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, The Netherlands
[2]Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
[3]Language and Computation in Neural Systems Group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Contemporary AI 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 (ANNs): *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 ANNs and cognition, brains, or behaviour (Guest & Martin, 2023), we use metatheoretical analysis techniques (Guest, 2024), including formal logic, to characterise two distinct tendencies within connectionism that we dub classical and modern, with divergent properties, e.g. goals, mechanisms, 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 ANNs, our field can obviate the cycle of AI winters and summers, which need not be inevitable.

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

> 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
>
> Margaret A 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 boom in artificial intelligence (AI) research (Hamilton, 1998; also see 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) wondrously are as applicable now as when they were written 35 years ago. In this paper, we aim to critique and juxtapose modern connectionist stances with those around prior to 2010 when it can be argued the classical, pre-deep learning, (cf. Dechter, 1986) pre-widespread GPU 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 non-scientific engineering-oriented AI uses, e.g. face recognition to unlock a smartphone).

The analysis presented in this paper is centred 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 $\mathfrak{C}$, and post-2010, which we call modern connectionism and $\mathfrak{M}$ (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, e.g. the rise of using ANNs as models of the brain and cognition, (Kriegeskorte, 2015a; 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 rigour. To this end, we construct a *metatheoretical calculus* (viz. 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 parallel distributed processing, neural network modeling, or connectionism[. A] term introduced by Donald Hebb in the

Olivia Guest https://orcid.org/0000-0002-1891-0972
Acknowledgements: We would like to thank Nils Donselaar, Iris van Rooij, R. Mellema, Laura van de Braak, Rineke Verbrugge, Todd Wareham, and the CCS group for discussions and feedback. We also thank Leonidas A. A. Doumas and Iris van Rooij for discussion of the universal approximation claims.
Correspondence: Olivia Guest, Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands. E-mail: olivia.guest@donders.ru.nl

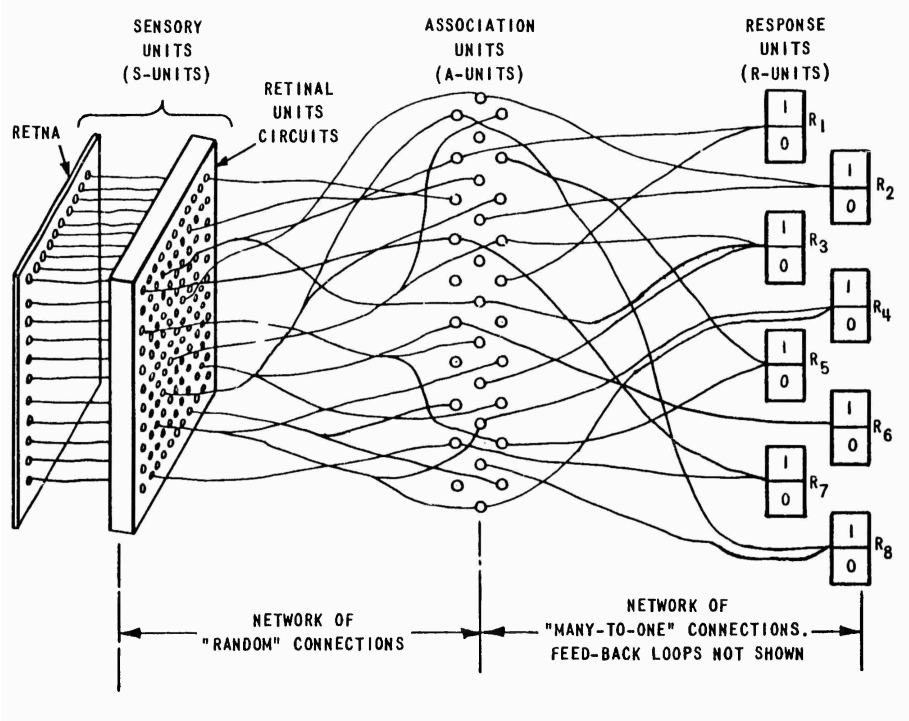

Figure 1: "Organization of the Mark I perceptron" (Hay et al., 1960, Figure 1, p. 13): a hardware ANN built by Frank Rosenblatt's Cognitive Systems Section at Cornell Aeronautical Laboratory Inc. (also see Rosenblatt, 1958, 1959, 1960) From their inception, ANNs have had a proposed parallel structure to biological neural systems and have been forwarded as a model of human cognition. In the figure above, 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 modelling, and phraseology like "microstructure of cognition" (Rumelhart et al., 1986) versus function approximation (viz. Egan, 2017; Guest et al., 2024; van Rooij & Baggio, 2021). Connectionism resides at the inflections of these often contrasting ideological positions (viz. Pasquinelli, 2017).

