Modularity is the Bedrock of Natural and Artificial Intelligence

Anonymous authors
Paper under double-blind review

Abstract

The astonishing performance showcased by AI systems in the last decade has been achieved through the use of massive amounts of data, computation, and, in turn, energy, which vastly exceed what human intelligence requires. This wide gap underscores the need for further research and points to leveraging brains as a valuable source of guiding principles. On the other hand, the No Free Lunch Theorem highlights that effective inductive biases must be problem-specific. This suggests designing architectures with specialized components that can solve subproblems — namely, modular architectures. Interestingly, modularity is an established principle of brain organization that is considered essential for supporting the efficient learning and strong generalization abilities consistently demonstrated by humans. However, despite its recognized importance in natural intelligence and the proven benefits it has shown across various seemingly unrelated AI research areas, modularity remains somewhat underappreciated in AI. In this work, we review several research threads in artificial intelligence and neuroscience through a conceptual framework that highlights the central role of modularity in supporting both artificial and natural intelligence. In particular, we examine what computational advantages modularity provides, how it has emerged as a solution in several AI research areas, which modularity principles the brain exploits, and how modularity can help bridge the gap between natural and artificial intelligence.

1 Introduction

Present-day AI systems can execute well-defined tasks with a proficiency that appeared unattainable just a few years back. Convolutional Neural Networks (CNNs) (He et al., 2016) classify images with superhuman accuracy (Russakovsky et al., 2015); Reinforcement Learning (RL) agents can defeat elite professional players in complex strategy games with intractable search spaces, such as Go (Silver et al., 2016); Large Language Models (LLMs) excel at natural language processing tasks (Brown et al., 2020), master professional and academic knowledge (Achiam et al., 2023), coding[1][2], and appear[3] to be able to perform abstract reasoning (Chollet, 2019).

However, these feats are accomplished with massive quantities of data, computation, and, in turn, energy, which vastly exceed what human intelligence necessitates. For example, CNNs can reach high classification accuracy only after being trained on thousands of examples for each class (LeCun et al., 1998; Russakovsky et al., 2015), whereas humans can do so even after being exposed to a single example (Lake et al., 2015). To make things worse, CNNs typically have to be presented with several variations of the same instance over different epochs through data augmentation. Similarly, RL agents, such as AlphaGo (Silver et al., 2016), can achieve superhuman performance only after being trained for at least three orders of magnitude more games than elite human players, who excel in a much wider array of tasks (Lake et al., 2017). Finally, the impressive skills of current LLMs require datasets of tens of trillions of tokens (Dubey et al., 2024), which would take an above-average reader (Brysbaert, 2019) 5-50 thousand years of continuous reading. This

[1] https://www.anthropic.com/news/claude-3-5-sonnet
[2] https://openai.com/index/learning-to-reason-with-llms/
[3] https://arcprize.org/blog/oai-o3-pub-breakthrough

poor sample efficiency is accompanied by poor energy efficiency: training a model such as GPT-3 (Brown et al., 2020) was estimated to consume 1287 megawatt hours (MWh), an amount of energy that would power over 100 average American households for a year — roughly three orders of magnitude more than the 3.15 MWh required to power a 20W human brain (Sokoloff, 1960) for 18 years. The energy consumption of next-generation LLMs, and thus their resulting carbon footprint, is likely to worsen further over the coming years as more recent models are rumored to have at least 10X more parameters[4] and rely heavily on additional test-time computations (Yao et al., 2024; Snell et al., 2024; Guan et al., 2025) due to the emerging performance gains[3].

Despite their reliance on enormous datasets and energy resources, current AI systems typically still struggle with out-of-distribution (OOD) generalization (Geirhos et al., 2020; Yuan et al., 2023; Mayilvahanan et al., 2024), have poor compositional skills (Lake & Baroni, 2018; Hupkes et al., 2020; Dziri et al., 2024), and suffer from catastrophic forgetting and limited positive transfer (French, 1999; Luo et al., 2023; Huang et al., 2024), while humans, typically, do not (Lake et al., 2017; Lake & Baroni, 2018; Ito et al., 2022). How can we move beyond such limitations? The No Free Lunch Theorem (Wolpert et al., 1995; Wolpert & Macready, 1997) suggests that no inductive bias can lead to the design of a general-purpose architecture that is a universal problem solver. Thus, good inductive biases need to be problem-specific (Sinz et al., 2019). It follows that rather than attempting to build monolithic architectures with good inductive biases, we should focus on building problem-specific modules. Therefore, tackling some of the limitations of current AI systems by leveraging *modularity principles* — whereby specialized modules work synergistically to solve complex tasks requiring new combinations of learned skills — appears to be a particularly promising research avenue.

Modular architectures were indeed proposed early on (Grossberg, 1976; Rueckl et al., 1989; Jacobs et al., 1991a;b; Happel & Murre, 1994) as a way to both capture the modular design of brains and speed up learning, and they were shown to perform competently while reproducing selected features of brain activation patterns. Currently, however, the usage of modular designs is driven by two main motivations: the growing need to efficiently reuse large pre-trained models across different domains while minimizing the costs associated with fine-tuning, and the ambition to design architectures capable of performing multi-task and, ideally, continual learning (Pfeiffer et al., 2023; Yadav et al., 2024). Nevertheless, a closer look reveals that modular architectures have also emerged as a powerful solution to diverse challenges across various AI research areas, suggesting a broader convergence toward modular designs (Andreas et al., 2016; Rusu et al., 2016; Guo et al., 2025). Moreover, modularity principles are virtually omnipresent in modern AI systems, as deep neural networks (DNNs) have an inherently modular structure composed of stacked layers — an inductive bias that grants them significant computational advantages (Poggio et al., 2017). Interestingly, as we will illustrate in § 4, brains are also modular (Fodor, 1983; Simon, 1962; Meunier et al., 2009), and modularity is believed to be core to the learning efficiency and robust generalization abilities humans possess (Zhang et al., 2024).

