Neither Neural Networks nor the Language of Thought alone Make a Complete Game

Iris Oved
Independent Scholar
911 Central Ave; San Francisco, CA 94115.
520-730-3530
irisoved@gmail.com
irisoved@paradoxlab.org

Nikhil Krishnaswamy
Colorado State University
279 Computer Science Building
1873 Campus Delivery
Fort Collins, Colorado 80523-1873
nikhil.krishnaswamy@colostate.edu
https://www.nikhilkrishnaswamy.com/

James Pustejovsky
Brandeis University
415 South Street MS-018
Brandeis University
Waltham, MA 02454 USA
1-781-736-2709
jamesp@cs.brandeis.edu
https://jamespusto.com/

Joshua Hartshorne
Boston College
Department of Psychology and Neuroscience
McGuinn 300
140 Commonwealth Ave
Chestnut Hill, MA 02467
617-552-0463
joshua.hartshorne@bc.edu
http://l3atbc.org/index.html

Word Counts: Abstract (58), Body (998), References (271), Total (1,477)

Abstract.

Cognitive Science has evolved since early disputes between Radical Empiricism and Radical Nativism. The authors are reacting to the revival of Radical Empiricism spurred by recent successes in Deep Neural Network (NN) models. We agree that language-like mental representations (LoTs) are part of the best game in town, but they cannot be understood independent of the other players.

Main Body

Quilty-Dunn, Porot, & Mandelbaum (QDPM) have done a service in summarizing major lines of empirical data supporting a role for symbolic, language-like representations (an LoT, construed broadly) in theories of cognition. This overview is particularly pressing for audiences of the overly-hyped popular press on Deep Neural Networks. However, QDPM have done a disservice to LoT by setting unfavorable terms for the debate. In particular, they 1) overlook the fact that an LoT is necessarily part of a larger system and thus its effects should rarely be cleanly observed, and 2) do not address well-known concerns about LoTs.

QDPM note that statements in an LoT are formed of discrete constituents and denote functions from possible worlds to truth values. Consider a situation in which Alice beats Bart at tug-of-war. This might be represented in an LoT as BEAT(ALICE, BART, TUG-OF-WAR). Fodor (1975) argued that constituents (BEAT, ALICE, BART, TUG-OF-WAR) are "atomic" (unstructured) pointers to metaphysically real entities: events (beating), properties (Aliceness, Bartness), kinds (tug-of-war), etc.

Unfortunately, such entities do not appear to exist; nature is not so easily carved at its joints. An alternative – implicit in many Bayesian models – is to treat the symbols as reifications of some distribution in the world: There are some features that are reliably (if probabilistically) encountered in combination, and we use (for example) "Alice" to refer to one such combination (or a posited essence that explains the combination; see Oved, 2015). This straightforwardly allows for recognition, for example through a Neural Network classifier (Pustejovsky & Krishnaswamy, 2022; Wu et al., 2015). Thus, the LoT sentence *BEAT(ALICE, BART, TUG-OF-WAR*) means that in observing the referred-to scene, we would recognize (our NN classifier would identify) an Alice, a Bart, a beating, and tug-of-war, and that these entities would be arranged in the appropriate way (see also Pollock & Oved, 2005). (For readers familiar with possible worlds semantics, the proposition picks out the set of possible worlds where all those recognitions would happen.)

This approach explains, for instance, why we tie ourselves in knots trying to decide whether a cat with the brain of a skunk is a cat or a skunk, or whether the first chicken egg preceded or followed the first chicken. In the LoT, SKUNK, CAT, and CHICKEN are reified abstractions tied to recognition procedures. The world is messier, and the recognition procedures sometimes gum up. Note further that different methods for identifying skunks and cats, etc., (NNs, prototypes, inverse graphics, etc.) have characteristic imprecisions if not outright hallucinations. The predictions of any LoT theory cannot be separated from the manner in which the symbols map onto the world.

