How Can Deep Neural Networks Inform Theory in Psychological Science?

Sam Whitman McGrath

MASTER'S THESIS

Submitted in partial fulfilment of the requirements for the degree of Master of Science in the Department of Cognitive and Psychological Sciences at Brown University

Supervised by:

Roman Feiman

Examining Committee:

Michael Frank, Ellie Pavlick

BROWN UNIVERSITY
THE GRADUATE SCHOOL

Name:

Division/Department/Program/School: Department of Cognitive and Psychological Sciences

Previous degrees:

Title of thesis: How Deep Neural Networks Can Inform Theory in Psychological Science

Thesis approved for department by: Roman Feiman

Approval of thesis recommended to Graduate Council by Committee: Roman Feiman, Ellie Pavlick, Michael Frank

Date of Final Thesis Presentation: April 7, 2025

Votes on recommending degree (to be signed by examining committee):

Aye Nay

Roman Feiman
Ellie Pavlick
Michael Frank

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
1–9

DOI: 10.1177/09637214241268098
www.psychologicalscience.org/CDPS
Sage

How Can Deep Neural Networks Inform Theory in Psychological Science?

Sam Whitman McGrath[1,2], Jacob Russin[2,3], Ellie Pavlick[3,4], and Roman Feiman[2,4]

[1]Department of Philosophy, Brown University; [2]Department of Cognitive and Psychological Sciences, Brown University; [3]Department of Computer Science, Brown University; and [4]Program in Linguistics, Brown University

Abstract

Over the last decade, deep neural networks (DNNs) have transformed the state of the art in artificial intelligence. In domains such as language production and reasoning, long considered uniquely human abilities, contemporary models have proven capable of strikingly human-like performance. However, in contrast to classical symbolic models, neural networks can be inscrutable even to their designers, making it unclear what significance, if any, they have for theories of human cognition. Two extreme reactions are common. Neural network enthusiasts argue that, because the inner workings of DNNs do not seem to resemble any of the traditional constructs of psychological or linguistic theory, their success renders these theories obsolete and motivates a radical paradigm shift. Neural network skeptics instead take this inability to interpret DNNs in psychological terms to mean that their success is irrelevant to psychological science. In this article, we review recent work that suggests that the internal mechanisms of DNNs can, in fact, be interpreted in the functional terms characteristic of psychological explanations. We argue that this undermines the shared assumption of both extremes and opens the door for DNNs to inform theories of cognition and its development.

Keywords

deep learning, neural networks, large language models, interpretability, psycholinguistics, cognitive development, philosophy of cognitive science

Contemporary deep neural networks (DNNs) can recognize and label images and videos, write poems, and plan travel itineraries—tasks that, just a few years ago, only humans could accomplish. Can understanding how artificial neural networks acquire and exercise such abilities help psychologists understand the corresponding abilities in humans?

Although modern DNNs are in their adolescence, this question is approaching its 40th anniversary. It first arose with the "connectionist" networks of the 1980s (McClelland et al., 1986; Smolensky, 1988). Even as contemporary models have progressed beyond these predecessors (Smolensky et al., 2022), they continue to work according to many of the same principles. At a fundamental level, all neural networks are made up of layers of interconnected units ("nodes" or "neurons"), each of which calculates a weighted combination of its inputs and applies a (usually nonlinear) function to the result, which it then passes to further nodes. The network's weights are initially random and their computations meaningless because the designer does not specify by hand what functions they should compute. Meaningful behavior emerges only as the network learns by gradually adjusting its matrices of weights so that its outputs more closely match specified targets (e.g., the numerical encoding of the next word in a text).

Because of their architectural similarities, contemporary attempts to link DNNs to cognitive theories face the same challenge that confronted earlier connectionists: Neural networks are difficult to interpret in higher level terms. Even once they have been trained to produce meaningful behavior, DNNs can seem like "black boxes"—amorphous webs of indistinguishable nodes, their internal activity an inscrutable cascade of matrix multiplications that bears no obvious relationship to

Corresponding Author:
Roman Feiman, Department of Cognitive and Psychological Sciences and Program in Linguistics, Brown University
Email: roman_feiman@brown.edu

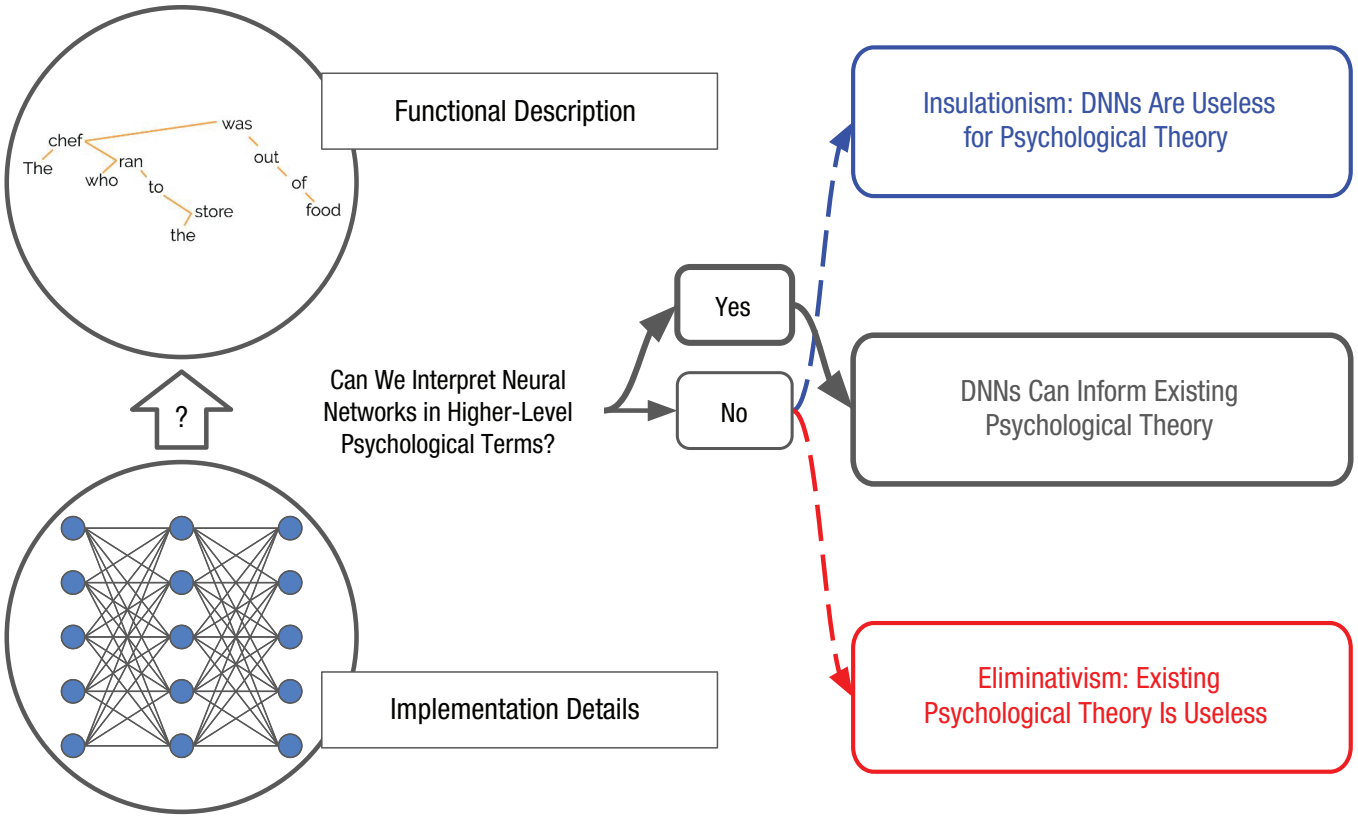

Fig. 1. Neural networks and psychological theory. The computations in a neural network model (bottom left) consist of matrix operations. Although these operations fully specify a given network, they are difficult to interpret in terms of the functional characterizations (e.g., tree structures; Manning et al., 2020) used in traditional psychological explanations (top left). The view that these networks are not interpretable has led to two extreme reactions: Eliminativists (red) conclude that higher level interpretations are unnecessary, whereas insulationists (blue) conclude that networks cannot be relevant to psychological science. However, recent work suggests that it is possible to interpret the inner workings of neural networks in succinct, functional terms, opening up the possibility of informing cognitive theory with DNNs (gray). DNN = deep neural network.

the theoretical constructs of psychology or linguistics. Where in a DNN is a *belief*, or a *concept*, or a *rule*?

Past debates about the ability of neural networks to inform psychological theory lost steam as older networks ran up against limitations in their ability to model key aspects of human behavior. What has come as a surprise to neural network enthusiasts and skeptics alike is that simply making the models bigger—both in architecture (e.g., the number of layers) and the amount of training data they are exposed to—has dramatically improved their capabilities (Brown et al., 2020). For the first time, DNNs can solve problems of the scale and complexity actually faced by humans rather than just the highly restricted tasks common in earlier work. Moreover, the fact that modern DNNs are already "pretrained" before being applied to a new task makes them a closer analogue to human participants, who come into any experimental context equipped with relevant experience. These improvements dovetail with recent insights into how to more fairly compare human and machine performance (Buckner, 2023; Firestone, 2020) and motivate a reexamination of whether and how DNNs can inform psychological theory.

