Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Review

Reconciling symbols and neural networks: evidence for a quasi-symbolic view

Thomas McCoy[1,2], Abigail L Tenenbaum[1], Zhang Enyan[3] and Zhenghao Herbert Zhou[1]

Many domains of cognition, such as language and reasoning, have traditionally been modeled with symbolic systems. However, in both biological and artificial intelligence, some of the most advanced systems in these domains are neural networks, despite the fact that neural networks appear fundamentally different from symbolic systems. What, then, is the relationship between neural computation and symbolic behavior? In this paper, we review recent literature that analyzes the internal mechanisms used by neural networks from AI. We conclude that this line of work supports a quasi-symbolic view of cognition, in which the mind uses computation that is approximately symbolic but with some important non-symbolic properties that derive from the mind's neural underpinnings.

Addresses
[1] Department of Linguistics, Yale University, 37 Hillhouse Avenue, New Haven, CT 06511, USA
[2] Wu Tsai Institute, Yale University, 100 College Street, New Haven, CT 06511, USA
[3] Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06511, USA

Corresponding author: McCoy, Thomas (tom.mccoy@yale.edu),
McCoy, Thomas (𝕏 @RTomMcCoy),
Enyan, Zhang (𝕏 @EnyanZhang),
Zhou, ZhenghaoHerbert (𝕏 @zh_herbert_zhou)

Current Opinion in Behavioral Sciences 2026, **72**:101686

This review comes from a themed issue on **AI and Psychology**

Edited by

Available online xxxx

Received: 20 May 2026; Revised: 14 August 2026;
Accepted: 30 August 2026

https://doi.org/10.1016/j.cobeha.2026.101686

Introduction

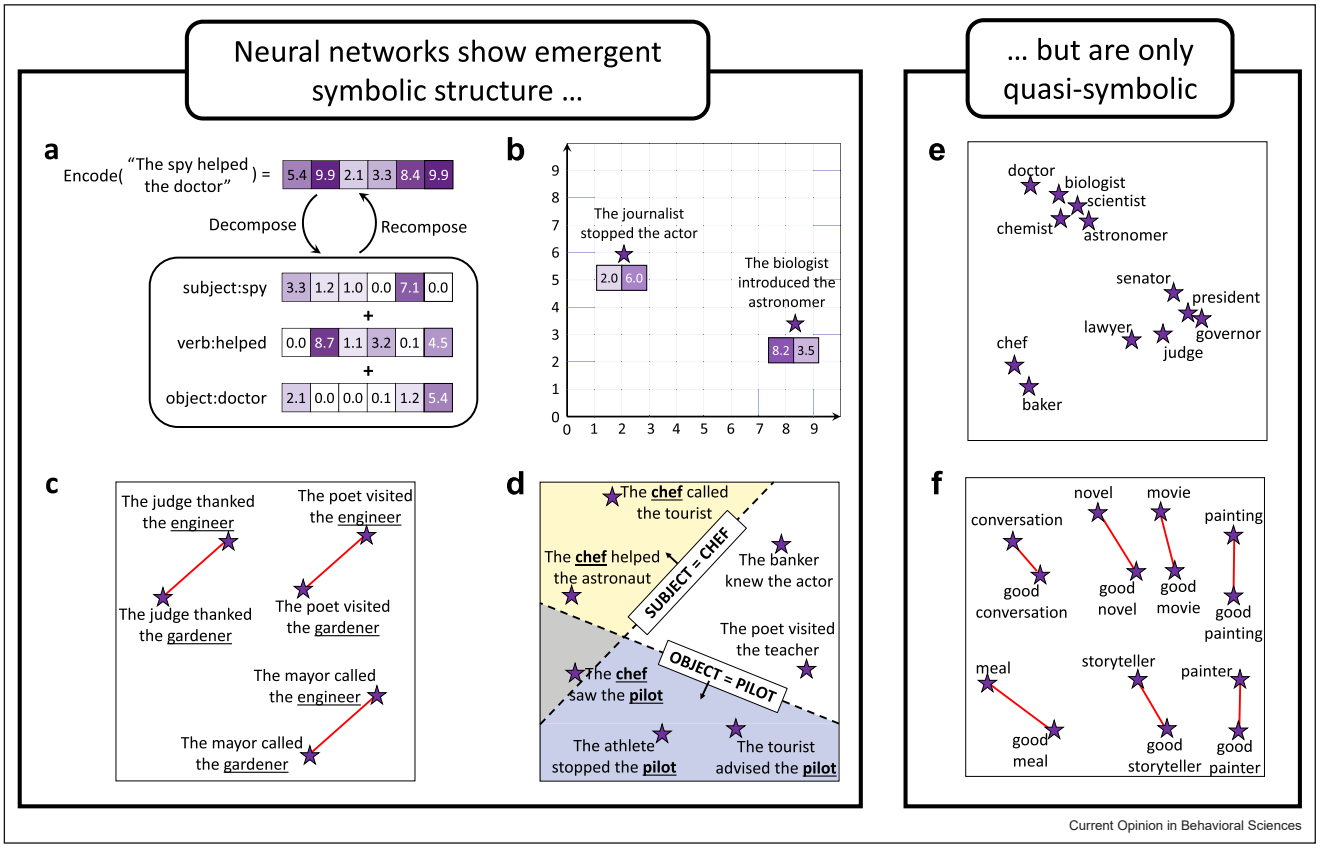
A central task for cognitive science is determining what type of computational system the mind is. One long-standing proposal is that the mind is a symbolic system [1], in which discrete units are combined in structured ways (e.g. combining logical primitives to form propositions), potentially augmented with probabilities [2]. An alternative proposal is to model the mind with artificial neural networks, in which information is represented in continuous vectors [3]. Though they seem qualitatively different from symbolic systems, neural networks are the state-of-the-art in artificial intelligence (AI) for many domains long thought to require symbolic representations, such as language [4], board games [5], and math problems [6].