Thus, it appears that modularity principles have yet to receive the attention they deserve in AI. One of the aims of this survey is to highlight their importance across both artificial and natural intelligence, and to suggest that they could serve as a core design principle in AI systems and a valuable focus for future research. Furthermore, because modular architectures can be difficult to design effectively, we propose that insights from the brain may help identify the fundamental functions that modules should specialize in, offering a promising avenue for bridging the gap between natural and artificial intelligence.

In the remainder of this work, we first discuss why modularity is a fundamental design principle in engineering (§ 2.1) with clear robustness advantages. We then present evidence of its widespread presence in complex natural systems (§ 2.2). Next, we introduce a formalism to describe modular AI frameworks (§ 2.3), review influential work that sheds light on the computational advantages they provide (§ 3.1,§ 3.2), and examine their growing adoption across several AI research areas (§ 3.3). Finally, we survey key modularity principles the brain is thought to exploit (§ 4), discuss alternative perspectives and open research questions (§ 5), and summarize the main takeaways (§ 6).

[4] `https://en.wikipedia.org/wiki/GPT-4`

2 Modularity Principles

A system is modular if it is composed of subsystems whose structural elements are strongly connected among themselves and weakly connected to elements of other subsystems (Rumelhart et al., 1986; Baldwin & Clark, 1999). In a typical modular system, the modules are *specialized* — i.e., excel at performing specific functions — *sparsely interacting* — i.e., can exchange information when needed — and *largely autonomous* — i.e., do not rely on other modules to perform their functions. This organization naturally implements a divide-and-conquer strategy, whereby a complex problem is decomposed into sub-problems that modules are specialized in solving. This decomposition is achieved through *information factorization*, which fosters both specialization and robustness: since the modules receive only information that is relevant to perform their function, they are invariant to (and thus robust against) irrelevant information and can effectively specialize in processing their inputs (Merel et al., 2019).

2.1 Modularity in Engineering

Modularity principles are widely applied in engineering (Suh, 1990), particularly in software (Booch et al., 2008) and hardware (Baldwin & Clark, 1999) design, and are regarded as foundational to building scalable and robust systems (Lipson et al., 2007). A good overview of the properties underlying their success in engineering is provided by the six fundamental *operators* identified by Baldwin & Clark (1999): splitting, substituting, augmenting, excluding, inverting, and porting. *Splitting* allows breaking a complex problem into simpler, module-specific, sub-problems. *Substituting* enables replacing one module with an upgraded version. *Augmenting* allows adding new modules to the system (providing either new functionalities or enhanced robustness through the introduction of redundancy). *Excluding* allows removing unnecessary modules. *Inverting* facilitates the creation of better design rules that leverage the existing modules more efficiently. *Porting* allows reusing existing modules in new systems. These unique properties of modular systems are often considered fundamental to the evolution of technology, fostering a continuous cycle of refinement where existing components are incrementally improved or new ones are seamlessly integrated into the systems.

2.2 Modularity in Nature

Modularity principles, as we have argued in the previous section, are widespread in artificial systems as they foster the design of scalable, robust, and increasingly sophisticated systems. Is there evidence of similar principles also in the natural world? It turns out that most complex systems[5] such as biological, physical, social, and symbolic systems are also widely regarded as being modular, with modules often arranged in hierarchies (Simon, 1962; Callebaut & Rasskin-Gutman, 2005; Newman, 2006). For example, vertebrates can be understood as organ systems, where each system performs a specialized function. These systems are composed of organs, which carry out sub-functions. Moving further down the hierarchy, we encounter tissues, cells, organelles, proteins, polypeptides, amino acids, etc. At every level of this hierarchy, we observe systems composed of modules, each often dedicated to specialized functions. It has been theorized that the reason why most complex systems are hierarchically modular is because this organization promotes the formation of relatively stable, long-lived, intermediate units — the modules — that mitigate the natural tendency toward disorder acting on their constituent parts (Schrödinger & Penrose, 1992; Simon, 1962). As these intermediate units typically exhibit emergent functionalities and will also tend to aggregate into stable super-units, systems will tend toward greater complexity and sophistication.

The modularity of biological systems (Wagner et al., 2001; 2007) has been typically studied also in relation to evolution and has been hypothesized to favor evolvability: the ability to flexibly adapt to new environments. According to this theory, modular designs allow selective pressure to optimize each module separately without interference (Hansen, 2003). Influential simulation studies clarified this mechanism further and suggested that modularity in biological networks arises in response to changing environments (Lipson et al., 2002), that this effect is particularly strong when the environment changes in a modular manner (i.e., keeping sub-requirements constant Kashtan & Alon (2005)), and that it significantly speeds up adaptation (Kashtan et al., 2007). Interestingly, this finding may explain why more recent studies have found that multi-task

[5]Simon (1962) considered complex all the systems composed of multiple interacting parts interacting in non-trivial ways

learning promotes modularity in CNNs (Dobs et al., 2022) and RNNs (Yang et al., 2019). Additional studies highlighted that modularly changing environments are not the only potential determinant of modular networks, as the minimization of connection costs in constant environments also leads to modular solutions (Clune et al., 2013).

Finally, we note that while modular networks have been studied in the context of evolution, they have also served as powerful models for understanding the computational principles of natural intelligence. For instance, studies suggest that a bias toward short connections — which offer clear energetic benefits — naturally leads to modular networks. These networks decompose tasks into subtasks (Jacobs & Jordan, 1992), exhibit greater resilience to catastrophic forgetting in continual learning (Ellefsen et al., 2015), demonstrate brain-like mixed selectivity and low average activation patterns (Achterberg et al., 2023), and form sparse information streams while reusing useful features (Liu et al., 2023). Taken together, these findings highlight some of the computational advantages that brains may rely on by virtue of their modular organization (Meunier et al., 2009).