Making Sense of the Deep Learning Revolution

With some notable exceptions (Smolensky, 1988), the debate about what neural networks mean for psychological science has largely been shaped by two extreme, opposing views (Fig. 1). One view is that the success of neural networks makes much of psychological and linguistic theory obsolete. This view holds that the rise of DNNs is like other cases of scientific elimination in psychology, in which once promising theoretical constructs, such as Freud's id or ego, were discarded in favor of more empirically successful ones. From this perspective, DNNs are taken to constitute theories (Piantadosi, 2023) that win out empirically over alternatives that appeal to symbolic concepts or rules. In the extreme

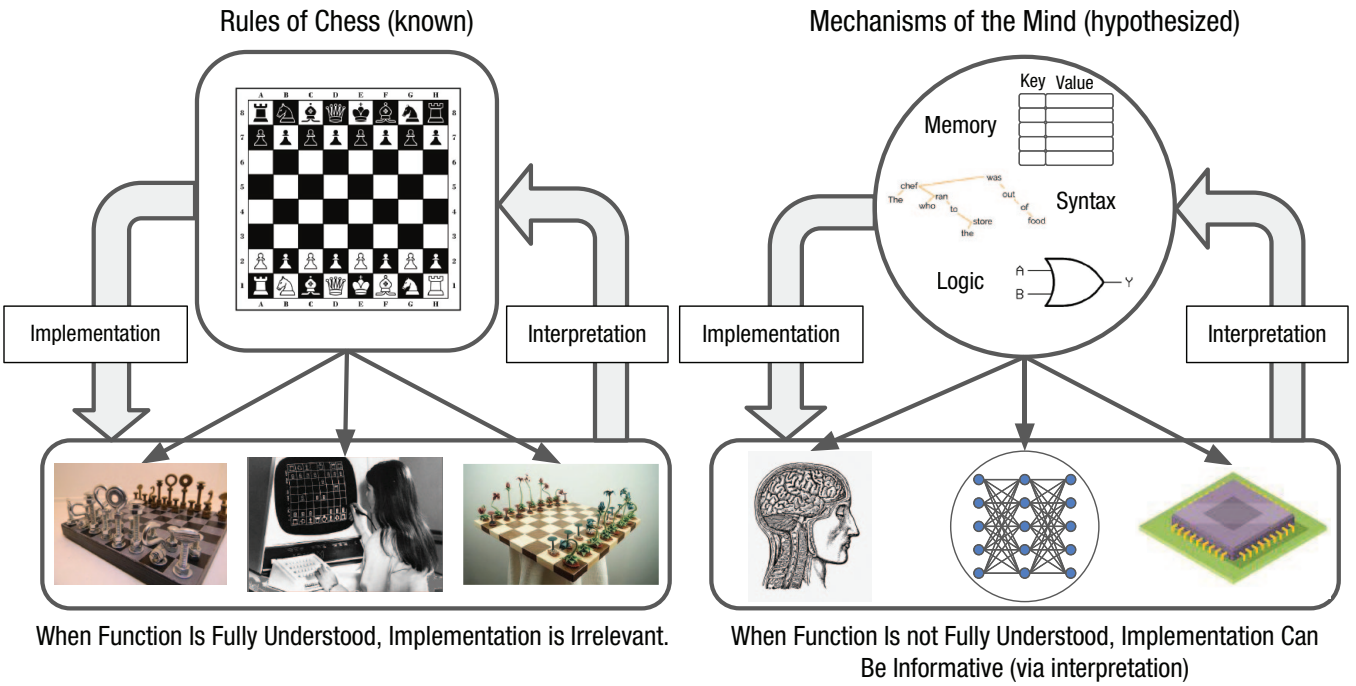

Fig. 2. Implementation and interpretation. To get a sense of the intuition behind the highly influential insulationist objection (Fodor & Pylyshyn, 1988)—and to see what is wrong with it—consider the game of chess (left). Chess can be implemented in multiple physical formats. The higher level functional rules that define the game cannot be reduced to lower level physical properties of any one implementation. For a theorist interested in the rules or strategy of chess, the implementation—whether it is played on a board or on a screen, what the pieces are made of, and so on—is entirely irrelevant. However, this is true of chess only because we know the rules in advance. If instead we had to figure out the rules by watching people play, then implementation-level details, such as all the pawns being the same shape and size, would offer informative, highly relevant indicators (furnishing evidence, in this case, that the pawns will all move in the same way). This is the situation we find ourselves in with respect to the mind (right). We start off not knowing how the mind works and need to reverse-engineer it via a complex process of abductive inference. Implementation-level facts form a key part of this inferential base and may well prove useful and informative in discerning the relevant higher level facts. Even if DNNs are taken to be implementation-level models, they can still feature in this abductive process and prove informative for cognitive science, provided that their representations and mechanisms are interpretable. DNN = deep neural network.

case, this *eliminativist* reaction threatens not just specific constructs but all theory not articulated in the lower level vocabulary of nodes and weights.

This radical reaction is driven by the combination of DNNs' impressive success at mimicking human behavior and the apparent impossibility of aligning their matrix multiplications with the higher level constructs of cognitive theory (e.g., beliefs, concepts, rules). According to this view, if DNNs do not afford interpretations in these terms, all the worse for cognitive theory—the relevant causal structure of behavior can, apparently, be captured in lower level terms anyway. The rise of DNNs is thus taken to signal a paradigm shift for psychology, overturning long-standing theoretical commitments and reshaping what counts as a satisfying explanation of cognition.

At the same time, skeptics of neural networks have relied on the same black-box assumption to argue for the opposite extreme: To the extent that DNNs cannot be put into any meaningful contact with existing higher level characterizations of cognition, they are simply irrelevant to psychological theory (Bowers et al., 2023). If higher level properties are not discernible in the weights and activations of trained networks, how can we even be sure that they offer incompatible cognitive theories, as the eliminativist claims? For all we know, DNNs' successes may be the result of merely implementing the psychological constructs of existing higher level theories (Fig. 2; Fodor & Pylyshyn, 1988; Quilty-Dunn et al., 2023). According to this view, no matter how impressive DNNs' behavior becomes, psychological science is effectively insulated from their influence. Even if the rise of deep learning represents a tectonic technological shift, the *insulationist* holds that it will barely register as a scientific tremor.

Challenging the Black-Box Assumption

Both extremes—eliminativism and insulationism—rest on the assumption that DNNs are black boxes that are not interpretable in the vocabulary of higher order theories of the mind. Recent empirical work in interpretability[1] challenges this assumption because researchers in artificial intelligence (AI) have

increasingly uncovered high-level, functional explanations of DNN behavior.

For example, the ability of large language models (LLMs) to generate long strings of grammatical text (Linzen & Baroni, 2021) spurred research on correspondences between aspects of LLMs' internal states and specific constructs in linguistic theory, such as parts of speech and syntax trees (Manning et al., 2020; Tenney et al., 2019). The overwhelming consensus from such work is that LLMs' internal states are neither unstructured nor inscrutable but exhibit geometric regularities that can be aligned with these linguistic constructs. For example, the way that sentence information is organized across the layers of these models reflects the traditional language-processing pipeline, with earlier layers representing parts of speech and parsing grammatical relations and later layers encoding semantic roles (who did what to whom) and tracking when different terms refer to the same entity (Tenney et al., 2019). It is even possible to reconstruct the syntactic parse trees postulated by traditional linguistic theory from models' internal representations of particular sentences (Manning et al., 2020).

Further work has suggested not only that these higher level descriptions are useful for interpreting model components but also that they play a causal role in the models' grammatical behavior. For instance, Chen et al. (2024) found that reconstructing syntactic parse trees first becomes possible during a specific window of training in which models show a sudden increase in the grammatical capacities that these same representations should enable. They also demonstrated a more direct causal link—targeted interventions designed to prevent the model from learning these high-level representations obstruct its acquisition of the corresponding grammatical capabilities. Other evidence suggests that higher level descriptions also play a causal role once a model is trained. Manipulating the nodes and weights that correspond to higher level interpretations elicits the expected changes in the models' behavior. For example, editing the weights that a model has used to produce "The Eiffel Tower is in Paris" to instead produce "The Eiffel Tower is in Rome" has related downstream effects, such as causing the edited model to give directions that refer to the "nearby" St. Peter's Basilica (Meng et al., 2022). Related techniques have been applied to logical inference (Geiger et al., 2021) and syntactic structure (Tucker et al., 2021). These findings suggest that higher order constructs identified by interpretability methods can be causally efficacious rather than merely epiphenomenal.

Interpretability research is uncovering not only the representations that DNNs use but also the algorithms that operate on these representations. For example, some of the most impressive, flexible reasoning behaviors of LLMs, such as their ability to learn new tasks given just a few examples (Brown et al., 2020), may be explained by specialized components called "induction heads" (Olsson et al., 2022) that emerge over the course of training. As the model predicts the next word that follows a word *w*, induction heads search the current context for any previous occurrence of *w* and copy the subsequent word. This mechanism appears to be relatively content-independent, performing the same function regardless of the identity of *w*, thus providing the model with computational machinery that supports the ability to "reason by induction," predicting that a pattern will be repeated (Olsson et al., 2022).