What should cognitive science conclude from the recent advances in neural networks used for AI? The seemingly straightforward conclusion is that these systems refute symbolic conceptions of the mind. However, there is an important alternative possibility: Perhaps state-of-the-art AI models achieve their successes by internally implementing symbolic computation. It has long been known that symbolic systems can be implemented inside neural networks [7–9]. Indeed, proponents of symbolic cognitive theories have argued that the mind must be leveraging such a 'mere implementation' of symbolic systems inside the biological neural network that is the brain [10,11]. Although current AI models have not been explicitly designed to implement symbols, they might nonetheless do so implicitly, in which case their successes would support rather than refute symbolic cognitive theories.

Of course, there is no guarantee that the mind operates in ways similar to current AI systems. Nonetheless, for many domains, artificial neural networks provide strikingly strong fits to human behavior [12] and human brain data [13]. Thus, understanding what internal mechanisms are being used by these systems could form the basis for a promising theory of human cognition [14]. A major challenge for this goal is that neural networks are famously difficult to understand, but the field has made substantial progress on this problem in recent years.

In this paper, we review the available evidence about how artificial neural networks perform seemingly symbolic tasks, and we argue that this evidence supports a *quasi-symbolic* view of cognition (Figure 1). Under the quasi-symbolic view, which was proposed by Smolensky in 1988 [15], neural networks that perform structured tasks use processing that is approximately but not fully symbolic. The processing is partially symbolic because it operates

Figure 1



Quasi-symbolic representations in neural networks. The specific examples in this figure are hypothetical, but all except for *f* are based on well-recognized properties of neural networks. **(a–d)** Symbolic structure in network representations. ***a.*** Neural networks encode information, such as sentences, in continuous vectors. Such vectors can often be decomposed into sums of vectors encoding basic representational units. ***b.*** Vectors can also be thought of as points in space; for example, the vector [2.0, 6.0] corresponds to the point with coordinates (2.0, 6.0) on a 2D plane. This geometric perspective provides other windows into symbolic structure in neural network vector encodings: ***c.*** Systematic relationships between structures are often realized via systematic offsets in vector space. ***d.*** Representations can often be classified along symbolic dimensions (e.g. identifying what a sentence's subject is) via linear partitions. **(e–f)** Ways in which neural networks might deviate from pure symbolic structure. ***e.*** Distance in vector space can capture gradations of semantic similarity. ***f.*** Representations in vector space can handle domains that are only partially systematic. For instance, the relationship between *NOUN* and *good NOUN* is broadly similar across nouns but is not uniform; for example, a good novel is one that is interesting, while a good meal is one that is tasty. This situation can hypothetically be reflected via offsets in vector space that vary somewhat depending on the noun.

over representations made of structured compositions of symbols. However, it is not fully symbolic because the symbols are implemented in a continuous way that produces deviations from the behavior expected in a 'mere implementation' of a symbolic system. We argue that the quasi-symbolic view can explain how neural networks (whether artificial or biological) perform well in symbolic domains while still deviating in important ways from strictly symbol-system-like behavior [16,17].

Marr's levels and neural network interpretability

Our argument can be situated in Marr's proposal that information-processing systems, such as the mind, can be understood at three levels [18]. The computational level characterizes the abstract function that the system performs; the algorithmic level characterizes the algorithms and representations that the system uses to perform that function; and the implementational level characterizes how the algorithm is realized. We now consider how this framework applies to neural networks that display seemingly symbolic behavior. Griffiths et al. [19] note that all known learning procedures that produce such symbolically behaving networks involve training them on data from symbolic domains. Since the training data characterize the function that the system is optimized to perform, these networks can be viewed as symbolic at Marr's computational level. Meanwhile, if we take the network's explicit architecture as its implementational level, it is clear that the implementational level is non-symbolic,

since standard neural architectures operate over continuous vectors rather than symbols. But what should we make of the intermediate algorithmic level?

Unfortunately, it is notoriously difficult to understand what sorts of algorithms and representations drive a given neural network's behavior. This issue has inspired a discipline called *interpretability* or *mechanistic interpretability*, which aims to understand the inner workings of neural networks. In the following sections, we discuss evidence from interpretability about whether neural networks implicitly make use of symbolic algorithms and representations.

Symbolic representations in neural networks

We consider a representational system to be symbolic if it uses a finite vocabulary of discrete units (symbols) that form representations by being combined in structured ways. For instance, Arabic numerals (e.g. *28*) are symbolic representations; the basic symbols are the digits 0 through 9, which are combined by placing the ones digit in the rightmost position, the tens digit in the second-rightmost position, etc.