2.3 Definition

A more formal way to define a modular model in machine learning is the following. Given an input x, a model $f(x) : \mathbb{R}^i \to \mathbb{R}^o$ is modular with modules $\mathcal{M} = \{m_{\mu_i}(x)\}_{i=1}^M$ if it can be rewritten as $f(x) = \phi(m_{\mu_1}(x), m_{\mu_2}(x), ..., m_{\mu_M}(x))$, where $m_{\mu_i}(x)$ is the i-th module with parameters μ_i. Typically, modular models have a *routing function* $r_\rho(x) : \mathbb{R}^i \to 2^{\mathcal{M}}$ with parameters ρ, which determines which modules are active, and an *aggregation function* $g_\gamma(x) : 2^{\mathcal{M}} \to \mathbb{R}^o$ with parameters γ, which determines how the modules are aggregated. Thus, we can rewrite a modular model more compactly as $f(x) = g_\gamma(r_\rho(x))$[6]. Typically, in modular networks different tasks elicit distinct network behavior, so the input x often includes task information (e.g., it might be a concatenation of input features and task embeddings). Also, note that modules may operate on different input projections. Here, we assume these projections are extracted within the modules; however, they might also be extracted by the routing function.

Routing can be *hard* — when the modules are either active or inactive — or *soft* — when all modules are active with probability p_i. Critically, hard routing leads to sparse models, which are particularly efficient during inference as signals only need to propagate through selected modules. However, these models cannot be trained end-to-end via gradient descent and require specialized training techniques such as reinforcement learning (e.g., Rosenbaum et al. (2017)), evolutionary algorithms (e.g., Fernando et al. (2017)), or stochastic parametrization (e.g., Sun et al. (2020)). On the other hand, models with soft routing can be trained end-to-end via gradient descent but are not sparse and thus can be fairly computationally expensive.

Aggregation functions often define simple operations, such as the weighted summation: $g_\gamma(x) = \sum_{j \in r_\rho(x)} \alpha_j \, m_{\mu_j}(x)$ (as in Jacobs et al. (1991b); Shazeer et al. (2017)); however, in some other cases, they can also define more complex, attention-based operations, for example, with the input x as a query, and the active modules' outputs Z as key and values: $g_\gamma(x) = \text{Attn}(xQ, ZK, ZV)$, where Q, K, and V are matrices of learned parameters (e.g., as in Pfeiffer et al. (2020a)).

Finally, we note that, in some contexts, such as efficient transfer learning (Pfeiffer et al. (2023), cf. § 3.3.3), modules are employed to fine-tune pre-trained models; in this scenario, they are interspersed between the layers of the base model to modify their behavior. In this case, one needs to also define a *modifier* function $d_\delta^{(f)}(\cdot)$ to specify how the modular model f modifies the behavior of the base network's layer $l_\lambda(x)$. In this case, modifier functions tend define simple operations such as summations: $d_\delta^{(f)}(l_\lambda(x)) = l_\lambda(x) + f(x)$ (as in Ansell et al. (2021)), and function compositions $d_\delta^{(f)}(l_\lambda(x)) = f(l_\lambda(x))$ (as in Rebuffi et al. (2017)) in order to minimally alter the base network's behavior.

We refer the reader to Pfeiffer et al. (2023) for a more in-depth overview of the routing, aggregation, and modifier functions used in modular models.

[6]For more complex inter-module interactions, one can introduce multiple layers $y = f_{l+1}(f_l(x))$ or recurrence $x_{t+1} = f(x_t)$

3 Modularity in Artificial Intelligence

In the previous sections, we have presented evidence that modularity is a fundamental feature of complex natural systems, conferring numerous advantages. We have also shown how this insight has inspired engineers to formalize the benefits of modularity principles and harness them to design more robust artificial systems. Here, we demonstrate that these principles have also strongly influenced AI.

In fact, modularity was an essential design feature of early-stage connectionist (Jacobs et al., 1991a), cybernetic (Wiener, 2019) and symbolic (Kautz, 2022) AI systems, which has more recently reemerged as a central theme in several AI research areas such as Continual Learning, Transfer Learning, LLMs, RL, and Autonomous Agents. This trend is motivated by the recognition that modularity principles offer a promising means of increasing the efficiency and capability of current deep-learning-based AI systems while minimizing inter-task interference. Thus, modularity principles have been adopted, advocated, and reviewed in several influential publications, which we will discuss in this section.

We structure this section around three fundamental aspects of modularity: implicit, emergent, and architectural. *Implicit modularity* refers to the core property of Deep Neural Networks (DNNs): their hierarchical arrangement of layers, each receiving unique transformations of the raw input and learning to compute specialized functions. *Emergent modularity* refers to the phenomenon that is often observed in trained networks: the organization of their units into structural or functional modules. Finally, *architectural modularity* is the design property of architectures that are explicitly designed with modularity priors that encourage the architectures to leverage separate computational building blocks that perform specialized functions.

3.1 Implicit modularity

Deep Neural Networks (DNNs) are composed of a stack of non-linear layers, where each layer provides a processed version of its input to its downstream layer. Thus, in essence, DNNs have a particular kind of modular architecture, a hierarchical architecture with layers — that is, its modules — arranged hierarchically. Thus, this design encourages information factorization (Merel et al., 2019): each layer tends to only receive the information that is relevant to compute its output and can specialize in learning this mapping. Empirically, it has been observed that hierarchical designs facilitate the extraction of progressively more complex, invariant, and abstract features along the hierarchy. For example, moving down the layers of a CNN, (Zeiler & Fergus, 2014; Olah et al., 2017), it is typical to find layers whose neurons are strongly activated by increasingly more global and abstract image features, going from edges, simple shapes, and textures to patterns, object parts, and entire objects[7].

