Psychological Review

A Metatheory of Classical and Modern Connectionism

Olivia Guest and Andrea E. Martin

Online First Publication, October 16, 2025. https://dx.doi.org/10.1037/rev0000591

CITATION

Guest, O., & Martin, A. E. (2025). A metatheory of classical and modern connectionism. *Psychological Review*. Advance online publication. https://dx.doi.org/10.1037/rev0000591

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

© 2025 American Psychological Association
ISSN: 0033-295X

https://doi.org/10.1037/rev0000591

THEORETICAL NOTE

A Metatheory of Classical and Modern Connectionism

Olivia Guest[1, 2] and Andrea E. Martin[1, 3]
[1] Donders Institute for Brain, Cognition, and Behaviour, Radboud University
[2] Department of Cognitive Science and Artificial Intelligence, Radboud University
[3] Language and Computation in Neural Systems Group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Contemporary artificial intelligence models owe much of their success and discontents to connectionism, a framework in cognitive science that has been (and continues to be) highly influential. Herein, we analyze artificial neural networks: (a) when used as scientific instruments of study and (b) when functioning as emergent arbiters of the zeitgeist in the cognitive, computational, and neural sciences. Building on our previous work with respect to analogizing between artificial neural networks and cognition, brains, or behavior (Guest & Martin, 2023), we use metatheoretical analysis techniques (Guest, 2024), including formal logic, to characterize two distinct tendencies within connectionism that we dub classical and modern, with divergent properties, for example, goals, mechanisms, and scientific questions. We also demonstrate how we, as a field, often fail to follow important lines of argument to their end—this results in a paradoxical praxis. By engaging more deeply with (meta)theory surrounding artificial neural networks, our field can obviate the cycle of artificial intelligence winters and summers, which need not be inevitable.

Keywords: artificial neural network, cognitive neuroscience, computational modelling, connectionism, metatheortical calculus

> Connectionism was conceived almost three centuries ago (in the 1740s), but had a long gestation. In its embryonic form, it was merely biologized introspection: David Hartley's view that thinking is grounded in associative mechanisms in the brain. (Boden, 2006, p. 885)

According to Boden (2006), connectionism has been around in some form or another for almost three centuries. Even a more conservative estimate still places connectionism's beginnings in the 1940s, making it much older than the current technology industry-driven hype cycle in artificial intelligence (AI) research (Hamilton, 1998; also see Dhaliwal et al., 2024; Wilson, 2016). Connectionism has gone through boom and bust cycles, so-called summers and winters; thus statements such as "we attend today an explosive infatuation for this once old style but now new fashioned view of cognition" (Bersini, 1989, p. 472) wondrously are as applicable now as when they were written 36 years ago. In this article, we aim to critique and juxtapose modern connectionist stances with those around prior to 2010 when it can be argued the classical, predeep learning (cf. Dechter, 1986), and prewidespread graphics processing unit use, era ended (Schmidhuber, 2015; Sevilla et al., 2022; Thompson, 2021).

Importantly, contemporary AI models owe much of their success and discontents to connectionism. This historical and present friction between connectionism and the rest of the fields it touches on, or draws inspiration from, is worthy of examination. Thus, we present a nuanced critical perspective on artificial neural networks (ANNs) when used as scientific instruments of study (i.e., as computational models of the brain and behavior, e.g., as used in cognitive computational neuroscience; Guest & Martin, 2023). This use is in contrast to when ANNs are used as statistical and engineering methodologies (i.e., in nonscientific engineering-oriented AI uses, e.g., face recognition to unlock a smartphone). Notwithstanding, there are important overlaps between the technology sector, which is driven by profits and engineering, and

Peter D. Kvam served as action editor.

Olivia Guest https://orcid.org/0000-0002-1891-0972

Andrea E. Martin https://orcid.org/0000-0002-3395-7234

Andrea E. Martin is supported by a Lise Meitner Research Group "Language and Computation in Neural Systems" from the Max Planck Society and by the Netherlands Organization for Scientific Research, VIDI Grant 016.Vidi.188.029 and Aspasia Grant 015.014.013, and by an ERC Consolidator Grant from the European Research Council (DYNALANG; ERC-2024-COG-101170162). Because this research was funded in whole, or in part, by the Netherlands Organisation for Scientific Research, for the purpose of open access, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

The authors thank Nils Donselaar, Iris van Rooij, R. Mellema, Laura van de Braak, Rineke Verbrugge, Todd Wareham, and the Computational Cognitive Science group for discussions and feedback. The authors thank Leonidas A. A. Doumas and Iris van Rooij for discussion of the universal approximation claims.

Olivia Guest played a lead role in conceptualization, formal analysis, writing–original draft, and writing–review and editing. Andrea E. Martin played a supporting role in conceptualization, formal analysis, and writing–review and editing.

Correspondence concerning this article should be addressed to Olivia Guest, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands. Email: olivia.guest@donders.ru.nl

science, such as that funding flows from private industry to science and that many models' codebases and training sets are proprietary. This results in important undesirable contradictions and in conflicts of interest (e.g., Forbes & Guest, 2025; Gerdes, 2022; Guest et al., 2025; Liesenfeld & Dingemanse, 2024; Liesenfeld et al., 2023).

The analysis presented in this article is centered on the idea that, both critics and advocates, we as a field must follow important lines of argumentation to their logical conclusions. We examine the effects of the converse: when we take defensive rhetorical positions too far in discussing the scientific and engineering contributions of and purported capacities of ANNs. To do this, we propose a bisection of the connectionist tendency into roughly pre-2010, what we dub classical connectionism and abbreviate to ℭ, and post-2010, which we call modern connectionism and 𝔐 (see Table 1). Such a distinction accommodates a variety of related scientific events occurring as a function of so-called deep ANNs becoming computationally feasible and accessible to many scientists around the world, for example, the rise of using ANNs as models of the brain and cognition (Kriegeskorte, 2015; Schmidhuber, 2015; Sevilla et al., 2022; Thompson, 2021), as a result of successes (such as Cireşan et al., 2010; Hinton, 2012; Krizhevsky et al., 2012).

Building on our previous work (Guest & Martin, 2021, 2023), we will unpack where and how the (meta)theoretical positions with respect to connectionism appear to lack rigor. To this end, we construct a *metatheoretical calculus* (Guest, 2024; Guest & Martin, 2023) for connectionist tendencies: a description of the adjudication over theories, models, and scientific contributions that is carried out within and between this framework.

But before we can do any of that, what is *connectionism*? According to Rumelhart et al. (1986), it is "the notion that intelligence emerges from the interactions of large numbers of simple processing units" (p. ix). "This framework has been variously called

Table 1

A Collection of Perhaps Contradictory (Meta)Theoretical Claims or Commitments Between Older, Classical Versus Newer, Modern Connectionist Tendencies

Property	ℭ-connectionism: classical, pre-2010	𝔐-connectionism: modern, post-2010
Goal	The goal is understanding the repercussions of our theories, that is "models [are] tools for exploring the implications of ideas" (McClelland, 2009, p. 12). Models are used to understand the theories within connectionism, which themselves are about understanding brain, cognition, and behavior. Additionally, a "good fit never means that a model can be declared to provide the true explanation for the observed data" (McClelland, 2009, p. 12).	"The goal of the science is to be able to predict what systems are going to do. These artificial neural networks get us closer to that goal in neuroscience" (Josh McDermott in Ananthaswamy, 2021). And so "when we say we understood a phenomenon, first and foremost it means that we can predict all of the explainable variance in the data for any input in the domain over which the model is claimed to hold" (Kubilius, 2018, p. 110).
Question	Can connectionist principles give rise to similar behavior and brain data or cognitive capacities as seen in humans (e.g., Elman et al., 1996; Rumelhart et al., 1986)? No specific prediction requirements are imposed on the models and anatomical mappings, if present, are baked-in. The model is forwarded as a way to explore theory (McClelland, 2009).	Can ANNs predict, here used to mean correlate with, behavioral or brain data? As such, they are used like inferential statistics, but framed like theoretical models (viz. Guest & Martin, 2023). "Not only did we get good predictions… but also there's a kind of anatomical consistency" (Daniel Yamins in Ananthaswamy, 2021).
Theory	Theory is implemented by the model; the model is not a stand-in for theory. For example, "we consider a simple computational implementation of the theory, in which visual representations of objects and perceptual representations of verbal statements about these objects interact with one another by means of an intermediating semantic system" (Rogers et al., 2004, p. 206).	Theory is the model, for example, "theory [is] instantiated in task-performing computational models" (Kriegeskorte & Douglas, 2018, p. 4). Additionally, theorizing is (often) inspired by engineered systems, not nature directly, for example, "current computational neuroscience practice [looks to] AI [which] has historically provided a fund of ideas for biological theories" (Gershman, 2024, p. 4).
Mechanism	Mechanisms are proposed, which the model embodies, and experiments are done to show proof of concept, that is can connectionist principles give rise to phenomena and/or capacities of interest? "The [ANN] allows us not just to probe the response to a given test stimulus[, but to also] ask questions about the nature of the existing representations the model has learned" (Althaus et al., 2020, p. 5).	The model is assumed to be equivalent in some way to a cognitive or neural system, and experiments are done to support this assumption. "The core idea is to 'treat [an ANN] as a participant in a psychology experiment,' in order to tease out the system's mechanisms of decision making, reasoning, cognitive biases, and other important psychological traits" (Shiffrin & Mitchell, 2023, p. 1).
Brain	Brain regions, if related to models, are presented as being modeled—not as uncovered correlationally. Theory, or some knowledge of the to-be-modeled, to-be-understood, system, comes first and these ideas are placed into the ANN model purposefully (Guest et al., 2020; Rogers et al., 2004).	"Computational models can help infer the function of brain regions by linking model and brain activity. Multilayer models […] are particularly promising in this regard because their layers can be systematically mapped to brain regions" (Sexton & Love, 2022, p. 3).
Training	There is often explicit awareness of the possibility for a behaviorist or associationist stance and the load placed on the training regime and set, which in the case of ANNs is statistics in the input, for example, "[d]on't pre-wire structure into your mechanism if it can get it for free from the environment" (Plunkett, 2001, p. 193).	Claims about statistics in the inputs, the proverbial ghost in the machine (viz. Ryle, 1949), are downplayed. The model's depth or architecture generally is taken as the important factor. The training set is not implicated in argumentation, except to say it comprises realistic stimuli, for example, photographs (e.g., Jozwik et al., 2017; Storrs et al., 2021).

Note. We do not assume that this dichotomy characterizes all connectionist work, but we propose it functions as a simplifying lens—to display points on a continuum of beliefs—through which to understand the differences within this broader research program. ANN = artificial neural network; AI = artificial intelligence.

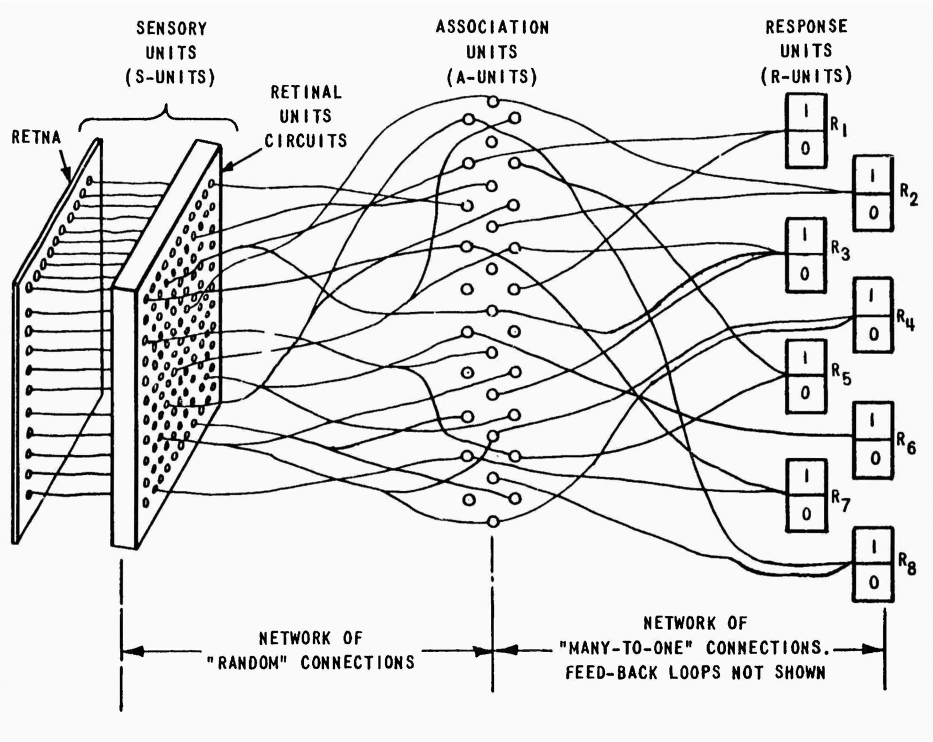
parallel distributed processing, neural network modeling, or connectionism[. A] term introduced by Donald Hebb in the 1940's" (Elman et al., 1996, p. xii). Most broadly, connectionism is the drawing of parallels between ANN models and the brain and cognition, and specifically using them to model neurocognitive systems or phenomena. Worthy of underlining here is "that connectionist networks typically do not bear a transparent relation to the neurological structures that realize them, [and so] the description of connectionist networks as 'neural nets' is somewhat misleading." (Egan, 1995, p. 184).