Neural networks encode information in continuous vectors, which seem qualitatively different from symbolic structures. However, several papers have found evidence that, when neural networks are trained on symbolic tasks, they implicitly implement symbolic representations in vector space. These papers are based on the Tensor Product Representation (TPR) formalism, which provides a way to realize symbolic representations inside fixed-size vectors [20]. When neural networks are trained to perform symbolic tasks (e.g. reversing a list), their internal representations can be closely approximated by TPRs, at least in the cases that have been investigated [21]; similar results have been found for larger-scale models trained on natural language [22]. Further, TPR-based analyses enable a network's vector representations to be edited in targeted ways to change the model's behavior [23]. For instance, in a model trained to translate a command into an action sequence (e.g. `jump twice` → `JUMP JUMP`), the vector that encodes `jump twice` can be edited by removing the vector component that encodes `twice` and inserting the vector component predicted to encode `thrice`; when the edited vector is fed to the network, its output changes from `JUMP JUMP` to `JUMP JUMP JUMP`.

Many other techniques exist for analyzing neural network representations; these techniques are not usually framed around symbols, but they often point to symbol-like units in network representations. First, symbol-like features and concepts — for example, parts of speech — can often be decoded straightforwardly (i.e. linearly) from neural network representations using simple decoding networks called probes [24–26]. Second, network encodings often support meaningful vector arithmetic [27,28]; for example, in a system trained to encode words, the vector for *king* minus the vector for *man* plus the vector for *woman* is approximately equal to the vector for *queen* [29]; such results are evidence that the vectors are structured in ways that capture meaningful concepts (though see Ref. [30] for complications in interpreting these results). Third, work with sparse autoencoders [31] shows that network representations can be decomposed into sums of distinct features, which often capture discrete conceptual meanings such as the Golden Gate Bridge [32]. In fact, [33] demonstrated that all three of these methods — linear decoding, additive analogies, and sparse autoencoders — can be derived from the TPR formalism; thus, the results discussed in this paragraph and the previous one provide converging evidence that neural networks naturally develop symbolic, TPR-like representations when they are trained on structured domains.

Symbolic algorithms in neural networks

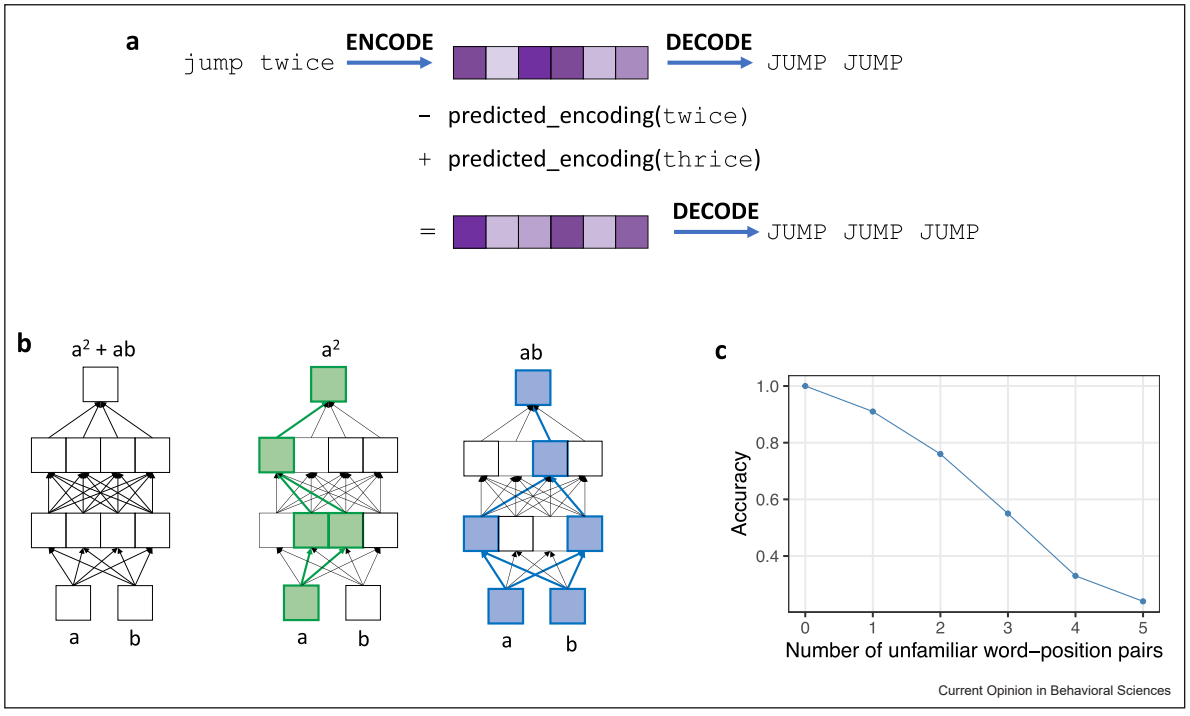
We consider two (non-mutually-exclusive) definitions of what it means for an algorithm to be symbolic:

(1) Symbolic algorithm (causal definition): An algorithm is symbolic if its input-output behavior is causally driven by compositional symbolic structure in its representations.
(2) Symbolic algorithm (structural definition): An algorithm is symbolic if it is a structured composition of distinct operations over symbolic representations.

If symbolic representations have been identified in a system, one can check whether the system satisfies definition (1) by performing causal interventions, in which one edits the symbolic structure of a representation and checks if the system's output changes accordingly. The previous section already gave one example of a causal intervention, in which a network's internal representations were modified in ways that changed the output of `JUMP JUMP` to `JUMP JUMP JUMP`. Many other papers have also demonstrated successful symbolic causal interventions [35–38].