Such specialized mechanisms do not operate in isolation but appear to interface with one another to form task-specific circuits (Elhage et al., 2021; Olsson et al., 2022), for example, for processing the relative magnitudes of numbers (Hanna et al., 2023) or identifying syntactic constituents (Wang et al., 2022). Further evidence shows that some components are reused across circuits for different tasks (Merullo et al., 2023), suggesting partly content-independent operations that can be flexibly recombined.

Taken together, such findings open a path beyond the extreme views discussed in the previous section. The correspondence between LLMs' internal states and existing theory (e.g., theories of syntax, algorithms from traditional symbolic computing) challenges the eliminativist claim that the success of DNNs obviates higher level explanations. To the contrary, the fact that DNNs seem to reinvent structures similar to those postulated in existing theory (Manning et al., 2020; Tenney et al., 2019) means that existing explanations can help to understand how DNNs work and to causally intervene in their operation. At the same time, the fact that we can often interpret the black box of the neural network in higher level functional terms challenges the insulationist view that the success of DNNs is irrelevant to psychological theory. To the contrary, our growing understanding of the internal mechanisms of DNNs—currently the only models capable of reproducing a wide range of human-like behaviors in diverse settings—suggests that they may offer valuable insights into these same abilities in humans.

How Deep Neural Networks Can Inform Theory in Psychological Science

This recent progress suggests a broader strategy for leveraging interpretations of DNNs to inform theories of both the architecture of human cognition and its development. Where a DNN's performance matches human behavior, the key to evaluating whether it can

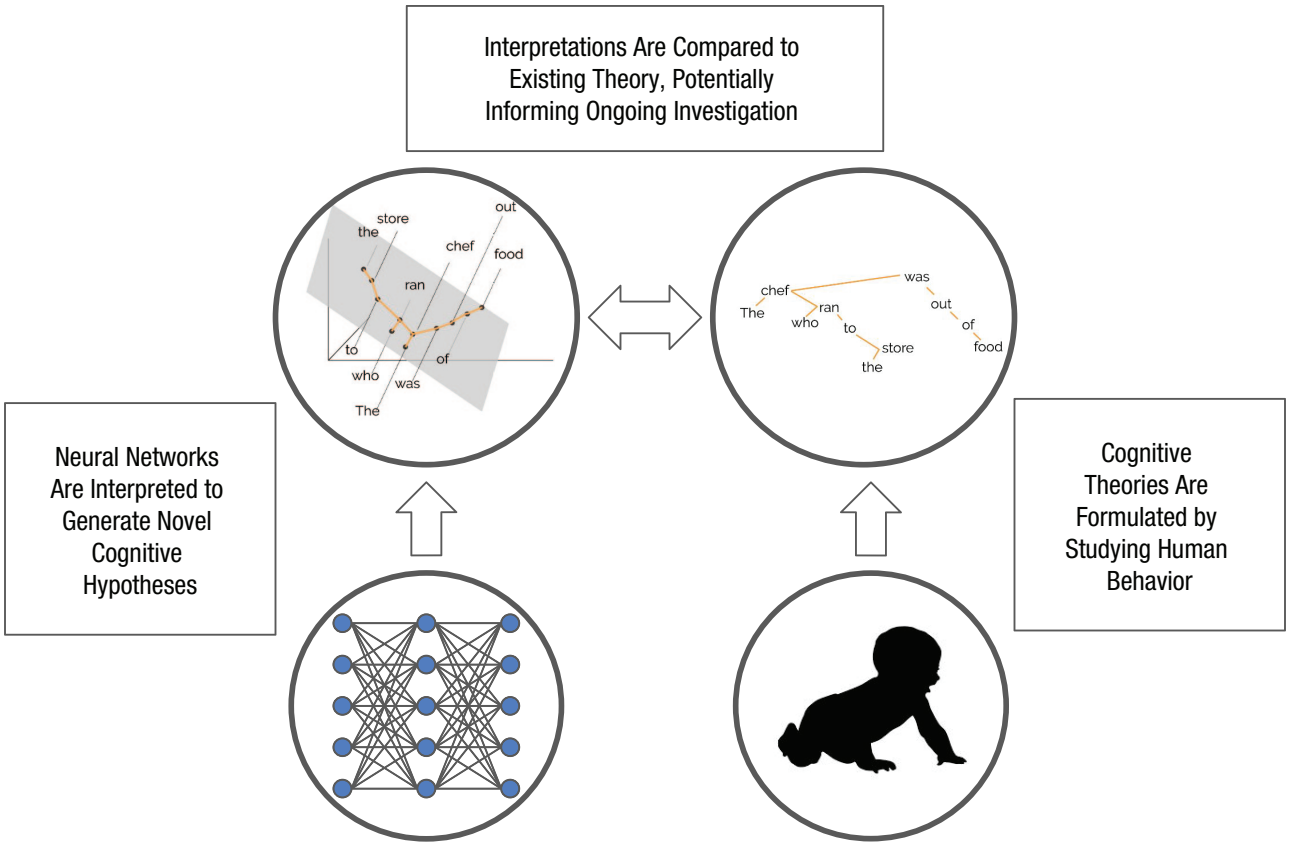

Fig. 3. Comparing DNN interpretations and cognitive theory. Recent advances in interpreting neural networks may allow DNNs to inform theory in cognitive science. When the low-level mechanisms of a trained neural network can be interpreted in the high-level functional terms familiar to cognitive psychology using new techniques (Manning et al., 2020), direct comparisons with human behavior and existing cognitive theory become possible. DNN = deep neural network.

inform psychological science will be to extract functional characterizations of its competencies to compare with existing cognitive theory (Fig. 3). If these functional characterizations precisely align with some existing theory, we may conclude that the model is a "mere implementation." In that case, finding that a DNN independently converged on a previously hypothesized cognitive solution would bolster claims of its optimality and generalizability (analogous to interpretations of convergent evolution in biology).

On the other hand, where we find divergences from existing theory, there are two possibilities. One is that the model has hit on an alien solution—that its internal representations and algorithms are fundamentally different from those used by humans. The other is that these divergent functional characterizations can serve as genuine alternative hypotheses about how human cognition works. Given that researchers are rarely sure that they have exactly the right cognitive theory in the first place (Fig. 2), taking this possibility seriously may motivate revision, although not necessarily wholesale replacement, of existing theory. Here, DNNs will neither eliminate nor merely implement existing psychological constructs. Instead, they can offer *informative implementations*, generating new insights and hypotheses about human representations for psychologists to evaluate. Work along these lines is ongoing, and there are as of yet no uncontroversial instances of contemporary DNNs modifying psychological theory. Below, we highlight significant examples and promising directions.

Deriving psychological insights from neural network models is not a new aspiration but one of the original goals of the connectionist movement. For example, influential early work by McClelland et al. (1986) aimed to use neural networks to model the psychological processes involved in converting verb stems (e.g., "write," "slight") into their past tense forms (e.g., "wrote," "slighted"). They focused on a feature of the English past tense that preexisting rule-based theories struggled to explain: quasiregularity (for a review, see Seidenberg & Plaut, 2014). Although most English verbs conform to a "regular" *add -ed* rule, striking subregularities exist among the "irregular" verbs as well (e.g., "write" → "wrote," "smite" → "smote"). Neural networks could

capture these quasiregular patterns with distributed representations that can simultaneously encode regularities at multiple levels of granularity. This implementation-driven insight was then incorporated back into higher order psychological theories, even by theorists who continued to posit a separate rule-based mechanism for regular verbs (Pinker & Prince, 1988). Today, DNNs have overcome many of the empirical challenges that Pinker and Prince raised, matching human performance better than any prior model (Kirov & Cotterell, 2018). This means they now have the potential to shed light on an unresolved question from the earlier debate: Is the past tense for regular and irregular verbs formed via a common mechanism? It may be tempting to assume that a successful DNN must reflect a single mechanism, but applying emerging interpretive techniques to these models could allow researchers to investigate both alternatives; it is possible that the new networks' success rests on circuits that form associations among irregulars separately from those that apply a discrete rule, such as adding *-ed*, to the regulars.

These new methods are already being used to interpret networks in other cognitive domains, such as vision, in which DNNs have transformed inquiry in cognitive psychology and neuroscience (e.g. Yamins et al., 2014). It is well established that DNNs trained on images learn representations that are predictive of human neural responses to visual stimuli (Schrimpf et al., 2020). Researchers have begun to interpret these networks to investigate cognitive hypotheses, such as the extent to which the human visual system can be understood as performing generative inference (Golan et al., 2020), or the extent to which human visual representations are shaped by top-down pressure from explicit category labels (Konkle & Alvarez, 2022). Neural network models have also been argued to explain the emergence of cognitive maps, one of the key psychological phenomena that motivated the cognitive revolution against behaviorism. When trained on spatial or relational tasks, these networks end up capturing the neural response patterns associated with place cells and grid cells, which have been hypothesized to underlie human spatial navigation and relational reasoning abilities (Whittington et al., 2020).