ANNs are mathematical objects implemented on digital computers that involve banks of units, artificial neurons, grouped into layers that propagate activations to other banks of units (or recurrently back to themselves) through matrix multiplication and some type of nonlinear squashing function, for example, hyperbolic tangent. Learning is achieved, for example, using backpropagation, by changing the numerical value of the connection weights, artificial synapses, between and within layers as a function of the difference between the network's output behavior and some target state, for example, in supervised learning: the output units. For a schematic of a hardware ANN, see Figure 1: the Mark I perceptron (Hay et al., 1960; Rosenblatt, 1958). Even in a nascent stage, such models resemble modern connectionist framings with respect to, for example, drawing parallels between input units and retinal cells (as seen on the leftmost part of Figure 1).

We will next analyze the relationships within and between connectionist tendencies, using Guest (2024) as a way to tease out (meta)theoretical properties. First, the Metaphysical Commitment section, we explore the clarity of how the two branches of connectionism that we propose herein, $\mathfrak{C}$ and $\mathfrak{M}$ (see Table 1), differentiate themselves as unique stances, as unique sets of assumptions. So within the broader framework of connectionism we differentiate two tendencies, both from the rest of the relevant fields' offerings (the Identity: What Characterizes Connectionism? section) and from each other (the Separation: What Differentiates Types of Connectionism? section). This exercise provides the building blocks for the formal accounts given in later sections.

Figure 1

Organization of the Mark I Perceptron


SENSORY UNITS (S-UNITS)
RETNA
RETINAL UNITS CIRCUITS
ASSOCIATION UNITS (A-UNITS)
RESPONSE UNITS (R-UNITS)
1
0
R1
R2
R3
R4
R5
R6
R7
R8
NETWORK OF "RANDOM" CONNECTIONS
NETWORK OF "MANY-TO-ONE" CONNECTIONS.
FEED-BACK LOOPS NOT SHOWN

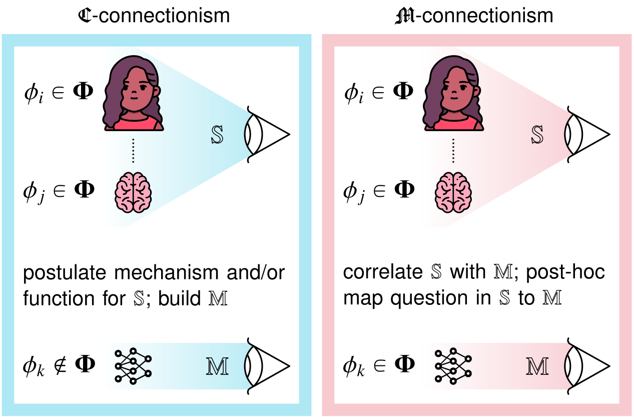
Note. "Organization of the Mark I Perceptron" (Hay et al., 1960, Figure 1, p. 13): a hardware ANN built by Frank Rosenblatt's Cognitive Systems Section at Cornell Aeronautical Laboratory Incorporated (also see Rosenblatt, 1958, 1959, 1960). From their inception, ANNs have had a proposed parallel structure to biological neural systems and have been forwarded as a model of human cognition. In this figure, it is demonstrated that hardware architectures along these principles were seen as important—this requirement is now relaxed with GPUs and other modern hardware, which undergird modern ANN models, not having proposed brain-like components. This blurs the lines between mechanistic and functional modeling, and phraseology like "microstructure of cognition" (Rumelhart et al., 1986) versus function approximation (Egan, 2017; Guest et al., 2025; van Rooij & Baggio, 2021). Connectionism resides at the inflections of these often contrasting ideological positions (Pasquinelli, 2017). GPU = graphics processing unit; ANNs = artificial neural networks. Adapted from *Mark I Perceptron Operators' Manual* (Report No. VG-1196-G-5) (p. 13), by J. C. Hay, B. E. Lynch, and D. R. Smith, 1960, Cornell Aeronautical Lab (https://apps.dtic.mil/sti/tr/pdf/AD0236965.pdf). In the public domain.

Second, the Discursive Survival section, we investigate how both $\mathfrak{C}$- and $\mathfrak{M}$-connectionism are discussed by the broader fields they are embedded in (neuro-, psychological, and cognitive sciences). For example, we know that connectionism, and specifically as a modeling strategy and as a methodology, has been subject to targeted attacks, for example, claims about inability to compute certain functions, like exclusive or (XOR), for dealing with nonlinearly separable data. We analyze this along two broad lines: the ability to craft a coherent reaction to attacks (the Argumentation section), and the ability to tell a coherent story about the scientific theory (the Narration section). We formally describe the two types of connectionist reasoning using modal and doxastic logic in Proposition 1 for $\mathfrak{C}$-connectionism and Proposition 2 for $\mathfrak{M}$-connectionism.

Finally, the Empirical Interface section, we elaborate on how, or even if, $\mathfrak{C}$- and $\mathfrak{M}$-connectionist models successfully mediate between theory and data. We provide an experimental typology in Figure 2 to document and formalize how practitioners within connectionist tendencies reason about their experimental manipulations and their modeling praxis.

Figure 2

Classical Connectionism Versus Modern Connectionism


$\mathfrak{C}$-connectionism
$\mathfrak{M}$-connectionism
$\phi_i \in \mathbf{\Phi}$
$\mathbb{S}$
$\phi_j \in \mathbf{\Phi}$
postulate mechanism and/or function for $\mathbb{S}$; build $\mathbb{M}$
$\phi_k \notin \mathbf{\Phi}$
$\mathbb{M}$
$\phi_i \in \mathbf{\Phi}$
$\mathbb{S}$
$\phi_j \in \mathbf{\Phi}$
correlate $\mathbb{S}$ with $\mathbb{M}$; post-hoc map question in $\mathbb{S}$ to $\mathbb{M}$
$\phi_k \in \mathbf{\Phi}$
$\mathbb{M}$

Note. A cartoon depiction of the simplified differences between $\mathfrak{C}$- and $\mathfrak{M}$-connectionism with respect to postulating mechanisms and building models (collectively $\mathbb{M}$) and relating them to the cognitive and neural systems (collectively $\mathbb{S}$). On the left, in blue and in the top panel, we see an eye, which represents the scientist working under $\mathfrak{C}$-connectionism, looking at human behavior and cognition (phenomena ϕ_i) and the brain (phenomena ϕ) all of which comprise $\mathbb{S}$, the system under study. All the observed phenomena here are very explicitly in Φ, the set of phenomena that neuro-, cognitive, and psychological sciences care about. On the left, in blue and in the bottom panel, we see what $\mathfrak{C}$-connectionism does to model ϕ_i and ϕ_j above: create a model that embodies what they think connectionism can do, that is build $\mathbb{M}$ as a proof of concept, to capture the phenomena. On the right, in pink, we see the same top panel as before: $\mathfrak{M}$-connectionists document, witness, research ϕ_i, and ϕ_j of $\mathbb{S}$. However, under $\mathfrak{M}$-connectionism there is little to no engagement with the mechanistic and/or functional theoretical positions about $\mathbb{S}$ when building $\mathbb{M}$, shown in the bottom panel on the right. In fact, $\mathbb{M}$ is proposed as the theory itself—not built under a theory, and trained to perform tasks that $\mathbb{S}$ is also subject to. What happens is data from $\mathbb{M}$ is correlated with data from $\mathbb{S}$, such that statistical relationships, for example, goodness-of-fit, are what make $\mathbb{M}$ a useful scientific model. Thus, scientific questions (Table 1) in $\mathbb{S}$ are matched post hoc to aspects of $\mathbb{M}$. Indeed, behaviors of $\mathbb{M}$ are often seen as worthy of scientific investigation as if ϕ_k from the perspective of connectionism, potentially subverting the field's goals. (Icons licensed from Victoruler—Flaticon.) See the online article for the color version of this figure.

Metaphysical Commitment

> Almost everyone who is discontent with contemporary cognitive psychology and current "information processing" models of the mind has rushed to embrace "the Connectionist alternative." (Fodor & Pylyshyn, 1988, p. 4)

We propose that connectionism can usefully be split into two tendencies: $\mathfrak{C}$- versus $\mathfrak{M}$-connectionism. The proposed differentiating factors are those found in Table 1, which characterize the two positions' scientific goals and questions, their beliefs about theory, and what constitutes a theory, their mechanistic assumptions or proposals, their interface with the brain, and their framing of what the models' training sets provide to reasoning about the models' successes and failures. Because indeed we find these dimensions provide evidence of difference—although we by no means preclude further differences within what we dub $\mathfrak{C}$- and $\mathfrak{M}$-connectionism—we present them side-by-side in Table 1 to further understanding and heighten awareness of these changes in connectionist (meta)theorizing. Where we use words like theory, mechanism, and model we do so as a function of authors' use, and so hold up a mirror to what they may mean, and as labels or placeholders for a wide variety of possible use cases since scientists may have idiosyncratic as well as field- or framework-level meanings they deploy in different contexts.

Importantly, the authors strive to avoid imposing on the literature and reader their own definitions of terms, like theory, because they hang in relation to each other as variables where often only the relationships between them or their properties are articulated, such as in the examples we provide in Table 1 and in the rest of the article. Notwithstanding, if a description of what each of these terms could mean is useful, the reader can lean on the following: *Theory* is "a scientific proposition—described by a collection of natural-language sentences, mathematics, logic, and figures—that introduces causal relations with the aim of describing, explaining, and/or predicting a set of phenomena" (Guest & Martin, 2021, p. 794; also Guest, 2024). *Mechanism*

> in neuroscience [can be] synonymous to [neurobiological] substrate (e.g., Darden, 2006; Lisman et al., 2017), but under computationalism a substrate will not, cannot, cut it as a mechanistic analysis [therefore many practitioners] completely rule out using substrate and mechanism interchangeably. (Guest & Martin, 2025, Box 1; Guest et al., 2025)

Connectionism has a well-documented history of straddling the boundaries between neurobiology, on the one hand, and AI and psychology, on the other hand, and so substrates (e.g., individual neurons or brain areas) and, for example, constitutive mechanisms (Guest & Martin, 2023; Krickel, 2024; Ross & Bassett, 2024) are seen to converge. However, these two types of mechanistic underpinning can also be entirely separate, for example, representations that emerge in the ANN model can be described as not related to any substrate detail, especially in $\mathfrak{C}$ (Egan, 1995). Finally, *model* herein is the term we use for the role of the ANN, however specified and implemented (Guest & Martin, 2023), characterized by the mediation it provides between theory, formal or verbal, and the relevant observations, data, and phenomena (Morgan & Morrison, 1999)

On a related note, we do not mean that all work that involves ANNs can be easily classified into either one type of connectionism over another. We also do not mean this distinction is purely temporal and thus explicitly include examples of ℭ-connectionism that are post-2010 to highlight this. We merely propose that seeing these high-level strands of difference is informative.

To presage the coming analysis, the difference between the two types of connectionism can be boiled down to, on the one hand, ℭ has the goal to show *that* ANNs learn or behave like the neurocognitive system because they are held to share important structural or mechanistic properties—a proof of concept, a framework for housing theories. On the other hand, 𝔐 has the goal to show *how* ANNs learn or behave like the neurocognitive system—a totalizing view of method as theory and of map as territory—and not as a function of verbal theory, formal specification, and other modeler choices.

Identity: What Characterizes Connectionism?