Although DNNs took time to gain traction due to the lack of efficient training algorithms, large datasets, and efficient hardware (Goodfellow, 2016), the computational advantages of hierarchical architectures have long been known (Håstad, 1986). Influential work clarified that, although shallow architectures such as 1-hidden layer neural networks and kernel machines are also universal function approximators, they are inefficient learners of *highly varying functions* compared to DNNs due to their reliance on local estimators (Bengio & LeCun, 2007). More recent work (Poggio et al., 2017) has narrowed down the class of functions that DNNs excel at capturing to those with a compositional structure — that is, functions that can be written as a function of functions $f(x) = h_L(h_{L-1}(...h_1(x)))$. Specifically, DNNs are proven to avoid the curse of dimensionality when the function they are trained to approximate is compositional. Conversely, shallow neural networks can achieve the same approximation error only with a number of parameters that grows exponentially with the number of inputs. For example, compositional functions of $n = 8$ variables and smoothness m with a binary tree-like computational graph $f(x_1, ..., x_8) = h_{1:8}(h_{1:4}(x_1, ...,x_4),h_{5:8}(x_5, ...,x_8))$ with $h_{1:4}(x_1, ...,x_4) = \phi_{1:4}(h_{1:2}(x_1,x_2),h_{3:4}(x_3,x_4))$ and $h_{5:8}(x_5, ...,x_8) = \phi_{5:8}(h_{5:6}(x_5,x_6),h_{7:8}(x_7,x_8))$ can be approximated with an accuracy of at least ϵ by a shallow neural network with $N_{shallow} = \mathcal{O}(\epsilon^{-n/m})$ units, or by a deep network with $N_{deep} = \mathcal{O}((n-1)\epsilon^{-2/m}$ units[8].

[7]Note that hierarchical processing is gradual, with adjacent layers performing similar functions especially in very deep networks (Lad et al., 2024; González et al., 2025)

[8]This statement is true as long as the graph defining $f(\cdot)$ is a subgraph of the graph defining the DNN

3.2 Emergent modularity

Recent studies have investigated emergent modularity, that is, the emergence of modular structures in DNNs that were not imposed by design. This property is of particular interest as the presence of modules helps in understanding the computational mechanism the networks use to solve the task. Ideally, by studying the activation patterns of the modules, one can identify the fundamental subfunctions or rules that are learned and exploited by the networks to perform the task. Importantly, when the networks learn compositionally — that is, when they learn the fundamental atomic rules underlying the target tasks and can combine them in arbitrary new ways (Fodor & Pylyshyn, 1988; Hupkes et al., 2020) — these subfunctions are systematically reused whenever the corresponding subtasks need to be performed. For example, a network that learns to classify objects compositionally can correctly label an image of a white cube even if, during training, it has only seen white spheres and red cubes, and, moreover, it does so using separate color and shape modules.

Traditionally, modules are identified by clustering connectivity (Watanabe et al., 2018; Casper et al., 2022) or activation (Watanabe, 2019; Lange et al., 2022) statistics, which makes understanding the functional role of a module very hard, and studying the compositionality of the network even harder. More recent studies (Csordás et al., 2020; Lepori et al., 2023a) were able to identify modules responsible for specific subtasks by training binary masks on the full network. Specifically, this approach brought to light subfunction-specific subnetworks. However, further analyses (Csordás et al., 2020) showed that the identified subnetworks were not reused in different contexts where the same rules had to be applied in a different combination, providing evidence of the lack of compositionality (Lake & Baroni, 2018; Barrett et al., 2018; Hupkes et al., 2020). Thus, DNNs might generally struggle to identify similarities between subtasks, an issue related to the more general problem of information binding (Greff et al., 2020). Interestingly, the more recent *mechanistic interpretability* studies focused on identifying circuit components that are reused across tasks with similar structures tend to find such components in pretrained LLMs (Wang et al., 2022; Olsson et al., 2022; Conmy et al., 2023; Merullo et al., 2023; Lepori et al., 2023b); this suggests that some compositional abilities might emerge in such models, especially as they are scaled up (Xu et al., 2024).

3.3 Architectural modularity

3.3.1 Compositional generalization

Several studies have shown that modular architectures possess better generalization ability (Clune et al., 2013; Andreas et al., 2016; Kirsch et al., 2018; Chang et al., 2018; Goyal et al., 2019; Mittal et al., 2022) and sample efficiency (Bahdanau et al., 2018; Purushwalkam et al., 2019; Khona et al., 2023; Boopathy et al., 2025) than their monolithic counterparts. A potential explanation for these advantages is their superior compositional learning ability — i.e., their ability to learn and combine atomic rules in arbitrary new ways (Fodor & Pylyshyn, 1988; Hupkes et al., 2020). This skill, which humans excel at, is a significant challenge for AI systems (Lake & Baroni, 2018; Keysers et al., 2019) that modularity principles can help tackle (c.f., § 3.2). This has led to research work aimed at understanding how to best design modular architectures that can generalize compositionally.

In-depth analyses have demonstrated that modular networks can indeed achieve strong compositional generalization abilities (Andreas et al., 2016; Bahdanau et al., 2018); however, this only happens when the task structure is known and can be used to assign modules to the constituent subtasks they can specialize in (Bahdanau et al., 2018; Béna & Goodman, 2021; Mittal et al., 2022). However, it has been observed that when simulating real-world settings more closely, where the task structure is unknown, modular networks often do not specialize and do not exhibit a consistent performance boost. Thus, to fully leverage the capabilities of modular architectures, it is critical to develop suitable inductive biases and learning algorithms that can automatically discover the latent task structure (Boopathy et al., 2025) and filter out the non-compositional features (Jarvis et al., 2024).

Some recent studies took important steps in this direction by identifying useful inductive biases with carefully designed experiments. For example, (Bahdanau et al., 2018) highlighted that a perfect one-to-one mapping between subtasks and modules is not always the best design strategy: architectures composed of modules that specialize in performing groups of related subtasks can perform better, likely because they can learn

to identify and leverage commonalities between subtasks. Similarly, (Béna & Goodman, 2021) showed that strong inter-module connection sparsity and resource constraints (measured by the number of units in a module) facilitate module specialization.

