



Products and Services



Discover Content



Cambridge Core

Home > Journals > Behavioral and Brain Sciences > Volume 46

> Even deeper problems with neural network models of...

Access

English | Français

In response to: **Deep problems with neural network models of human vision**

[Related commentaries \(29\)](#) [Author response](#)



Behavioral and Brain Sciences

Article contents


Abstract

Even deeper problems with neural network models of language

Published online by Cambridge University Press: **06 December 2023**

In response to: **Deep problems with neural network models of human vision**

[Related commentaries \(29\)](#) [Author response](#)

Thomas G. Bever , Noam Chomsky, Sandiway Fong and Massimo Piattelli-Palmarini




Commentary Related commentaries Metrics

Competing interest

References

Abstract

We recognize today's deep neural network (DNN) models of language behaviors as engineering achievements. However, what we know intuitively and scientifically about language shows that what DNNs are and how they are trained on bare texts, makes them poor models of mind and brain for language organization, as it interacts with infant biology, maturation, experience, unique principles, and natural law.

Type	Open Peer Commentary
Information	Behavioral and Brain Sciences, Volume 46, 2023, e387 DOI: https://doi.org/10.1017/S0140525X23001619 
Copyright	Copyright © The Author(s), 2023. Published by Cambridge University Press

There is a long tradition of taking current engineering devices as models of how nature itself works. Starting in the eighteenth century, new techniques to control motion in practical uses were adapted to create lifelike models of animals and humans, often populating royal gardens with marvelous lifelike creatures. It was tempting to suggest that the brain itself operates with its own instantiation of such mechanistic principles. Early on came the analogy with a clock, then the telephone network, then the digital computer. Today's "deep neural networks" (DNNs) are sometimes taken as models, since they achieve fantastic performance accuracy in human mental activities, discriminating objects, or producing normal language.

Bowers et al. question the utility of using DNNs and related methodology as models of vision with optical object recognition and categorization as a rubric: They note a range of empirical failures, experimental flaws, and principled reasons why DNNs fail to include other vision facts.

Vision in large part organizes the independent physical world, but human language lies at the internal extreme – almost completely created by the human mind/brain. Accordingly, investigations of language necessarily start with study of the internal knowledge itself. What such investigations reveal about language is inconsistent with DNNs that eschew linguistic theories.

- (a)** The poverty of the stimulus for children – limited experience, and little explicit training result in sophisticated language ability.
- (b)** The immediate role in early child language of hierarchical categories and computational constraints (e.g., on anaphoric relations).
- (c)** Structural elements of language syntax are discrete, the number of combinations is infinite.
- (d)** The distinction between grammar (aka *competence*) and behavior (aka *performance*) (note: DNNs are intentionally dependent on actual language behaviors).
- (e)** The role of maximum computational simplicity underlying nature (Einstein's *Miracle Creed*; Chomsky, in press; McDonough, 2022).

Bowers et al. note the notorious flaw of DNNs: "...state-of-the-art DNNs of natural language processing receive training that far exceeds any human experience.... This highlights how these DNNs are missing key human inductive biases that facilitate the learning of natural languages but impair the learning of unstructured languages (something akin to a human language acquisition device)" (target article, sect. 5, para. 2).

The term "inductive biases" reflects an assumption that the "poverty of the stimulus" can be overcome by a list of built-in "priors" which increase the speed of gradual inductive learning: Yet decades of research show that that language emerges without any such general induction process. Rather the evidence indicates an available universal grammar, which defines limited structural options for all languages. Children quickly latch onto particular options of their native language from "signature sentences" (e.g., Gleitman & Landau, 2013; Guasti, 2002; Yang, 2011).

The study of vision and language do share some features. For example, in both domains, the structure is often clarified by subtle cues, illuminating its critical properties. In vision, an image of a panda plus a little visual noise is reported as a gibbon for trained DNNs (Goodfellow et al., 2015).

Correspondingly in language research, "minimal pairs" (sentences that vary slightly), can result in strong and reliable differences in structure, interpretation, and grammaticality. Language's discrete infinity property ensures an endless supply of such examples. Thus, large-scale DNN systems, despite unlimited storage, and vast amounts of language

data, do not reliably match human performance: Imitation without the human language faculty.