> The model … represents an essentially empiricist approach to perception [with] an optical input, and a printer or set of signal lights as an output. [A]fter a period of training, the system will exhibit capabilities for discrimination, association, and stimulus generalization. [T]his is the first time that a set of theoretical principles will have been clearly proven to generate a perceptual capability, in a system of completely known structure. (Rosenblatt, 1959, pp. 296–297)

ℭ-connectionism bases its identity in part on showing that connectionist models can account for phenomena that do, or did, not appear to be easily composable into computations carried out by smaller units, such as the so-called neurons and their connection weights in ANNs (see Table 1, especially rows *Goal*, *Question*, and *Training*). ℭ "has demonstrated that a great deal of information is latent in the environment and can be extracted using simple but powerful learning rules" (Elman et al., 1996, pp. xii–xiii). In many ways, this is a theoretical point about the nature of what the modeled organism is doing, that is humans are able to learn statistical regularities from the environment, which ℭ-connectionism rightly takes to mean that the resulting learning is by virtue of the training set and of the learning algorithm (Guest et al., 2020). That is, they set their scientific sights on understanding if their connectionism can, given the ANNs they build and the training sets they train their model on, give rise to what they see people do. ℭ-connectionists, like Elman et al. (1996) and many others, explicitly reacted to claims of innateness (however construed) of capacities, asserting and showing (according to their standards of evidence) that appealing to innate properties of neurocognitive systems is not required.[1]

On the other hand, what characterizes 𝔐-connectionism is that its identity is based on (a) deep ANN models, which are framed as human-like[2] because they score highly on typically unrelated (or at least different) tasks to those being neurocognitively researched and because they accept as input the same types of stimulus files that can be used in experiments with people, that is realistic training and testing sets (see rows *mechanism* and *training*, in Table 1). Contrast this with ℭ perspectives on larger models: "increasing complexity creates a tension with the primary goals of modeling—simplification and understanding" (M. S. C. Thomas & McClelland, 2008, p. 50). And this tension reaches a climax when one considers that if the model is treated as a black box, often is produced by the technology industry for their purposes and is likely closed source *and* the training regimen and stimuli are outside the scientists' control (Jain et al., 2024; Liesenfeld et al., 2023; cf. Sullivan, 2022): what is left for the scientist to contribute other than uncover correlations between a largely preset model and a phenomenon?[3] This relegates in many ways the ANN to no more than a statistical model which provides scientists with a value of model-to-person or model-to-brain matching (e.g., "brain-score," Pasquinelli, 2017; Schrimpf et al., 2020).

On this point, 𝔐's identity is also based on (b) the ANNs displaying (statistical) prediction capabilities with respect to correlation to observations from brain and behavior (see rows *goal* and *question*, in Table 1). Additionally, (c) also unique in 𝔐-connectionism is the stance that the model embodies the theory—not that the model is imbued or enriched by the scientist's theoretical positions, nor that the model mediates between theory and data, nor that the model helps test a theory, nor that the model as in ℭ allows us to debug our thinking, but that the model constitutes theory as such (see row *theory*, in Table 1).

Finally, (d) 𝔐-connectionism implicates brain areas, and neuroscience generally, more often than ℭ-connectionism and with an agenda entangled with so-called neuro-/bio-plausibility (see row *brain*, in Table 1). Let us contrast again with a typical ℭ stance on this: "Neural plausibility should not be the primary focus for a consideration of connectionism" (M. S. C. Thomas & McClelland, 2008, pp. 28–29, also Smolensky, 1988, but cf. Elman et al., 1996; McLaughlin & Warfield, 1994; Stinson, 2020). Additionally, appealing to innateness and nativism—that is that "aspect[s] of cognition […] must be innate, or, (at the very least) subject to powerful biological constraints" (Elman et al., 1996, p. 240)—is not ruled out by 𝔐-connectionism. In fact, such constraints are appealed to, or seen as imperative to include in models, using phrases such as "inductive bias" (a concept which has been around in machine learning for a while, including with respect to ANNs; Gordon & Desjardins, 1995; Pavlick, 2023). Inductive biases like innate capacities are "not *learned*" (Goldberg, 2008; see footnote 1). And as such stand in stark contrast to requests from ℭ practitioners to not "pre-wire structure into [the] mechanism if it can [be obtained] for free from the environment" (Plunkett, 2001, p. 193), that is the training set.

In both types of connectionism, although especially in 𝔐, there emerge similar entanglements between observations and the so-called "bridging" of levels, which proposes that, for example,

[1] As Elman et al. (1996) and others acknowledge, innateness and nativism are not clear-cut concepts and the polarized, or even wrong framings of the nature/nurture discussions in science and society at large are likely more damaging than useful. Notwithstanding, it is not the case that "innateness" as discussed or used by linguists is the same (concept) as "innateness" in genetics or biology.

[2] This characterization of human-likeness is the case only in complex or dubious ways (e.g., Leivada, Günther, & Dentella, 2024). Also, the concept of deeper models being more human-like is consistent with the classical dualist perspective: "The difference between machines and natural objects is simply one of degree for Descartes[. M]achines are works of nature differing from other natural objects only by degree of complexity" (Hattab, 2009, p. 85).

[3] As in other manifestations of the current AI hype, the deskilling force of ANNs raises its head in the fields that use them for their scientific endeavors. Threatening, albeit with the seeming consent of their creators, to replace and/or deskill the very scientists who use these models (Forsythe, 1993; Pfaffenberger, 1988; Rich et al., 2021; van Rooij et al., 2024).

findings about neurons can help constrain theories about psychology or vice versa (e.g., Griffiths et al., 2012; Love, 2015; Mok & Love, 2023; cf. Elgin, 2009; Nagel, 1979). Such discussions in the literature indicate a misunderstanding of Marr's (1982) levels, that is a confusion between his levels of analysis versus broader levels of description of the world (cf. Blokpoel, 2018; Chirimuuta, 2018; Rich et al., 2020; van Rooij & Baggio, 2021). Stinson (2018, p. 126) explained how "[c]onnectionists talk about taking constraints from both physiology and psychology, as though they are employing an inferential pincer movement[, but really this is just words to little effect since] there are no halting conditions" in the search for such bridges between levels (Guest & Martin, 2023; Sejnowski et al., 1988). "In the cognitive sciences and the philosophy of mind, […] appropriate bridge laws will not be forthcoming" (Arkoudas, 2008, p. 471).

This situation is noteworthy because if models get caught up in computationalist confusions, they possess no protection against issues like formal intractability (van Rooij, 2008; van Rooij et al., 2024), limiting not only their usefulness as models, but ensuring severe implications for their related theory if its evidential basis depends on the model. It is also notable because such reductionist maneuvers are intriguing in of themselves when studying how a set of scholars reason metatheoretically, especially in relation to what empirical evidence is privileged, and in quasidefiance of Marr (1982). Taken together, what this means is that certain types of evidence are seen as primal, like behavioral or neuroimaging data sets, while others, such as mathematical truths about given computational specifications are less important, even though the latter directly affect the properties of ANNs and other models (van Rooij et al., 2024).

Separation: What Differentiates Types of Connectionism?

> According to. … Ali Rahimi and others, [ANNs] and deep learning techniques are based on a collection of tricks, topped with a good dash of optimism, rather than systematic analysis. Modern engineers … assemble their codes with the same wishful thinking and misunderstanding that the ancient alchemists had when mixing their magic potions. (Dijkgraaf, 2021, n.p.)

Connectionism can be seen as a dramatically divergent way of doing science. In the earlier argumentation regarding ℭ-connectionism, it was explicitly stated that the goal was to show a proof of concept (see rows *goal* and *question*, in Table 1). In other words, human cognition was seen as driven by explicit rules or requiring rule-like knowledge, and connectionism reacted to that by asking if certain cognitive capacities could be captured through statistical learning mechanisms. For example,

> the rules of English pronunciation are complex and highly variable, and have been difficult to model with traditional Artificial Intelligence techniques. But neural networks can be taught to read out loud simply by being exposed to very large amounts of data. (Sejnowski & Rosenberg, 1987; Elman et al. 1996, p. 5)

While this is not only true in more recent incarnations of ℭ-connectionism, older versions still, all the way back to McCulloch and Pitts (1943) have highly divergent takes on framing cognition (also see Abraham, 2002; Aizawa, 1992; Boden, 1991; Chirimuuta, 2021; Gefter, 2015):

> To psychology, however defined, specification of the net would contribute all that could be achieved in that field even if the analysis were pushed to ultimate psychic units or "psychons," for a psychon can be no less than the activity of a single neuron. Since that activity is inherently propositional, all psychic events have an intentional, or "semiotic," character. The "all-or-none" law of these activities, and the conformity of their relations to those of the logic of propositions, insure that the relations of psychons are those of the two-valued logic of propositions. Thus in psychology, introspective, behavioristic or physiological, the fundamental relations are those of two-valued logic. (McCulloch & Pitts 1943, p. 131)

In other words, "[McCulloch] had been inspired by the *Principia*, in which Russell and Whitehead tried to show that all of mathematics could be built from the ground up using basic, indisputable logic" (Gefter, 2015, p. 96; also see Abraham, 2002, 2012; LePage-Richer, 2024). Ironically, of course, *Principia Mathematica* failed in its stated goal "to solve the paradoxes which […] have troubled students of symbolic logic" (Whitehead & Russell, 1910, p. 1) as shown by Gödel's incompleteness theorems (Gödel, 1992, originally published in German in 1931). A further irony is uncovered when one considers that 1980s and 1990s connectionism was embroiled in fierce debates against symbolic style modeling and perspectives on cognition (Aizawa, 1992; Boden, 1991).

Perhaps the real spirit of connectionisms of all types is the divergence from methodological status quos with meagre or minimally weak theoretical commitments. For example, notice above how in McCulloch and Pitts (1943) connectionism intervenes in psychological theorizing to propose propositional logic-based neurons as a universal substrate, for example, "most connectionist researchers are really committed to ultimate neural plausibility, which is more than you can say for most other approaches" (Elman et al., 1996, pp. 49–50). These forms of reasoning over ANNs help to set the stage for future deep learning.

Separating 𝔐-connectionism from the rest of the literature it is embedded in involves again highlighting its nature as a methodology and not a theory, as discussed above where it is used as proof of concept for a framework. However, 𝔐-connectionism also has features present in mathematical psychology (cf. Navarro, 2021) wherein models are fit to the data directly, and in this case extremely large data sets. Often this is in lieu of frequentist statistical models and through the use of similar techniques to achieve correlations (such as representational disimilarity matrices; cf. Dujmović et al., 2020; Guest & Martin, 2023; Pasquinelli, 2017). This stands in contrast to a lot of other cognitive modeling techniques used in the relevant fields connectionism typically touches on (see the 𝔐-connectionism column of Table 1). As such, 𝔐-connectionism embodies a methodology-as-theory approach, something that is rarely so brazen even in the theory-light hypothesis-driven cognitive sciences like mainstream experimental psychology (Flis & van Eck, 2018; Guest & Martin, 2021; van Rooij & Baggio, 2021).[4]

[4] Notably, Flis and van Eck (2018) surveys all the last half of the previous century and the word "theory" appears nowhere prominently in his analyses. While Guest and Martin (2021) and van Rooij and Baggio (2021) discuss what to do with with such a continuing lack of theory in a world of so-called replication and other crises.

In conclusion, $\mathfrak{C}$-connectionism allows and deploys a metatheoretical calculus that contains:

> If we believe ANNs can p: learn an input–output mapping.
>
> Then, we create ANNs to check p.

On the other hand, $\mathfrak{M}$-connectionism allows and deploys:

> If ANNs appear to perform tasks at a human level.
>
> Then, ANNs have human capacities.

This second type of reasoning has been analyzed in depth in Guest and Martin (2023), but we will return to both in the following sections.

Discursive Survival

> Only very recently has connectionism resurfaced in AI. And, according to some, with revival has come reversal: Seemingly the tables have turned. Many people today can be heard announcing that GOFAI [good old-fashioned AI; symbolic approaches] is utterly discredited. [O]nly connectionist theories can explain the mind—so, at least, we are told. (Boden, 1991, p. 11)

We propose connectionism has a survival strategy honed by decades of proverbial summers and winters; engineered high and low funding periods (Boden, 2006; Dreyfus, 1965; Haigh, 2023; Lighthill, 1972; McCorduck, 2004; Olazaran, 1996). We will analyze such historical narratives, but first we will examine how the metatheorical calculus of $\mathfrak{M}$-connectionism is—perhaps unexpectedly given its problematic structure—robust and transmissible. Before that, we explain the uptake and deployment of the more mainstream calculus of $\mathfrak{C}$-connectionism.

Argumentation

Typical Science

> [C]onnectionists have paid far too much attention to the successes of connectionist modelling in AI and far too little attention to theoretical issues concerning the nature of cognition. (McLaughlin & Warfield, 1994, p. 382)

In the case of $\mathfrak{C}$-connectionism, we see a typical scientific framing for the modeling endeavor. In the most zoomed out case, they posit $\mathbb{M}_p$, which they believe to be the case, for example, "ANNs can learn this input–output mapping," where $\mathbb{M}$ is the model and p is the mapping, task, or capacity being modeled; see Figure 2. In fact, this is part of their identity (normatively) as connectionists, but they also clearly state they believe it. $\mathfrak{C}$-connectionism commits to a metatheoretical calculus that permits reasoning such that, if one observes or infers a phenomenon or capacity in a neurocognitive system, and one believes it can be modeled using connectionist methodologies, then one constructs an ANN such that it does; formally:

$$\mathcal{O}(\mathbb{S}_{p'}) \wedge \mathscr{B}(p' \sim p) \wedge \mathscr{B}(\mathbb{M}_p) \rightarrow \mathcal{C}(\mathbb{M}_p), \tag{1}$$

where $\mathbb{S}_{p'}$ is the system under study observed as performing p', which appears equivalent to, or is formalized or modeled as p our target phenomenon or capacity. So, $\mathcal{O}(\mathbb{S}_{p'})$ is the observation that p' occurs in the system under study (see Figure 2). $\mathscr{B}(p' \sim p)$ is the belief that constitutes the scientific mediation between model, which has a relationship to (e.g., produces) p, and phenomenon under study p', provided by our practice and cognition as scientists.[5] $\mathscr{B}(\mathbb{M}_p)$ represents our belief in "ANNs can learn some input–output mapping," or in "ANNs can show behavior similar to that seen in humans doing a task," or in "ANNs can compute internal representations that are theoretically useful." So $\mathscr{B}(\mathbb{M}_p)$ is the explicitly stated belief that p is modelable by $\mathbb{M}$. $\mathcal{C}(\mathbb{M}_p)$ is the process of checking whether there exists an accessible possible world that is truth making for $\mathbb{M}_p$ (see Barcan Marcus, 1961, 1967, 1990, 1997).