There is also evidence that such networks satisfy the structural definition in (2). At least in many cases, networks can be decomposed into subnetworks (also called circuits) that perform meaningful sub-components of the network's overall function [39–41]. For instance, a network trained to take in two numbers a and b and return $a^2 + ab$ was found to incorporate internal subnetworks computing a^2 and ab [34] (note that all operations mentioned here are in modulus 7). There has also been work mapping entire neural networks to the complete structure of symbolic algorithms [42], and work that 'decompiles'

Figure 2



Quasi-symbolic algorithms in neural networks. ***a.*** As evidence that there is causal symbolic structure in neural network processing, the compositional structure of neural network representations can be causally implicated in network processing via targeted edits to representations [23] ***b***. Neural networks performing structured tasks have been shown to develop subnetworks computing sub-components of the broader task, such as having subnetworks for a^2 mod 7 and ab mod 7 inside a network computing $a^2 + ab$ mod 7 [34] ***c***. Neural network algorithms differ from purely symbolic algorithms in that they display graceful degradation as inputs stray farther from the network's training domain [16].

neural networks by reading out a symbolic program from the network's weights [43,44]. Thus, whether symbolic algorithms are defined in causal terms or structural terms, there is evidence that neural networks that perform structured tasks do so (at least in many cases) by recruiting symbolic algorithms (Figure 2).

Symbolic learning in neural networks

In addition to representations and algorithms, learning provides a third area where neural networks and symbolic systems seem qualitatively different. Models based around symbolic representations can often learn from small numbers of examples, whereas neural network learning typically requires far more training data. However, recent work has shown that a technique called meta-learning can give symbolic inductive biases to neural networks [45,46], where inductive biases are the factors that guide how a learner generalizes. The resulting systems can learn symbolic rules from small numbers of examples, paralleling the learning behavior of systems based on explicit symbolic representations.

What enables meta-learned neural networks to learn in ways that traditionally required symbolic processing? [47] has found evidence that these networks impose symbolic structure on their hypothesis space and that this structure is connected to the networks' rapid learning ability (Figure 3, left). Thus, it appears that rapid, symbol-like learning in neural networks relies on implicit symbolic structure.

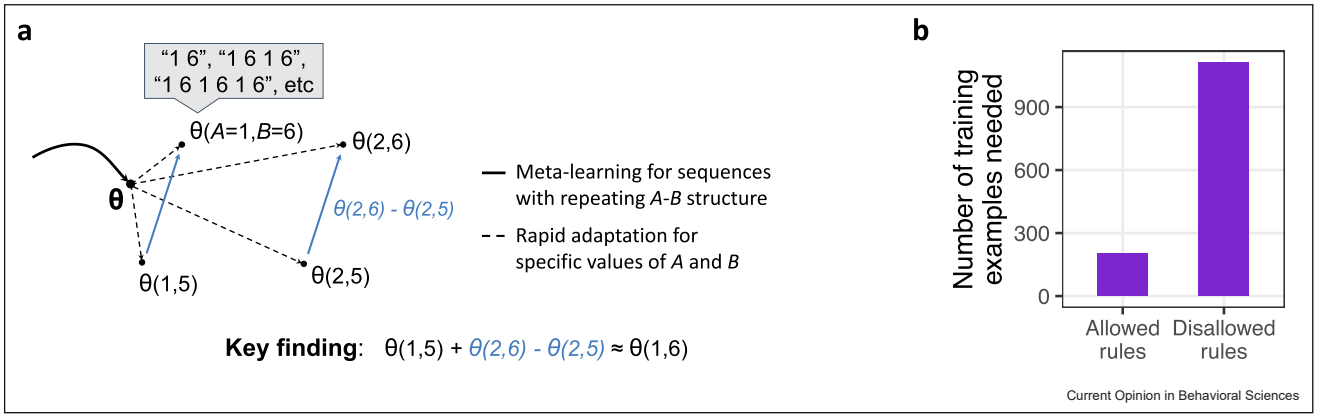
Evidence that neural networks are quasi-symbolic rather than symbolic

We have argued that, when neural networks display symbol-system-like behavior, they make use of internal processing that is also symbol-system-like. In this section, we discuss why the networks in question should be viewed as only quasi-symbolic rather than fully symbolic.

First, symbolic approximations of neural networks are usually imperfect. For instance, causal interventions often have accuracy that is high but not 100%. These imperfections suggest that symbolic idealizations of neural networks capture much of how a network operates but still miss some aspects.

A second reason to view these networks as only quasi-symbolic is behavioral: they typically do not fully behave as they would be expected to if they were precisely symbolic. For example, neural networks often struggle

Figure 3



Quasi-symbolic learning in neural networks. ***a.*** The technique of meta-learning is used to find an initial parameter setting θ for a neural network; from this initial setting, it can rapidly learn any function within a domain of functions. For instance, consider networks that can rapidly learn sequence patterns governed by a repeating *A B A B* structure; for example, one function in this domain would have $A = 2$ and $B = 5$, allowing sequences such as 2 5, 2 5 2 5, 2 5 2 5 2 5, etc.; this function is captured by starting at θ and then rapidly learning the parameter setting $\theta(2, 5)$. [47] found that these meta-learned networks induce systematic structure in the space of network parameters. ***b.*** When a neural network meta-learns over a symbolic domain, hard constraints from the symbolic domain turn into soft constraints in the neural network. For instance, in a meta-learned network that could learn symbolic linguistic rules, certain rules that were impossible in the symbolic definition of the rule space instead became merely difficult to learn, but not impossible, in a meta-learned network [48].