3.3.2 Continual learning

While regularization-based approaches (Kirkpatrick et al., 2017; Zenke et al., 2017) and replay-based approaches (Shin et al., 2017; Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019) are popular solutions for tackling catastrophic interference (McCloskey & Cohen, 1989) in continual learning — where learning signals from different tasks interfere with one another — modular architectures are inherently structured to provide a solution to this problem (Parisi et al., 2019; Hadsell et al., 2020; Wang et al., 2024). In fact, architectures with task-specific parameters that are only trained with their corresponding, task-specific learning signals cannot suffer from interference by construction. As a result, several flavors of modularity approaches have been explored and proven valuable in continual learning settings. Many approaches rely on dynamic architectures that learn to perform new tasks based on a partially trainable, shared network — which is trained on previous tasks — and a fully trainable, task-specific module that is dynamically added to the network. Some approaches keep the shared network completely frozen and add fixed-capacity modules (Rusu et al., 2016). Other solutions allow selective fine-tuning of some parameters of the shared network and add modules with a capacity that depends on how much the task differs from the ones that were previously encountered (Yoon et al., 2017); Finally, alternative approaches frame the learning problem as one of selecting, training, and freezing the modules on a task-specific path through a fixed high-capacity network (Fernando et al., 2017). For a recent review of the latest modular and non-modular approaches, we refer the reader to (Wang et al., 2024).

3.3.3 Transfer learning and large language models

Modular approaches are also widely used in transfer learning settings as a way to boost cross-task positive transfer while reducing the number of parameters to fine-tune. This is because it has been found that fine-tuning the entire network on a downstream task is often unnecessary for good performance; instead, training only the last layers or small, task-specific *adapter* modules interspersed within the pre-trained network is typically sufficient (Pan & Yang, 2009; Collobert et al., 2011; Donahue et al., 2014; Zeiler & Fergus, 2014; Rebuffi et al., 2017; Houlsby et al., 2019). This strategy can be adopted, for example, to reuse the feature extractors a CNN learned with extensive training on a large-scale dataset in a new, data-limited domain (Donahue et al., 2014; Zeiler & Fergus, 2014; Rebuffi et al., 2017). More impressively, it can even be used for multi-task cross-lingual transfer, where a pre-trained LLM is repurposed to perform new target tasks in new languages by carefully designing and combining task- and language-specific adapter modules (Pfeiffer et al., 2020b). For a careful review of modular approaches in transfer learning, we refer the reader to Pfeiffer et al. (2023).

Modularity is also the essential design feature in the emerging LLM frameworks of *Model MoErging* and *Augmented Language Models*. Model MoErging — reviewed in Yadav et al. (2024) — is a new framework aimed at building general-purpose AI systems with emergent capabilities through the effective composition of independently pre-trained expert models; the compositions are achieved by careful choice of pre-trained models as well as routing and aggregation functions. Augmented Language Models — reviewed in Mialon et al. (2023) — are systems composed of LLMs and task-specialized modules, and thus also, clearly, modular. These models can be trained or instructed to leverage external *tools* to solve specific subtasks: for example, they can use a calculator tool to perform arithmetic operations, a retriever tool to retrieve information from document collections, a web browser tool to perform web searches, or a code interpreter tool to execute Python code.

Finally, we note that state-of-the-art LLMs have been increasingly leveraging (Jiang et al., 2024; Guo et al., 2025) or are rumored to leverage[9] Mixture-of-Experts (MoE) layers (Jacobs et al., 1991b) as an efficient way to expand model capacity. One of the first successful applications of MoE in LLMs was the Sparsely-Gated MoE Layer, introduced by Shazeer et al. (2017), which enabled a 1000x increase in model capacity with only

[9]https://openai.com/index/gpt-4-research/

a modest rise in computational cost. Building on this, Fedus et al. (2022) proposed the Switch Transformer, which replaces standard feedforward network layers in the Transformer architecture (Vaswani, 2017) with a sparser MoE routing strategy. This approach routes each token to a single expert, achieving a 7x increase in training speed. Finally, recent advances in retrieval techniques (Lample et al., 2019) have led to the development of the Parameter-Efficient Expert Retrieval (PEER) layer, a new MoE variant that offers an improved compute-performance tradeoff and promises to improve training speed further (He, 2024).

3.3.4 Autonomous agents

Similarly, LLM-based agents — autonomous agents that use LLMs as a reasoning engine to flexibly make decisions to interact with their environment — are also, by design, modular. As a matter of fact, LLM-based agents (reviewed in Sumers et al. (2023)) often feature not only an LLM module as a main reasoning engine and external tools, but also additional specialized modules. For example, they are often endowed with additional short- and long-term memory modules — e.g., to keep track of their state and past experiences — learning modules — e.g., to select episodes or insights worth storing — and evaluator modules — e.g., to refine the decisions of the main LLM reasoning module.

Highly modular is also the influential architecture proposed by (LeCun, 2022) for autonomous agents that can learn and exploit world models to reason and plan at multiple levels of abstraction. At the core of the architecture are six interacting, fully differentiable modules: a perception module, a world model module, a short-term memory module, an actor module, a cost module, and a configurator module. The modules can interact in two main working modes, which mirror Kahneman's Dual-Process Theory of Cognition (Kahneman, 2011). During Mode-1, the actor directly computes an action and sends it to the effectors based on inputs from the perception, short-term memory, and configurator modules; this mode is purely reactive and does not involve planning or world model predictions. During Mode-2, the actor infers a minimum-cost action sequence through an iterative optimization procedure involving the world model module — which predicts the likely world state sequence resulting from the proposed action sequence — and the cost module — which computes the costs of the predicted world state sequence. Mode-2 is thus a form of reasoning-based planning that engages the entire system; it is thus computationally expensive and only used when encountering novel tasks. After a Mode-2 engagement episode, the experienced world state sequence and the minimum-cost action sequence can be reused to train the actor to perform amortized action inference. This mechanism ensures that the agent will rely on the cheaper, reactive Mode-1 to deal with future occurrences of the states that led to Mode-2 engagement.