Proposition 1 can be read as: If we observe p' in a cognitive system, and we believe the p' can be related to a process, p, in our model, and we believe our model can give rise to this process, then we endeavor to create such a model. In all cases, p is not typically general, but a specific example of conceptualizing a subset of human cognition, such as a capacity, series of behavioral tasks, a disorder, and so forth. The referent of p is much broader than only the set of observations that can be written to datafiles; it can involve scientific entities like cognitive capacities. In other words, $\mathfrak{C}$ involves checking if we can create an ANN model that indeed can capture the given constraints, be they input–output mappings as above, or some configuration of internal states, or both, etc. This is not controversial or unusual science, wherein we go from a theoretical position that we entertain for any number of reasons (including belief), to looking for/at models that can capture our beliefs.

To repeat Proposition 1, if we observe a system $\mathbb{S}$ performing p', which can include capacities, (quasi)theoretical constructs, as well as phenomena, ϕ, such as those in those depicted in Figure 2, then we believe we can build a model $\mathbb{M}$ that performs p, which is the modeled variant or stand-in for p'. Based on this belief, we go off and search for $\mathbb{M}_p$. This practice is taken to be robust, primarily because it demonstrates that connectionism can indeed account for many phenomena or capacities through this way of modeling. Purely because of ANNs' high expressive power, such models can seemingly model everything—or at least every task or capacity that $\mathfrak{C}$-connectionists engineer stimulus sets for. This had not been the case previously with the XOR problem during the Perceptrons Controversy (Olazaran, 1996). But since the advent of contemporary big data combined with large models, which has improved performance on some benchmarks—likely due to the *neural scaling law* (Bahri et al., 2024), whose implication for the use of ANNs as arbiters of theory is alas out of the scope of this article—it has been too tempting to simply add more layers, *just one more* hidden pool of units. Rather than contemplating how to better model causal structures of phenomena or of cognitive capacities with particular and specific specifications beyond additional layers or units, the focus is on increasing the size of models and strengthening correlations between models and empirical data (Guest et al., 2025).

So discursive survival here lies in the prima facie sensible nature of believing something could be modeled with ANNs and then demonstrating to the rest of the interested scientific world that it can

[5] We do not propose it is unique to $\mathfrak{C}$-connectionism, but a common framing of how we work (Guest & Martin, 2021; Morgan & Morrison, 1999). What $\sim$ captures here is this mediation; and if it is replaced with $\equiv$ we could diagnose a transition to $\mathfrak{M}$-connectionism, or a confusion between map, p, and territory, p'.

indeed be done. $\mathfrak{C}$-connectionism as a framework can model such a vast swathe of cognitive and psychological findings that it appears to be—not a failure any more, but—a coherent modeling paradigm through which to test ideas and hypotheses. As mentioned, we will return to this, as what might appear discursively to be the case need not always indeed be the case.

Putting the Con in Connectionism

> Despite there being academic reasoning for how Egyptians built the pyramids — AKA the pulleys — people still believe aliens built or at least instructed the building and go on to make YouTube videos about the conspiracy, which then perpetuates the belief for a whole new generation of skeptics and extraterrestrial enthusiasts. We'll never know the full story, so people fill in the gaps with the narrative they believe the most—which, for lots of people, goes back to aliens. (R. Thomas, 2021, n.p.)

As we will explicate herein, connectionism in its modern incarnation can be seen as often applying conspiratorial or otherwise pseudoscientific thinking to scientific reasoning (Guest & Martin, 2023; Spanton & Guest, 2022). This also harks back to the neurobiological origins of ANNs. For example, LePage-Richer (2024) documented that

> neural networks were first introduced as a neuroanatomical approach to racial difference [as well as constituting part of] an experimental ethos that systematically involved organizing, managing, and disciplining human bodies while devaluing the practical, local, and contextual knowledge they hold. (p. 21)

For $\mathfrak{M}$, such cases have a slippage that we have also warned against (Guest & Martin, 2023): a confusion between "the outward appearance and the essence of things [such that it can be possible we believe that the two have] directly coincided" (Marx, 1894, p. 592). The warning when the way things behave, or seem to appear, is held as a stand-in for how things work or are, is not only that "science would be superfluous" (Marx, 1894, p. 592) in the general case, but also that we risk severe fallacies minimally and crimes against humanity maximally (Guest, 2024). In the former case, we can confuse the sun appearing to rise with evidence for geocentrism, because it certainly looks that way; and in the latter case we can confuse, for example, superficial differences in skin tone for humanity existing on a sliding scale (Andrews et al., 2024; LePage-Richer, 2024; Saini, 2019).

In past work, we have boiled this reasoning down to: "If the model correlates with human behavioral and/or neuroimaging data, then the model does what humans do" (Guest & Martin, 2023, p. 216). Which is to say that when we, as a field, see models behaving consistently with our experimental participants, we allow conclusions regarding equivalence of the two systems (engineered model and phenomenon under study) in terms of their mechanistic proprieties or otherwise important core aspects. As we shall see through some worked examples below, this perpetuates reasoning that depends on a false analogy—just because two systems appear similar, these conclusions are not warranted and are in fact harmful in the ways described.

To analogize, the pyramids—whose construction techniques are subject to research and lay discussion, not only because they are beautiful feats of architecture and engineering but also because they appear to be impossible without the use of modern tools—are subject to similar conspiratorial thinking as are ANNs. Some go so far as to state that because of their appearance, that it is indeed impossible for them to have been built millennia ago by humans, attributing their existence to extraterrestrial aliens. But even outside canonical conspiratorial thought we see that even the simple pyramid gives rise to incredibly many proposed explanations and understandings—which is to say that, at least if outside pancomputationalism (cf. Dodig-Crnkovic, 2023), the Pyramids are not computing anything.

To take the worst case of (racist) conspiratorial thought, as found in pseudoarcheology, we get (via modus ponens):

> If the Pyramids appear to require modern techniques, then they were built by aliens.
>
> They appear to require modern techniques.
>
> Therefore, they were built by aliens.

Bearing in mind, this does not change even if we allow for more realistic scenarios: The point still stands that more superficially sensible scenarios still implicate a huge swathe of potential options to choose from, for example, with respect to ramp types attached onto the Pyramids to aid in construction (multiple realizability; Chirimuuta, 2018, 2021; Egan, 2017; Figdor, 2010; Guest & Martin, 2023; Guest et al., 2025; Hardcastle, 1995, 1996; Litch, 1997; Polger & Shapiro, 2016; Ross, 2020):

> If the Pyramids appear to require modern techniques, then they were built using modern techniques.

Even in this less conspiratorial case, some argue that—even though we have found evidence of quarrying on the stones themselves and have good candidate quarries—human-made stones cast from concrete-like substances are part of the Pyramids (see Folk & Campbell, 1992, for these kinds of claims). To reiterate, we do not even need to take a stand on the fact of the matter of construction techniques of the Pyramids, we need only to scrutinize the relationships between statements to understand the conclusion does not follow (in Guest & Martin, 2023, we criticize the order of each statement within the conditional). This is what makes the search for extraterrestrial intelligence (e.g. *SETI Institute*, 1984) typical science, while the above claims are pseudoarcheology. The Ancient Egyptians may well have possessed the knowledge of concrete, but we cannot safely conclude this (solely) from the way things look. Applying this analysis to $\mathfrak{M}$-connectionism, we get,

> if ANN behaviors appear to be cognition, then they have human capacities.

And since it is the case that ANNs display quite readily human(-like) behaviors, modus ponens can be applied as above. Placed into doxastic logic, we can express the calculus this way:

$$\mathcal{O}(\mathbb{S}_{p'}) \wedge \mathscr{B}(p' \equiv p) \wedge \mathcal{O}(\mathbb{M}_p) \rightarrow \mathscr{B}(\mathbb{M}_{p'}), \qquad (2)$$

where each symbol is as before; $\mathcal{O}(\mathbb{S}_{p'})$ is the observation that p' occurs in the system under study, $\mathscr{B}(p' \equiv p)$ expresses that $\mathfrak{M}$-connectionists believe that p', the phenomenon or capacity people do is the same as p, the behavior of the model; $\mathcal{O}(\mathbb{M}_p)$ is the observation that the ANN can perform p, recall Figure 2; and $\mathscr{B}(\mathbb{M}_{p'})$ is the belief that $\mathbb{M}$ does do p', which is what the neurocognitive system does. The belief $p' \equiv p$ is (minimally tacitly, maximally opportunistically, and purposefully) a direct confusion between map and territory, explanandum and explanans, model and phenomenon, and theory and capacity.

In both cases, aliens visiting Earth, and matrix multiplications being brains or cognition, our thinking jumps from seeing what something appears to be, looks like, to what something is, and in so doing selects a, or even the, most unlikely and complex solution. In our case, going from a model correlating with neurocognitive data to proposing that mechanistic or functional properties are somehow importantly shared, can be seen to violate what we know about computationalist principles, like multiple realizability (Chirimuuta, 2018, 2021; Egan, 2017; Figdor, 2010; Guest & Martin, 2021, 2023; Guest et al., 2025; Hardcastle, 1995, 1996; Litch, 1997; Polger & Shapiro, 2016; Ross, 2020; van Rooij et al., 2024). For example, the multiple realizability of a given function by an infinite set of possible codebases means that the inferences on the implementation level between codebases about shared mechanisms is limited. On the other hand, if one rejects that cognition can be multiply realized, then what mechanistic or otherwise understanding do models provide under such an assumption? If one accepts multiple realization is at play, which connectionists typically explicitly do, then one must bite the bullet and negotiate the inferential and mathematical complexities that come with it (Guest & Martin, 2025; Guest et al., 2025). A mirror—not even an AI mirror—is not what it reflects (Vallor, 2024). These jumps in logic can be found in many jocular stories, for example, when a person first encounters a television, would they assume the device contained small people? Or would a person assume somebody was trapped inside a telephone because it emits human voices? 𝔐-connectionist thought follows along these lines when it formulates its metatheoretical calculus to allow for these types of beliefs to follow from these types of observations. This all stems from the fact that inputs, outputs, of people and models and even the internals of a model are not a specification and not a theory (Guest & Martin, 2021). Landing back in Egypt for a moment, we see another parallel:

> Everything that I have found convinces me more and more that indeed it is this society that built the Sphinx and the pyramids. Everytime I go back to Giza my respect increases for those people and that society, that they could do it. You see, to me it's even more fascinating that they did this. … Rather than just saying, you know copping out and saying, there's no way they could have done this. I think that denigrates the people whose evidence we actually find. (Mark Lehner as quoted in NOVA, 1997, n.p.)

Thus, in much the same way, confusing the ANN map for the neurocognitive territory ends up underestimating the system under study, derailing our science, and plays fast and loose with our experimental participants' humanities (Erscoi et al., 2023; Forbes & Guest, 2025; LePage-Richer, 2024; van der Gun & Guest, 2024). Nonetheless, it keeps surviving as a rhetorical strategy in our scientific endeavors.

Narration

In a Deep World, We Need to Go Deeper Still

> There are a lot of people out there who are deeply annoyed by the outlandish claims being made in some quarters about the accomplishments and power of connectionism. (Smolensky, 1988, p. 67)

While the authors think that ideologies consistent with 𝔐-connectionism are rhetorically flawed and allow themselves to be easily attacked (Guest & Martin, 2023), we propose that connectionism generally, and especially ℭ-connectionism, is occasionally robust in the face of direct attacks, for example, the discussions involving symbolic AI. Notwithstanding, it must not be forgotten that connectionism has had long periods of being extremely scientifically unfashionable, for example, the Perceptrons Controversy (Olazaran, 1996).

