Received: 14 November 2023 | Revised: 22 April 2024 | Accepted: 24 April 2024

DOI: 10.1002/wcs.1684

PERSPECTIVE

WIREs COGNITIVE SCIENCE WILEY

Philosophy of cognitive science in the age of deep learning

Raphaël Millière

Department of Philosophy, Macquarie University, Sydney, Australia

Correspondence
Raphaël Millière, Department of Philosophy, Macquarie University, Sydney, Australia.
Email: raphael.milliere@mq.edu.eu

Edited by: Mohamad El Haj, Editor

Abstract

Deep learning has enabled major advances across most areas of artificial intelligence research. This remarkable progress extends beyond mere engineering achievements and holds significant relevance for the philosophy of cognitive science. Deep neural networks have made significant strides in overcoming the limitations of older connectionist models that once occupied the center stage of philosophical debates about cognition. This development is directly relevant to long-standing theoretical debates in the philosophy of cognitive science. Furthermore, ongoing methodological challenges related to the comparative evaluation of deep neural networks stand to benefit greatly from interdisciplinary collaboration with philosophy and cognitive science. The time is ripe for philosophers to explore foundational issues related to deep learning and cognition; this perspective paper surveys key areas where their contributions can be especially fruitful.

This article is categorized under:

Philosophy > Artificial Intelligence

Computer Science and Robotics > Machine Learning

KEYWORDS

artificial intelligence, connectionism, deep learning, neural networks, philosophy

1 | INTRODUCTION

Deep learning has enabled major breakthroughs in virtually every area of artificial intelligence over the past decade—including computer vision, game playing, robotics, speech recognition, and natural language processing. In most of these domains, deep neural networks (DNNs) have matched or exceeded human performance on long-standing challenges. For example, DNNs surpass humans on standard image classification benchmarks (He et al., 2016), beat world champions at chess and Go (Silver et al., 2016, 2017), achieve top scores on many tests including medical and law exams (OpenAI, 2023), and generate text often indistinguishable from human writing (Jones & Bergen, 2023; Schwitzgebel et al., 2024).

The history of artificial neural networks is deeply interwined with theoretical and empirical research exploring their adequacy as computational models of human cognition. For much of its history, this research program yielded only modest empirical results, with neural networks often comparing disfavorably to concurrent symbolic approaches. The recent achievements of DNNs on real-world challenges stand in stark contrast to the limited success of older neural network models. Yet they are often perceived as mere engineering feats, increasingly enabled by product-oriented research from technology companies rather than academia.

While the main focus of deep learning research may not be on understanding human cognition, it would be naive to discount the potential contribution of engineering advances to scientific research. Technical breakthroughs can open up new research directions that catalyze scientific progress, sometimes in unexpected ways. Academic researchers have been quick to leverage advances in deep learning to build better models in computational linguistics, psychology, and neuroscience. In return, deep learning researchers are increasingly borrowing from the interdisciplinary toolkit of cognitive science—if not to build models, at least to evaluate them.

This calls for a reappraisal of the place of modern artificial neural networks within the project of cognitive science. What is the relevance of the progress of deep learning for cognitive science? Conversely, what is the relevance of cognitive science to deep learning research? This paper will provide an opinionated perspective on these questions through the lens of the philosophy of cognitive science.[1] Section 2 provides a very brief outline of the history of neural network research leading to deep learning Section 3 examines whether and how recent developments in deep learning can inform research on human cognition. Finally, Section 4 discusses how insights from cognitive science and philosophy can help address ongoing methodological issues with the evaluation of DNNs and human–machine comparisons.

2 | THE COMING OF AGE OF CONNECTIONISM

Cognitive science has always been centrally informed by theoretical concepts from computer science (Boden, 2008; Miller, 2003). In its early days, the dominant research program sought to explain human cognition through computations over structured symbolic representations, analogous to rule-based programs executed by digital computers (Newell & Simon, 1976). Artificial neural networks (ANNs), also known as connectionist models, have come to play an increasingly important role in challenging and transforming this classical research program, with a few significant milestones leading to modern DNNs.