3.3.5 Hybrid architectures

Modular designs are also a natural choice for the hybrid architectures typically used for multi-modal AI systems (e.g., Achiam et al. (2023); Anil et al. (2023); Liu et al. (2024a) reviewed in Song et al. (2024); Liang et al. (2024)) as well as for neuro-symbolic architectures (e.g., Mao et al. (2019); Guan et al. (2025), reviewed in (Garcez & Lamb, 2023; Chaudhuri et al., 2021)). Multi-modal models typically learn modality-invariant representations by aligning modality-specific representations of multi-modal data. In these models, the modality-specific representations are often computed with dedicated encoder models (Radford et al., 2021; Alayrac et al., 2022), which can be pre-trained. Similarly, neuro-symbolic architectures — which attempt to combine the strengths of deep learning and symbolic approaches — are also modular by design. These architectures often feature deep learning modules to extract abstract, high-level features from raw, input data and symbolic approaches to perform high-level reasoning using symbolic tools like heuristic search, automated deduction, and program synthesis. This organization — often interpreted as emulating the fast, reactive System-1 thinking and the slow, analytical System-2 thinking (Kahneman, 2011) — is often shown to boost abstract reasoning, interpretability, and safety.

A great example of a multi-modal, neuro-symbolic architecture is the one proposed by Mao et al. (2019), which performs image question answering by leveraging deep learning-based components to learn embeddings of images and questions and symbolic components to perform reasoning on the latent input representations. Specifically, the architecture features three main modules: a perception module, a semantic parser module, and a program executor module. The perception module leverages two specialized CNN architectures (He et al., 2017; 2016) to extract latent representations of the visual attributes of all the objects in the scene.

The semantic parser module translates the input question into an executable program in a target domain specific language (DSL) using two feed-forward and two recurrent GRU networks (Cho et al., 2014). Finally, the program executor module generates the final answer by executing the executable program on the image embeddings.

4 Modularity in Brains

In the previous sections, we illustrated why brains, like most complex systems, are modular (§ 2.2), and what advantages this property provides (§ 2.1). Here, we discuss different, selected perspectives on brain modularity. In recent years, one of the most influential modular decompositions of the brain's cognitive abilities in AI has been Kahneman's Dual-Process Theory of Cognition (Kahneman, 2011; LeCun, 2022; Goyal & Bengio, 2022), which posits the existence of two complementary systems underlying human cognition: a fast, reactive system underlying intuition and a slow, analytical system underlying reasoning. However, this is only part of the picture. Brains are hierarchically modular (Meunier et al., 2009), that is, modular at different spatial scales and levels of abstraction. While there is no consensus on which analysis level best captures the fundamental principles underlying human intelligence, each perspective provides valuable insights. Here, we review several influential perspectives spanning these different levels.

At the lowest spatial scale, we have *neurons*. While neurons in standard feedforward networks compute a simple weighted sum of their inputs followed by a nonlinearity (McCulloch & Pitts, 1943), biological neurons exhibit far greater complexity (Beniaguev et al., 2021). Each biological neuron operates as a multi-state, nonlinear dynamical system (Hodgkin & Huxley, 1952) that generates binary signals, or spikes, whose precise timing carries behaviorally relevant information (Maass, 1997). Moreover, rather than merely summing inputs linearly, biological neurons process incoming signals nonlinearly along their dendritic branches. In fact, a single neuron typically receives multiple synaptic connections from each presynaptic source, with dendritic integration introducing additional layers of potentially different nonlinear processing before signals reach the soma (London & Häusser, 2005; Jones & Kording, 2020). This suggests that a single biological neuron can be thought of as a recurrent, highly nonlinear, multilayer network in itself — one endowed with inductive biases that enable it to extract rich, structured features from its inputs. Interestingly, recent work (Liu et al., 2024b) has demonstrated promising results on AI tasks by introducing greater flexibility in the nonlinear functions that neurons use to process their inputs.

Moving up the hierarchy, we encounter *canonical microcircuits* (Harris & Shepherd, 2015). Anatomical studies have shown that the cortex is systematically organized into six layers (or laminae), labeled L1 through L6, running parallel to the skull. Electrophysiological analyses have further revealed the existence of stereotyped synaptic connectivity patterns, which induce recurrent loops between neurons in these layers. These network motifs are ubiquitous across the cortex, encompassing motor, visual, somatosensory, and auditory areas (Douglas & Martin, 2004). One of the most well-established loops begins in the thalamus, which provides input to L4. From there, information is relayed to L2/3, which in turn excites L5/6 of the same cortical area. L5/6 neurons follow two main pathways: one loops back to L4, forming an inner recurrent circuit, while the other projects to subcortical regions, including the thalamus, forming an outer feedback loop. Additionally, an inter-area loop has been identified, induced by L2/3 neurons projecting to L4 of adjacent cortical areas, facilitating cross-regional communication. Canonical microcircuits are thought to play a crucial role in integrating sensory inputs relayed through the thalamus with contextual information from other cortical areas, enabling context-dependent decision-making (Haeusler & Maass, 2006). These circuits have been linked to predictive coding theories (Bastos et al., 2012), which propose that the brain computes prediction errors to refine future inferences and minimize surprise. They are also the fundamental module in the Thousand Brains Theory of Intelligence (Hawkins, 2021), which suggests that cortical columns — comprising preferentially connected neurons spanning all six cortical layers within a cylindrical region — learn sensory-input-dependent models for the objects we interact with. However, it is important to note that many other network motifs and recurrent loops exist throughout the brain, some of which are relatively frequent but remain less well understood (Shepherd & Yamawaki, 2021). Do these stereotypical connectivity patterns confer any computational advantages? Recent influential work (Chen et al., 2022) suggests that they may enhance out-of-distribution generalization. By initializing a recurrent network's weights based on

connectivity patterns observed in the primary visual cortex (Billeh et al., 2020), the study demonstrated significantly improved OOD generalization compared to both feedforward and recurrent CNNs.