On the robustness point, however, ℭ-connectionism adapts to critique. For example, connectionism started evaluating the so-called internal representations in ANNs (rows *theory* and *mechanism* in Table 1) before 𝔐-connectionism appeared on the scene. Criticisms such as, connectionist models are only theoretically useful "if one can interpret the *internal activity* of the simulation that the simulation increases our knowledge; that is it is only then that the simulation is to be considered a scientific *theory* worthy of consideration" (Green, 1999, p. 143) were obviously taken to heart since connectionists took to investigating so-called internal representations, which means looking at the values of units in the hidden layers, that is any units not involved in direct input and output. Thus, incorporating looking at the hidden units became standard methodological practice. "After a computer model has been trained to generate a behavior which is of interest to us, we can inspect its internal representations, vary subsequently the input to it, alter the way it processes the input, and so forth" (Elman et al., 1996, p. 45).

Another point of successful scientific presentation of ℭ-connectionism comprises acknowledging and understanding fit-to-data is not enough (rows *goal* and *question* in Table 1). "Now all cognitive connectionists will agree that simulation alone is not explanation" (M. S. C. Thomas, 1998, n.p.). However, he followed with, "Connectionist models must have constraints, and those constraints must be supported by empirical data" (n.p.) with no reference to theory as a factor. Notwithstanding, ℭ-connectionism presents a coherent research program: "we want […] to use the model to help develop a theory about the internal processing which gives rise to […] behavior, rather than just implementing a theory we already hold" (Elman et al., 1996, p. 56). This is an appealing rhetorical frame in which they present connectionist models as tools to refine scientific thinking and theorizing, and in which they concede fitting the data is a red herring. This presages 𝔐-connectionism's theoretical stance, or lack thereof.

The story ℭ-connectionists tell is that they keep surviving being unfashionable because they are indeed onto something. And their main way of surviving, post-XOR fiasco, was to argue strongly their case. Aspects of learning problems like nonlinear separability remain underexplored to this day, or minimally underdiscussed (Baayen & Hendrix, 2017; Olazaran, 1996).

More so than previous incarnations, 𝔐-connectionism as a movement in the neuro-, psychological, and cognitive sciences is caused by and causes a scorching hot AI summer. As such, discursive survival is granted, we propose, since the media and public and private funding provide, minimally financial, protection from exposure to critique or improvement attempts (Kindig, 2024). Also, because of this cover, statements about the 𝔐 framework can be made without the normal scientific standards of citation, for example, such as that ANNs have reached human-level capabilities (for more on analyzing this rhetoric, see Guest & Martin, 2023;

Titus, 2024). Furthermore, the abilities of models are rarely questioned, or the questioning is ignored, for example, it is not well-known that the MNIST data set is linearly separable (Just & Ghosal, 2019). This is important—at least rhetorically—because restricted Boltzmann machines trained on MNIST were part of the AI spring prior to this summer (Hinton et al., 2006). If they could have been undercut in the past, rhetorically as before with the perceptrons controversy, by investigating potential oversights in the training data, such doors are now truly closed. The thermal runaway reaction of the current AI summer was by 2010s in full force.

Thus, it could be said, that having learned from the perceptrons controversy (Olazaran, 1996), that adding more layers to networks is (or at least was until recently) seen as the mantra for fixing (m)any problems (Dawson, 2013; Medler, 1998). By the same token, apparently serious problems with (understanding) overfitting, overparametrization, or the behavior of these models generally (e.g., Belkin et al., 2019; Gardner, 1988; Nichani et al., 2020; Richter et al., 2021; Zhang et al., 2016, 2017) are seen as anything from quaint to nonexistent in the communities that use these models for brain, behavior, and cognition.

Getting Past Past-Tense

> [ANNs] are not perfect: they are not really explainable, they are not pliable, i.e., they cannot be easily modified to correct any errors observed, and they are not efficient due to the overhead of decoding. In contrast, rule-based methods are more transparent to subject matter experts; they are amenable to having a human in the loop through intervention, manipulation and incorporation of domain knowledge; and further the resulting systems tend to be lightweight and fast. (Chiticariu et al. 2023, p. iii)

In what is known in the literature as the past-tense debate (e.g., Elman et al., 1996; Pinker & Ullman, 2002), cognition and its underpinning substrates were discussed in terms of whether hard-wired capacities, such as grammatical rules for English past-tense formation, are encoded in the genes or otherwise without learning. Furthermore, claims were made about connectionist systems, such as, ANN "models cannot deal with languages such as Hebrew, where regular and irregular nouns are intermingled in the same phonological neighborhoods" (Pinker & Ullman, 2002, p. 459). While it may have been true for models at the time that certain data sets were unlearnable, or specific nondeep ANNs had limited learning abilities due to their architecture or training set or regimen, this both does not hold in the present day for certain data sets (discussed below) and continues to hold in the sense that there are data sets that are inaccessible to modeling endeavors using ANNs (see proof in van Rooij et al., 2024). Work such as Zhang et al. (2016, 2017) can serve to neutralize the claim that ANNs might struggle with certain unstructured data sets, for example, "where regular and irregular nouns are intermingled" (Pinker & Ullman, 2002, p. 459), by demonstrating that ANNs can learn utterly random mappings between inputs and outputs. Of course, such a finding about ANNs is also problematic to $\mathfrak{C}$-connectionists, who propose that in many cases similar input–output pairs are represented similarly inside the model's learned internal representations. And in return, anticonnectionists will and do explain that therefore connectionist models are overly powerful, "reducing connectionism to a universal statistical approximation technique rather than a source of empirical predictions" (Pinker & Ullman, 2002, p. 474). This is perhaps prescient; compare this to the *Goal* row in Table 1. The reality is complex because it is both the case that ANNs can learn an infinite set of impressive input–output mappings—hence all the hype—but it is *not* the case, and formally so, that they can learn *any* such mapping (van Rooij et al., 2024). We unpack this below.

Rehashing the past-tense debate is not useful (for our purposes), but learning from the mistakes and pitfalls of past rhetoric is useful to the practitioners who wish to carry out connectionist modeling. On the one hand, it may not come as a surprise to some that even at the birth of $\mathfrak{M}$-connectionism (circa 2010; Table 1) and to this day, the past-tense "veritable brouhaha" (Kirov & Cotterell, 2018) was and is discussed by practitioners (e.g. Corkery et al., 2019; Kohli et al., 2020; X. Ma & Gao, 2022; Oh et al., 2011; Seidenberg & Plaut, 2014; Westermann & Ruh, 2012).

On the other hand, ANNs, on the cusp of $\mathfrak{M}$-connectionism, are far from their days of being framed as flawed for being unable to compute XOR. They are now seemingly impervious to critique and in fact an old theoretical weakness is now coopted, reframed as a strength—these models are now upgraded to universal function approximators:

> According to the universal function approximation theorem, any sufficiently deep and sufficiently large network, given sufficient training data, learns to approximate any (continuous) function from input to output arbitrarily well. (Cybenko, 1989; Hornik, 1991)—Ma and Peters (2020, p. 7)

> Connectionism underwent a revival in the mid-1980s, primarily triggered by the development of back propagation, a learning algorithm that could be used in multilayer networks (Rumelhart et al., 1986). This advance dramatically expanded the representational capacity of connectionist models to the point where they were capable of approximating any function to arbitrary precision, bolstering hopes that paired with powerful learning rules any task could be learnable (Hornik et al., 1989). This technical advance led to a flood of new work as researchers sought to show that neural networks could reproduce the gamut of psychological phenomena, from perception to decision making to language processing. (McClelland et al., 1986; Rumelhart et al., 1986)—Jones and Love (2011, p. 172)

Notably, these statements do not follow one way or another. If a model is indeed a universal approximator for any function, why would scientists need to "show that neural networks could reproduce the gamut of psychological phenomena"? On the contrary, this is given if they are indeed so powerful (hence the critique above by Pinker & Ullman, 2002). To analyze this properly, as many miscommunications abound with respect to this period (Olazaran, 1996; Schmidhuber, 2015), what is proven by results such as Cybenko (1989), Hornik (1991), and Hornik et al. (1989) are not that ANNs can *find* a function approximation for any input–output mapping, but that in principle a model that looks like an ANN, that is could be built up of ANN components, can stand in for any function from a given class of functions.

First, this has nothing to do with backpropagation, as the learning algorithm is not implicated in the universal approximation proofs cited (Cybenko, 1989; Hornik, 1991; Hornik et al., 1989)—only relevant is the idea of multiple hidden unit layers, which was known at the time of the perceptrons controversy and proponents repeated the claim that multiple layers are the way forwards, asking for funding to develop such networks (Boden, 2006; McCorduck, 2004; Olazaran, 1996). And this property of ANNs was only proven to be the case by Ismailov (2023), so decades later, and for ANNs with two hidden unit layers for approximating continuous and discontinuous functions. Also,

> models with several successive nonlinear layers of neurons date back at least to the 1960s … and 1970s. [Additionally, a]n efficient gradient descent method for teacher-based Supervised Learning (SL) in discrete, differentiable networks of arbitrary depth called backpropagation (BP) was developed in the 1960s and 1970s[.] (Schmidhuber, 2015, p. 86)

So the retelling by connectionists (e.g. Kriegeskorte, 2015, and other examples above) is not entirely faithful to what is found in the literature, but does heavily figure in the narration patterns we describe herein. For example, the addition of more layers being perceived as pivotal even though such a property had preexisted ANNs falling out of fashion (Schmidhuber, 2015).

Second and more importantly, these are not equivalent claims: just because any arbitrary distribution can be (in principle) captured by a mixture of Gaussian functions, does not mean this mixture is easy to find in practice; just because for any given traveling salesperson problem there exists an optimal solution, does not mean we have this solution handy, that it is easy to find. Such confusions are the same as confusing P for NP in theoretical computer science. Perhaps the authors know this, but statements such as "connectionist models [are] capable of approximating any function to arbitrary precision" (Jones & Love, 2011, p. 170) allows for either interpretation (as do others, e.g., Kriegeskorte & Douglas, 2018). Importantly, ANN models are *not* "capable of approximating any function," if "capable of approximating" means they are able to find an approximation using their learning algorithm for the inputs and outputs given. Training ANNs with hidden units using backpropagation is NP-hard: the solutions might be out there, but there is no guarantee we can find them (Šíma, 1996; also see Colbrook et al., 2022; van Rooij et al., 2024). These kinds of, purposeful or otherwise, confusions or unclarities with respect to what can be computed with ANNs, have been present since their inception, for example, "anything that can be completely and unambiguously put into words, is ipso facto realizable by a suitable finite neural network" (originally presented in 1948, Von Neumann, 1988, p. 310; also see Boden, 1991; Skinner, 2012).

In the present landscape, $\mathfrak{M}$-connectionist stances might fall into dramatically different traps to those in the past-tense debate, but they also recapitulate some of the core tensions. In such contexts, not only are historical facts muddied, distorted, or fabricated (Olazaran, 1996), but also strengths become weaknesses and vice versa. To be set on more solid and consistent scientific footings, we must remain vigilant and weary of such trends which appear to repeat down the ages when it comes to connectionisms of all stripes. In other words, the canonical weakness-turned-strength as expressed by Pinker and Ullman (2002) and many others that ANNs are extremely expressive causes problems to this day. And so it is either (a) backpropagation is computationally implausible—why parallel it to humans?—or (b) ANNs are so statistically expressive as to be useless experimentally—why train and test them on data? This aspect is the double-edged sword that connectionists, and anticonnectionists, must carefully wield because there are formal constraints on what connectionism can and cannot do and what we can and cannot conclude (meta)theoretically.

Empirical Interface

> [W]e now use the [brain] itself, as its own [model], and I assure you it does nearly as well. (Carroll, 1893, p. 169)

The experimental typology shown in Figure 2 caricatures the two possible ways connectionists carry out their modeling endeavors. In other words, and in line with the (hyper)empiricism found in modern incarnations of cognitive, neuro-, and psychological sciences, connectionist empirical work is the primary way in which models are defended as useful or valid scientific accounts of brain, behavior, and cognition. The way in which both types of connectionism interface with observation is perhaps uncontroversial, given this, that is both require correlation between the models and the human data. However, the difference between the two has important repercussions for scientific inference within each of the $\mathfrak{C}$- and $\mathfrak{M}$-connectionisms.

On the left side of Figure 2 in blue is a simplified version of how $\mathfrak{C}$-connectionism interfaces with observation. $\mathfrak{C}$-connectionists observe phenomena, denoted by $\phi_i, \phi_j \in \Phi$ which they relate to neurocognitive systems, denoted by $\mathbb{S}$ (Equation 1). Scientists within $\mathfrak{C}$-connectionism also postulate mechanisms and/or functions for $\mathbb{S}$—they do this based on their reading of the literature, their own theoretical commitments about neurocognitive capacities, and so on. Using their theoretical commitments, they then build a connectionist model, $\mathbb{M}$. That is to say, $\mathbb{M}$ embodies an attempt to capture what is relevant about $\mathbb{S}$ in compliance with $\mathfrak{C}$-connectionism, recall left column of Table 1. When it comes to evaluating the scientific properties of $\mathbb{M}$, the methods are typical frequentist inferential statistics, as used when analyzing the data for $\mathbb{S}$, for example, to show differences between or within groups, and so on, as well as qualitative comparisons (e.g., Guest et al., 2020; Rogers et al., 2004; Tyler et al., 2000).