ANNs consist of simple neuron-like units connected into networks via weighted connections. The units are segregated into an input layer that receives data to be processed, an output layer to produce adequate responses, and one or more hidden layers in between that learn to represent features and perform computations that map inputs to outputs. McCulloch and Pitts laid the groundwork for connectionism by introducing a simplified mathematical model of neuron functioning, and proving that networks of such neuron-like units can in principle compute any logical function (McCulloch & Pitts, 1943). Building on this pioneering work, Rosenblatt's perceptron later demonstrated that an ANN with trainable connection weights could learn to classify patterns, lending key support to connectionist explanations of learning (Rosenblatt, 1958).

Despite the theoretical dominance of the symbolic approach, connectionism was revived in the 1980s with innovations such as distributed representations, multilayer neural networks with hidden layers allowing nonlinear decision boundaries, and the backpropagation algorithm for training weights across multiple layers (Rumelhart et al., 1986, 1987). Connectionist models showed promise in explaining psychological phenomena that had eluded traditional symbolic approaches, including aspects of learning, memory, categorization, language, reasoning, and vision (Cohen et al., 1990; Elman, 1990; Fukushima, 1980; Kruschke, 1992). These initial results challenged the assumption that the mind performs serial computations over discrete symbolic representations.

The advent of deep learning from the mid-2000s onward enabled much larger and more complex ANN architectures to be trained effectively (LeCun et al., 2015). This is due to a combination of factors—including better training techniques, the availability of much larger datasets assembled from internet data, and major increases in computational power. DNNs differ from older connectionist models in several significant ways that account for their superior performance on a broad range of challenging tasks (Buckner, 2019). Their depth allows them to hierarchically compose complex concepts from simpler features across many hidden layers with increasing levels of abstraction. They also make use of sophisticated architectures with heterogenous components to promote specific inductive biases. For example, convolutional neural networks used in computer vision have convolutional filters tuned to detect specific features in images regardless of their position, while transformers used in LLMs have self-attention layers that track specific dependencies between lexical items. Other tricks such as sparse activations and regularization techniques prevent overfitting despite the incredibly large number of trainable parameters in these networks.

Thanks to these innovations, DNNs are much better than previous neural networks architectures at efficiently learning useful abstract representations and inducing computations that can generalize beyond the specific distribution of their training data. In this respect, they can be seen as bringing the connectionist program to fruition, by demonstrating that neural networks that are no longer bounded by scarse computing resources can overcome, in practice, many of the putative theoretical limitations leveled against connectionism in the previous decades.

3 | DEEP LEARNING MODELS AS COGNITIVE MODELS

The maturation of the connectionist program in the guise of modern DNNs is eminently relevant to several long-standing debates in (the philosophy of) cognitive science.[2] The most influential criticism of connectionism as a theory of cognition came from proponents of the language of thought hypothesis, in the form of a dilemma: either connectionist models are fundamentally inadequate accounts of cognition, or they merely implement classical symbol manipulation (Fodor & McLaughlin, 1990; Fodor & Pylyshyn, 1988; Pinker & Prince, 1988). If connectionist models are viewed as genuine alternatives to classical architectures, classicists argued they fail to capture core structure-sensitive properties of cognition, like productivity and systematicity, because they lack genuinely compositional representations. But if they can be viewed as implementations of classical systems, then connectionists models are just showing how symbols and rules could be realized in neural hardware. Either way, classicists concluded that connectionist models come up short compared to symbolic architectures in explaining core cognitive capacities like language and reasoning. This dilemma has been revived in the age of deep learning: if DNNs can match human performance on a broad spectrum of perceptual and cognitive tasks, they must do so by implementing core features of language of thought architectures (Mandelbaum et al., 2022; Marcus, 2018; Quilty-Dunn et al., 2023).

