though they leave open how humans actually discover such simplified representations. Building on this perspective, recent work has argued that open-world cognition proceeds by constructing bespoke, locally coherent probabilistic models on demand (Wong et al., 2025). At the process level, other theories explore how representations may arise from the incremental discovery of problem constraints (Chen et al., 2026), or through the caching of abstract, compositional structures that can be flexibly transferred across tasks (Nagy et al., 2025).

In artificial intelligence, tasks set in more naturalistic environments—such as video games—have become prominent performance benchmarks (Justesen et al., 2019). Advances in deep learning and world modeling have enabled artificial agents to exhibit increasingly flexible forms of problem solving. A recurring lesson is that success often hinges on learning an appropriate representation of the environment, frequently in the form of an abstract world model that supports prediction and planning (Hafner et al., 2025; Schrittwieser et al., 2020). A range of additional theoretical motifs have gained traction, including option learning, state abstractions, predicate invention, and program synthesis (Ahmed et al., 2025; Ellis et al., 2021; Silver et al., 2022, 2023). Adapting these formalisms to the study of human problem solving remains a fruitful research direction (Toussaint et al., 2018).

Goals

Despite renewed interest in studying human problem representations, relevant research remains fragmented across disciplines. By convening researchers in cognitive science, machine learning, and philosophy, the workshop seeks to clarify shared assumptions, identify unifying theoretical motifs, and articulate future directions for studying how representations are constructed, adapted, and reused in complex problem-solving environments. Some outstanding questions include:

1. **Normative models:** To what extent do human problem representations approximate resource-rational objectives?
2. **Methodology:** Which insights about problem representation generalize from controlled tasks to naturalistic environments? How can we best probe latent representations?
3. **Construction vs. reuse:** How do agents decide when to construct new representations versus reusing them across tasks?
4. **Crosstalk:** What do current cognitive models reveal about the frame problem, and what can be learned from modern systems such as large language models?

Building on recent community efforts—including CogSci workshops on abstractions (2023), compositionality (2024), and meta-reasoning (2025), as well as calls to study cognition in more naturalistic domains (Carvalho & Lampinen, 2025;

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Wise et al., 2024)—the workshop shifts focus from how representations support downstream inference to how they are constructed in the first place.

Approach and Workshop Structure

The proposed in-person workshop will span a full day (7 hours plus 1-hour lunch break), consisting of three thematic sessions of invited talks and one panel discussion. The four activities each last 90 minutes, aligning with CogSci's designated refreshment and lunch breaks.

Thematic sessions

Each **thematic session** will feature two **invited talks**. The first session, *Normative frameworks and cognitive architectures*, will examine how problem representations *ought* to be constructed, drawing on the philosophical frame problem as well as recent efforts that integrate Bayesian cognitive modeling with automated program synthesis. The second session, *Process-level models of human problem representation*, will focus on how such normative principles may be instantiated in human cognition, with an emphasis on empirical and computational accounts of how people expand, structure, and revise their representations during problem solving. The third session, *Methodology and Experiments*, will explore how recent advances in large-scale learning systems may inform the study of problem representation construction in naturalistic environments, through both enabling sophisticated experiment designs and automated data analysis. Each talk will last 20 minutes, followed by 5 minutes of audience questions and 5 minutes of speaker-to-audience discussion—a total of 30 minutes for each invited speaker.

In each 90-minute session, the two speakers together occupy 60 minutes. The remaining 30 minutes will be dedicated to introduction and concluding remarks by the organizing team members, as well as **group-based activities**—guided by the session's theme, participants will self-organize into groups to pursue questions that were not fully addressed by the talks. These more focused and informal discussions are intended to facilitate deeper exchange across disciplines and to foster collaboration opportunities emerging from shared interests.

Panel discussion

The **panel discussion** will be organized around open questions in the study of problem representation and abstraction. Prior to the conference, we will solicit discussion questions from the broader research community via an online poll. During the workshop, both speakers and audience members will be invited to vote on a subset of these questions, which will form the basis of a one-hour moderated discussion. The panel will begin with a brief introduction on the poll results and activity format, then approximately 60 minutes of discussion among the speakers, followed by audience questions and open debate. Example topics include the generalizability of insights from controlled tasks to naturalistic settings, methodologies for probing and manipulating latent representations, and the future roles of theory-driven modeling versus reverse-engineering in understanding human problem representations.