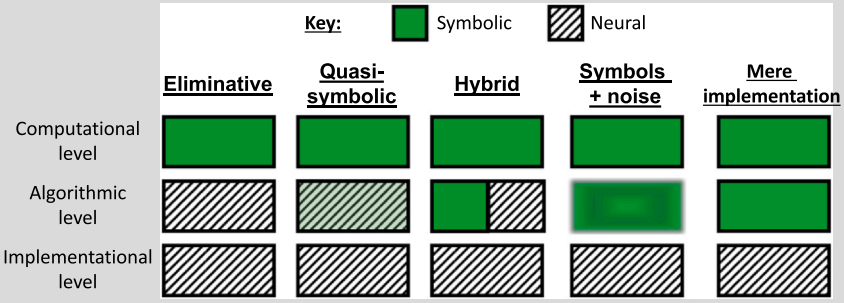
to generalize to novel types of examples, such as examples involving symbols used in novel positions [49–51]; these types of generalization should follow naturally from algorithms implemented in fully symbolic ways. Some of these behavioral trends are also observed in humans [52].

If we accept that symbolically-behaving neural networks are quasi-symbolic, what are the consequences of this fact? Below is a list of properties that would be likely to arise in a quasi-symbolic system but not a symbolic one.

1. *Capturing the gradient nature of similarity*. Neural networks represent units of information (e.g. words) in continuous vectors. Such vectors allow these elements to have a graded degree of similarity [16] by operationalizing similarity as closeness in vector space; for example, a *dog* might be more similar to a *wolf* than to a *cat*. In contrast, if these units were encoded with atomic symbols, gradient similarity would not follow in a straightforward way because of the all-or-nothing character of symbols. Gradient similarity might be useful for enabling the generalization of knowledge across similar entities. Effects of graded similarity also arise in human cognition, such as in interference effects — that is, more similar elements cause greater interference than less similar elements [53].
2. *Graceful degradation*. When a symbolic system encounters an input that is outside the domain in which it is defined, it fails completely on that input. In neural networks, however, performance instead tends to gradually decline as inputs become more unusual [3]. Thus, we would expect a quasi-symbolic approach to similarly display gradual decreases in performance as difficulty increases rather than sharp dropoffs.
3. *Softness of constraints*. Symbolic systems often involve strict constraints; for example, in symbolic grammars, certain types of sentences are strictly outside the grammar. In contrast, quasi-symbolic systems likely inherit the property of neural networks that constraints are rarely absolute, meaning that restrictions that are hard constraints in a symbolic system might instead be realized in quasi-symbolic systems as soft constraints that can be violated if there is sufficient reason to do so. As an example, see Figure 3 (right).
4. *Flexibility in learning*. A symbolic learner is constrained to represent hypotheses in symbolic terms, which might make it challenging to smoothly move between hypotheses over the course of learning. In contrast, it may be that the continuous nature of quasi-symbolic systems affords more flexible learning than occurs in systems based on fully symbolic representations [46].
5. *Handling of partially-symbolic tasks*. Many naturalistic domains seem best described in ways that are only partially characterizable by symbolic rules; for example, natural language involves rich symbolic structure, yet human language processing is also highly sensitive to continuous usage statistics. By combining symbolic structure and gradience, quasi-symbolic approaches may be better equipped to handle such domains than fully symbolic approaches. As one example of how quasi-symbolic structure enables a unified treatment of continuous and discrete features, Ref. [54] found that neural network language models encode a continuous linguistic property

Box 1 Approaches for reconciling vectors and symbols.

The image below sketches out a variety of ways in which symbolic systems and neural networks might be reconciled. Under our interpretation of these proposals, they all agree that the mind involves symbols at Marr's computational level (though in reality the computational level is less universally agreed-upon than what we show here) and that it is neural at Marr's implementational level, but they disagree on the handling of the algorithmic level:



We argue for the quasi-symbolic view, in which the algorithmic level is simultaneously neural and (approximately) symbolic. The eliminative view is that the algorithmic level is purely neural, with no symbolic character; we argue against this view based on evidence that artificial neural networks naturally develop approximate internal symbolic computation. Conversely, the mere implementation view is that the algorithmic level is purely symbolic; we argue against this view due to the plausibly algorithmic-level aspects of cognition that seem to depend on gradience (e.g. effects of gradient similarity). One way in which a mere implementation could capture graded similarity is by finer-grained symbols, such as encoding each word with a vector of binary features (rather than as an atomic unit) and quantifying similarity by counting how many features are shared; however, at least for the case of word meanings, it has been empirically very challenging to break down concepts into discrete, primitive subsymbols [56].

The hybrid view models cognition with some neural components and some symbolic components, but as discrete aspects of the model (e.g. [57]). However, hybrid approaches seem less well-poised than quasi-symbolic ones to handle aspects of cognition that plausibly involve gradience *within symbolic computation, such as graceful degradation and softness of constraints. Finally, the symbols + noise view is that cognition might arise from representations and algorithms that are fully symbolic but with some random noise added. Random noise might be able to capture graceful degradation and softness of constraints, but it does not naturally account for gradient similarity, since a noise-focused account would predict that deviations from pure symbolic structure would be random rather than producing structured similarity relationships.*

(verb bias) with a strategy similar to the strategy found by previous papers to be used for encoding discrete features (such as whether a word is singular vs. plural). Quasi-symbolic structure can even provide useful explanatory power in areas of language that seem much less statistical and much more discrete, such as grammar [55].

