How does the mind discover useful abstractions?

Marcelo Mattar[1], Wai Keen Vong[1], Lionel Wong[3], and Judith E. Fan[2,4]

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

Overview and Motivation

Abstraction enables humans to distill a cascade of sensory experiences into a useful format for making sense of the world and generalizing to new contexts. For example, **visual abstraction** allows us to generalize from a single instance of a flower to the set of all flowers (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; Kemp, Perfors, & Tenenbaum, 2007) and effortlessly parse visual scenes into their constituent parts and relations (Ji et al., 2022; Chen et al., 2022; Hamrick et al., 2018; Ichien et al., 2021). **Relational abstractions** allow us to recognize distinctions between objects, such as whether they are the "same" or "different", regardless of the specific entities being compared (Walker & Gopnik, 2014; Webb, Sinha, & Cohen, 2021; Geiger, Carstensen, Frank, & Potts, 2022). Abstraction also plays a crucial role in reinforcement learning and memory, where concepts such as **state** and **temporal abstractions** (Ho, Abel, Griffiths, & Littman, 2019) can be leveraged to support sophisticated forms of mental simulation and planning in complex environments (Hamrick, 2019; Ha & Schmidhuber, 2018), or segmenting continuous experience into discrete events (Baldassano et al., 2017; Leshinskaya, Nguyen, & Ranganath, 2022) Finally, **linguistic abstraction** enables us to compose words in innumerable ways (von Humboldt, 1836) to generate new meanings and support effective communication (Gentner & Hoyos, 2017; Spelke, 2017; Wong, Ellis, Tenenbaum, & Andreas, 2021).

In developing theories of how these various forms of abstraction are discovered and used, cognitive scientists have proposed a multitude of representational formats with different properties to capture behavioral data and neural activity. These representational formats include clusters (Anderson, 1991; Love, Medin, & Gureckis, 2004; Gershman, Blei, & Niv, 2010), schemas (McClelland, 2013; Bein, Reggev, & Tompary, 2018), trees (Chomsky, 1957; Kemp & Tenenbaum, 2009; Xu & Tenenbaum, 2007), neural networks (LeCun, Bengio, & Hinton, 2015; Geiger, Lu, Icard, & Potts, 2021) or programs (Bramley, Schulz, Xu, & Tenenbaum, 2018; Ellis et al., 2021; Tavares, Koppel, Zhang, Das, & Solar-Lezama, 2021). However, because abstraction manifests in human cognition and behavior in so many ways, individual communities within cognitive science have generally studied these specific forms of abstraction under domain-specific representational assumptions.

The goal of this workshop is to facilitate the search for unifying principles governing how humans learn, discover, and use abstractions in different domains, by providing a venue for the exchange of theoretical and empirical insights between research communities. Such unifying principles may help to explain how and why certain abstractions emerge, perhaps as the product of functional pressures such as predictive coding and information compression (Friston, 2010) or cultural transmission via pedagogy and teaching (Tomasello, 2009; Chopra, Tessler, & Goodman, 2019). Moreover, the project of developing a more unified theory of how the human mind discovers useful abstraction may also produce insights that advance the development of artificially intelligent systems that learn and think more like humans.

Approach and Schedule

Because our workshop seeks unifying approaches to understanding abstraction in humans and machines, we aim to bring together perspectives from machine learning, neuroscience, and cognitive science, and spanning the role of abstraction in perception, memory, causality, and language. Our goal is to synthesize theories and empirical evidence from these distinct fields and domains to move towards a more cohesive and integrated theory of abstraction.

To maximize our impact and reach, this workshop will be held in a flipped format, with both a virtual and in-person component. The virtual component will happen before the main conference and will include four virtual seminars, each focused on abstraction in a distinct domain: Causal Abstractions, Language, Perception and Motor Control, and Memory. Within each of these seminars, invited speakers will be asked to present flash talks around a unifying question: "In this domain, what evidence must a computational model of abstraction explain?" We will then hold a panel discussion with prepared questions shared across all four seminars. These virtual seminars will be open to the public and recorded.

The in-person component will be held at the CogSci conference in July and will be focused on unifying approaches from the four cognitive domains. We will begin with a recap of the virtual seminars with an edited video that summarizes common themes from the recorded seminars. We will then hold a live, hybrid panel with our invited speakers, focused on further integrating ideas across fields and domains. Finally, in-person participants will be invited to participate in small-group discussions and to report back to the audience with summaries of their discussions.

After the event, we intend to synthesize major themes from the virtual seminars and in-person discussions in a review article, the writing of which will be spearheaded by the workshop organizers in collaboration with the speakers and workshop participants. More broadly, we hope that this workshop will also motivate novel integrative work across cognitive domains, facilitating the development of unified theories of abstraction in cognitive science and artificial intelligence.

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

Invited Speakers & Organizers

The organizers share a deep interest in computational theories of abstraction that unite evidence across diverse domains.

Wai Keen Vong (Organizer) is a Research Scientist in the Human and Machine Learning Lab at New York University. His research focuses on computational models of concept and language acquisition from naturalistic multimodal input, and understanding how language shapes the mind.

Lionel Wong (Organizer) is a PhD Candidate in Brain and Cognitive Sciences at MIT. Their research focuses on computational and cognitive models that integrate structured conceptual reasoning with language, and that learn new concepts and abstractions from language.

Marcelo G. Mattar (Organizer) is an Assistant Professor of Psychology and Neural Science at New York University. His lab studies memory and decision-making using a combination of theoretical and human behavioral/imaging approaches, with a particular interest in reinforcement learning.

Judith E. Fan (Organizer) is an Assistant Professor of Psychology at Stanford. Research in her lab focuses on the use of physical representations of thought, including sketches and other objects, during learning, communication, and problem solving.