Invited Speakers

Thomas Icard is a Professor of Philosophy and Computer Science at Stanford University, focusing on resource rationality, causal inference, and abstract representations.
Lio Wong is a postdoctoral researcher at Stanford University, studying language-induced representations, probabilistic models of cognition, and program synthesis.
Marcelo Mattar is an Assistant Professor of Psychology and Neural Science at New York University, studying how memory supports planning and decision making.
Frederick Callaway is a postdoctoral researcher at NYU and an incoming Assistant Professor at Dartmouth College studying human metareasoning, planning, and resource rationality.
Samuel Cheyette is a postdoctoral researcher at MIT (CoCoSci) studying human reasoning, information seeking, and problem solving under limited cognitive resources.
Kelsey Allen is a Senior Research Scientist at DeepMind studying human tool creation and physical problem solving from the perspective of cognitive science and robotics.

Organizers

Organizers will give talks between sessions to connect themes:
David G. Nagy is a postdoc at TU Darmstadt and the Max Planck Institute for Biological Cybernetics focusing on computational modeling of memory and abstraction in humans.
Hanqi Zhou is a PhD student at University of Tübingen and International Max Planck Research School for Intelligent Systems (IMPRS-IS), studying human resource rationality.
Shuze Liu is a PhD student at Harvard University studying resource-rational decision making and cognitive processes underlying naturalistic problem solving.
Tracey Mills is a PhD student at MIT studying human problem solving and metacognition.
Tony Chen is a PhD student at MIT studying human problem solving and planning.
Zergham Ahmed is a PhD student at Harvard University studying scalable approaches to Theory-Based RL.
Mark K. Ho is an assistant professor in the Department of Psychology at New York University. His group studies the computational principles that underpin abstraction in individual and group problem solving.

Acknowledgements D.G.N. and H.Z. acknowledge support from the Hessian research funding programme LOEWE and the DFG under Germany's Excellence Strategy, EXC-3057 and EXC 3066/1 "The Adaptive Mind".

References

Abel, D. (2022). A theory of abstraction in reinforcement learning. *arXiv preprint arXiv:2203.00397*.

Abel, D., Arumugam, D., Lehnert, L., & Littman, M. (2018). State abstractions for lifelong reinforcement learning. *International Conference on Machine Learning*, 10–19.

Ahmed, Z., Tenenbaum, J. B., Bates, C. J., & Gershman, S. J. (2025). Synthesizing world models for bilevel planning. *arXiv preprint arXiv:2503.20124*.

Carvalho, W., & Lampinen, A. (2025). Naturalistic computational cognitive science: Towards generalizable models and theories that capture the full range of natural behavior. *arXiv preprint arXiv:2502.20349*.

Chen, T., Cheyette, S., Allen, K., Tenenbaum, J., & Smith, K. (2026). " just in time" world modeling supports human planning and reasoning. *arXiv preprint arXiv:2601.14514*.

De Groot, A. D. (2008). *Thought and choice in chess*. Amsterdam University Press.

Dennett, D. C. (1990). Cognitive wheels: The frame problem of ai. *The philosophy of artificial intelligence*, *147*, 170.

Dunbar, K. (2017). Problem solving. *A companion to cognitive science*, 289–298.

Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Morales, L., Hewitt, L., Cary, L., Solar-Lezama, A., & Tenenbaum, J. B. (2021). Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation*, 835–850.

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. *Nature*, *640*(8059), 647–653. https://doi.org/10.1038/s41586-025-08744-2

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Icard III, T. F., & Goodman, N. D. (2015). A resource-rational approach to the causal frame problem. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*.

Justesen, N., Bontrager, P., Togelius, J., & Risi, S. (2019). Deep learning for video game playing. *IEEE Transactions on Games*, *12*(1), 1–20.

Konidaris, G. (2019). On the necessity of abstraction. *Current opinion in behavioral sciences*, *29*, 1–7.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? evidence from tower of hanoi. *Cognitive psychology*, *17*(2), 248–294.

Nagy, D. G., Shen, T., Zhou, H., Wu, C. M., & Dayan, P. (2025). Analogy making as amortised model construction. *arXiv preprint arXiv:2507.16511*.

Newell, A., Simon, H. A., et al. (1972). *Human problem solving* (Vol. 104). Prentice-hall Englewood Cliffs, NJ.

Ohlsson, S. (2012). The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm. *The Journal of problem solving*, *5*(1), 7.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.

Silver, T., Athalye, A., Tenenbaum, J. B., Lozano-Pérez, T., & Kaelbling, L. P. (2022). Learning neuro-symbolic skills for bilevel planning. *arXiv preprint arXiv:2206.10680*.

Silver, T., Chitnis, R., Kumar, N., McClinton, W., Lozano-Pérez, T., Kaelbling, L., & Tenenbaum, J. B. (2023). Predicate invention for bilevel planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(10), 12120–12129.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial intelligence*, *4*(3-4), 181–201.

Toussaint, M. A., Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2018). Differentiable physics and stable modes for tool-use and manipulation planning.

Wise, T., Emery, K., & Radulescu, A. (2024). Naturalistic reinforcement learning. *Trends in Cognitive Sciences*, *28*(2), 144–158.

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*.