In $\mathfrak{C}$-connectionism, importantly, patterns of data found in the model are not taken to be scientifically relevant to understanding $\mathbb{S}$ in and of themselves, denoted by explicitly stating $\phi_k \notin \Phi$, where Φ is the set of phenomena of interest. What this means is that if a pattern of results not found in people is found in the model, it is not taken to mean anything. No claims of similarity are postulated between $\mathbb{S}$ and $\mathbb{M}$ *other* than behavioral (or otherwise) data on modeled tasks, on simulated experimental manipulations. If $\mathbb{M}$ is seen to perform in ways in which no evidence exists either way for human participants, more experiments can be run to check, but that is not because $\mathfrak{C}$-connectionists believe they are identical as systems, but because the model can be seen as a way to generate novel ideas for experiments or test the implications of our ideas (e.g., McClelland, 2009; Tyler et al., 2000). No ϕ_k, that is anything that $\mathbb{M}$ does, is seen as relevant to people *other* than as a model of simulated behaviors or patterns of neurocognitive data. What this means is that if the ANN displays behaviors not known to exist in, for example, participants' experimental data, no default assumptions about this possibly errant behavior are made with respect to the cognitive system. The model is not explored or tested on input–output mappings conceptualized as different to the tasks or capacities it was designed for; such unexpected or underdefined behavior therefore has no bearing on the model's standing one way or another. The model is just an implementation of a theory. It is not imbued with any extra properties, such as being an instance of the phenomenon under study or of a cognitive capacity.

In contrast, on the right side of Figure 2 in pink is a simplified version of how $\mathfrak{M}$-connectionism interfaces with observation. While things may appear the same, there are some deep differences between the two types of connectionism with respect to empirical interfacing, the attempt at mediation between theory and data (Equation 2). At the top of both panels, both types observe $\mathbb{S}$ —the similarity ends here, as in $\mathfrak{M}$ different principles are deployed to relate $\mathbb{S}$ and $\mathbb{M}$ and ϕ_k to Φ.

A typical $\mathfrak{M}$-connectionist will take some deep ANN off-the-shelf, that is often not building it from scratch, but adapting, fine-tuning (e.g., Demszky et al., 2023; Schrimpf et al., 2020, cf. Liesenfeld et al., 2023; Pasquinelli, 2017) an existing ANN created by machine learning researchers to be their $\mathbb{M}$. This practice means that the "computational mechanism" available to $\mathfrak{M}$-connectionism is definitionally always the same, regardless of the question, hypothesis, or conclusion, and as such, $\mathfrak{M}$-connectionism misses out on the chance to pick out causal structures via modeling (Guest et al., 2025). This is in contrast to $\mathfrak{C}$-connectionism, where typically bespoke models are handcrafted, including the inputs and outputs used to train and test the model. $\mathfrak{M}$-connectionist claims about the relationship between $\mathbb{S}$ and $\mathbb{M}$ are based not on properties woven into the model's design, but correlations over data extracted from $\mathbb{S}$ and $\mathbb{M}$. Furthermore and recalling Proposition 2, $\mathbb{S}$ and $\mathbb{M}$ are both seen as of the same kind, resulting in a map-territory merger (also see: Guest & Martin, 2023). In Figure 2, this is expressed by $\phi_k \in \Phi$, which is to say the behaviors expressed by, the phenomena seen in, the ANN are seen as qualitatively equivalent to those found in people, for all intents and purposes of equal standing.[6] When data from ϕ_k correlates with data from ϕ_i and/or ϕ_j, the model is seen to provide a theory for these neurocognitive phenomena and/or their related capacities. This slippage between model, theory, and phenomena is often seamless and argued for on the basis of statements such as both the human system $\mathbb{S}$ and ANN system $\mathbb{M}$ are so-called black boxes or otherwise hard to prima facie understand (cf. Sullivan, 2022); recall the *goal*, *question*, and *mechanism* rows in Table 1.

This all being said, we have described an idealized empirical interface for $\mathfrak{M}$-connectionism. Breaks in this mediation are not uncommon in the literature when claims such as "we have artificial models performing complex cognitive tasks at human performance level" (Perconti & Plebe, 2020, p. 2) are presented with neither critical thought on the success-to-truth inference (Guest & Martin, 2023; Titus, 2024) nor with literature references. Others have also noticed this:

> Hu et al. (2024) argued that "LLMs show strong and human-like grammatical generalization capabilities." Yet, as also noted in Leivada, Günther, and Dentella (2024), this claim is not backed up with human data. (Leivada, Dentella, & Günther 2024, p. 4)

Leivada, Dentella, and Günther (2024) noted how such statements break from the empirical grounding otherwise appealed to by such connectionists (Guest & Martin, 2023, for other such instances).[7] This appears worrisome for $\mathfrak{M}$-connectionism.

The Matrix Multiplication of Domination

> [E]ven if a connectionist system manifests intelligent behavior, it provides no understanding of the mind because its workings remain as inscrutable as those of the mind itself. (Shepard, 1988, p. 52)

In the current climate—"connectionist AI is 'drought-inducing computing'" (McQuillan, 2023)— ANNs appear to be an unstoppable force with direct implications to both the daily lives of cognitive and/or computational (neuro)scientists and people outside these fields (e.g., Adams et al., 2023; Andrews et al., 2024; Bender et al., 2021; Gebru & Torres, 2024; Li et al., 2023; McQuillan, 2022; Ovalle et al., 2023; Urai & Kelly, 2023; van Rooij et al., 2024).[8] These models, even if setting aside the harmful implications they have outside science, pose serious questions about the quality of our work and our (meta)theoretical reasoning within science. Herein, we offer a serious reimagining along formal and historical lines of the connectionist tendency within the cognitive sciences: a bifurcation into, modern post-2010 $\mathfrak{M}$-connectionism, and classical pre-2010, $\mathfrak{C}$-connectionism. This analysis, our metatheoretical calculus (embodied in Table 1, in modal and doxastic logic, and in Figure 2), serves to investigate connectionist rhetorical framings, bringing to light what and how cognitive science is done when ANNs are implicated or connectionism is appealed to. Ultimately, our work aims to foster critical thinking about how we do our science, allowing us to question if such forms of scientific reasoning are desirable to us, if connectionism as a framework should remain in its current modern form.

To recapitulate our main points, we trace the current framings in modern connectionism, to statements from classical connectionists such as "We wish to replace the 'computer metaphor' as a model of mind with the 'brain metaphor' as model of mind" (Rumelhart et al., 1986, p. 75). As well as, "Don't pre-wire structure into your mechanism if it can get it for free from the environment" (Plunkett, 2001, p. 193). And the portentous: "connectionism […] might lead to a different form of cognitive theory[, away from seeing] the human mind as rule-governed [because] certain phenomena […] can be neatly and economically dealt with by connectionist theories" (M. S. C. Thomas, 1998, n.p.). During this transitional period, wherein so-called symbolic cognitive scientists urge themselves to think deeply about what so-called rules might govern cognition, we see that connectionists in contrast may have avoided stopping to think "a different form of theory" (M. S. C. Thomas, 1998) might be no theory at all.

The technoscientific embedding of ANNs, the ideological commitments of connectionism, and the current and projected usage of such frameworks and resulting theories or computational models deserve critical engagement (as does all of cognitive science; Birhane & Guest, 2021; Carbajal et al., 2024; Prather et al., 2022). In this article, we have analyzed these factors with an emphasis on the scientific reasoning that

[6] This is the case provided the data extracted from $\mathbb{S}$ and $\mathbb{M}$ correlate. For more analysis on this also see section *Inference Rules in (Mis)use* in Guest and Martin (2023), which explains what happens when the data from model and phenomenon do not correlate.

[7] Quote modified to cite papers previously referenced as "in press."

[8] Section heading is inspired by Collins (1990): "a matrix of domination contains few pure victims or oppressors. Each individual derives varying amounts of penalty and privilege from the multiple systems of oppression which frame everyone's lives" (p. 229).

connectionism deploys in two proposed flavors, what we dubbed classical (pre-2010) and modern (post-2010) connectionisms (recall Table 1 and Figure 2), culminating in—a formal description of what we take as the beliefs of practitioners, the adjudication over theories they carry out—a metatheoretical calculus for each flavor.

Setting aside: the problem of induction and the underdetermination of theory by data (true for all science, of course, but often un(der) acknowledged here); the eschewing of multiple realizability within computationalism (of which connectionism is a fellow traveler; Chirimuuta, 2018, 2021; Egan, 2017; Figdor, 2010; Guest & Martin, 2021, 2023; Guest et al., 2025; Hardcastle, 1995, 1996; Litch, 1997; Polger & Shapiro, 2016; Ross, 2020; van Rooij et al., 2024); the gross ethics violations and the slippage into pseudoscience when the science and technology sector are intertwined (from polluting the environment to harming people through data breaches and pseudoscience or human rights violations; Andrews et al., 2024; Bender et al., 2021; Birhane & Guest, 2021; Dhaliwal et al., 2024; Forbes & Guest, 2025; Gebru & Torres, 2024; Guest, 2024; Liesenfeld et al., 2023; Li et al., 2023; McQuillan, 2023; Pasquinelli, 2017; Urai & Kelly, 2023)—so setting aside this heavy baggage—how can connectionism obtain a justifiable scientific purpose?[9] In other words, if we as a field allow ourselves to be unduly charitable to connectionism, what do we risk? We have argued that we risk both scientific rigor and theoretical substance. But it need not be that way.

Is connectionism redeemable? For one, connectionism can return to a $\mathfrak{C}$ form, and limit itself to checking whether clear-cut and transparent aspects of models facilitate functional and/or mechanistic explanatory accounts of cognitive phenomena (e.g., Guest et al., 2020; Tyler et al., 2000). For example, by clearly discussing externalist theoretical and modeling commitments—that is do details encoded in the simulated stimuli, the model of the environment, drive patterns of behavior seen in the simulation?—and internalist theoretical and modeling commitments—that is do different types of connectivity, or other internal model features, account for these patterns? Connectionism can also explicitly avoid the issues we outline as troubling for $\mathfrak{M}$-connectionism from a scientific lens (Table 1; also see Guest & Martin, 2023), in addition to those that are broader and have societal consequences.

Relatedly, and most importantly perhaps for the practitioners themselves as individuals, we must cultivate an understanding that models of this nature do not constitute theories as such nor do they constitute the phenomena we study in cognitive science (Figure 2; also see Guest, 2024; Guest & Martin, 2021, 2023, 2025; van Rooij et al., 2024). The confusion between ANNs and the phenomenon, that is the system under study, as well as the theory, that is the scientific understanding we are attempting to obtain, is a dangerous rhetorical circumstance. Computational models in cognitive science serve as mediators between the world of theory—our verbal and formal descriptions and explanations—and the empirical world—experiments, phenomena, brains, and people (also see Guest, 2024; Morgan & Morrison, 1999). Confusions between these three: model, theory, and system under study, are thoroughly unscientific and need to be addressed head on. If taken seriously, what has been outlined above forces us to contend with problems within modern connectionist thought. These three requirements set out the bare minimum for a realigning of scientific goals, for modelers and theoreticians within this framework, and specifically with our own stated scientific values and practice.

Finally, connectionism, and cognitive science generally, can rid ourselves of the hidden conflicts of interest inherent in taking industry funding to build and use such models (Forbes & Guest, 2025; Gerdes, 2022; Liesenfeld & Dingemanse, 2024; Liesenfeld et al., 2023). This is possible by requesting that we and our fellow practitioners disclose such conflicts during and at the point of publication. Relatedly, we need to acknowledge that such relationships to industry effectively bend our metatheoretical positions towards un-, or minimally a-, scientific reasoning that we are under obligation to keep in check if not at bay (also see Andrews et al., 2024; Bender et al., 2021; Birhane & Guest, 2021; Forbes & Guest, 2025; Gerdes, 2022; Guest, 2024; Spanton & Guest, 2022). Ultimately, it is up to us, theoreticians and modelers alike, to decide on the fate of our own fields and on the basis on which we create, understand, and reason about and over our models. Connectionism can be perhaps be redeemed, but it requires us to: sacrifice superficial understanding of what role models play and what they constitute; halt the "anything goes" antiscientific dictum of industry funding; and become aware of what follows from our reasoning when we engage mechanistic and/or functional explanations; and if done carelessly, we risk being incoherent or self-undermining. Snatching defeat from the jaws of victory seems to be connectionists' speciality, however the only difference may be that, this time round the stakes are higher both for science specifically and society at large.