Invited Seminar Speakers

For each virtual seminar, speakers will present short talks followed by a structured panel discussion. We anticipate the seminars taking place between June and early July 2023.

Causal Abstractions Seminar

- **Thomas Icard** is an Associate Professor of Philosophy and Computer Science at Stanford University. He studies theoretical and computational models of logical, probabilistic, and causal reasoning.
- **Zenna Tavares** is a Research Scientist at the the Zuckerman Institute and Data Science Institute at Columbia University. His research focuses on algorithms and languages for Bayesian modeling and causal inference.
- **Caren Walker** is an Assistant Professor at the University of California, San Diego. Her research explores how children reason about the causal structure of the world, and acquire abstract causal representations.

Language Seminar

- **Jacob Andreas** is an Assistant Professor in Electrical Engineering and Computer Science at MIT. His research aims to build intelligent systems that communicate effectively with humans and learn from human guidance.
- **Alexandra Carstensen** is a Postdoctoral Scholar at the University of California, San Diego. Her research explores the nature of category systems across languages and over development, with a focus on space and relations.
- **Robert Hawkins** is a C.V. Starr Fellow at the Princeton Neuroscience Institute. He studies the cognitive mechanisms that allow people to coordinate on shared abstractions for communication and collaboration.
- **Alane Suhr** is an Assistant Professor in Electrical Engineering and Computer Science at UC Berkeley. Alane's research focuses on building systems that use and learn language through situated, collaborative human-agent interactions.

Perception and Action Seminar

- **Kelsey Allen** is a research scientist at DeepMind. She studies how people learn and use efficient, generalizable representations for physical problem solving.
- **Daniel Bear** is a Postdoctoral Scholar in Psychology at Stanford University. His research studies how animals create internal models of the world from sensory experience.
- **Mark Ho** is a Faculty Fellow in the NYU Center for Data Science. He studies how people solve problems, individually and interactively, by developing computational models of planning and social cognition.
- **Hongjing Lu** is a Professor in Psychology and Statistics at the University of California, Los Angeles. Her research investigates how people draw inferences from sparse data, with a focus on motion perception and causal learning.

Memory Seminar

- **Chris Baldassano** is an Assistant Professor of Psychology at Columbia University. He studies how prior knowledge about the temporal and spatial structure of the world influences how we recall complex experiences and events.
- **Anna Leshinskaya** is an Assistant Project Scientist in Neuroscience at the University of California, Davis. She studies the neural basis of semantic memory, and how causal principles influence memory and concept formation.
- **Alexa Tompary** is an Assistant Professor in Psychology at Drexel University. She studies how changes in the brain alter how we remember past events, and how we integrate new experiences with prior knowledge.
- **James Whittington** is a Postdoctoral Scholar at Stanford University and the University of Oxford. His research builds models and theories for understanding structured neural representations in brains and machines.

Acknowledgments

This workshop is supported by NSF CAREER award #2047191 to J.E.F.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, *98*(3), 409.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.

Bein, O., Reggev, N., & Tompary, A. (2018). Working with schemas, predicting with schemas. *Journal of Neuroscience*, *38*(7), 1608–1610.

Bramley, N., Schulz, E., Xu, F., & Tenenbaum, J. (2018). Learning as program induction.

Chen, H., Venkatesh, R., Friedman, Y., Wu, J., Tenenbaum, J. B., Yamins, D. L., & Bear, D. M. (2022). Unsupervised segmentation in real-world images via spelke object inference. In *Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part xxix* (pp. 719–735).

Chomsky, N. (1957). Syntactic structures.

Chopra, S., Tessler, M. H., & Goodman, N. D. (2019). The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *Cogsci* (pp. 226–232).

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

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2022). Relational reasoning and generalization using nonsymbolic neural networks. *Psychological Review*.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, *34*, 9574–9586.

Gentner, D., & Hoyos, C. (2017). Analogy and abstraction. *Topics in cognitive science*, *9*(3), 672–693.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological review*, *117*(1), 197.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, *29*, 8–16.

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., & Battaglia, P. W. (2018). Relational inductive bias for physical construction in humans and machines. *arXiv preprint arXiv:1806.01203*.

Ho, M. K., Abel, D., Griffiths, T. L., & Littman, M. L. (2019). The value of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 111–116.

Ichien, N., Liu, Q., Fu, S., Holyoak, K. J., Yuille, A., & Lu, H. (2021). Visual analogy: Deep learning versus compositional models. *arXiv preprint arXiv:2105.07065*.

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R. D., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes..

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307–321.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Leshinskaya, A., Nguyen, M., & Ranganath, C. (2022). Integration of event experiences to build relational knowledge in the human brain. *bioRxiv*, 2022–11.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological science*, *13*(1), 13–19.

Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development*, *13*(2), 147–170.

Tavares, Z., Koppel, J., Zhang, X., Das, R., & Solar-Lezama, A. (2021). A language for counterfactual generative models. In *International conference on machine learning* (pp. 10173–10182).

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard university press.

von Humboldt, F. W. H. A. (1836). *Über die verschiedenheit des menschlichen sprachbaues und ihrer einfluss auf die geistige entwickelung des menschengeschlechts*. Dümmler.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological science*, *25*(1), 161–169.

Webb, T. W., Sinha, I., & Cohen, J. (2021). Emergent symbols through binding in external memory. In *International conference on learning representations.*

Wong, C., Ellis, K. M., Tenenbaum, J., & Andreas, J. (2021). Leveraging language to learn program abstractions and search heuristics. In *International conference on machine learning* (pp. 11193–11204).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.