Connectionists typically resist this dilemma in two ways. The first is to suggest, on the basis of experimental results, that cognition is not as regimented and systematic as classicists take it to be, and that connectionist models should not be held to a higher standard than humans themselves (Johnson, 2004). The second is to argue that connectionist models can in fact account for the structure-sensitive properties of cognition and the constituent structure of mental representations *without* merely implementing a classical architecture (Smolensky, 1988). These two strategies are not exclusive; together, they suggest that connectionist models and human cognition can meet halfway between unstructured input–output mapping and idealized systematicity.

In recent years, DNNs have moved much closer to bridging the gap with human performance on structure-sensitive tasks. In particular, an extensive body of work building on insights from cognitive science has probed their capacity for systematic compositional generalization (Donatelli & Koller, 2023). To rule out confounds such as memorization of common compositional structures, DNNs can be trained from scratch on synthetic datasets and evaluated on held-out test samples (e.g., Kim & Linzen, 2020; Lake & Baroni, 2018). In these datasets, the train-test split is carefully designed such that high accuracy on test samples requires systematically recombining previously learned elements to map new inputs made up from these elements to their correct output. This line of research has shown that various parameters can have a significant impact on compositional generalization in DNNs, including their architectural features and training regime (Csordás et al., 2021; Kazemnejad et al., 2023; Ontanon et al., 2022; Qiu et al., 2022). When these parameters are selected appropriately, DNNs can achieve good performance on compositional generalization datasets without built-in compositional rules.

For example, Lake and Baroni (2023) show that a standard transformer-based neural network trained with meta-learning can achieve human-like systematic generalization in a controlled few-shot learning experiment, as well as state-of-the-art performance on systematic generalization benchmarks. Their meta-learning approach consists in training the network on a stream of artificial tasks, each based on an underlying "interpretation grammar" that specifies compositional mappings from instructions to output sequences. At test time, the model achieves human-like accuracy and error patterns, without the need for explicit compositional rules. While meta-learning from different tasks helps promote compositional generalization, recent research using a standard learning regime has also shown that simply training a network past the point where it achieves excellent accuracy on the training data can lead it to acquire more tree-structured computations, and generalize significantly better to held-out test data that require learning hierarchical rules (Murty et al., 2023).

In line with the second horn of the classicist dilemma, one might interpret these results as providing evidence that, given the right architecture and training regime, modern DNNs can account for the structure-sensitive properties of cognition by implementing a language of thought (Quilty-Dunn et al., 2023). However, this conclusion hinges on controversial assumptions about what the relevant notion of implementation ought to be, and what kinds of properties should be taken as specific evidence of implementing a language of thought (McGrath et al., 2023; Pavlick, 2023;

Smolensky, 1989). For example, one core feature of language of thought architectures is the ability to perform variable binding over discrete symbolic representations, where "roles" (variables) and "fillers" (values) are represented independently. Mechanistic interpretability research that seeks to reverse-engineer computations in trained DNNs does suggest they can acquire a mechanism for variable binding (Baroni, 2022; Davies et al., 2023; Elhage et al., 2021; Millière & Buckner, 2024b). However, this mechanism implements a "fuzzy" form of variable binding making use of vector subspaces that are not always functionally equivalent to discrete memory slots (Olsson et al., 2022); accordingly, role-filler independence in these networks is not absolute, but comes in degrees. This suggests that while modern DNNs can compute over compositional representations with real constituent structure, this structure is *non-classical* and should not be taken to reflect the core properties of a language of thought on pain of trivializing them. For the language of thought hypothesis to remain substantive, it must commit to specific claims about how representations are composed beyond pointing to their constituent structure. However, committing to such claims in light of the available evidence about modern DNNs may undermine the second horn of the dilemma—the view that their successful behavior is best explained by positing that they merely implement a language of thought architecture.