1940's" (p. xii Elman et al., 1996). 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 non-linear squashing function, e.g. hyperbolic tangent. Learning is achieved, e.g. 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 behaviour and some target state, e.g. 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, e.g. drawing parallels between input units and retinal cells (as seen on the leftmost part of Figure 1).

We will next analyse the relationships within and between connectionist tendencies, using Guest (2024) as a way to tease out (meta)theoretical properties. First, in section 2, **Metaphysical commitment**, 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 (subsection 2.1, **Identity: What characterises connectionism?**) and from each other (subsection 2.2, **Separation: What differentiates types of connectionism?**). This exercise provides the building blocks for the formal accounts given in later sections.

Second, in section 3, **Discursive survival**, 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 modelling strategy and as a methodology, has been subject to targetted attacks, e.g. claims about inability to compute certain functions, like XOR, for dealing with non-linearly separable data. We analyse this along two broad lines: the ability to craft a coherent reaction to attacks (subsection 3.1, **Argumentation**), and the ability to tell a coherent story about the scientific theory (subsection 3.1, **Narration**). We formally describe the two types of connectionist reasoning using modal and doxastic logic in Proposition 1 for $\mathfrak{C}$- and Proposition 2 for $\mathfrak{M}$-connectionism.

Finally, in section 4, **Empirical Interface**, 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 formalise how practitioners within connectionist tendencies reason about their experimental manipulations and their modelling praxis.

Table 1: A collection of perhaps contradictory (meta)theoretical claims or commitments between older, classical versus newer, modern connectionist tendencies. We do not assume that this dichotomy characterises 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 programme.

	$\mathfrak{C}$-connectionism: classical, pre-2010	$\mathfrak{M}$-connectionism: modern, post-2010
Goal	The goal is understanding the repercussions of our theories, i.e. "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 behaviour. 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)
Question	Can connectionist principles give rise to similar behaviour 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, behavioural 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, e.g. "theory [is] instantiated in task-performing computational models" (Kriegeskorte & Douglas, 2018, p. 4). Additionally, theorising is (often) inspired by engineered systems, not nature directly, e.g. "current computational neuroscience practice [looks to] AI [which] has historically provided a fund of ideas for biological theories." (Gershman, 2023, p. 4)
Mechanism	Mechanisms are proposed, which the model embodies, and experiments are done to show proof of concept, i.e. 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 modelled — not as uncovered correlationally. Theory, or some knowledge of the to-be-modelled, -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 behaviourist or associationist stance and the load placed on the training regime and set, which in the case of ANNs is statistics in the input, e.g. "[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, e.g. photographs (e.g. Jozwik et al., 2017; Storrs et al., 2021).

2 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".
>
> Jerry A Fodor and Zenon W 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 characterise 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 feel compelled to present them side-by-side to further understanding and heighten awareness of these changes in connectionist (meta)theorising. 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 — 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 totalising view of method as theory and of map as territory — and not as a function of verbal theory, formal specification, and other modeller choices.

2.1 Identity: What characterises 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.
>
> Frank 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 modelled organism is doing, i.e. 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 (viz. 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 characterises 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, i.e. 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 modelling — 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 -brain matching, (e.g. 'brain-score', Schrimpf et al., 2020; viz Pasquinelli, 2017).