Finally, at the highest spatial scale, we find *cortical areas* and *cortical networks*. Cortical areas are identified by parcellating the cortex into clusters of adjacent neurons with shared common properties such as histological characteristics, connectivity patterns, spatial tuning, and functional tuning (Van Essen & Glasser, 2018; Petersen et al., 2024). These parcellations are typically obtained from data collected using invasive methods and processed with varying assumptions and algorithms. Consequently, estimates of the number of cortical areas vary widely, generally ranging from 100 to 200 regions, with no universal consensus. However, a recent semi-automated, multimodal parcellation has gained traction, identifying 180 areas per hemisphere (Glasser et al., 2016). An alternative emergent modular decomposition approach leverages non-invasive functional magnetic resonance imaging (fMRI) recordings. These data are used to identify *functional networks*, which consist of multiple cortical regions that tend to be coactivated during the execution of cognitive tasks or while at rest. Unlike anatomy-based parcellations, these networks can include spatially distant regions. Although there is no universally accepted functional parcellation, influential studies have decomposed the cortex into 7 to 20 large-scale networks (Yeo et al., 2011; Power et al., 2011). Importantly, a recent meta-analysis (Uddin et al., 2019) examined the commonalities among functional networks identified in multiple resting-state and task-based fMRI studies and identified six core functional networks, each linked to distinct cognitive functions. These comprise: (1) the *occipital* network, involved in visual processing; (2) the *pericentral* network, supporting somatomotor functions; (3) the *dorsal frontoparietal* network, mediating attentional control; (4) the *lateral frontoparietal* network, regulating executive control; (5) the *midcingulo-insular* network, controlling salience; (6) the *medial frontoparietal* network, responsible for the default mode of brain activity.

So far, we have discussed modular decompositions of human intelligence based on direct neural recordings. An alternative approach relies on the analysis of large scale cognitive test results (Kaufman, 2018). The most influential framework emerging from this approach is the Cattell-Horn-Carroll (CHC) theory (Schneider & McGrew, 2012), which models intelligence as a three-tier hierarchical structure based on the analysis of correlation patterns among test scores. At the lowest level are *narrow* abilities—specialized cognitive skills applied across multiple tasks. At the intermediate level are *broad* abilities—general cognitive functions that encompass multiple narrow abilities. For instance, Kaufman (2018) identified 17 broad abilities. At the highest level is the *g-factor* (Jensen, 1998), which contributes to all broad abilities and has been associated with brain properties such as size, energy efficiency, nerve conduction velocity, and inter-node path length (Deary et al., 2010). Computational problem-solving, for example, is believed to depend on three broad abilities: fluid reasoning, perceptual processing, and short-term memory (Román-González et al., 2017; Ambrósio et al., 2014). Fluid reasoning, in turn, comprises three narrow skills: *inductive reasoning* — the ability to infer underlying patterns or rules from observed data; *general sequential reasoning* — the ability to apply learned rules sequentially to solve problems; and *quantitative reasoning* — the ability to use mathematical relationships and operations to reason about numerical quantities.

Finally, we note that all the approaches discussed thus far attempt to identify brain modules based on direct or indirect measurements of the brain's current state. However, an alternative perspective argues that a true modular decomposition of the brain requires integrating phylogenetic data to study its evolutionary history (Cisek, 2019). This approach aims to infer the sequence of modifications that transformed primitive feedback control mechanisms—originally implemented by single cells in multicellular organisms to maintain homeostasis—into the complex decision-making systems of the mammalian brain. The central idea is that neural modules emerged and evolved incrementally, adapting to an ever-changing environment by enabling progressively more sophisticated behaviors. Consequently, understanding the present function of a neural module requires examining the role it played at the time of its evolutionary emergence.

5 Discussion

In the previous sections, we have reviewed a substantial body of research that highlights the critical role of modularity in designing intelligent systems. However, it may also be argued that modular architectures and architectural constraints could, in fact, hinder AI system performance and that research should instead

prioritize scaling model and data size. While this view seemingly contradicts classical results like the No Free Lunch Theorem (Wolpert et al., 1995), it aligns with the *Bitter Lesson* (Sutton, 2019), which suggests that progress is driven by scaling rather than handcrafted design. Scaling laws (Kaplan et al., 2020) further support this perspective, showing that performance consistently improves with larger models and datasets. While this finding is supported by a large body of work, a scale-centric approach has significant downsides. Expanding model and dataset size amplifies financial costs and environmental impact, both of which are already pressing concerns (Bender et al., 2021). Moreover, state-of-the-art AI systems have nearly exhausted publicly available internet data, raising concerns about data limitations. While synthetic data generation shows promise (Singh et al., 2024), it risks leading to model collapse (Shumailov et al., 2024). Given these constraints, prioritizing data quality over sheer quantity is becoming increasingly important. In fact, high-quality data can reduce dataset size and training time while matching (Eldan & Li, 2023) or even surpassing larger models trained on less curated data (Gunasekar et al., 2023)

Given the impressive capabilities of current AI systems, does it still make sense to look to the brain as a source of inspiration for improving AI? Although it is sometimes argued that AI systems have already surpassed human capabilities, such achievements have so far been limited to narrow, well-defined tasks. General intelligence—characterized by the ability to reason, plan, and act flexibly across diverse domains—remains out of reach. For instance, while large language models (LLMs) may rival—or even surpass—human performance in *formal* language skills, they continue to fall short in *functional* linguistic abilities (Mahowald et al., 2024). Additionally, human intelligence encompasses far more than language processing: humans can seamlessly integrate perception, reasoning, and action within a single, adaptable system. Moreover, as noted in § 1, current AI models also require orders of magnitude more data and computation than the human brain to achieve comparable performance. Taken together, these observations underscore that human intelligence remains a meaningful benchmark for AI, and that biological brains continue to provide valuable insights for AI research, as they have historically. The development of AI and deep learning has, in fact, long been intertwined with efforts to understand and emulate human cognition (Appendix A), and several brain-inspired principles have already contributed to significant advances in the field (e.g., LeCun et al. (1998); Kirkpatrick et al. (2017); Rolnick et al. (2019)).