A related issue concerns the content-specificity of computations performed by DNNs. Neural networks are traditionally assumed to learn many specific input–output mappings. On this view, each layer-to-layer transformation deals with a given input in a way that depends on the particular content of that input, rather than general computational principles applied across inputs. In other words, ANNs are generally taken to perform only *content-specific computations*, by contrast with classical architectures. However, there is compelling evidence that modern DNNs can also perform *non-content-specific computations*, as argued by Shea (2023). For example, episodic deep reinforcement learning models apply non-content-specific similarity computations to stored memory representations. When a new state is encountered, the system retrieves all previously stored memories and calculates their similarity to the current state using the same similarity algorithm, regardless of what the specific contents of those memories are. Transformer-based LLMs also induce a broad repertoire of non-content-specific computations, including domain-general "induction head" mechanisms that implement the aforementioned capacity for variable binding (Olsson et al., 2022).

One of the notable feature of DNNs, in contrast with classical systems, is that their architecture does not delineate a strict distinction between *content-specific* and *non-content-specific* computations. Rather, as I previously alluded to when discussing role-filler independence, the content-specificity of DNN computations is a matter of degrees. This is significant if we take DNNs seriously as cognitive models that provide genuine alternatives to classical architectures, rather than mere implementations. For example, it dovetails with recent findings about the behavioral convergence between DNNs and humans on various classical reasoning tasks. Indeed, LLMs show similar accuracy overall to humans on various reasoning problems, including natural language inference, syllogism validity, or the Wason selection task. Moreover, both humans and LLMs exhibit "content effects" on reasoning tasks; that is, they tend to perform more accurately when the content of a reasoning problem is familiar and plausible (Dasgupta et al., 2023). The best LLMs also match human performance across a range of verbal and non-verbal analogy tasks requiring inductive reasoning about abstract relations, such as Raven's progressive matrices or letter string analogies, among other abstract reasoning tasks (Geiger et al., 2023; Han et al., 2023; Mirchandani et al., 2023; Webb et al., 2023).

These empirical results highlight two significant points about the relevance of deep learning to cognitive science. First, modern DNNs have fulfilled the promise of older connectionist models in matching human performance on many tasks probing core aspects of cognition. Second, emerging evidence from interventional studies suggests that DNNs achieve human-like performance on these tasks through mechanisms that differ in nontrivial ways from those postulated by classical architectures (Millière & Buckner, 2024b). Whether these mechanisms are similar to those of human cognition remains an open question that should be explored by experimentalists and philosophers of cognitive science in tandem. At the very least, these findings suggest that the classicist alternative to connectionism is no longer the "only game in town" (Fodor, 1975)—if it ever was.

The progress of DNNs has important implications for many other ongoing issues in philosophy and cognitive science; I will briefly highlight two that have attracted a lot of attention recently. The first pertains to the "grounding problem" originally coined by Harnad (1990): How can symbol-manipulating systems have representations that are intrinsically connected to the worldly referents of the symbols they manipulate? Without securing such connection, it seems that computational models of cognition would have difficulty escaping the "merry-go-round" of symbols and connecting to the real world. While the grounding problem originally targeted classical symbolic systems, it applies mutatis mutandis to neural networks used in natural language processing, such as LLMs, that manipulate linguistic tokens. However, Mollo and Millière (2023) argue in light of philosophical theories of representation that LLMs can in fact acquire world-involving functions that secure norms of representational correctness relative to the referents of linguistic items.

Another hotly debated issue concerns the relevance of LLMs to theoretical linguistics and theories of language acquisition. A wealth of evidence from targeted experiments in computational linguistics suggests that LLMs acquire sophisticated syntactic knowledge (Linzen & Baroni, 2021; Pavlick, 2022). This knowledge is reflected in the overall convergence of their predictions with human grammaticality judgments regarding minimal pairs of sentences that differ only with respect to some target linguistic property (Warstadt et al., 2020). LLMs' representations of syntactic features can also be linearly decoded from the activations of the network and manipulated with predictable effects on behavior (Belinkov, 2022; Hao & Linzen, 2023; Ravfogel et al., 2021). These results call for an examination of the potential for deep learning to inform linguistic theory (Baroni, 2022; Dupre, 2021; Linzen, 2019). In particular, it has been argued that LLMs challenge core tenets of generative linguistics, on which statistical approaches to language modeling relying on linear string order cannot account for the hierarchical structure dependence of syntactic competence (Chomsky, 1957; Contreras Kallens et al., 2023; Everaert et al., 2015; Piantadosi, 2023). Most LLMs are exposed to a vastly greater quantity of words than children during their learning phase (Frank, 2023a), which typically limits their relevance to debates about linguistic nativism. Nonetheless, ongoing efforts to train LLMs in developmentally plausible learning scenarios may vindicate their usefulness as model learners that can constrain theories of language acquisition (Millière, forthcoming; Warstadt & Bowman, 2022).