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 behaviour (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 and McClelland, 2008, pp. 28–29, also Smolensky, 1988, but cf. Elman et al., 1996; McLaughlin and Warfield, 1994) Additionally, appealing to innateness and nativism — i.e. 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*' (viz. Goldberg, 2008).[1] Similar entanglements with neural and behavioural data and observation emerge with the so-called 'bridging' of levels, a newer catchphrase (e.g. Griffiths et al., 2012; Love, 2015; Mok and Love, 2023; cf. Elgin, 2009; Nagel, 1979). This idea emerges in both types of connectionisms, but especially in 𝔐. And indeed such discussions implicate and betray a misunderstanding of Marr's levels, i.e. a confusion between his levels of analysis versus broader levels of description of the world (cf. Blokpoel, 2018; Rich et al., 2020; van Rooij & Baggio, 2021). Catherine Stinson (2018, p. 126) explains how "[c]onnectionists talk about taking constraints from both physiology and psychol-

[1] As Elman et al. (1996) and others acknowledge, innateness and nativism are not clear-cut concepts and the polarised, or even completely 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 characterisation of human-likeness is the case only in complex or dubious ways (e.g. Leivada, Günther, et al., 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 very field(s) that use them for their scientific endeavours. 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).

ogy, as though they are employing an inferential pincer movement[, but really this is just words to little effect since] there are no halting conditions" for the search for such bridges between levels (viz. 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)

2.2 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.
>
> Robbert Dijkgraaf (2021, n.p.)

Connectionism can be seen as a dramatically divergent way of doing science. In the earlier argumentation regarding $\mathfrak{C}$-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).
>
> Jeffrey L. Elman et al. (1996, p. 5)

While this is not only true in more recent incarnations of $\mathfrak{C}$-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.
>
> Warren S. McCulloch and Walter 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) 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 modelling and perspectives on cognition (viz. 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 theorising to propose propositional logic-based neurons as a universal substrate, e.g. "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 $\mathfrak{M}$-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, $\mathfrak{M}$-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 datasets. 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 modelling techniques used in the relevant fields connectionism typically touches on (see the $\mathfrak{M}$-connectionism column of Table 1). As such, $\mathfrak{M}$-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 (e.g. see Guest & Martin, 2021).

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 analysed in depth in Guest and Martin (2023), but we will return to both in the following sections.

3 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.
>
> Margaret A Boden (1991, p. 11)

We propose connectionism has a survival strategy honed by decades of proverbial summers and winters. We will analyse 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.

3.1 Argumentation

3.1.1 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.
>
> Brian P McLaughlin and Ted A Warfield (1994, p. 382)

In the case of $\mathfrak{C}$-connectionism, we see a typical scientific framing for the modelling endeavour. In the most zoomed out case, they posit $\mathbb{M}_p$, which they believe to be the case, e.g. 'ANNs can learn this input-output mapping,' where $\mathbb{M}$ is the model and p is the mapping, task, or capacity being modelled; 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 modelled using connectionist methodologies, then one constructs an ANN such that it does; formally:

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

where $\mathbb{S}_{p'}$ is the system under study observed as performing p', which appears equivalent to, or is formalised or modelled 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). $\mathcal{B}(p' \sim p)$ 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.[4] $\mathcal{B}(\mathbb{M}_p)$ represents our belief in 'ANNs can learn some input-output mapping,' or in 'ANNs can show behaviour similar to that seen in humans doing a task,' or in 'ANNs can compute internal representations that are theoretically useful. So $\mathcal{B}(\mathbb{M}_p)$ is the explicitly stated belief that p is modellable 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, 1991, 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 endeavour to create such a model. In all cases p is not typically general, but a specific example of conceptualising a subset of human cognition, such as a capacity, series of behavioural tasks, a disorder, etc. 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, $\mathcal{C}$ 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 modelled 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 modelling. Purely because of ANN's 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 (viz. Olazaran, 1996), but once nothing was stopping them, all they need to do is add more layers, or minimally one hidden pool of units.

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

3.1.2 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.
>
> Rhys Thomas (2021, n.p.)

Connectionism in its modern incarnation can be seen as often applying conspiratorial-like thinking to scientific reasoning. To analogise, 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

[4] We do not propose it is unique to $\mathfrak{C}$-connectionism, but a common framing of how we work (viz. 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'.

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 sensible scenarios still implicate a huge swathe of potential options to choose from, e.g. with respect to ramp types attached onto the Pyramids to aid in construction (recall multiple realisability, Chirimuuta, 2018, 2021; Egan, 2017; Figdor, 2010; Guest & Martin, 2023; 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 scrutinise the relationships between statements to understand the conclusion does not follow (recall how in Guest and Martin, 2023, we criticise 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 behaviours appear to be cognition, then they have human capacities.