Compositionality, the ability to recombine familiar linguistic or cognitive components into novel combinations, has posed arguably the deepest challenge to neural networks as models of human linguistic and psychological capacities (Fodor & Pylyshyn, 1988). Even here, however, recent neural networks have made major breakthroughs, reproducing impressive human-like compositional generalization behaviors (Lake & Baroni, 2023). Although it would be premature to draw concrete conclusions about the implications of this work for theories of human cognition, interpreting networks capable of replicating the behavioral signatures of compositionality offers a novel approach to questions that have been central to cognitive science since its inception.

In addition to informing cognitive theories of adult competence, DNNs can also inform theories of cognitive development—a second avenue of influence on psychological science. Of course, it would be simplest to treat DNNs as models of human learners and then test how changes in model components or input data change learning trajectories and outcomes. As has been widely discussed, however, there are significant barriers to this approach. LLMs have been trained on orders of magnitude more language data than children could ever observe (Warstadt & Bowman, 2022). They learn primarily through next-word prediction, with feedback about each prediction's accuracy. And no current LLM interacts with the world like a human child—they cannot follow a parent's gaze or learn that light bulbs are hot by touching one. Creating DNNs that learn like humans along these dimensions is, at best, a long-term goal.

Critically, however, DNNs can still indirectly inform theories of cognitive development even if they do not learn like humans by serving as powerful tools in the study of learnability. Learnability is not the study of how humans actually learn but of what is in principle possible to learn under which conditions, by which kinds of learners. This can guide the study of actual learning by constraining plausible hypotheses, helping to identify what may or may not be special about human minds. Here, DNNs join the researcher's roster of model organisms, which range from nonhuman animals to classical computers.

For instance, because DNNs start with no content-specific knowledge, they can be used to evaluate "poverty of the stimulus" arguments, assessing whether particular knowledge is learnable by any agent both in principle and from the amount of data plausibly available to a child (Wilcox et al., 2022). In a recent example of this approach, Vong et al. (2024) showed that a generic DNN trained on visual and language data that was recorded from a single child's experience learned the right associations between words and images, suggesting that some aspects of linguistic meaning can be acquired without prior knowledge. A related approach is "controlled rearing," in which researchers manipulate a learner's input (often in ways that would be impossible or unethical to do with humans) and evaluate how this changes what is learned. Misra and Mahowald (2024) leveraged the possibility of precisely controlling DNN training sets to show that certain rare grammatical phenomena (article + adjective + numeral + noun constructions such as "a beautiful five days") are learnable even when all instances of these constructions are

deleted from the training corpus. This suggests that it is possible to generalize to these rare constructions from similar but less rare ones (e.g., "a few days"), undermining the need to posit an innate syntactic constraint. Extending this approach to more of what adults know—from theorized syntactic parameters to conceptual primitives—can further constrain hypotheses about what is innate in humans.

Beyond questions of what is learnable, DNNs can be used to test theoretical accounts of "learning dependencies." In language acquisition, for example, Gleitman (1990) argued that the reason toddlers learn nouns before verbs is that knowing nouns makes the meanings of accompanying verbs easier to identify, whereas early nouns are easy enough to learn in syntactic isolation. By experimentally manipulating the curriculum of training data (perhaps in extensions of paradigms such as the one used by Vong et al., 2024), DNNs could be used to evaluate such hypotheses. In cognitive development, Feiman et al. (2022) proposed that even a concept as basic as logical negation could be constructed out of precursors that are only implicitly logical, such as constraints on representations of physical objects (if an object is in one place, it cannot be in another). Traylor et al. (2023) subsequently tested this proposal with different models. They found evidence for implicitly logical inferences in a symbolic model of intuitive physics but not in a DNN trained to track and reidentify objects through occlusion. This suggests that although intuitive physics might be a sufficient base for implicit logic (evidence in favor of one learning dependency), object tracking alone might not be sufficient to learn intuitive physics (evidence against another dependency).

Finally, it is worth reiterating that DNNs can inform our understanding of learning dependencies even when they differ from human learners. Indeed, the less a DNN resembles human learners the more informative it would be if it recapitulated a similar developmental trajectory. That would suggest that the trajectories observed in humans are determined by the structure of their input rather than specific features of their cognitive architecture. Conversely, if DNNs do not need to learn information in the same order that humans do, that would indicate that human learners bring something importantly different to the task.

Conclusion

Recent advances in interpreting the internal processes of DNNs have identified functionally specialized internal components, emerging from cascades of matrix multiplications that once seemed inscrutable. The development of these methods of interpretation means that models capable of producing sophisticated human-like behaviors can be analyzed and may serve as informative implementations of the corresponding cognitive processes in humans, generating novel insights and providing new ways to evaluate theories of both mature cognition and its development. This points to new research directions beyond the extreme reactions that have shaped debates about the significance of neural networks and opens the door to a collaborative, mutually informative relationship between deep learning research and psychological science.

Recommended Reading

Fodor, J. A., & Pylyshyn, Z. W. (1988). (See References). One of the most influential critical discussions from the original connectionist debates; contests the relevance of neural networks to psychological science.

Linzen, T., & Baroni, M. (2021). (See References). Detailed review of the syntactic abilities of DNNs and discussion of their relation to theoretical linguistics.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). (See References). Develops and deploys methods for identifying linguistic structure in artificial neural networks.

Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., & Gao, J. (2022). (See References). Provides an accessible review of recent advances in DNNs, with a similar emphasis on the need to bridge the neural and high-level symbolic perspectives and on interpreting DNNs.

Traylor, A., Feiman, R., & Pavlick, E. (2023). (See References). Examines the logical reasoning capabilities of neural networks, focusing on implicit logical operators, a hypothesized developmental precursor to full-fledged logical concepts.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Author Contributions
S. W. McGrath and J. Russin contributed equally to this manuscript. All of the authors approved the final manuscript for submission.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Sam Whitman McGrath https://orcid.org/0009-0000-5383-1885

Note

1. The term "interpretability" has become commonplace in machine learning research (Elhage et al., 2021), but its exact definition varies slightly with the goals of different research programs (e.g., AI safety, trust, bias mitigation). Here, by "interpretation" we mean the characterization of the internal processes

of a trained neural network in the higher level terms typical of psychological explanation.

References

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, Article e385. https://doi.org/10.1017/S0140525X22002813

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165v4

Buckner, C. (2023). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *British Journal for the Philosophy of Science*, *74*(3), 681–712. https://doi.org/10.1086/714960

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. (2024). *Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs*. arXiv. https://doi.org/10.48550/arXiv.2309.07311

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., . . . Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Feiman, R., Mody, S., & Carey, S. (2022). The development of reasoning by exclusion in infancy. *Cognitive Psychology*, *135*, Article 101473. https://doi.org/10.1016/j.cogpsych.2022.101473

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences, USA*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. In M. Ranzato et al. (Eds.), *Advances in neural information processing systems 34 (NeurIPS 2021)* (pp. 9574–9586). Curran Associates, Inc.

Gleitman, L. (1990). The structural sources of verb meaning. *Language Acquisition*, *1*(1), 3–55.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences, USA*, *117*(47), 29330–29337. https://doi.org/10.1073/pnas.1912334117

Hanna, M., Liu, O., & Variengien, A. (2023). *How does GPT-2 compute greater-than? Interpreting mathematical abilities in a pre-trained language model*. arXiv. https://doi.org/10.48550/arXiv.2305.00586

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665. https://doi.org/10.1162/tacl_a_00247

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*, Article 491. https://doi.org/10.1038/s41467-022-28091-4

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*, 115–121. https://doi.org/10.1038/s41586-023-06668-3

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, USA*, *117*, 30046–30054. https://doi.org/10.1073/pnas.1907367117

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2: Psychological and biological models*. MIT Press.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In S. Koyejo et al. (Eds.), *Advances in neural information processing systems 35 (NeurIPS 22)* (pp. 17359–17372). Curran Associates, Inc.

Merullo, J., Eickhoff, C., & Pavlick, E. (2023). *Circuit component reuse across tasks in transformer language models*. arXiv. https://doi.org/10.48550/arXiv.2310.08744

Misra, K., & Mahowald, K. (2024). *Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs*. arXiv. https://doi.org/10.48550/arXiv.2403.19827

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., . . . Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

Piantadosi, S. (2023). *Modern language models refute Chomsky's approach to language*. LingBuzz. https://lingbuzz.net/lingbuzz/007180

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, Article e261. https://doi.org/10.1017/S0140525X22002849

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). *Brain-score: Which artificial neural network for object recognition is most brain-like?* bioRxiv. https://doi.org/10.1101/407007

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, *38*(6), 1190–1228. https://doi.org/10.1111/cogs.12147

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23. https://doi.org/10.1017/S0140525X00052432

Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., & Gao, J. (2022). *Neurocompositional computing: From the central paradox of cognition to a new generation of AI systems*. arXiv. https://doi.org/10.48550/arXiv.2205.01128

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT rediscovers the classical NLP pipeline*. arXiv. http://arxiv.org/abs/1905.05950

Traylor, A., Feiman, R., & Pavlick, E. (2023, May 1–5). *Can neural networks learn implicit logic from physical reasoning?* [Conference session]. Eleventh International Conference on Learning Representations, Kigali, Rwanda. https://openreview.net/forum?id=HVoJCRLByVk

Tucker, M., Qian, P., & Levy, R. (2021). What if this modified that? Syntactic interventions with counterfactual embeddings. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 862–875). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.76

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). *Interpretability in the wild: A circuit for indirect object identification in GPT-2 small*. arXiv. https://doi.org/10.48550/arXiv.2211.00593

Warstadt, A., & Bowman, S. R. (2022). *What artificial neural networks can tell us about human language acquisition*. arXiv. http://arxiv.org/abs/2208.07998

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.e23. https://doi.org/10.1016/j.cell.2020.10.024

Wilcox, E. G., Futrell, R., & Levy, R. (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*. https://doi.org/10.1162/ling_a_00491

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, USA*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
1–9

DOI: 10.1177/09637214241268098
www.psychologicalscience.org/CDPS
Sage

How Can Deep Neural Networks Inform Theory in Psychological Science?