Reasoning presents additional complications. Most people infer from *Alice beat Bart at tug-of-war* that Alice is stronger, that both are humans not platypodes, are not quadriplegic, and played tug-of-war in a gym or field not while flying. While none of these inferences necessarily hold, keeping a completely open mind about them requires wilful obtuseness. Critically, such graded, probabilistic inferences have been the bane of symbolic reasoning theories, including LoTs. A promising avenue is to treat LoT statements as conditions on probable worlds generated from a generative model of the world (Goodman et al., 2014; Hartshorne et al., 2019). That is, one considers all possible worlds in which Alice beat Bart at tug-of-war. Because the prior probability of aerial quadriplegic tadpoles playing tug-of-war is low, we discount those possibilities (barring additional evidence).

Since we cannot do a census of possible worlds, this process requires an internal model of the world. Thus, the exact inferences one gets depend on not just the LoT but on what one believes about the world. They also depend on the nature of the model. In some domains, symbolic generative models seem to capture human intuitions, whereas in others we seem to use analog simulations (Jara-Ettinger et al., 2016; Ullman et al., 2017). For example, when imagining Alice beating Bart at tug-of war, we

might use abstract causal beliefs about tug-of-war (Hartshorne et al., 2019), or we might simulate Alice pulling the rope and Bart dragging along the ground in her direction; the latter is more sensitive to physical properties of the players and the field. Moreover, as a practical matter, one must marginalize out ("average over") irrelevant parts of one's world-model (e.g., who Bart's parents are and what he plans to eat after the match). Determining what is relevant is tricky and substantially affects inferences. Indeed, Bass and colleagues (2021) show that some "cognitive illusions" may be explained by biases in how relevance is determined.

Note that if the above approach is right, the categorical behavior often taken as emblematic of LoTs is likely to be masked by the probabilistic, graded natures of the grounding procedure and the model of the world.

So far, we've followed the Fodorian atomic treatment of constituents, but this is controversial. Linguists note that words tend to have many distinct meanings: one can throw a book (the physical object) or like a book (usually the content conveyed by the book, not the physical object). One can beat Bart or the bell, but in fundamentally different ways. There are many reasons not to treat these different meanings as homophones (a single word that refers to many unrelated concepts), one of the most obvious being that you end up needing an enormous (potentially unbounded) conceptual library. Perhaps we do, but linguists have noted that there are systematic correspondences between the various meanings, and that this can only be explained if the symbols Fodor takes to be atomic in fact have structure that contributes to meaning and governs their resulting conceptual combination and composition (Jackendoff, 1990; Pustejovsky, 1995). These solutions can be debated, but the problems have to be solved somehow.

QDPM provide a useful description of LoTs. Testing LoT theories, however, requires looking beyond the LoT to how it is used within a larger cognitive system. This, in almost all cases, will involve complex tradeoffs and interactions with graded, distributed, and analog systems of representation and processing.

References

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. Cognitive Neuropsychology, 38(7-8), 413-424.

Fodor, J. A. (1975). The language of thought (Vol. 5). Harvard University Press.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. Center for Brains, Minds and Machines (CBMM).

Hartshorne, J. K., Jennings, M. V., Gerstenberg, T., & Tenenbaum, J. (2019). When circumstances change, update your pronouns. In CogSci (p. 3472).

Jackendoff, R. S. (1992). Semantic Structures. MIT Press, Cambridge.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. Trends in cognitive sciences, 20(8), 589-604.

Oved, I. (2015). Hypothesis formation and testing in the acquisition of representationally simple concepts. Philosophical Studies 172 (1):227-247.

Pollock, J. & Oved, I. (2005). Vision, Knowledge, and the Mystery Link. Philosophical Perspectives, 19. Epistemology. 309-351.

Pustejovsky, J. (1995). The Generative Lexicon. MIT Press, Cambridge.

Pustejovsky, J., & Krishnaswamy, N. (2022, June). Multimodal Semantics for Affordances and Actions. In Human-Computer Interaction. Theoretical Approaches and Design Methods: Thematic Area, HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I (pp. 137-160). Cham: Springer International Publishing.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. Trends in cognitive sciences, 21(9), 649-665.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. Advances in neural information processing systems, 28.

Acknowledgements Statement: We thank members of the BabyBAW team, Mengguo Jing, andWei Li for valuable discussion.

Competing Interests: None.

Funding Statement: Funding was provided by NSF 2033938 and NSF 2238912 to JKH, NSF 2033932 to JP, and ARO W911NF-23-1-0031 to NK.