Note that all of the above properties could, in principle, be realized by a fully symbolic system. Indeed, most neural networks used today are run on digital computers, which could be viewed as symbolic because they operate over binary zeroes and ones; therefore, anything that these networks do is, in some sense, being done by a symbolic system. Our claim that the above properties favor quasi-symbolic analyses over fully symbolic ones is not about what is possible but rather about what is straightforward. While the above effects could in principle be realized by inherently-discrete symbolic systems, we believe that they can be more straightforwardly explained by processing that is inherently continuous.

Defining the quasi-symbolic view and its alternatives

The evidence that we have presented in favor of the quasi-symbolic view is far from conclusive. The field does not by any means fully understand how advanced neural networks operate, and the fact that many of these networks use quasi-symbolic processing does not entail that all of them do. Further, many existing analyses of neural networks analyze only a simplified version of the network or a constrained space of stimuli, and insights from such partial analyses do not always capture the full complexity of the network's internal mechanisms [58,59]; as a result, these analyses might lead to an overestimation of the importance of simple symbolic mechanisms.

In the hope of facilitating future work that continues to test the quasi-symbolic view, here we define this view in more detail and discuss how it could be falsified. A quasi-symbolic system is one that possesses all three of the following traits:

i. Symbols (that are neurally realized) play an important, causal role in the system's representations and algorithms.
ii. The system's internal symbolic components deviate from their fully-discrete idealizations.
iii. The deviations mentioned in (ii) are responsible for important properties of the system.

There are several ways in which the quasi-symbolic view could be falsified for a given neural network. Each falsification would point to a different proposal about how symbols and neural networks would interact; these

alternative proposals are italicized in this section and discussed in more detail in Box 1. First, if symbols do not play an important, causal role in the network, condition (i) would be violated, supporting an *eliminative* view. Alternatively, if the system uses symbolic components, but with all such components being precisely symbolic, condition (ii) would be violated, resulting in a *mere implementation* view or a *hybrid* view. Finally, a system would violate (iii) if all of its deviations from purely-symbolic processing are judged to be unimportant (e.g. because they fall within a noise threshold), which would point to a *symbols + noise* view (Box 1).

Teasing apart these possibilities is non-trivial. It will require philosophical progress on some key questions: What does it mean for an aspect of a system to be 'important,' and how far can a symbol deviate from its precise idealization before it stops being a symbol? (We have assumed that some such deviation should be allowed, but even that assumption may not be uncontroversial). Future work should also investigate the extent to which quasi-symbolic structure is inevitable. It might be that a neural network trained on symbolic tasks is virtually guaranteed to be quasi-symbolic; perhaps symbolic task structure strongly encourages internal analogs of symbols to develop, and perhaps neural networks, due to their continuous implementation, will always have properties that a purely-symbolic system would not. If that line of reasoning holds, it would have important implications for the inferences that we could draw about human cognition based on the structure of the brain and the types of tasks it learns to perform.

Conclusion

Smolensky argued in 1988 that a quasi-symbolic view is the 'proper treatment of connectionism' [15]. We have presented evidence supporting this claim: When artificial neural networks carry out tasks in symbolic domains, they seem to do so by implicitly using quasi-symbolic representations and algorithms. Given the fact that such systems provide strong fits to human behavior and brain data, this conclusion further suggests that human cognition might have a similarly quasi-symbolic nature. Much further work remains to answer several critical questions: In what precise ways do cognitive algorithms deviate from purely symbolic processing? What are the behavioral consequences of these deviations? To what extent are these deviations a source of power vs. a source of errors? Answering these questions will help us understand the systematicity, the flexibility, and the limits of natural and artificial minds.

Data Availability

No data were used for the research described in the article.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

For helpful discussion, we are grateful to Tyler Brooke-Wilson, Robert Frank, Andrew Lampinen, Kanishka Misra, L.A. Paul (who suggested the term *quasi-symbolic*), Paul Smolensky, Paul Soulos, Kaya Stechly, and the members of the Computational Linguistics at Yale lab (CLAY). Any errors are our own.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Newell A: **Physical symbol systems**. *Cogn Sci* 1980, **4**:135-183.
2. Griffiths TL, Chater N, Tenenbaum JB: **Bayesian Models of Cognition: Reverse Engineering the Mind**. MIT Press; 2024.
3. McClelland JL, Rumelhart DE, Hinton GE: **The Appeal of Parallel Distributed Processing**. MIT Press; 1986:3-44.
4. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al: **Language Models are Few-Shot Learners**. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. *Advances in Neural Information Processing Systems*; 2020: vol. 33:1877-1901. Available from: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
5. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, *et al.*: **Mastering the game of Go with deep neural networks and tree search**. *Nature* 2016, **529**:484-489.
6. Glazer E, Erdil E, Besiroglu T, Chicharro D, Chen E and Gunning A, FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI, *arXiv preprint*, 2024, arXiv:2411.04872, doi:https://doi.org/10.48550/arXiv.2411.04872.
7. McCulloch WS, Pitts W: **A logical calculus of the ideas immanent in nervous activity**. *Bull Math Biophys* 1943, **5**:115-133.
8. Siegelmann HT, Sontag ED: **On the computational power of neural nets**. In: *Proceedings of the fifth annual workshop on Computational learning theory*; 1992.440-449.
9. • Strobl L, Merrill W, Weiss G, Chiang D, Angluin D: **What formal languages can transformers express? A survey**. *Trans Assoc Comput Linguist* 2024, **12**:543-561.
This paper provides a recent survey of work about what types of formal symbolic systems can be realized in neural networks.
10. Fodor JA, Pylyshyn ZW: **Connectionism and cognitive architecture: a critical analysis**. *Cognition* 1988, **28**:3-71.
11. Marcus GF: **Rethinking eliminative connectionism**. *Cogn Psychol* 1998, **37**:243-282.
12. Goodkind A, Bicknell K: **Predictive power of word surprisal for reading times is a linear function of language model quality**. In: *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*; 2018.10-18.
13. Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, *et al.*: **The neural architecture of language: Integrative modeling converges on predictive processing**. *Proc Natl Acad Sci USA* 2021, **118**:e2105646118.
14. •• McGrath SW, Russin J, Pavlick E, Feiman R: **How can deep neural networks inform theory in psychological science?** *Curr Dir Psychol Sci* 2024, **33**:325-333.
This paper discusses how understanding the internal operations of deep neural networks is critical for determining how these systems should inform debates in psychology.
15. Smolensky P: **On the proper treatment of connectionism**. *Behav Brain Sci* 1988, **11**:1-23.

