What we can (and can't) learn from artificial neural networks about biological neural networks.

Uri Hasson
Princeton University, Princeton, NJ, USA
https://hassonlab.princeton.edu/

One of the ultimate goals of our collective research endeavor in human neuroscience is to model and understand how the brain supports dynamic, context-dependent behaviors in the real world. Perhaps the most distinctly human behavior—and the focus of this talk—is our capacity for using language to communicate our thoughts to others during free, open-ended conversations.

Historically, cognitive neuroscientists have confronted the problem of how our brain navigates a complex, multidimensional social world using an incremental divide-and-conquer strategy. Individual labs use clever experimental manipulations to isolate a particular slice of language processing—for example, parametrically varying the syntactic complexity of isolated sentences—and measure the corresponding brain activity (e.g., Friederici, 2011; Price, 2012). The implicit aspiration behind this collective effort is to, one day, aggregate all of these piecemeal studies into a coherent neurocomputational model of natural language processing. While this paradigm has produced many foundational results, it has become increasingly clear that there is no easy way to synthesize these findings into a holistic understanding of real-world language processing.

After decades of research, there is increasing awareness of the gap between controlled laboratory experiments and the natural world's complexity (Nastase et al., NeuroImage, 2020). Models and theories developed in a particular experimental context often fail to generalize to other, more ecological contexts while accounting for only a minuscule proportion of variance in real-world behavior and brain activity. No matter how much we improve the sophistication and replicability of our laboratory findings, there is no guarantee these findings will have sufficient explanatory power or relevance for real-world behavior. To make matters even more challenging, language and communication are spontaneous, dynamic, and fundamentally contextual, unamenable to many core tenets of experimental design (e.g., repetition, trial averaging; Ben-Yakov et al., NeuroImage, 2012). Even for those who recognize this tension, moving from the laboratory to the real world seems daunting: most psychologists and neuroscientists are

trained to design experiments and do not have the tools to collect and analyze real-world data at scale. A central question at the core of this talk will be: How can we develop new theories and computational methods to model the underlying neural basis of natural language processing and communication in real-world contexts?

Deep learning provides a unified computational framework that can serve as an alternative approach to natural language processing in the human brain (Hasson et al., 2020; Richards et al., 2019). Leveraging principles from statistical learning theory and using vast real-world datasets, deep learning algorithms can reproduce complex natural behaviors in visual perception, speech analyses, and even human-like conversations. With the recent emergence of large language models (LLMs), we are finally beginning to see explicit computational models that respect and reproduce the context-rich complexity of natural language and communication. Remarkably, these models learn from much the same shared space as humans: from real-world language generated by humans. LLMs rely on simple self-supervised objectives (e.g., next-word prediction) to learn to produce context-specific linguistic outputs from real-world corpora—and, in the process, implicitly encode the statistical structure of natural language into a multidimensional embedding space (Manning et al., 2020; Linzen & Baroni, 2021; Pavlick, 2022).

These breakthroughs, however, have been driven mainly by industry-scale engineering, often neglecting biological plausibility (most are entirely disembodied) and lacking meaningful socio-environmental interaction. That is, state-of-the-art deep models rely on biologically implausible architectures and learning rules and are trained on unrealistic amounts of non-ecological training data. For example, a state-of-the-art language model is trained on ~500 billion words — and it would take a human baby ~6,000 years to process that many words. In sharp contrast, most children learn their first language within a few years by interacting with a small social network and relying on perceptual data that are not textual but spoken, multimodal, embodied, and immersed in social actions.

In the talk, I wish to ask whether we can adopt the new standards provided by the recent success in deep learning and build a new family of deep models that will respect the cognitive, embodied, and social constraints of the human brain. For example, can we build human-centric computational models of child development? Models that can

respect and emulate the progression of known developmental milestones (Sinha et al. 2006; Smith 2015; Frank 2023).

In the talk, I will review our recent attempts to build cognitive models of child development that focus primarily on the following critical dimensions: 1. Ecological and Child-Centered Data: Our models will be trained using realistic and ecological data gathered during the first 1000 days of a child's life. This contrasts large language models trained on textual corpora scraped from the internet. 2. Biologically Constrained Models: Our modeling approach is heavily influenced by our knowledge of the human brain and body. We choose not to use transformer-based architectures as they do not have biological plausibility. Instead, we opt for recurrent neural networks, which are more biologically feasible. In addition, instead of analyzing speech sounds using engineering tools, we intend to use models of the human ear to convert acoustic signals into model input. Similarly, we intend to use models of the articulatory system to create a child's speech rather than relying on sophisticated speech synthesizers. 3. Embodied Models: We treat each model as an active agent that uses reinforcement learning (RL) principles to learn language skills. Like children, this ability to act gives the RL agent an embodied way to interact with its environment as it shapes its ability to speak. 4. Multi-Modal Models: Unlike text-based large language models, our models integrate information from various modalities, including vision, speech, gesture, action, and touch. 5. Social Learning: Our RL-agent framework offers the opportunity to introduce a caregiver RL-agent. This additional agent can supply external guidance and feedback to the child RL-agent, infusing a social learning element into our modeling framework. 6) Developmental stages: Our learning approach will be modeled and evaluated following children's developmental stages.

Background Materials.

1. Hasson, U., Nastase, S. A. & Goldstein, (2020) **A. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks**. *Neuron* **105,** 416–434 https://www.sciencedirect.com/science/article/pii/S089662731931044X

2. Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Schain, M., Nastase, S. A., Zada, Z., Ham, E., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Reichart, R., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A. & Hasson, U. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat Commun* **15,** 2768 https://www.nature.com/articles/s41467-024-46631-y