Sam Whitman McGrath[1,2], Jacob Russin[2,3], Ellie Pavlick[3,4], and Roman Feiman[2,4]

[1]Department of Philosophy, Brown University; [2]Department of Cognitive and Psychological Sciences, Brown University; [3]Department of Computer Science, Brown University; and [4]Program in Linguistics, Brown University

Abstract

Over the last decade, deep neural networks (DNNs) have transformed the state of the art in artificial intelligence. In domains such as language production and reasoning, long considered uniquely human abilities, contemporary models have proven capable of strikingly human-like performance. However, in contrast to classical symbolic models, neural networks can be inscrutable even to their designers, making it unclear what significance, if any, they have for theories of human cognition. Two extreme reactions are common. Neural network enthusiasts argue that, because the inner workings of DNNs do not seem to resemble any of the traditional constructs of psychological or linguistic theory, their success renders these theories obsolete and motivates a radical paradigm shift. Neural network skeptics instead take this inability to interpret DNNs in psychological terms to mean that their success is irrelevant to psychological science. In this article, we review recent work that suggests that the internal mechanisms of DNNs can, in fact, be interpreted in the functional terms characteristic of psychological explanations. We argue that this undermines the shared assumption of both extremes and opens the door for DNNs to inform theories of cognition and its development.

Keywords

deep learning, neural networks, large language models, interpretability, psycholinguistics, cognitive development, philosophy of cognitive science

Contemporary deep neural networks (DNNs) can recognize and label images and videos, write poems, and plan travel itineraries—tasks that, just a few years ago, only humans could accomplish. Can understanding how artificial neural networks acquire and exercise such abilities help psychologists understand the corresponding abilities in humans?

Although modern DNNs are in their adolescence, this question is approaching its 40th anniversary. It first arose with the "connectionist" networks of the 1980s (McClelland et al., 1986; Smolensky, 1988). Even as contemporary models have progressed beyond these predecessors (Smolensky et al., 2022), they continue to work according to many of the same principles. At a fundamental level, all neural networks are made up of layers of interconnected units ("nodes" or "neurons"), each of which calculates a weighted combination of its inputs and applies a (usually nonlinear) function to the result, which it then passes to further nodes. The network's weights are initially random and their computations meaningless because the designer does not specify by hand what functions they should compute. Meaningful behavior emerges only as the network learns by gradually adjusting its matrices of weights so that its outputs more closely match specified targets (e.g., the numerical encoding of the next word in a text).

Because of their architectural similarities, contemporary attempts to link DNNs to cognitive theories face the same challenge that confronted earlier connectionists: Neural networks are difficult to interpret in higher level terms. Even once they have been trained to produce meaningful behavior, DNNs can seem like "black boxes"—amorphous webs of indistinguishable nodes, their internal activity an inscrutable cascade of matrix multiplications that bears no obvious relationship to

Corresponding Author:
Roman Feiman, Department of Cognitive and Psychological Sciences and Program in Linguistics, Brown University
Email: roman_feiman@brown.edu

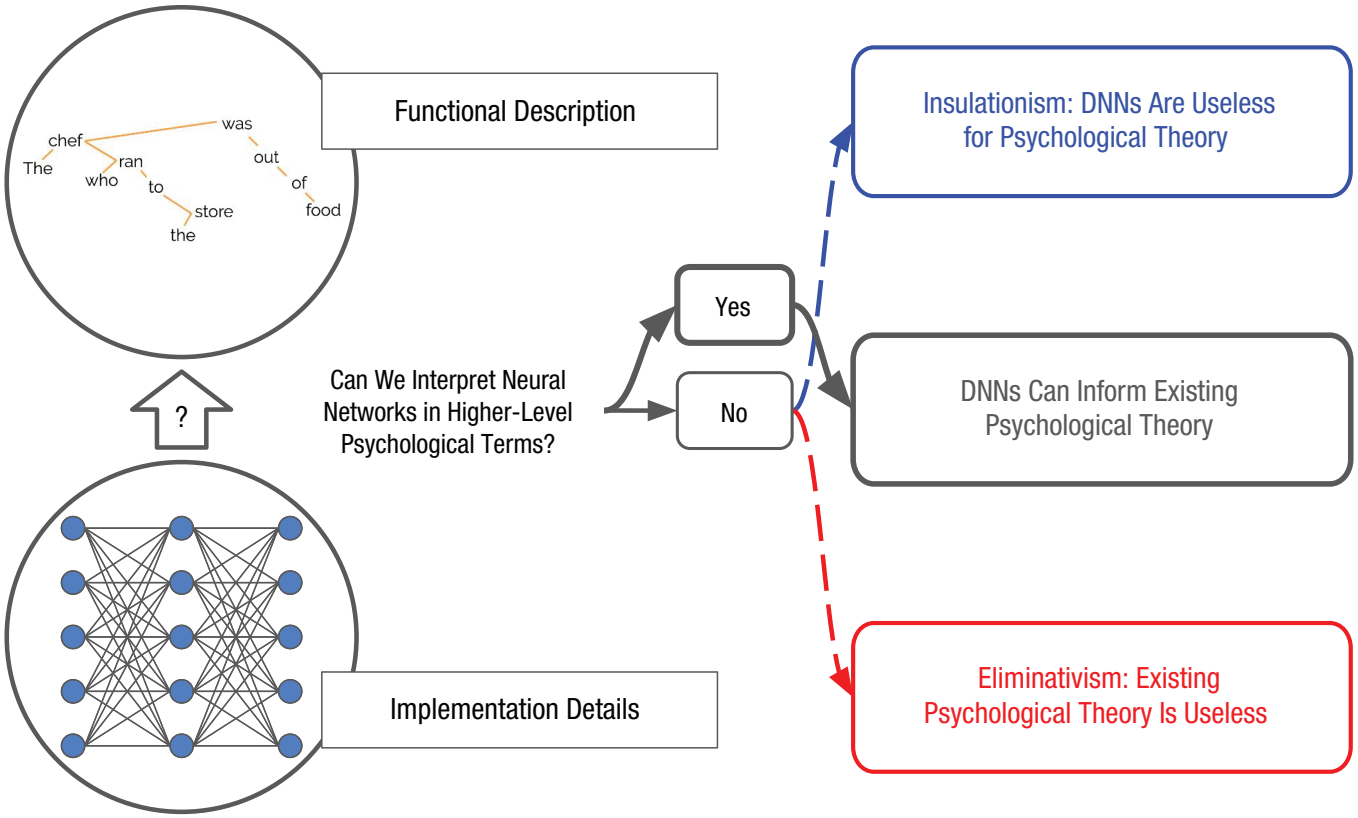

Fig. 1. Neural networks and psychological theory. The computations in a neural network model (bottom left) consist of matrix operations. Although these operations fully specify a given network, they are difficult to interpret in terms of the functional characterizations (e.g., tree structures; Manning et al., 2020) used in traditional psychological explanations (top left). The view that these networks are not interpretable has led to two extreme reactions: Eliminativists (red) conclude that higher level interpretations are unnecessary, whereas insulationists (blue) conclude that networks cannot be relevant to psychological science. However, recent work suggests that it is possible to interpret the inner workings of neural networks in succinct, functional terms, opening up the possibility of informing cognitive theory with DNNs (gray). DNN = deep neural network.

the theoretical constructs of psychology or linguistics. Where in a DNN is a *belief*, or a *concept*, or a *rule*?

Past debates about the ability of neural networks to inform psychological theory lost steam as older networks ran up against limitations in their ability to model key aspects of human behavior. What has come as a surprise to neural network enthusiasts and skeptics alike is that simply making the models bigger—both in architecture (e.g., the number of layers) and the amount of training data they are exposed to—has dramatically improved their capabilities (Brown et al., 2020). For the first time, DNNs can solve problems of the scale and complexity actually faced by humans rather than just the highly restricted tasks common in earlier work. Moreover, the fact that modern DNNs are already "pretrained" before being applied to a new task makes them a closer analogue to human participants, who come into any experimental context equipped with relevant experience. These improvements dovetail with recent insights into how to more fairly compare human and machine performance (Buckner, 2023; Firestone, 2020) and motivate a reexamination of whether and how DNNs can inform psychological theory.