16. Smolensky P, McCoy R, Fernandez R, Goldrick M, Gao J:
•• **Neurocompositional computing: from the central paradox of cognition to a new generation of AI systems**. *AI Mag* 2022, **43**:308-322.
This paper describes ways in which a synthesis of neural computation and symbolic computation creates systems that have advantages over either type of computation on its own.

17. Piantadosi ST, Muller DC, Rule JS, Kaushik K, Gorenstein M, Leib
•• ER, *et al*.: **Why concepts are (probably) vectors**. *Trends Cogn Sci* 2024, **28**:844-856.
Using concepts as an area of focus, this paper describes how a vector-based theory can capture both the types of properties traditionally captured with symbolic theories as well as more gradient properties that are not well-handled by symbolic theories.

18. Marr D: **Vision**. W.H. Freeman and Company; 1982.

19. Griffiths TL, Lake BM, McCoy RT, Pavlick E, Webb TW: **Whither**
•• **symbols in the era of advanced neural networks?** *Trends Cogn Sci* 2025,.
This paper surveys the ways in which recent neural networks have been able to capture properties previously believed to require symbolic systems, and it discusses the implications of this fact.

20. Smolensky P: **Tensor product variable binding and the representation of symbolic structures in connectionist systems**. *Artif Intell* 1990, **46**:159-216.

21. McCoy RT, Linzen T, Dunbar E, Smolensky P: **RNNs implicitly implement tensor-product representations**. In: *International Conference on Learning Representations*; 2019. Available from: https://openreview.net/forum?id=BJx0sjC5FX.

22. McCoy RT, Soulos P, Linzen T, Smolensky P, The emergent symbolic structure of artificial neural networks, *arXiv preprint*,2026, arXiv:2608.29530, doi:https://doi.org/10.48550/arXiv.2608.29530.

23. Soulos P, McCoy RT, Linzen T, Smolensky P: **Discovering the Compositional Structure of Vector Representations with Role Learning Networks**. Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics; 2020:238-254. Available from: https://aclanthology.org/2020.blackboxnlp-1.23.

24. Ettinger A, Elgohary A, Resnik P: **Probing for semantic evidence of composition by means of simple classification tasks**. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*; 2016:134-139.

25. Adi Y, Kermany E, Belinkov Y, Lavi O, Goldberg Y: **Fine-grained analysis of sentence embeddings using auxiliary prediction tasks**. In: *International Conference on Learning Representations*; 2017. Available from: https://openreview.net/pdf?id=BJh6Ztuxl.

26. Ivanova AA, Hewitt J, Zaslavsky N: **Probing artificial neural networks: Insights from neuroscience**. *ICLR 2021 Workshop: How Can Findings About The Brain Improve AI Systems*?; 2021.

27. Merullo J, Eickhoff C, Pavlick E: **Language models implement simple word2vec-style vector arithmetic**. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; 2024:5030-5047.

28. Todd E, Li M, Sen Sharma A, Mueller A, Wallace B, Bau D: **Function vectors in large language models**. In: *International Conference on Learning Representations*; 2024:17282-17333.

29. Mikolov T, Yih Wt, Zweig G: **Linguistic regularities in continuous space word representations**. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2013:746-751.

30. Linzen T: **Issues in evaluating semantic spaces using word analogies**. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*; 2016:13-18.

31. Cunningham H, Ewart A, Riggs L, Huben R, Sharkey L: **Sparse autoencoders find highly interpretable features in language models**. In: *The Twelfth International Conference on Learning Representations*; 2024. Available from: https://openreview.net/forum?id=F76bwRSLeK.

32. Templeton A, Conerly T, Marcus J, Lindsey J, Bricken T, Chen B, et al: **Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet**. *Transformer Circuits Thread*; 2024. Available from: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

33. Zhang E, McCoy RT: **A unifying perspective on language model representations: from filler-role structure to mechanistic interpretability, *arXiv preprint*,** 2026, arXiv:2608.29034, doi: https://doi.org/10.48550/arXiv.2608.29034.

34. Zhang E, Lepori MA and Pavlick E, Instilling inductive biases with subnetworks, *arXiv preprint*, 2023, arXiv:2310.10899, doi: https://doi.org/10.48550/arXiv.2310.10899.