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

$$\mathcal{O}(\mathbb{S}_{p'}) \wedge \mathcal{B}(p' \equiv p) \wedge \mathcal{O}(\mathbb{M}_{p}) \rightarrow \mathcal{B}(\mathbb{M}_{p'}) \tag{2}$$

where each symbol is as before; $\mathcal{O}(\mathbb{S}_{p'})$ is the observation that p' occurs in the system under study, $\mathcal{B}(p' \equiv p)$ expresses that $\mathfrak{M}$-connectionists believe that p', the phenomenon or capacity people do is the same as p, the behaviour of the model; $\mathcal{O}(\mathbb{M}_{p})$ is the observation that the ANN can perform p, recall Figure 2; and $\mathcal{B}(\mathbb{M}_{p'})$ 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, 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. These jumps in logic can be found in many jocular stories, e.g. 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? $\mathfrak{M}$-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:

> 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. Nonetheless, it keeps surviving as a rhetorical strategy in our scientific endeavours.

3.2 Narration

3.2.1 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.
>
> Paul Smolensky (1988, p. 67)

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

On the robustness point, however, $\mathfrak{C}$-connectionism adapts to critique. For example, connectionism started evaluating the so-called internal representations in ANNs (recall rows *Theory* and *Mechanism* in Table 1) before $\mathfrak{M}$-connectionist 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; i.e. 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, i.e. 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 𝔈-connectionism comprises acknowledging and understanding fit-to-data is not enough (recall rows *Goal* and *Quenstion* in Table 1). "Now all cognitive connectionists will agree that simulation alone is not explanation." (M. S. C. Thomas, 1998, n.p.) However, M. S. C. Thomas (1998, n.p.) follows with: "Connectionist models must have constraints, and those constraints must be supported by empirical data" with no reference to theory as a factor. Notwithstanding, 𝔈-connectionism presents a coherent research programme: "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 theorising, and in which they concede fitting the data is a red herring. This presages 𝔐-connectionism's theoretical stance, or lack thereof.

The story 𝔈-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 very strongly their case. Aspects of learning problems like non-linear separability remain under-explored 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 protection from exposure to critique or improvement attempts. 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 analysing 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 dataset 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 (viz. Olazaran, 1996), that adding more layers to networks is (or at least was until very recently) seen as the mantra for fixing (m)any problems (viz. Dawson, 2013; Medler, 1998). By the same token, apparently serious problems with (understanding) overfitting, over-parametrisation, or the behaviour 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 non-existent in the communities that use these models for brain, behaviour, and cognition.

3.2.2 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.
>
> Laura Chiticariu et al. (2023, p. iii)

In the past-tense debate (e.g. Elman et al., 1996; Pinker & Ullman, 2002), cognition and its underpinning substrates were discussed in terms of whether hardwired 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) Work such as Zhang et al. (2016, 2017) can serve to neutralise this claim by demonstrating that ANNs can learn utterly random mappings between inputs and outputs. Of course, such a finding about ANNs is also problematic to 𝔈-connectionists, who propose that in many cases similar input-output pairs are represented similarly inside the model's learned internal representations. And in return, anti-connectionists 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.

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 modelling. On the one hand, it may not come as a surprise to some that even at the birth of 𝔐-connectionism (circa 2010; recall 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 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).
>
> Wei Ma and Benjamin 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).
>
> Matt Jones and Bradley C. 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 analyse this properly, as many miscommunications abound with respect to this period, (viz. Olazaran, 1996; Schmidhuber, 2015) what is proven by results such as Cybenko (1989), Hornik (1991), and Hornik et al. (1989) is not that ANNs can *find* a function approximation for any input-output mapping, but that in principle a model that looks like an ANN, i.e. 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 (viz. Boden, 2006; McCorduck & Cfe, 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[.]
>
> Jürgen Schmidhuber (2015, p. 86)

So the retelling by connectionists (e.g. Kriegeskorte, 2015b, and other examples above) is not entirely faithful to what is found in the literature, but does heavily figure the narration patterns we describe herein. For example, the addition of more layers being perceived as pivotal even though such a property had pre-existed 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 travelling 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, e.g. "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 anti-connectionists, must carefully wield because it has formal constraints on what connectionism can and cannot do and what we can and cannot conclude (meta)theoretically.