Making Sense of the Deep Learning Revolution

With some notable exceptions (Smolensky, 1988), the debate about what neural networks mean for psychological science has largely been shaped by two extreme, opposing views (Fig. 1). One view is that the success of neural networks makes much of psychological and linguistic theory obsolete. This view holds that the rise of DNNs is like other cases of scientific elimination in psychology, in which once promising theoretical constructs, such as Freud's id or ego, were discarded in favor of more empirically successful ones. From this perspective, DNNs are taken to constitute theories (Piantadosi, 2023) that win out empirically over alternatives that appeal to symbolic concepts or rules. In the extreme

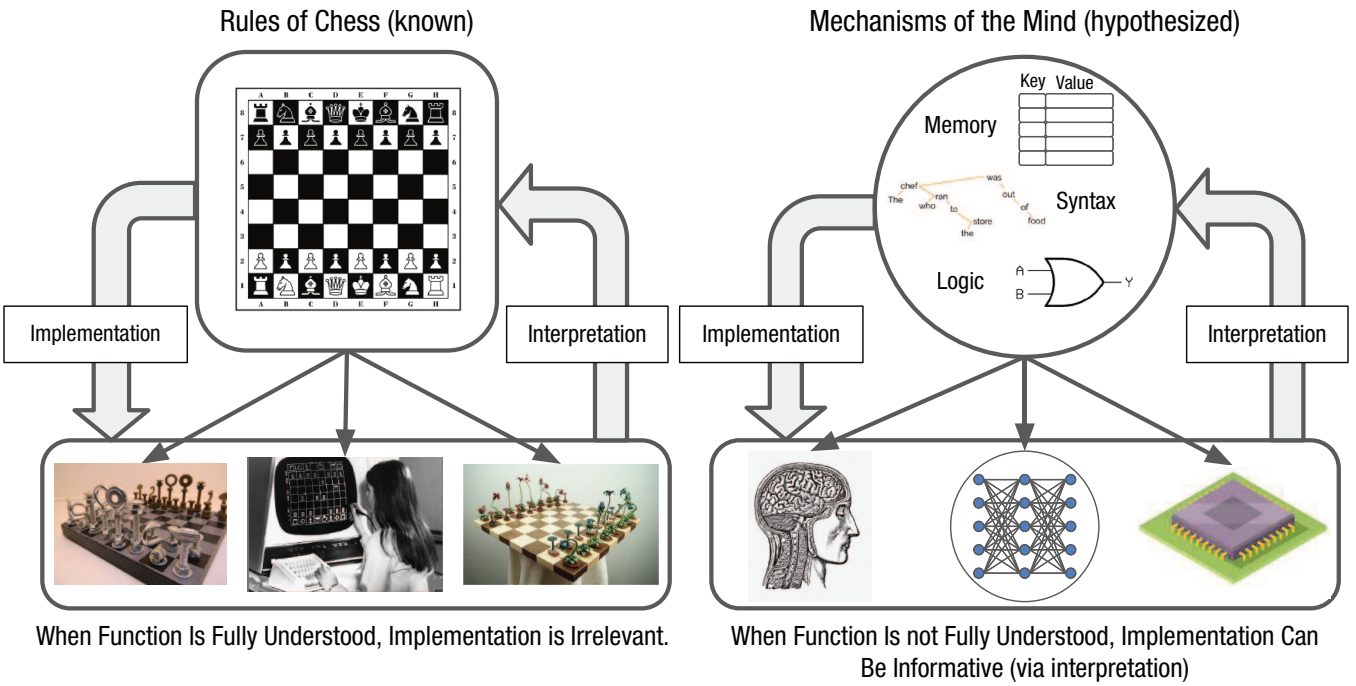

Fig. 2. Implementation and interpretation. To get a sense of the intuition behind the highly influential insulationist objection (Fodor & Pylyshyn, 1988)—and to see what is wrong with it—consider the game of chess (left). Chess can be implemented in multiple physical formats. The higher level functional rules that define the game cannot be reduced to lower level physical properties of any one implementation. For a theorist interested in the rules or strategy of chess, the implementation—whether it is played on a board or on a screen, what the pieces are made of, and so on—is entirely irrelevant. However, this is true of chess only because we know the rules in advance. If instead we had to figure out the rules by watching people play, then implementation-level details, such as all the pawns being the same shape and size, would offer informative, highly relevant indicators (furnishing evidence, in this case, that the pawns will all move in the same way). This is the situation we find ourselves in with respect to the mind (right). We start off not knowing how the mind works and need to reverse-engineer it via a complex process of abductive inference. Implementation-level facts form a key part of this inferential base and may well prove useful and informative in discerning the relevant higher level facts. Even if DNNs are taken to be implementation-level models, they can still feature in this abductive process and prove informative for cognitive science, provided that their representations and mechanisms are interpretable. DNN = deep neural network.

case, this *eliminativist* reaction threatens not just specific constructs but all theory not articulated in the lower level vocabulary of nodes and weights.

This radical reaction is driven by the combination of DNNs' impressive success at mimicking human behavior and the apparent impossibility of aligning their matrix multiplications with the higher level constructs of cognitive theory (e.g., beliefs, concepts, rules). According to this view, if DNNs do not afford interpretations in these terms, all the worse for cognitive theory—the relevant causal structure of behavior can, apparently, be captured in lower level terms anyway. The rise of DNNs is thus taken to signal a paradigm shift for psychology, overturning long-standing theoretical commitments and reshaping what counts as a satisfying explanation of cognition.

At the same time, skeptics of neural networks have relied on the same black-box assumption to argue for the opposite extreme: To the extent that DNNs cannot be put into any meaningful contact with existing higher level characterizations of cognition, they are simply irrelevant to psychological theory (Bowers et al., 2023). If higher level properties are not discernible in the weights and activations of trained networks, how can we even be sure that they offer incompatible cognitive theories, as the eliminativist claims? For all we know, DNNs' successes may be the result of merely implementing the psychological constructs of existing higher level theories (Fig. 2; Fodor & Pylyshyn, 1988; Quilty-Dunn et al., 2023). According to this view, no matter how impressive DNNs' behavior becomes, psychological science is effectively insulated from their influence. Even if the rise of deep learning represents a tectonic technological shift, the *insulationist* holds that it will barely register as a scientific tremor.

Challenging the Black-Box Assumption

Both extremes—eliminativism and insulationism—rest on the assumption that DNNs are black boxes that are not interpretable in the vocabulary of higher order theories of the mind. Recent empirical work in interpretability[1] challenges this assumption because researchers in artificial intelligence (AI) have

increasingly uncovered high-level, functional explanations of DNN behavior.

For example, the ability of large language models (LLMs) to generate long strings of grammatical text (Linzen & Baroni, 2021) spurred research on correspondences between aspects of LLMs' internal states and specific constructs in linguistic theory, such as parts of speech and syntax trees (Manning et al., 2020; Tenney et al., 2019). The overwhelming consensus from such work is that LLMs' internal states are neither unstructured nor inscrutable but exhibit geometric regularities that can be aligned with these linguistic constructs. For example, the way that sentence information is organized across the layers of these models reflects the traditional language-processing pipeline, with earlier layers representing parts of speech and parsing grammatical relations and later layers encoding semantic roles (who did what to whom) and tracking when different terms refer to the same entity (Tenney et al., 2019). It is even possible to reconstruct the syntactic parse trees postulated by traditional linguistic theory from models' internal representations of particular sentences (Manning et al., 2020).

Further work has suggested not only that these higher level descriptions are useful for interpreting model components but also that they play a causal role in the models' grammatical behavior. For instance, Chen et al. (2024) found that reconstructing syntactic parse trees first becomes possible during a specific window of training in which models show a sudden increase in the grammatical capacities that these same representations should enable. They also demonstrated a more direct causal link—targeted interventions designed to prevent the model from learning these high-level representations obstruct its acquisition of the corresponding grammatical capabilities. Other evidence suggests that higher level descriptions also play a causal role once a model is trained. Manipulating the nodes and weights that correspond to higher level interpretations elicits the expected changes in the models' behavior. For example, editing the weights that a model has used to produce "The Eiffel Tower is in Paris" to instead produce "The Eiffel Tower is in Rome" has related downstream effects, such as causing the edited model to give directions that refer to the "nearby" St. Peter's Basilica (Meng et al., 2022). Related techniques have been applied to logical inference (Geiger et al., 2021) and syntactic structure (Tucker et al., 2021). These findings suggest that higher order constructs identified by interpretability methods can be causally efficacious rather than merely epiphenomenal.

Interpretability research is uncovering not only the representations that DNNs use but also the algorithms that operate on these representations. For example, some of the most impressive, flexible reasoning behaviors of LLMs, such as their ability to learn new tasks given just a few examples (Brown et al., 2020), may be explained by specialized components called "induction heads" (Olsson et al., 2022) that emerge over the course of training. As the model predicts the next word that follows a word *w*, induction heads search the current context for any previous occurrence of *w* and copy the subsequent word. This mechanism appears to be relatively content-independent, performing the same function regardless of the identity of *w*, thus providing the model with computational machinery that supports the ability to "reason by induction," predicting that a pattern will be repeated (Olsson et al., 2022).

Such specialized mechanisms do not operate in isolation but appear to interface with one another to form task-specific circuits (Elhage et al., 2021; Olsson et al., 2022), for example, for processing the relative magnitudes of numbers (Hanna et al., 2023) or identifying syntactic constituents (Wang et al., 2022). Further evidence shows that some components are reused across circuits for different tasks (Merullo et al., 2023), suggesting partly content-independent operations that can be flexibly recombined.