[9] We do not mean to say to set aside these issues which relate to connectionism for all intents and purposes, but to assume they can be addressed by, for example, cautioning against less ideal forms of scientific reasoning (Guest & Martin, 2023, 2025; Guest et al., 2025), using smaller models (Dingemanse & Liesenfeld, 2022; Jain et al., 2024), and other relevant adjustments.

References

Abraham, T. H. (2002). (Physio)logical circuits: The intellectual origins of the Mcculloch-Pitts neural networks. *Journal of the History of the Behavioral Sciences*, *38*(1), 3–25. https://doi.org/10.1002/jhbs.1094

Abraham, T. H. (2012). Transcending disciplines: Scientific styles in studies of the brain in mid-twentieth century America. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(2), 552–568. https://doi.org/10.1016/j.shpsc.2012.02.001

Adams, C. J., Crary, A., & Gruen, L. (Eds.). (2023). *The good it promises, the harm it does: Critical essays on effective altruism*. Oxford University Press.

Aizawa, K. (1992). Connectionism and artificial intelligence: History and philosophical interpretation. *Journal of Experimental & Theoretical Artificial Intelligence*, *4*(4), 295–313. https://doi.org/10.1080/09528139208953753

Althaus, N., Gliozzi, V., Mayor, J., & Plunkett, K. (2020). Infant categorization as a dynamic process linked to memory. *Royal Society Open Science*, *7*(10), Article 200328. https://doi.org/10.1098/rsos.200328

Ananthaswamy, A. (2021). Deep neural networks help to explain living brains. *Quanta Magazine*. https://www.quantamagazine.org/deep-neural-networks-help-to-explain-living-brains-20201028/

Andrews, M., Smart, A., & Birhane, A. (2024). The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, *5*(9), Article 101027. https://doi.org/10.1016/j.patter.2024.101027

Arkoudas, K. (2008). Computation, hypercomputation, and physical science. *Journal of Applied Logic*, *6*(4), 461–475. https://doi.org/10.1016/j.jal.2008.09.007

Baayen, R. H., & Hendrix, P. (2017). Two-Layer networks, nonlinear separation, and human learning. In M. Kroon, G. van Noord, & G. Bouma (Eds.), *From semantics to dialectometry: Festschrift in honor of John Nerbonne* (Vol. 32, pp. 13–22). College Publications.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., & Sharma, U. (2024). Explaining neural scaling laws. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(27), Article e2311878121. https://doi.org/10.1073/pnas.2311878121

Barcan Marcus, R. (1961). Modalities and intensional languages. *Synthese*, *13*, 303–322. https://doi.org/10.1007/BF00486629

Barcan Marcus, R. (1967). Essentialism in modal logic. *Noûs*, *1*(1), 91–96. https://doi.org/10.2307/2214714

Barcan Marcus, R. (1990). Some revisionary proposals about belief and believing. *Philosophy and Phenomenological Research*, *50*, 133–153. https://doi.org/10.2307/2108036

Barcan Marcus, R. (1997). Are possible, non actual objects real? *Revue Internationale de Philosophie*, *51*, 251–257. https://www.jstor.org/stable/23954465

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Bersini, H. (1989). Connectionism vs Gofai: A brief critical analysis. In A. S. Jovanović, K. F. Kussmaul, A. C. Lucia, & P. P. Bonissone (Eds.), *Expert systems in structural safety assessment. Lecture notes in engineering* (Vol. 53, pp. 472–493). Springer. https://doi.org/10.1007/978-3-642-83991-7_25

Birhane, A., & Guest, O. (2021). Towards decolonising computational sciences. *Kvinder, Køn & Forskning*, *29*(1), 60–73. https://doi.org/10.7146/kkf.v29i2.124899

Blokpoel, M. (2018). Sculpting computational-level models. *Topics in Cognitive Science*, *10*(3), 641–648. https://doi.org/10.1111/tops.12282

Boden, M. A. (1991). Horses of a different color? In W. Ramsey, D. E. Rumelhart, & S. P. Stich (Eds.), *Philosophy and connectionist theory* (pp. 3–19). Psychology Press.

Boden, M. A. (2006). *Mind as machine: A history of cognitive science two-volume set*. Oxford University Press.

Carbajal, I., Moore, E., Cabrera Martinez, L., & Hunt, K. (2024). Critical cognitive science: A systematic review towards a critical science. *Journal of Social Issues*, *80*(1), 100–123. https://doi.org/10.1111/josi.12597

Carroll, L. (1893). Chapter 11: The man in the moon. *Sylvie and Bruno concluded*. Macmillan and Co. https://etc.usf.edu/lit2go/211/sylvie-andbruno-concluded/4652/chapter-11-the-man-in-the-moon/

Chirimuuta, M. (2018). Marr, Mayr, and MR: What functionalism should now be about. *Philosophical Psychology*, *31*(3), 403–418. https://doi.org/10.1080/09515089.2017.1381679

Chirimuuta, M. (2021). Your brain is like a computer: Function, analogy, simplification. In F. Calzavarini & M. Viola (Eds.), *Neural mechanisms. Studies in brain and mind* (Vol. 17, pp. 235–261). Springer. https://doi.org/10.1007/978-3-030-54092-0_11

Chiticariu, L., Hahn-Powell, G., Frietag, D., Riloff, E., Morrison, C. T., Sharp, R., Valenzuela-Escárcega, M., Surdeanu, M., & Noriega-Atala, E. (2023). *Proceedings of the 2nd workshop on pattern-based approaches to NLP in the age of deep learning*. https://aclanthology.org/2023.pandl-1.0.pdf

Cireşan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, *22*(12), 3207–3220. https://doi.org/10.1162/NECO_a_00052

Colbrook, M. J., Antun, V., & Hansen, A. C. (2022). The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. *Proceedings of the National Academy of Sciences of the United States of America*(12), Article e2107151119. https://doi.org/10.1073/pnas.2107151119

Collins, P. H. (1990). Knowledge, consciousness, and the politics of empowerment. In K. Deaux, M. M. Ferree, & V. Sapiro (Eds.), *Black feminist thought: Knowledge, consciousness, and the politics of empowerment* (Vol. 138, pp. 221–238). Unwin Hyman.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). *Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection*. arXiv.

Cybenko, G. V. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*, 303–314. https://doi.org/10.1007/BF02551274

Darden, L. (2006). Discovering mechanisms in neurobiology: The case of spatial memory with Carl F. Craver. *Reasoning in biological discoveries: Essays on mechanisms, interfield relations, and anomaly resolution* (pp. 40–64). Cambridge University Press. https://doi.org/10.1017/CBO9780511498442

Dawson, M. R. W. (2013). New powers of old networks. In C. Houlihan (Ed.), *Mind, body, world—Foundations of cognitive science* (p. 490). Athabasca University Press.

Dechter, R. (1986). *Learning while searching in constraint-satisfaction problems* [Conference session]. AAAI-86 Proceedings, Philadelphia, Pennsylvania, United States.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Dhaliwal, R. S., LePage-Richer, T., & Suchman, L. (2024). *Neural networks*. Meson Press.

Dijkgraaf, R. (2021). *The uselessness of useful knowledge*. https://www.quantamagazine.org/science-has-entered-a-new-era-of-alchemy-good-20211020/

Dingemanse, M., & Liesenfeld, A. (2022). From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. *1*, pp. 5614–5633).

Dodig-Crnkovic, G. (2023). *Computational natural philosophy: A thread from presocratics through turing to ChatGPT*. arXiv.

Dreyfus, H. L. (1965). *Alchemy and artificial intelligence*. https://www.rand.org/content/dam/rand/pubs/papers/2006/P3244.pdf

Dujmović, M., Malhotra, G., & Bowers, J. (2020). *What do adversarial images tell us about human vision?* bioRxiv.

Egan, F. (1995). Folk psychology and cognitive architecture. *Philosophy of Science*, *62*(2), 179–196. https://doi.org/10.1086/289851

Egan, F. (2017). Function-theoretic explanation and the search for neural mechanisms. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science* (pp. 145–163). Oxford University Press.

Elgin, C. Z. (2009). Construction and cognition. *THEORIA: Revista de Teoría, Historia Y Fundamentos de la Ciencia*, *24*(2), 135–146. https://doi.org/10.1387/theoria.439

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. The Massachusetts Institute of Technology Press.

Erscoi, L., Kleinherenbrink, A. V., & Guest, O. (2023). *Pygmalion displacement: When humanising AI dehumanises women*. https://doi.org/10.31235/osf.io/jqxb6

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philosophy of Science*, *77*(3), 419–456. https://doi.org/10.1086/652964

Flis, I., & van Eck, N. J. (2018). Framing psychology as a discipline (1950–1999): A large-scale term co-occurrence analysis of scientific literature in psychology. *History of Psychology*, *21*(4), 334–362. https://doi.org/10.1037/hop0000067

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

Folk, R. L., & Campbell, D. H. (1992). Are the pyramids of Egypt built of poured concrete blocks? *Journal of Geological Education*, *40*(1), 25–34. https://doi.org/10.5408/0022-1368-40.1.25

Forbes, S. H., & Guest, O. (2025). To improve literacy, improve equality in education, not large language models. *Cognitive Science*, *49*(4), Article e70058. https://doi.org/10.1111/cogs.70058

Forsythe, D. E. (1993). Engineering knowledge: The construction of knowledge in artificial intelligence. *Social Studies of Science*, *23*(3), 445–477. https://doi.org/10.1177/0306312793023003002

Gardner, E. (1988). The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, *21*(1), Article 257. https://doi.org/10.1088/0305-4470/21/1/030

Gebru, T., & Torres, É. P. (2024). *The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence*. First Monday.

Gefter, A. (2015). The man who tried to redeem the world with logic. *Nautilus*, *21*, 106–154. https://www.aaas.org/sites/default/files/Amanda%20Gefter%20(4).pdf

Gerdes, A. (2022). The tech industry hijacking of the AI ethics research agenda and why we should reclaim it. *Discover Artificial Intelligence*, *2*(1), Article 25. https://doi.org/10.1007/s44163-022-00043-3

Gershman, S. J. (2024). What have we learned about artificial intelligence from studying the brain? *Biological Cybernetics*, *118*, 1–5. https://doi.org/10.1007/s00422-024-00983-2

Gödel, K. (1992). *On formally undecidable propositions of principia mathematica and related systems*. Dover Publications.

Goldberg, A. E. (2008). Universal grammar? Or prerequisites for natural language? *Behavioral and Brain Sciences*, *31*(5), 522–523. https://doi.org/10.1017/S0140525X0800513X

Gordon, D. F., & Desjardins, M. (1995). Evaluation and selection of biases in machine learning. *Machine Learning*, *20*(1–2), 5–22. https://doi.org/10.1023/A:1022630017346

Green, C. D. (1999). Are connectionist models theories of cognition? *Challenges to Theoretical Psychology Selected/Edited Proceedings of the Seventh Biennial Conference of the International Society for Theoretical Psychology*. Captus Press.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. https://doi.org/10.1177/0963721412447619

Guest, O. (2024). What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*, *7*, 508–522. https://doi.org/10.1007/s42113-023-00193-2

Guest, O., Caso, A., & Cooper, R. P. (2020). On simulating neural damage in connectionist networks. *Computational Brain & Behavior*, *3*(3), 289–321. https://doi.org/10.1007/s42113-020-00081-z

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*, 213–227. https://doi.org/10.1007/s42113-022-00166-x

Guest, O., & Martin, A. E. (2025). *Are neurocognitive representations 'small cakes'?* https://philsci-archive.pitt.edu/24834/

Guest, O., Scharfenberg, N., & van Rooij, I. (2025). *Modern alchemy: Neurocognitive reverse engineering*. https://philarchive.org/rec/GUEMAN

Haigh, T. (2023). There was no 'first AI winter'. *Communications of the ACM*, *66*(12), 35–39. https://doi.org/10.1145/3625833

Hamilton, S. N. (1998). Incomplete determinism: A discourse analysis of cybernetic futurology in early cyberculture. *Journal of Communication Inquiry*, *22*(2), 177–204. https://doi.org/10.1177/0196859998022002005

Hardcastle, V. G. (1995). Computationalism. *Synthese*, *105*, 303–317. https://doi.org/10.1007/BF01063561

Hardcastle, V. G. (1996). *How to build a theory in cognitive science*. State University of New York Press.

Hattab, H. (2009). *Descartes on forms and mechanisms*. Cambridge University Press.