4 | METHODOLOGICAL ISSUES

With the performance of DNNs improving across a range of linguistic and cognitive tasks, the need for robust methods to evaluate DNNs and compare them with humans under similar conditions is becoming more pressing. Methodological insights from the philosophy of cognitive science can inform evaluation practices in deep learning.

Behavioral evaluations based on benchmarks face challenges that increasingly limit their usefulness in the age of LLMs. New benchmarks tend to saturate very rapidly, although the best-performing models may still exhibit failures modes in the target domain (Kiela et al., 2021; Ott et al., 2022). The perverse incentive of "SOTA-chasing"—pursuing *state-of-the-art* status on benchmark leaderboards—can lead to the exploitation of proxy metrics that diverge from the underlying evaluation targets, in accordance with Goodhart's law (Gururangan et al., 2018; Manheim & Garrabrant, 2018). Because LLMs are trained on internet-scale data, benchmark contamination is also a common issue: test samples can easily leak into the training data, leading to misleading improvements on standard evaluation metrics (Zhou et al., 2023). Finally, the connection between latent theoretical constructs and operational variables measured through benchmarks is not always explicit or well supported.

These challenges can be addressed through hypothesis-driven experiments that incorporate best practices inspired by cognitive science. For example, researchers can use novel stimuli to avoid data contamination; use minimal stimuli (such as minimal pairs of sentences) to avoid confounds; control the training data (by analogy with controlled rearing experiments, e.g. Lee et al. (2021)); use multiple tasks to test the same capacity; collect multiple model responses for each test item; and use appropriate control conditions (Frank, 2023b). Consider for example the question whether LLMs are capable of acquiring a Theory of Mind (ToM). Using a classic false-belief task, Kosinski (2023) suggests that GPT-3 exhibits ToM reasoning comparable to 9-year-old children. However, performance success on a single task does not provide robust evidence of the underlying competence. Indeed, Ullman (2023) shows that minor conceptual task variations, which maintain the core demands for false belief inference, reveal the model's lack of abstract reasoning about mental states. Patterns of performance can only constrain inferences about competence given well-supported background assumptions about the measuring instrument, measuring conditions, and target system.

The distinction between performance and competence cuts both ways (Firestone, 2020). When comparing performance across humans and DNNs, it is crucial to create adequately matched testing conditions. For example, Lakretz et al. (2022) suggest that transformer models like GPT-2 are fundamentally limited in their capacity to process long-range recursive nesting compared to humans. In their experiment, however, human subjects received substantial training with examples, instructions, and feedback—while GPT-2 was tested "zero-shot" without equivalent context. Lampinen (2023) shows that after adding context analogous to human training, LLMs actually perform *better* than humans even on the most challenging conditions. These examples highlight how careful one should be in interpreting both success and failure modes of DNNs on tasks originally designed for humans. Methodological principles from comparative and developmental psychology can help mitigate comparative biases in experimental design and analysis (Buckner, 2021).

5 | CONCLUSION

The progress of deep learning over the past decade has been more significantly driven by engineering achievements than by theoretical insights from cognitive science. This certainly does not mean that it is irrelevant to cognitive science; nor does it mean that cognitive science has nothing to contribute to deep learning research in return. Modern DNNs do not merely mark incremental improvements over older neural networks models, but represent a turning point for the connectionist program. While they still fall short of human cognitive competence in various ways, and show noteworthy dissimilarities with human biases and developmental trajectories, they also demonstrate an unprecedented convergence with human performance on many long-standing challenges—many of which were once widely thought to be hard limitations of connectionist architectures. More than ever, neural networks show promise as *cognitive models* that can be systematically studied and manipulated by scientists in carefully controlled conditions to enable surrogative reasoning about core aspects of cognition. More than ever, the need for rigorous evaluations of neural networks requires interdisciplinary insights, particularly when it comes to theoretically-informed comparisons with humans on linguistic and cognitive tasks. Philosophers of cognitive science have much to contribute to both theoretical and methodological issues raised by deep learning.