Taken together, such findings open a path beyond the extreme views discussed in the previous section. The correspondence between LLMs' internal states and existing theory (e.g., theories of syntax, algorithms from traditional symbolic computing) challenges the eliminativist claim that the success of DNNs obviates higher level explanations. To the contrary, the fact that DNNs seem to reinvent structures similar to those postulated in existing theory (Manning et al., 2020; Tenney et al., 2019) means that existing explanations can help to understand how DNNs work and to causally intervene in their operation. At the same time, the fact that we can often interpret the black box of the neural network in higher level functional terms challenges the insulationist view that the success of DNNs is irrelevant to psychological theory. To the contrary, our growing understanding of the internal mechanisms of DNNs—currently the only models capable of reproducing a wide range of human-like behaviors in diverse settings—suggests that they may offer valuable insights into these same abilities in humans.

How Deep Neural Networks Can Inform Theory in Psychological Science

This recent progress suggests a broader strategy for leveraging interpretations of DNNs to inform theories of both the architecture of human cognition and its development. Where a DNN's performance matches human behavior, the key to evaluating whether it can

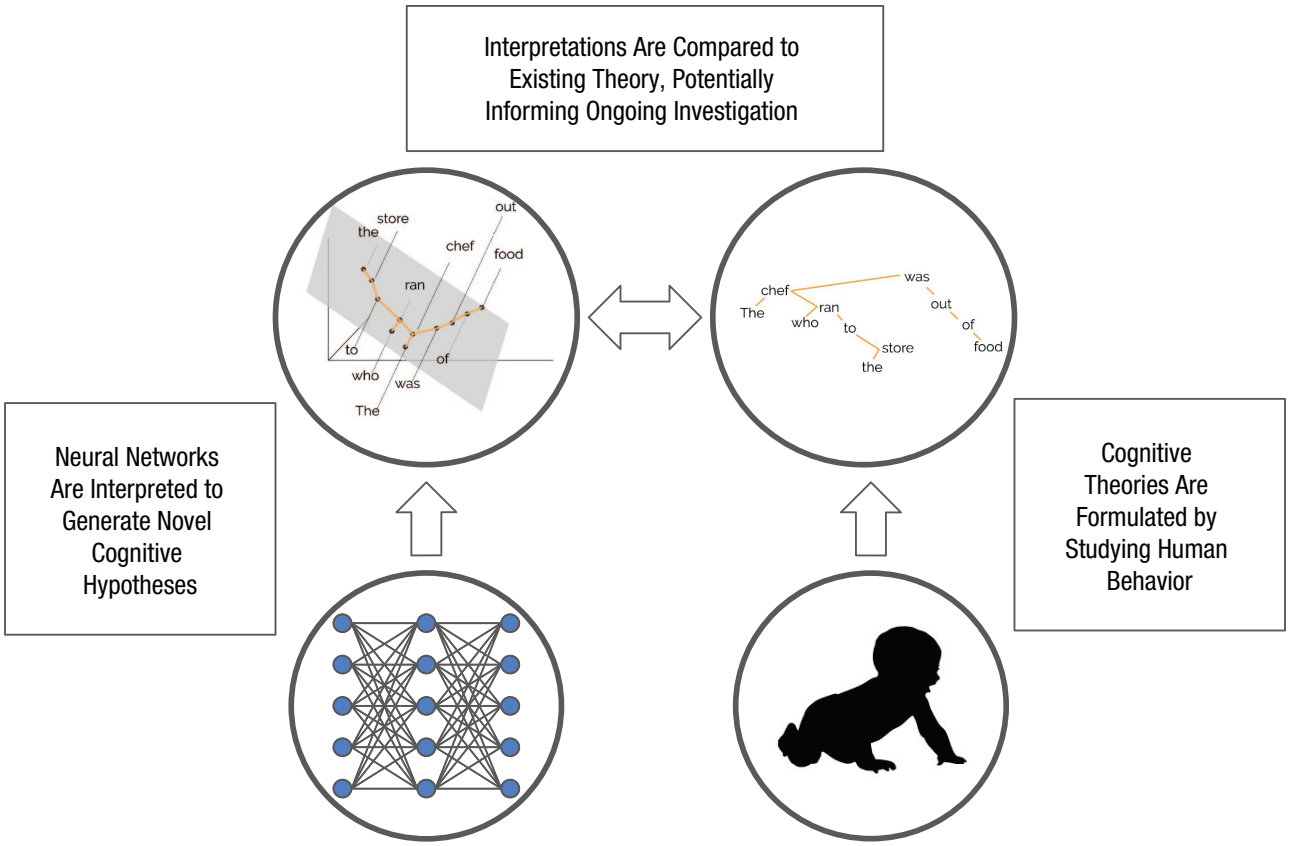

Fig. 3. Comparing DNN interpretations and cognitive theory. Recent advances in interpreting neural networks may allow DNNs to inform theory in cognitive science. When the low-level mechanisms of a trained neural network can be interpreted in the high-level functional terms familiar to cognitive psychology using new techniques (Manning et al., 2020), direct comparisons with human behavior and existing cognitive theory become possible. DNN = deep neural network.

inform psychological science will be to extract functional characterizations of its competencies to compare with existing cognitive theory (Fig. 3). If these functional characterizations precisely align with some existing theory, we may conclude that the model is a "mere implementation." In that case, finding that a DNN independently converged on a previously hypothesized cognitive solution would bolster claims of its optimality and generalizability (analogous to interpretations of convergent evolution in biology).

On the other hand, where we find divergences from existing theory, there are two possibilities. One is that the model has hit on an alien solution—that its internal representations and algorithms are fundamentally different from those used by humans. The other is that these divergent functional characterizations can serve as genuine alternative hypotheses about how human cognition works. Given that researchers are rarely sure that they have exactly the right cognitive theory in the first place (Fig. 2), taking this possibility seriously may motivate revision, although not necessarily wholesale replacement, of existing theory. Here, DNNs will neither eliminate nor merely implement existing psychological constructs. Instead, they can offer *informative implementations*, generating new insights and hypotheses about human representations for psychologists to evaluate. Work along these lines is ongoing, and there are as of yet no uncontroversial instances of contemporary DNNs modifying psychological theory. Below, we highlight significant examples and promising directions.

Deriving psychological insights from neural network models is not a new aspiration but one of the original goals of the connectionist movement. For example, influential early work by McClelland et al. (1986) aimed to use neural networks to model the psychological processes involved in converting verb stems (e.g., "write," "slight") into their past tense forms (e.g., "wrote," "slighted"). They focused on a feature of the English past tense that preexisting rule-based theories struggled to explain: quasiregularity (for a review, see Seidenberg & Plaut, 2014). Although most English verbs conform to a "regular" *add -ed* rule, striking subregularities exist among the "irregular" verbs as well (e.g., "write" → "wrote," "smite" → "smote"). Neural networks could

capture these quasiregular patterns with distributed representations that can simultaneously encode regularities at multiple levels of granularity. This implementation-driven insight was then incorporated back into higher order psychological theories, even by theorists who continued to posit a separate rule-based mechanism for regular verbs (Pinker & Prince, 1988). Today, DNNs have overcome many of the empirical challenges that Pinker and Prince raised, matching human performance better than any prior model (Kirov & Cotterell, 2018). This means they now have the potential to shed light on an unresolved question from the earlier debate: Is the past tense for regular and irregular verbs formed via a common mechanism? It may be tempting to assume that a successful DNN must reflect a single mechanism, but applying emerging interpretive techniques to these models could allow researchers to investigate both alternatives; it is possible that the new networks' success rests on circuits that form associations among irregulars separately from those that apply a discrete rule, such as adding *-ed*, to the regulars.

These new methods are already being used to interpret networks in other cognitive domains, such as vision, in which DNNs have transformed inquiry in cognitive psychology and neuroscience (e.g. Yamins et al., 2014). It is well established that DNNs trained on images learn representations that are predictive of human neural responses to visual stimuli (Schrimpf et al., 2020). Researchers have begun to interpret these networks to investigate cognitive hypotheses, such as the extent to which the human visual system can be understood as performing generative inference (Golan et al., 2020), or the extent to which human visual representations are shaped by top-down pressure from explicit category labels (Konkle & Alvarez, 2022). Neural network models have also been argued to explain the emergence of cognitive maps, one of the key psychological phenomena that motivated the cognitive revolution against behaviorism. When trained on spatial or relational tasks, these networks end up capturing the neural response patterns associated with place cells and grid cells, which have been hypothesized to underlie human spatial navigation and relational reasoning abilities (Whittington et al., 2020).

Compositionality, the ability to recombine familiar linguistic or cognitive components into novel combinations, has posed arguably the deepest challenge to neural networks as models of human linguistic and psychological capacities (Fodor & Pylyshyn, 1988). Even here, however, recent neural networks have made major breakthroughs, reproducing impressive human-like compositional generalization behaviors (Lake & Baroni, 2023). Although it would be premature to draw concrete conclusions about the implications of this work for theories of human cognition, interpreting networks capable of replicating the behavioral signatures of compositionality offers a novel approach to questions that have been central to cognitive science since its inception.