Hay, J. C., Lynch, B. E., & Smith, D. R. (1960). *Mark I perceptron operators' manual* (Report No. VG-1196-G-5). Cornell Aeronautical Lab. https://apps.dtic.mil/sti/tr/pdf/AD0236965.pdf

Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (pp. 599–619). Springer Berlin Heidelberg.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(36), Article e2400917121. https://doi.org/10.1073/pnas.2400917121

Ismailov, V. E. (2023). A three layer neural network can represent any multivariate function. *Journal of Mathematical Analysis and Applications*, *523*(1), Article 127096. https://doi.org/10.1016/j.jmaa.2023.127096

Jain, S., Vo, V. A., Wehbe, L., & Huth, A. G. (2024). Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, *5*(1), 80–106. https://doi.org/10.1162/nol_a_00101

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. https://doi.org/10.1017/S0140525X10003134

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, Article 1726. https://doi.org/10.3389/fpsyg.2017.01726

Just, J., & Ghosal, S. (2019). *Deep generative models strike back! Improving understanding and evaluation in light of unmet expectations for OoD data*. arXiv.

Kindig, B. (2024). *AI spending to exceed a quarter trillion next year*. Forbes. https://www.forbes.com/sites/bethkindig/2024/11/14/ai-spending-to-exceed-a-quarter-trillion-next-year/

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/10.48550/arXiv.1807.04783

Kohli, M., Magoulas, G. D., & Thomas, M. S. (2020). Evolving connectionist models to capture population variability across language development: Modeling children's past tense formation. *Artificial Life*, *26*(2), 217–241. https://doi.org/10.1162/artl_a_00316

Krickel, B. (2024). Different types of mechanistic explanation and their ontological implications. In J. L. Cordovil, G. Santos, & D. Vecchi (Eds.), *New mechanism. History, philosophy and theory of the life sciences* (Vol. 35, pp. 9–28). Springer. https://doi.org/10.1007/978-3-031-46917-6_2

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks* [Conference session]. Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, United States. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kubilius, J. (2018). Predict, then simplify. *NeuroImage*, *180*, 110–111. https://doi.org/10.1016/j.neuroimage.2017.12.006

Leivada, E., Dentella, V., & Günther, F. (2024). Evaluating the language abilities of large language models vs. humans: Three caveats. *Biolinguistics*, *18*, Article e14391. https://doi.org/10.5964/bioling.14391

Leivada, E., Günther, F., & Dentella, V. (2024). Reply to Hu et al.: Applying different evaluation standards to humans vs. large language models overestimates AI performance. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(36), Article e2406752121. https://doi.org/10.1073/pnas.2406752121

LePage-Richer, T. (2024). *Neural networks*. Meson Press.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making ai less" thirsty": Uncovering and addressing the secret water footprint of AI models*. arXiv.

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI act. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1774–1787).

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). *Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators* [Conference session]. Proceedings of the 5th International Conference on Conversational User Interfaces, Eindhoven, The Netherlands. https://doi.org/10.1145/3571884.3604316

Lighthill, J. (1972). *Artificial intelligence: A paper symposium*. Science Research Council. https://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm

Lisman, J., Buzsáki, G., Eichenbaum, H., Nadel, L., Ranganath, C., & Redish, A. D. (2017). Viewpoints: How the hippocampus contributes to memory, navigation and cognition. *Nature Neuroscience*, *20*(11), 1434–1447. https://doi.org/10.1038/nn.4661

Litch, M. (1997). Computation, connectionism and modelling the mind. *Philosophical Psychology*, *10*(3), 357–364. https://doi.org/10.1080/09515089708573225

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, *7*(2), 230–242. https://doi.org/10.1111/tops.12131

Ma, W., & Peters, B. (2020). *A neural network walks into a lab: Towards using deep nets as models for human behavior*. arXiv.

Ma, X., & Gao, L. (2022). *How do we get there? Evaluating transformer neural networks as cognitive models for English past tense inflection*. arXiv.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Massachusetts Institute of Technology Press.

Marx, K. (1894). *Capital: Volume III*. International Publishers.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

McClelland, J. L., Rumelhart, D. E., & Group, P. R. (1986). *Parallel distributed processing, volume 2: Explorations in the microstructure of cognition: Psychological and biological models*. Massachusetts Institute of Technology Press.

McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/CRC Press. https://en.wikipedia.org/wiki/CRC_Press

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*, 115–133. https://doi.org/10.1007/BF02478259

McLaughlin, B. P., & Warfield, T. A. (1994). The allure of connectionism reexamined. *Synthese*, *101*, 365–400. https://doi.org/10.1007/BF01063895

McQuillan, D. (2022). *Resisting AI: An anti-fascist approach to artificial intelligence*. Policy Press.

McQuillan, D. (2023). *Connectionist AI is "drought-inducing computing*." https://x.com/danmcquillan/status/1722562742031090094

Medler, D. A. (1998). A brief history of connectionism. *Neural Computing Surveys*, *1*, 18–72. https://people.engr.tamu.edu/rgutier/web_courses/cpsc636_s10/medler1998briefHistoryConnectionism.pdf

Mok, R. M., & Love, B. C. (2023). A multilevel account of hippocampal function in spatial and concept learning: Bridging models of behavior and neural assemblies. *Science Advances*, *9*(29), Article eade6903. https://doi.org/10.1126/sciadv.ade6903

Morgan, M. S., & Morrison, M. (1999). *Models as mediators*. Cambridge University Press.

Nagel, E. (1979). *The structure of science* (Vol. 411). Hackett Publishing Company.

Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, *16*(4), 707–716. https://doi.org/10.1177/1745691620974769

Nichani, E., Radhakrishnan, A., & Uhler, C. (2020). *Increasing depth leads to U-shaped test risk in over-parameterized convolutional networks*. arXiv.

NOVA. (1997). *Who built the pyramids?* https://www.pbs.org/wgbh/nova/pyramid/explore/builders.html

Oh, T. M., Tan, K. L., Ng, P., Berne, Y. I., & Graham, S. (2011). The past tense debate: Is phonological complexity the key to the puzzle? *NeuroImage*, *57*(1), 271–280. https://doi.org/10.1016/j.neuroimage.2011.04.008

Olazaran, M. (1996). A sociological study of the official history of the perceptrons controversy. *Social Studies of Science*, *26*(3), 611–659. https://doi.org/10.1177/030631296026003005

Ovalle, A., Subramonian, A., Gautam, V., Gee, G., & Chang, K.-W. (2023). Factoring the matrix of domination: A critical review and reimagination of intersectionality in AI fairness. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 496–511).

Pasquinelli, M. (2017). Machines that morph logic: Neural networks and the distorted automation of intelligence as statistical inference. *Glass Bead*, *1*(1), 1–17. https://www.glass-bead.org/article/machines-that-morph-logic/

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), Article 20220041. https://doi.org/10.1098/rsta.2022.0041

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition*, *203*, Article 104365. https://doi.org/10.1016/j.cognition.2020.104365

Pfaffenberger, B. (1988). Fetishised objects and humanised nature: Towards an anthropology of technology. *Man*, *23*, 236–252. https://doi.org/10.2307/2802804

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463. https://doi.org/10.1016/S1364-6613(02)01990-3

Plunkett, K. (2001). Connectionism today. *Synthese*, *129*, 185–194. https://doi.org/10.1023/A:1013099222414

Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford University Press.

Prather, R. W., Benitez, V. L., Brooks, L. K., Dancy, C. L., Dilworth-Bart, J., Dutra, N. B., Faison, M. O., Figueroa, M., Holden, L. R., Johnson, C., Medrano, J., Miller-Cotto, D., Matthews, P. G., Manly, J. J., & Thomas, A. K. (2022). What can cognitive science do for people? *Cognitive Science*, *46*(6), Article e13167. https://doi.org/10.1111/cogs.13167

Rich, P., Blokpoel, M., de Haan, R., & van Rooij, I. (2020). How intractability spans the cognitive and evolutionary levels of explanation. *Topics in Cognitive Science*, *12*(4), 1382–1402. https://doi.org/10.1111/tops.12506

Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). *How hard is cognitive science?* PsyArXiv.

Richter, M. L., Schöning, J., Wiedenroth, A., & Krumnack, U. (2021). Should you go deeper? Optimizing convolutional neural network architectures without training by receptive field analysis. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 964–971).

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235. https://doi.org/10.1037/0033-295X.111.1.205

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rosenblatt, F. (1959). A probabilistic model for visual perception. *Acta Psychologica*, *15*, 296–297. https://doi.org/10.1016/S0001-6918(59)80143-8

Rosenblatt, F. (1960). Perceptron simulation experiments. *Proceedings of the IRE*, *48*, 301–309. https://api.semanticscholar.org/CorpusID:51656509

Ross, L. N. (2020). Multiple realizability from a causal perspective. *Philosophy of Science*, *87*(4), 640–662. https://doi.org/10.1086/709732

Ross, L. N., & Bassett, D. S. (2024). Causation in neuroscience: Keeping mechanism meaningful. *Nature Reviews Neuroscience*, *25*(2), 81–90. https://doi.org/10.1038/s41583-023-00778-7

Rumelhart, D. E., McClelland, J. L., & the Parallel Distributed Processing Research Group. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT Press.

Ryle, G. (1949). *The concept of mind*. Barnes & Noble.

Saini, A. (2019). *Superior: The return of race science*. Beacon Press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

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

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational neuroscience. *Science*, *241*(4871), 1299–1306. https://doi.org/10.1126/science.3045969

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, *1*(1), 145–168.

SETI Institute. (1984). https://www.seti.org/history-seti-institute

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28), 1–9. https://doi.org/10.1126/sciadv.abm2219

Shepard, R. N. (1988). How fully should connectionism be activated? Two sources of excitation and one of inhibition. *Behavioral and Brain Sciences*, *11*(1), 52. https://doi.org/10.1017/S0140525X00052729

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(10), Article e2300963120. https://doi.org/10.1073/pnas.2300963120

Šíma, J. (1996). Back-propagation is not efficient. *Neural Networks*, *9*(6), 1017–1023. https://doi.org/10.1016/0893-6080(95)00135-2

Skinner, R. E. (2012). *Building the second mind: 1956 and the origins of artificial intelligence computing*. UC Berkeley.

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

Spanton, R. W., & Guest, O. (2022). *"Measuring trustworthiness or automating physiognomy? A comment on Safra, Chevallier, Grèzes, and Baumard (2020)."* arXiv.

Stinson, C. (2018). Explanation and connectionist models. In M. Sprevak & M. Colombo (Eds.), *The routledge handbook of the computational mind* (pp. 120–133). Routledge.

Stinson, C. (2020). From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence. *Philosophy of Science*, *87*(4), 590–611. https://doi.org/10.1086/709730

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044–2064. https://doi.org/10.1162/jocn_a_01755

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, *73*(1), 109–133. https://doi.org/10.1093/bjps/axz035

Thomas, M. S. C. (1998). Connectionism is a progressive research programme. *Psycoloquy*. https://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?9.36

Thomas, M. S. C., & McClelland, J. (2008). Connectionist models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 23–58). Cambridge University Press.

Thomas, R. (2021). *Why do so many people still think aliens built the pyramids?* https://www.vice.com/en/article/g5bnpm/why-do-so-many-people-still-think-aliens-built-the-pyramids

Thompson, J. A. F. (2021). Forms of explanation and under- standing for neuroscience and artificial intelligence. *Journal of Neurophysiology*, *126*(6), 1860–1874. https://doi.org/10.1152/jn.00195.2021

Titus, L. M. (2024). Does ChatGPT have semantic understanding? a problem with the statistics-of-occurrence strategy. *Cognitive Systems Research*, *83*, Article 101174. https://doi.org/10.1016/j.cogsys.2023.101174

Tyler, L., Moss, H., Durrant-Peatfield, M., & Levy, J. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, *75*(2), 195–231. https://doi.org/10.1006/brln.2000.2353

Urai, A. E., & Kelly, C. (2023). Rethinking academia in a time of climate crisis. *eLife*, *12*, Article e84991. https://doi.org/10.7554/eLife.84991

Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.

van der Gun, L., & Guest, O. (2024). Artificial intelligence: Panacea or non-intentional dehumanisation? *Journal of Human-Technology Relations*, *2*(1). https://doi.org/10.59490/jhtr.2024.2.7272

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*(6), 939–984. https://doi.org/10.1080/03640210801897856

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science.

Perspectives on Psychological Science, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, *7*, 616–636. https://doi.org/10.1007/s42113-024-00217-5

Von Neumann, J. (1988). *John von neumann*. American Mathematical Society.

Westermann, G., & Ruh, N. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review*, *119*(3), 649–667. https://doi.org/10.1037/a0028258

Whitehead, A. N., & Russell, B. (1910). *Principia mathematica* (Vol. 1). Cambridge University Press.

Wilson, E. A. (2016). *Neural geographies: Feminism and the microstructure of cognition*. Routledge.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). *Understanding deep learning requires rethinking generalization*. arXiv.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). *Understanding deep learning requires rethinking generalization* [Conference session]. International Conference on Learning Representations, Palais des Congrès Neptune, Toulon, France. https://openreview.net/forum?id=Sy8gdB9xx

Received November 25, 2024
Revision received June 11, 2025
Accepted July 30, 2025 ■