AUTHOR CONTRIBUTIONS

Raphaël Millière: Conceptualization (lead); writing – original draft (lead); writing – review and editing (lead).

ACKNOWLEDGMENT

Open access publishing facilitated by Macquarie University, as part of the Wiley - Macquarie University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Raphaël Millière https://orcid.org/0000-0001-6965-6073

RELATED WIREs ARTICLES

Connectionist perspectives on language learning, representation and processing

ENDNOTES

[1] While I will consider the progress of deep learning as a whole in relation to cognitive science, I will pay particular attention to large language models (LLMs), whose capabilities have invited more specific comparisons with human cognition (see Millière and Buckner (2024a) for discussion).

[2] While I will mainly focus on psychology here, DNNs also occupy an increasingly important place in cognitive neuroscience (e.g., Doerig et al., 2023; Lindsay, 2024; Richards et al., 2019).

REFERENCES

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language*. CRC Press.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, *48*(1), 207–219.

Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Clarendon Press.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), e12625.

Buckner, C. (2021). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *The British Journal for the Philosophy of Science*, *74*(3), 681–712.

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(3), 332–361.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256.

Csordás, R., Irie, K., & Schmidhuber, J. (2021). The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 619–634). Association for Computational Linguistics.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2023). *Language models show human-like content effects on reasoning tasks*. arXiv:2207.07051.

Davies, X., Nadeau, M., Prakash, N., Shaham, T. R., & Bau, D. (2023). *Discovering variable binding circuitry with desiderata*. arXiv:2307.03637.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450.

Donatelli, L., & Koller, A. (2023). Compositionality in computational linguistics. *Annual Review of Linguistics*, *9*(1), 463–481.

Dupre, G. (2021). (What) can deep learning contribute to theoretical linguistics? *Minds and Machines*, *31*(4), 617–635.

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.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Everaert, M. B. H., Huybregts, M. A. C., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571.

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

Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183–204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3–71.

Frank, M. C. (2023a). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990–992.

Frank, M. C. (2023b). *Large language models as models of human cognition*.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2023). Relational reasoning and generalization using nonsymbolic neural networks. *Psychological Review*, *130*(2), 308–333.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). *Annotation artifacts in natural language inference data*. arXiv:1803.02324.

Han, S. J., Ransom, K., Perfors, A., & Kemp, C. (2023). *Inductive reasoning in humans and large language models*. arXiv:2306.06548.

Hao, S., & Linzen, T. (2023). *Verb conjugation in transformers is determined by linear encodings of subject number*. arXiv:2310.15151.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE.

Johnson, K. (2004). On the systematicity of language and thought. *The Journal of Philosophy*, *101*(3), 111–139.

Jones, C., & Bergen, B. (2023). *Does GPT-4 pass the turing test?* arXiv:2310.20216.

Kazemnejad, A., Padhi, I., Ramamurthy, K. N., Das, P., & Reddy, S. (2023). *The impact of positional encoding on length generalization in transformers*. arXiv:2305.19466.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., & Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. arXiv: 2104.14337.

Kim, N., & Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9087–9105). Association for Computational Linguistics.

Kosinski, M. (2023). *Theory of mind might have spontaneously emerged in large language models*. arXiv:2302.02083.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2873–2882). PMLR.

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*, 1–7.

Lakretz, Y., Desbordes, T., Hupkes, D., & Dehaene, S. (2022). Can transformers process recursive nested constructions, like humans? In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3226–3232). Republic of Korea. International Committee on Computational Linguistics.