4 Empirical Interface

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

The experimental typology shown in Figure 2 caricatures the two possible ways connectionists carry out their modelling endeavours. 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, behaviour, and cognition. The way in which both types of connectionism interface with observation is perhaps completely uncontroversial, given this, i.e. 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 $\phi_i, \phi_j \in \mathbf{\Phi}$ which they relate to neurocognitive systems, denoted by $\mathbb{S}$. 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 com-

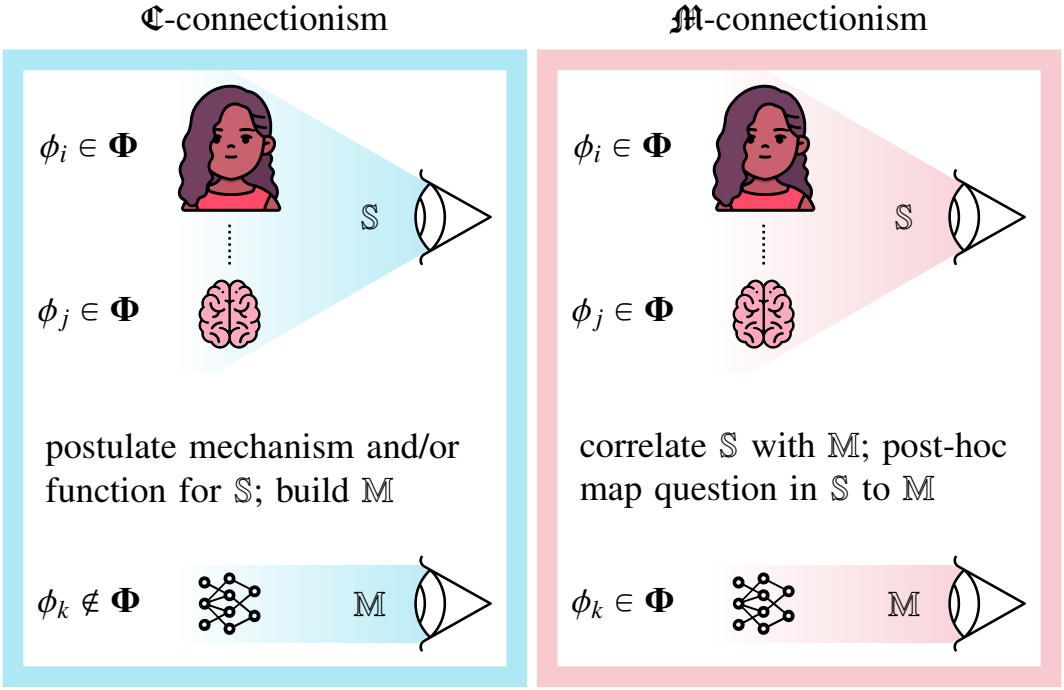

Figure 2: A cartoon depiction of the simplified differences between 𝔈- and 𝔐-connectionism with respect to postulating mechanisms and building models (collectively $\mathbb{M}$), and relating them to the cognitive and neural systems (collectively $\mathbb{S}$). On the left, in blue and in the top panel, we see an eye, which represents the scientist working under 𝔈-connectionism, looking at human behaviour and cognition (phenomena ϕ_i) and the brain (phenomena ϕ_j) all of which comprise $\mathbb{S}$, the system under study. All the observed phenomena here are very explicitly in $\mathbf{\Phi}$, 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 𝔈-connectionism does to model ϕ_i and ϕ_j above: create a model that embodies what they think connectionism can do, i.e. 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: 𝔐-connectionists document, witness, research ϕ_i and ϕ_j of $\mathbb{S}$. However, under 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, e.g. goodness-of-fit, are what make $\mathbb{M}$ a useful scientific model. Thus, scientific questions (recall Table 1) in $\mathbb{S}$ are matched post-hoc to aspects of $\mathbb{M}$. Indeed, behaviours of $\mathbb{M}$ are often seen as worthy of scientific investigation as if $\phi_k \in \mathbf{\Phi}$ from the perspective of connectionism, potentially subverting the field's goals.

mitments, 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 𝔈-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 analysing the data for $\mathbb{S}$, e.g. 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 𝔈-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 $\phi_k \notin \mathbf{\Phi}$, where $\mathbf{\Phi}$ 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 behavioural (or otherwise) data on modelled 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 𝔈-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, i.e. anything that $\mathbb{M}$ does, is seen as relevant to people *other* than as a model of simulated behaviours or patterns of neurocognitive data. 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 𝔐-connectionism interfaces with observation. While things may appear the same, there are some very deep differences between the two types of connectionism with respect to empirical interfacing, the attempt at mediation between theory and data, etc. At the top of both panels, both types observe $\mathbb{S}$ — the similarity ends here, as in 𝔐 completely different principles are deployed to relate $\mathbb{S}$ and $\mathbb{M}$.