In addition to informing cognitive theories of adult competence, DNNs can also inform theories of cognitive development—a second avenue of influence on psychological science. Of course, it would be simplest to treat DNNs as models of human learners and then test how changes in model components or input data change learning trajectories and outcomes. As has been widely discussed, however, there are significant barriers to this approach. LLMs have been trained on orders of magnitude more language data than children could ever observe (Warstadt & Bowman, 2022). They learn primarily through next-word prediction, with feedback about each prediction's accuracy. And no current LLM interacts with the world like a human child—they cannot follow a parent's gaze or learn that light bulbs are hot by touching one. Creating DNNs that learn like humans along these dimensions is, at best, a long-term goal.

Critically, however, DNNs can still indirectly inform theories of cognitive development even if they do not learn like humans by serving as powerful tools in the study of learnability. Learnability is not the study of how humans actually learn but of what is in principle possible to learn under which conditions, by which kinds of learners. This can guide the study of actual learning by constraining plausible hypotheses, helping to identify what may or may not be special about human minds. Here, DNNs join the researcher's roster of model organisms, which range from nonhuman animals to classical computers.

For instance, because DNNs start with no content-specific knowledge, they can be used to evaluate "poverty of the stimulus" arguments, assessing whether particular knowledge is learnable by any agent both in principle and from the amount of data plausibly available to a child (Wilcox et al., 2022). In a recent example of this approach, Vong et al. (2024) showed that a generic DNN trained on visual and language data that was recorded from a single child's experience learned the right associations between words and images, suggesting that some aspects of linguistic meaning can be acquired without prior knowledge. A related approach is "controlled rearing," in which researchers manipulate a learner's input (often in ways that would be impossible or unethical to do with humans) and evaluate how this changes what is learned. Misra and Mahowald (2024) leveraged the possibility of precisely controlling DNN training sets to show that certain rare grammatical phenomena (article + adjective + numeral + noun constructions such as "a beautiful five days") are learnable even when all instances of these constructions are

deleted from the training corpus. This suggests that it is possible to generalize to these rare constructions from similar but less rare ones (e.g., "a few days"), undermining the need to posit an innate syntactic constraint. Extending this approach to more of what adults know—from theorized syntactic parameters to conceptual primitives—can further constrain hypotheses about what is innate in humans.

Beyond questions of what is learnable, DNNs can be used to test theoretical accounts of "learning dependencies." In language acquisition, for example, Gleitman (1990) argued that the reason toddlers learn nouns before verbs is that knowing nouns makes the meanings of accompanying verbs easier to identify, whereas early nouns are easy enough to learn in syntactic isolation. By experimentally manipulating the curriculum of training data (perhaps in extensions of paradigms such as the one used by Vong et al., 2024), DNNs could be used to evaluate such hypotheses. In cognitive development, Feiman et al. (2022) proposed that even a concept as basic as logical negation could be constructed out of precursors that are only implicitly logical, such as constraints on representations of physical objects (if an object is in one place, it cannot be in another). Traylor et al. (2023) subsequently tested this proposal with different models. They found evidence for implicitly logical inferences in a symbolic model of intuitive physics but not in a DNN trained to track and reidentify objects through occlusion. This suggests that although intuitive physics might be a sufficient base for implicit logic (evidence in favor of one learning dependency), object tracking alone might not be sufficient to learn intuitive physics (evidence against another dependency).

Finally, it is worth reiterating that DNNs can inform our understanding of learning dependencies even when they differ from human learners. Indeed, the less a DNN resembles human learners the more informative it would be if it recapitulated a similar developmental trajectory. That would suggest that the trajectories observed in humans are determined by the structure of their input rather than specific features of their cognitive architecture. Conversely, if DNNs do not need to learn information in the same order that humans do, that would indicate that human learners bring something importantly different to the task.

Conclusion

Recent advances in interpreting the internal processes of DNNs have identified functionally specialized internal components, emerging from cascades of matrix multiplications that once seemed inscrutable. The development of these methods of interpretation means that models capable of producing sophisticated human-like behaviors can be analyzed and may serve as informative implementations of the corresponding cognitive processes in humans, generating novel insights and providing new ways to evaluate theories of both mature cognition and its development. This points to new research directions beyond the extreme reactions that have shaped debates about the significance of neural networks and opens the door to a collaborative, mutually informative relationship between deep learning research and psychological science.

Recommended Reading

Fodor, J. A., & Pylyshyn, Z. W. (1988). (See References). One of the most influential critical discussions from the original connectionist debates; contests the relevance of neural networks to psychological science.

Linzen, T., & Baroni, M. (2021). (See References). Detailed review of the syntactic abilities of DNNs and discussion of their relation to theoretical linguistics.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). (See References). Develops and deploys methods for identifying linguistic structure in artificial neural networks.

Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., & Gao, J. (2022). (See References). Provides an accessible review of recent advances in DNNs, with a similar emphasis on the need to bridge the neural and high-level symbolic perspectives and on interpreting DNNs.

Traylor, A., Feiman, R., & Pavlick, E. (2023). (See References). Examines the logical reasoning capabilities of neural networks, focusing on implicit logical operators, a hypothesized developmental precursor to full-fledged logical concepts.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Author Contributions
S. W. McGrath and J. Russin contributed equally to this manuscript. All of the authors approved the final manuscript for submission.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Sam Whitman McGrath https://orcid.org/0009-0000-5383-1885

Note

1. The term "interpretability" has become commonplace in machine learning research (Elhage et al., 2021), but its exact definition varies slightly with the goals of different research programs (e.g., AI safety, trust, bias mitigation). Here, by "interpretation" we mean the characterization of the internal processes

of a trained neural network in the higher level terms typical of psychological explanation.

References

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, Article e385. https://doi.org/10.1017/S0140525X22002813

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165v4

Buckner, C. (2023). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *British Journal for the Philosophy of Science*, *74*(3), 681–712. https://doi.org/10.1086/714960

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. (2024). *Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs*. arXiv. https://doi.org/10.48550/arXiv.2309.07311

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., . . . Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Feiman, R., Mody, S., & Carey, S. (2022). The development of reasoning by exclusion in infancy. *Cognitive Psychology*, *135*, Article 101473. https://doi.org/10.1016/j.cogpsych.2022.101473

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences, USA*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. In M. Ranzato et al. (Eds.), *Advances in neural information processing systems 34 (NeurIPS 2021)* (pp. 9574–9586). Curran Associates, Inc.

Gleitman, L. (1990). The structural sources of verb meaning. *Language Acquisition*, *1*(1), 3–55.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences, USA*, *117*(47), 29330–29337. https://doi.org/10.1073/pnas.1912334117

Hanna, M., Liu, O., & Variengien, A. (2023). *How does GPT-2 compute greater-than? Interpreting mathematical abilities in a pre-trained language model*. arXiv. https://doi.org/10.48550/arXiv.2305.00586

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665. https://doi.org/10.1162/tacl_a_00247

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*, Article 491. https://doi.org/10.1038/s41467-022-28091-4

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*, 115–121. https://doi.org/10.1038/s41586-023-06668-3

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, USA*, *117*, 30046–30054. https://doi.org/10.1073/pnas.1907367117

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2: Psychological and biological models*. MIT Press.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In S. Koyejo et al. (Eds.), *Advances in neural information processing systems 35 (NeurIPS 22)* (pp. 17359–17372). Curran Associates, Inc.

Merullo, J., Eickhoff, C., & Pavlick, E. (2023). *Circuit component reuse across tasks in transformer language models*. arXiv. https://doi.org/10.48550/arXiv.2310.08744

Misra, K., & Mahowald, K. (2024). *Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs*. arXiv. https://doi.org/10.48550/arXiv.2403.19827

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., . . . Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

Piantadosi, S. (2023). *Modern language models refute Chomsky's approach to language*. LingBuzz. https://lingbuzz.net/lingbuzz/007180

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, Article e261. https://doi.org/10.1017/S0140525X22002849

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). *Brain-score: Which artificial neural network for object recognition is most brain-like?* bioRxiv. https://doi.org/10.1101/407007

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, *38*(6), 1190–1228. https://doi.org/10.1111/cogs.12147

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23. https://doi.org/10.1017/S0140525X00052432

Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., & Gao, J. (2022). *Neurocompositional computing: From the central paradox of cognition to a new generation of AI systems*. arXiv. https://doi.org/10.48550/arXiv.2205.01128

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT rediscovers the classical NLP pipeline*. arXiv. http://arxiv.org/abs/1905.05950

Traylor, A., Feiman, R., & Pavlick, E. (2023, May 1–5). *Can neural networks learn implicit logic from physical reasoning?* [Conference session]. Eleventh International Conference on Learning Representations, Kigali, Rwanda. https://openreview.net/forum?id=HVoJCRLByVk

Tucker, M., Qian, P., & Levy, R. (2021). What if this modified that? Syntactic interventions with counterfactual embeddings. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 862–875). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.76

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). *Interpretability in the wild: A circuit for indirect object identification in GPT-2 small*. arXiv. https://doi.org/10.48550/arXiv.2211.00593

Warstadt, A., & Bowman, S. R. (2022). *What artificial neural networks can tell us about human language acquisition*. arXiv. http://arxiv.org/abs/2208.07998

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.e23. https://doi.org/10.1016/j.cell.2020.10.024

Wilcox, E. G., Futrell, R., & Levy, R. (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*. https://doi.org/10.1162/ling_a_00491

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, USA*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111