Humans recognize that “the chicken is ready to eat” exhibits structural ambiguity. DNN systems that explicitly compute parses, for example, Google Natural Language, do not recognize the ambiguity, preferring the sentential subject to be subject of “eat.” Generative artificial intelligence (AI) systems do not output parses, but we can still deduce underlying grammatical relations by appending a question. In the case of ChatGPT, we can ask for comment with “Is X an ambiguous sentence?” This line of questioning reveals that it assumes “chicken” is the object of “eat.” Swapping “children” for “chicken” reveals ambiguity that it reports quite disturbing. Context, for example, the relative proximity of discourses involving cannibals, the story of Hansel and Gretel, or hungry aliens, plays a relative role in ChatGPT's training.

In fact, ChatGPT uses a several thousand token context, potentially capturing discourse phenomena. Consider “The white rabbit jumped from behind the bushes. The animal looked around and then he ran away.” For both humans and ChatGPT, *he*, the *animal* and *white rabbit* are preferred to be the same. But if the sentences are reversed in order, only humans then treat “rabbit” as a different entity from “animal,” revealing a fundamental principle of anaphoric relations. If DNN is to be a useful model of human behavior, we must know which parameter out of the billions should be adjusted to correct such divergence: Within the statistical enterprise, such errors cannot be diagnosed nor fixed.

The authors briefly raise issues involving the evolution of vision as constraining it gradually over many species and eons. Most obvious, and important for vision science, cross-species analogies are multiple and detailed, but not available for language. The authors correctly say that, in spite of claimed success at learning languages, “DNNs will also happily learn [number agreement] in impossible languages with...structures that are not found within any natural languages and which humans struggle to process” (target article, sect. 5, para. 2).

This difference between real syntactic rules and impossible syntactic rules goes much deeper. Like DNNs, humans can master both kinds of rules. Yet in humans, this has underlying neurological correlates that reflect what we know independently about normal neurological processing of language (e.g., Musso et al., 2003). Learning a real language previously unknown to the subjects activates Broca's area: But the same task with an impossible syntactic rule (e.g., a rule that ignores hierarchical structure in favor of serial position) activates only brain areas normally activated during general problem-solving.

We have reviewed ways in which DNNs are empirically inadequate and discordant with theories of language in humans. Adequate or not, we have no idea how individual trained DNNs do what they do: For a DNN to be psychologically useful, we need a theory of the “psychological” innards of the DNN, which is either the same as the theory of human innards, or a unique theory of how initially random associations are compiled from actual

behaviors into a model that can be tested on humans (Bever, Fodor, & Garrett, 1968).

Why not focus on attempts to organize and constrain DNNs and other types of models so they comport with what we already know about language, language learning, language representations, and language behaviors? The answer for DNNs is also their touted practical virtue, they learn from actual text, free of hand tailored structural analysis. This engineering virtue pyrrhically underlies why they are doomed to be largely useless models for psychological research on language.



Acknowledgments

We thank Jay Keyser, Andrea Moro, and Robert Berwick for their advice.

Competing interest

None.

References

-  Bever, T. G., Fodor, J. A., & Garrett, M. (1968). A formal limitation of associationism. In Dixon, T. R. & Horton, D. L. (Eds.), *Verbal behavior and general behavior theory* (pp. 582–585). Prentice Hall. [Google Scholar](#)
-  Chomsky, N. (in press). The miracle creed and SMT. In Greco, M. & Mocci, D. (Eds.), *A Cartesian dream: A geometrical account of syntax: In honor of Andrea*

Moro. *Rivista di Grammatica Generativa/Research in Generative Grammar*. Lingbuzz Press. [Google Scholar](#)



Gleitman, L., & Landau, B. (2013). Every child an isolate: Nature's experiments in language learning. In Piattelli-Palmarini, M. & Berwick, R. C. (Eds.), *Rich languages from poor inputs* (pp. 91–106). Oxford University Press. [Google Scholar](#)



Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Bengio, Y. & LeCun, Y. (Eds.), *3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9*. [Google Scholar](#)



Guasti, M. T. (2002). *Language acquisition: The growth of grammar*. MIT Press. [Google Scholar](#)



McDonough, J. K. (2022). *A miracle creed: The principle of optimality in Leibniz's physics and philosophy*. Oxford University Press. [CrossRef](#) [Google Scholar](#)



Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Buechel, C., & Weiler, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, 6(7), 774–781. [CrossRef](#) [Google Scholar](#) [PubMed](#)



Yang, C. (2011). Learnability. In Roeper, T. & de Villiers, J. (Eds.) *Handbook of language acquisition* (pp. 119–154). Kluwer. [Google Scholar](#)