A typical 𝔐-connectionist will take some deep ANN off-the-shelf, i.e. 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 𝔐-connectionism is definitionally always the same, regardless of the question, hypothesis, or conclusion, and as such, 𝔐-connectionism misses out on the chance to pick out causal structures via modelling. This is in contrast to 𝔈-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, recalling what we unpack in Guest and Martin (2023) and Proposition 2, i.e. that because of these correlations, $\mathbb{S}$ and $\mathbb{M}$ are both of the same kind, a map-territory merger. In Figure 2, this is expressed by $\phi_k \in \mathbf{\Phi}$, which is to say the behaviours 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.[5] 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 and theory is often completely 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 idealised 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 the reason why no literature references are provided. Others have also noticed this:

> Hu et al. (2024) argue that "LLMs show strong and human-like grammatical generalization capabilities". Yet, as also noted in Leivada, Günther, et al. (2024), this claim is not backed up with human data.
>
> Evelina Leivada, Vittoria Dentella, et al. (2024, p. 4)[6]

Leivada, Dentella, et al. (2024) note how such statements break from the empirical grounding otherwise appealed to by such connectionists (viz. Guest & Martin, 2023, for other such instances). This appears worrisome for $\mathfrak{M}$-connectionism.

5 The matrix multiplication of domination[7]

> [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.
>
> Roger N 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; Bender et al., 2021; Gebru & Torres, 2024; McQuillan, 2022; Ovalle et al., 2023; Urai & Kelly, 2023; van Rooij et al., 2024). These models, even if setting aside the indubitably 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 transitionary 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, viz. Birhane & Guest, 2021; Carbajal et al., 2024; Prather et al., 2022). In this paper we have analysed these factors with an emphasis on the scientific reasoning that connectionism deploys in two proposed flavours, 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 flavour.

Setting aside[8] 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 realisability within computationalism (of which connectionism is a fellow traveller; Chirimuuta, 2018, 2021; Egan, 2017; Figdor, 2010; Guest & Martin, 2021, 2023; 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 vio-

[5] 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.

[6] Quote modified to cite papers previously referenced as 'in press'.

[7] This 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."

[8] We do not mean to say to set these aside for all intents and purposes, but to assume they can be addressed by, e.g. addressing less ideal forms of scientific reasoning, using smaller models, etc.

lations; Andrews et al., 2024; Bender et al., 2021; Birhane & Guest, 2021; Gebru & Torres, 2024; Guest, 2024; Liesenfeld et al., 2023; McQuillan, 2023; Pasquinelli, 2017; Urai & Kelly, 2023). Given this heavy baggage, how can connectionism obtain a justifiable scientific purpose? 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 rigour 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 modelling commitments — i.e. do details encoded in the simulated stimuli, the model of the environment, drive patterns of behaviour seen in the simulation? — and internalist theoretical and modelling commitments — i.e. 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 (recall 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 (recall Figure 2; also see Guest, 2024; Guest & Martin, 2021, 2023; van Rooij et al., 2024). The confusion between ANNs and the phenomenon, i.e. the systemunder study, as well as the theory, i.e. 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, people (also see Guest, 2024; Morgan & Morrison, 1999).Confusions between these three: model, theory, and system under study, are not only thoroughly unscientific, they are dangerous 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 modellers 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. 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 completely at bay (also see Andrews et al., 2024; Bender et al., 2021; Birhane & Guest, 2021; Guest, 2024). Ultimately, it is up to us, theoreticians and modellers 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' anti-scientific 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.

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.

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.

Adams, C. J., Crary, A., Gruen, L., Adams, C. J., Crary, A., & Gruen, L. (Eds.). (2023, May 12). *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.

Althaus, N., Gliozzi, V., Mayor, J., & Plunkett, K. (2020). Infant categorization as a dynamic process linked to memory. *Royal Society Open Science*, *7*(10), 200328.