Lampinen, A. K. (2023). *Can language models handle recursively nested grammatical structures? A case study on comparing models and humans*. arXiv:2210.15303.

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

Lee, D., Gujarathi, P., & Wood, J. N. (2021). *Controlled-rearing studies of newborn chicks and deep neural networks*. arXiv:2112.06106.

Lindsay, G. W. (2024). Grounding neuroscience in behavioral changes using artificial neural networks. *Current Opinion in Neurobiology*, *84*, 102816.

Linzen, T. (2019). What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, *95*(1), e99–e108.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212.

Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E. J., Harris, D., Kibbe, M. M., Kurdi, B., Mylopoulos, M., Shepherd, J., Wellwood, A., Porot, N., & Quilty-Dunn, J. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, *46*(12), e13225.

Manheim, D., & Garrabrant, S. (2018). *Categorizing variants of Goodhart's law*.

Marcus, G. (2018). *Deep learning: A critical appraisal*. arXiv:1801.00631 [cs, stat].

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

McGrath, S. W., Russin, J., Pavlick, E., & Feiman, R. (2023). Properties of LoTs: The footprints or the bear itself? *Behavioral and Brain Sciences*, *46*, e284.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144.

Millière, R. (forthcoming). Language models as models of language. In R. Nefdt, G. Dupre, & K. Stanton (Eds.), *The Oxford handbook of the philosophy of linguistics*. Oxford University Press.

Millière, R., & Buckner, C. (2024a). *A philosophical introduction to language models—Part I: Continuity with classic debates*. arXiv: 2401.03910.

Millière, R., & Buckner, C. (2024b). *A philosophical introduction to language models—Part II: The way forward*. arXiv:2405.03207.

Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., & Zeng, A. (2023). *Large language models as general pattern machines*. arXiv:2307.04721.

Mollo, D. C., & Millière, R. (2023). *The vector grounding problem*. arXiv:2304.01481.

Murty, S., Sharma, P., Andreas, J., & Manning, C. D. (2023). *Grokking of hierarchical structure in vanilla transformers*. arXiv:2305.18741.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126.

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.

Ontanon, S., Ainslie, J., Fisher, Z., & Cvicek, V. (2022). Making transformers solve compositional tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3591–3607). Association for Computational Linguistics.

OpenAI. (2023). *GPT-4 technical report*. arXiv:2303.08774.

Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., & Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, *13*(1), 6793.

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*(1), 447–471.

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220041.

Piantadosi, S. (2023). *Modern language models refute Chomsky's approach to language*.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193.

Qiu, L., Shaw, P., Pasupat, P., Nowak, P., Linzen, T., Sha, F., & Toutanova, K. (2022). Improving compositional generalization with latent structure and data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4341–4362). Association for Computational Linguistics.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, 1–21.

Ravfogel, S., Prasad, G., Linzen, T., & Goldberg, Y. (2021). Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 194–209). Association for Computational Linguistics.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Rumelhart, D. E., Mcclelland, J. L., & Group, P. R. (1987). Parallel distributed processing. In *Explorations in the microstructure of cognition: Foundations* (Vol. 1). MIT Press.

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a large language model of a philosopher. *Mind & Language*, *39*(2), 237–259.

Shea, N. (2023). Moving beyond content-specific computation in artificial neural networks. *Mind & Language*, *38*(1), 156–177.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). *Mastering chess and shogi by self-play with a general reinforcement learning algorithm*. arXiv: 1712.01815.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23.

Smolensky, P. (1989). Connectionism and constituent structure. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, & L. Steels (Eds.), *Connectionism in perspective*. Elsevier.

Ullman, T. (2023). *Large language models fail on trivial alterations to theory-of-mind tasks*. arXiv:2302.08399.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*. CRC Press.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*, 1–16.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., & Han, J. (2023). *Don't make your LLM an evaluation benchmark cheater*. arXiv:2311.01964.

How to cite this article: Millière, R. (2024). Philosophy of cognitive science in the age of deep learning. *WIREs Cognitive Science*, *15*(5), e1684. https://doi.org/10.1002/wcs.1684