35. Giulianelli M, Harding J, Mohnert F, Hupkes D, Zuidema W: **Under the hood: using diagnostic classifiers to investigate and improve how language models track agreement information**. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics; 2018:240-248. https://aclanthology.org/W18-5426 Available from:.

36. Ravfogel S, Prasad G, Linzen T, Goldberg Y: **Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction**. In: *Proceedings of the 25th Conference on Computational Natural Language Learning*; 2021:194-209.

37. Zhang F, Nanda N: **Towards best practices of activation patching in language models: metrics and methods**. In: *International Conference on Learning Representations*; 2024, vol. p 1651-1678.

38. Millière R, Buckner C: **Interventionist methods for interpreting deep neural networks**. Neurocognitive Foundations of Mind. Routledge; 2024:190-221.

39. Lepori M, Serre T, Pavlick E: **Break it down: evidence for structural compositionality in neural networks**. *Adv Neural Inf Process Syst* 2023, **36**:42623-42660.

40. Lindsey J, Gurnee W, Ameisen E, Chen B, Pearce A, Turner NL,
• *et al*.: **On the biology of a large language model**. *Transform Circuits Thread* 2025, Available from https://transformer-circuits.pub/2025/attribution-graphs/biology.html.
This paper provides extensive analyses of a large-scale neural network using a variety of state-of-the-art interpretability techniques, demonstrating ways in which interpretable symbolic algorithms can be extracted from such systems.

41. Yang Y, Campbell DI, Huang K, Wang M, Cohen JD, Webb TW:
• **Emergent Symbolic Mechanisms Support Abstract Reasoning in Large Language Models**. In: *Forty-second International Conference on Machine Learning*; 2025. Available from: https://openreview.net/forum?id=y1SnRPDWx4..
This paper gives a thorough account of how the architectural mechanisms of a standard type of neural network can support symbolic processing.

42. Geiger A, Wu Z, Potts C, Icard T, Goodman N: **Finding alignments between interpretable causal variables and distributed neural representations**. Causal Learning and Reasoning. PMLR; 2024:160-187.

43. Friedman D, Wettig A, Chen D: **Learning transformer programs**. *Adv Neural Inf Process Syst* 2023, **36**:49044-49067.

44. Huang X, Bakalova A, Bhattamishra S, Merrill W and Hahn M, Discovering Interpretable Algorithms by Decompiling Transformers to RASP, *arXiv preprint*, 2026, arXiv:2602.08857, doi: https://doi.org/10.48550/arXiv.2602.08857.

45. Lake BM, Baroni M: **Human-like systematic generalization through a meta-learning neural network**. *Nature* 2023, **623**:115-121.

46. McCoy RT, Griffiths TL: **Modeling rapid language learning by distilling Bayesian priors into artificial neural networks**. *Nat Commun* 2025, **16**:4676.

47. Tenenbaum A, McCoy RT: **Additive Analogies Reveal Compositional Structure in Neural Network Weights**. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*; 2025, vol. 47.

48. McCoy RT, Grant E, Smolensky P, Griffiths TL, Linzen T: **Universal linguistic inductive biases via meta-learning**. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*; 2020: 737-743.

49. Lake B, Baroni M: **Generalization without systematicity: on the compositional skills of sequence-to-sequence recurrent networks**. International Conference on Machine Learning. PMLR; 2018:2873-2882.

50. Kim N, Linzen T. **COGS: A compositional generalization challenge based on semantic interpretation**. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2020:9087-9105.

51. Lewis M, Mitchell M: **Evaluating the robustness of analogical reasoning in large language models**. *Trans Mach Learn Res* 2025,1-33 https://openreview.net/forum?id=t5cy5v9wph.

52. Lampinen AK, Dasgupta I, Chan SC, Sheahan HR, Creswell A, Kumaran D, *et al*.: **Language models, like humans, show content effects on reasoning tasks**. *PNAS Nexus* 2024, **3**:233.

53. Smith G, Franck J, Tabor W: **Encoding interference effects support self-organized sentence processing**. *Cogn Psychol* 2021, **124**:101356.

54. Zhou Z, McCoy RT, Frank R: **Causal interventions on continuous features in LLMs: a case study in verb bias**. In: *First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models*; 2025.

55. Smolensky P, Goldrick M: **Gradient symbolic representations in grammar: The case of French liaison**. *Rutgers Optim Arch* 2016, **1552**:1-37.

56. Fodor JA: **The present status of the innateness controversy**. Representations. Harvester Press; 1981:59-65.

57. Wong L, Collins KM, Ying L, Zhang CE, Weller A, Gerstenberg T, et al., Modeling open-world cognition as on-demand synthesis of probabilistic models, *arXiv preprint*, 2025, arXiv:2507.12547, doi: https://doi.org/10.48550/arXiv.2507.12547.

58. Friedman D, Lampinen AK, Dixon L, Chen D, Ghandeharioun A: **Interpretability illusions in the generalization of simplified models**. In: *Forty-first International Conference on Machine Learning*; 2024.Available from: https://openreview.net/forum?id=YJWlUMW6YP.

59. Rodriguez JD, Mueller A, Misra K: **Characterizing the role of similarity in the property inferences of language models**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for: Human Language Technologies (Volume 1: Long Papers)*. Edited by Chiruzzo L, Ritter A, Wang L. Association for Computational Linguistics; 2025:11515-11533. https://aclanthology.org/2025.naacl-long.574/ Available from.