Ananthaswamy, A. (2021). Deep neural networks help to explain living brains. *Quanta Magazine*, *25*. 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*, 101027.

Arkoudas, K. (2008). Computation, hypercomputation, and physical science. *Journal of Applied Logic*, *6*(4), 461–475.

Baayen, R. H., & Hendrix, P. (2017). Two-layer networks, nonlinear separation, and human learning. *From Semantics to Dialectometry. Festschrift in honor of John Nerbonne. Tributes*, *32*, 13–22.

Barcan Marcus, R. (1961). Modalities and intensional languages. *Synthese*, 303–322.

Barcan Marcus, R. (1967). Essentialism in modal logic. *Noûs*, 91–96.

Barcan Marcus, R. (1991). Some revisionary proposals about belief and believing. *Causality, Method, and Modality: Essays in Honor of Jules Vuillemin*, 143–173.

Barcan Marcus, R. (1997). Are possible, non actual objects real? *Revue internationale de philosophie*, 251–257.

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*, *116*(32), 15849–15854.

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*, 610–623.

Bersini, H. (1989). Connectionism vs gofai: A brief critical analysis. *Lecture Notes in Engineering*, 472–493.

Birhane, A., & Guest, O. (2021). Towards decolonising computational sciences. *Kvinder, Køn & Forskning*, (2), 60–73.

Blokpoel, M. (2018). Sculpting computational-level models. *Topics in cognitive science*, *10*(3), 641–648.

Boden, M. A. (1991). Horses of a different color? In *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, USA.

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.

Carroll, L. (1893). Chapter 11: The Man in the Moon. In *Sylvie and Bruno Concluded*. https://etc.usf.edu/lit2go/211/sylvie-and-bruno-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.

Chirimuuta, M. (2021). Your brain is like a computer: Function, analogy, simplification. *Neural mechanisms: New challenges in the philosophy of neuroscience*, 235–261.

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. *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*.

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.

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*, *119*(12).

Collins, P. H. (1990). Black feminist thought in the matrix of domination. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*, *138*(1990), 221–238.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. *arXiv preprint arXiv:1906.01280*.

Cybenko, G. V. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*, 303–314. https://api.semanticscholar.org/CorpusID:3958369

Dawson, M. R. W. (2013). New powers of old networks. In *Mind, Body, World — Foundations of Cognitive Science*. Athabasca University Press.

Dechter, R. (1986). Learning while searching in constraint-satisfaction problems. *AAAI-86 Proceedings*.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.

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

Dodig-Crnkovic, G. (2023). Computational Natural Philosophy: A Thread from Presocratics through Turing to ChatGPT.

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.

Egan, F. (2017). Function-theoretic explanation. *Explanation and integration in mind and brain science*, 145–163.

Elgin, C. Z. (2009). Construction and cognition. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *24*(2), 135–146.

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 MIT Press.

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philosophy of Science*, *77*(3), 419–456.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

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.

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

Gardner, E. (1988). The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, *21*(1), 257.

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.

Gershman, S. J. (2023). What have we learned about artificial intelligence from studying the brain?

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.

Gordon, D. F., & Desjardins, M. (1995). Evaluation and selection of biases in machine learning. *Machine Learning*, *20*(1–2), 5–22.

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*.

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.

Guest, O. (2024). What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*.

Guest, O., Caso, A., & Cooper, R. P. (2020). On simulating neural damage in connectionist networks. *Computational Brain & Behavior*, *3*(3), 289–321.

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science*, *0*(0), 1745691620970585.

Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*.

Guest, O., Scharfenberg, N., & van Rooij, I. (2024). Modern alchemy: Neurocognitive reverse engineering.

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

Hardcastle, V. G. (1995). Computationalism. *Synthese*, *105*, 303–317.

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. *Cornell Aeronautical Lab., Buffalo, NY, Rept. No. VG-1196-G-5*.

Hinton, G. E. (2012). A Practical Guide to Training Restricted Boltzmann Machines. In *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.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, *4*(2), 251–257.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

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*, *121*(36), e2400917121.

Ismailov, V. E. (2023). A three layer neural network can represent any multivariate function. *Journal of Mathematical Analysis and Applications*, *523*(1), 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.

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.

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*, 1726.

Just, J., & Ghosal, S. (2019). Deep Generative Models Strike Back! Improving Understanding and Evaluation in Light of Unmet Expectations for OoD Data. *arXiv preprint arXiv:1911.04699*.

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.

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.

Kriegeskorte, N. (2015a). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417–446.

Kriegeskorte, N. (2015b). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*(1), 417–446.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kubilius, J. (2018). Predict, then simplify. *NeuroImage*, *180*, 110–111.

Leivada, E., Dentella, V., & Günther, F. (2024). Evaluating the language abilities of large language models vs. humans: Three caveats. *Biolinguistics*, *18*.

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*, *121*(36), e2406752121.

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. *Proceedings of the 5th International Conference on Conversational User Interfaces*.

Litch, M. (1997). Computation, connectionism and modelling the mind. *Philosophical Psychology*, *10*(3), 357–364.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in cognitive science*, *7*(2), 230–242.

Ma, W., & Peters, B. (2020). A neural network walks into a lab: Towards using deep nets as models for human behavior. *ArXiv*, *abs/2005.02181*.

Ma, X., & Gao, L. (2022). How do we get there? Evaluating transformer neural networks as cognitive models for English past tense inflection. *arXiv preprint arXiv:2210.09167*.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

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* (Vol. 2). MIT press.

McCorduck, P., & Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/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.

McLaughlin, B. P., & Warfield, T. A. (1994). The allure of connectionism reexamined. *Synthese*, *101*, 365–400.

McQuillan, D. (2022). Resisting AI.

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

Medler, D. A. (1998). A brief history of connectionism. *Neural computing surveys*, *1*, 18–72.

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), eade6903.

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

Nagel, E. (1979). *The structure of science* (Vol. 411). Hackett publishing company Indianapolis.

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*, 174569162097476.

Nichani, E., Radhakrishnan, A., & Uhler, C. (2020). Increasing Depth Leads to U-Shaped Test Risk in Over-parameterized Convolutional Networks. *arXiv preprint arXiv:2010.09610*.

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.

Olazaran, M. (1996). A sociological study of the official history of the perceptrons controversy. *Social Studies of Science*, *26*(3), 611–659.

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*, 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.

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

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition*, *203*, 104365.

Pfaffenberger, B. (1988). Fetishised objects and humanised nature: Towards an anthropology of technology. *Man*, 236–252.

Pinker, S., & Ullman, M. T. (2002). The past-tense debate. *cognitive processing*, *5*, 7.

Plunkett, K. (2001). Connectionism today. *Synthese*, *129*, 185–194.

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).

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.

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

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)*, 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.

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

Rosenblatt, F. (1959). A probabilistic model for visual perception. *Acta Psychologica*, *15*, 296–297.

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.

Rumelhart, D. E., McClelland, J. L., & Group, P. R. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations* (Vol. 1). The MIT press.

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

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

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*. https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive science*, *38*(6), 1190–1228.

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational neuroscience. *Science*, *241*(4871), 1299–1306.

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)*, 1–8.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28).

Shepard, R. N. (1988). How fully should connectionism be activated? two sources of excitation and one of inhibition. *Behavioral and Brain Sciences*, *11*(1), 52–52.

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10).

Šíma, J. (1996). Back-propagation is not efficient. *Neural Networks*, *9*(6), 1017–1023.

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

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

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

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.

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*.

Thomas, M. S. C. (1998). Connectionism is a progressive research programme. *Psycoloquy*.

Thomas, M. S. C., & McClelland, J. (2008). Connectionist models of cognition. In R. Sun (Ed.), *The cambridge handbook of computational psychology*. 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 understanding for neuroscience and artificial intelligence. *Journal of Neurophysiology*, *126*(6), 1860–1874.

Titus, L. M. (2024). Does chatgpt have semantic understanding? a problem with the statistics-of-occurrence strategy. *Cognitive Systems Research*, *83*, 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.

Urai, A. E., & Kelly, C. (2023). Rethinking academia in a time of climate crisis. *eLife*, *12*.

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.

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*, 1–21.

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

Westermann, G., & Ruh, N. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review*, *119*(3), 649.

Whitehead, A. N., & Russell, B. (1910). *Principia mathematica volume 1* (1st). 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 preprint arXiv:1611.03530*.

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