Identifying and translating principles from neuroscience into useful inductive biases for AI systems remains a significant challenge. Should we attempt to replicate the brain's connectivity patterns, its activation dynamics, or even its reliance on spike-based communication? We suggest that modularity offers a promising framework for addressing this challenge. One potential path forward is to draw on neuroscientific findings to identify the functional modules that the brain employs to solve specific tasks and implement analogous modules in AI systems. For instance, it is well established that memory and language are supported by distinct systems in the brain (Squire, 2009). While conventional transformer-based LLMs do not include a dedicated memory module, even simple implementations of memory-augmented LLMs have demonstrated benefits, such as reduced hallucination rates and parameter counts (Shuster et al., 2021; Borgeaud et al., 2022; Lewis et al., 2020)—key factors behind the popularity of the retrieval-augmented generation (RAG) framework in industrial chatbots—and improved compositional generalization abilities (Lake, 2019). Similarly, emerging evidence suggests that language and thought are supported by distinct neural systems in the brain (Fedorenko et al., 2024; Mahowald et al., 2024). In contrast, current LLMs do not employ a dedicated reasoning module; rather, reasoning behavior appears to emerge implicitly from training procedures that encourage the model to generate more output tokens before producing a final answer (e.g., Guo et al. (2025)). This raises an important question: could an architecture that explicitly separates language and thought processes lead to improved reasoning capabilities? How might a dedicated reasoning module be designed, and which features of the brain's reasoning mechanisms would be most beneficial to emulate? These remain open and challenging questions. Moving toward a deeper alignment with the brain's structure—particularly at Marr's algorithmic and implementational levels (Marr, 2010)—is considerably more complex, and it remains uncertain whether such efforts would yield clear practical advantages. Nonetheless, as illustrated above, we believe that the computational level offers a viable framework for identifying and transferring functional modules—and, through them, core features of human cognition.

In summary, given the demonstrated benefits of incorporating modularity principles into AI architectures (§ 3) and the consistent evidence that such principles are exploited by the brain (§ 4), modularity appears

to offer an effective framework for integrating brain organization principles into AI system design—holding promise for enhancing the efficiency and performance of artificial systems.

6 Conclusion

Modular designs are ubiquitous across both natural and artificial systems and have been explored—often in isolation—within distinct subfields of biology and engineering. In this work, we brought together these diverse threads through a unified framework and argued that modularity is not merely a convenient structural choice or a biological epiphenomenon but a fundamental principle underlying both natural and artificial intelligence.

We examined the computational advantages modularity offers through multiple perspectives ranging from engineering and evolutionary biology to artificial intelligence and systems neuroscience. We reviewed how modular structures are implemented—explicitly or implicitly—in contemporary AI architectures and how modularity principles characterize the organization of the brain at several levels of abstraction, from micro-circuits to large-scale cognitive systems. Moreover, we discussed how modularity may serve as a viable framework for integrating brain-inspired principles into AI system design. For instance, one promising approach is to identify the functional modules that the brain leverages to solve specific tasks—such as the dedicated systems supporting memory and reasoning—and implement analogous components in AI architectures. This integration can mitigate some critical challenges current AI systems face, including high hallucination rates, poor sample efficiency, limited generalization, and high energy demands. The brain, shaped by hundreds of millions of years of evolution, offers an exceptionally efficient and adaptable model—one that continues to provide insights for improving artificial systems. Despite the promise, significant challenges remain. These include identifying the appropriate level of abstraction to extract functionally relevant brain modules, determining how to translate those modules into trainable or hardwired components in neural architectures, and designing mechanisms for effectively integrating their outputs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jascha Achterberg, Danyal Akarca, DJ Strouse, John Duncan, and Duncan E Astle. Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nature Machine Intelligence*, 5(12):1369–1381, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Ana Paula Ambrósio, Cleon Xavier, and Fouad Georges. Digital ink for cognitive assessment of computational thinking. In *2014 IEEE Frontiers in education conference (FIE) proceedings*, pp. 1–7. IEEE, 2014.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. *arXiv preprint arXiv:2110.07560*, 2021.

Bernard J Baars. In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4):292–309, 1997.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018.

Carliss Y Baldwin and Kim B Clark. *Design Rules: The Power of Modularity Volume 1*. MIT press, 1999.

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pp. 511–520. PMLR, 2018.

Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

Gabriel Béna and Dan FM Goodman. Dynamics of specialization in neural modules under resource constraints. *arXiv preprint arXiv:2106.02626*, 2021.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards ai. 2007. URL `https://api.semanticscholar.org/CorpusID:15559637`.

David Beniaguev, Idan Segev, and Michael London. Single cortical neurons as deep artificial neural networks. *Neuron*, 109(17):2727–2739, 2021.

Yazan N Billeh, Binghuang Cai, Sergey L Gratiy, Kael Dai, Ramakrishnan Iyer, Nathan W Gouwens, Reza Abbasi-Asl, Xiaoxuan Jia, Joshua H Siegle, Shawn R Olsen, et al. Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *Neuron*, 106(3):388–403, 2020.

Grady Booch, Robert A Maksimchuk, Michael W Engle, Bobbi J Young, Jim Connallen, and Kelli A Houston. Object-oriented analysis and design with applications. *ACM SIGSOFT software engineering notes*, 33(5):29–29, 2008.

Akhilan Boopathy, Sunshine Jiang, William Yue, Jaedong Hwang, Abhiram Iyer, and Ila R Fiete. Breaking neural network scaling laws with modularity. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047, 2019.

Werner Callebaut and Diego Rasskin-Gutman. *Modularity: understanding the development and evolution of natural complex systems*. MIT press, 2005.

Stephen Casper, Shlomi Hod, Daniel Filan, Cody Wild, Andrew Critch, and Stuart Russell. Graphical clusterability and local specialization in deep neural networks. In *ICLR 2022 Workshop on PAIR {\textasciicircum} 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

Michael B Chang, Abhishek Gupta, Sergey Levine, and Thomas L Griffiths. Automatically composing representation transformations as a means for generalization. *arXiv preprint arXiv:1807.04640*, 2018.

Swarat Chaudhuri, Kevin Ellis, Oleksandr Polozov, Rishabh Singh, Armando Solar-Lezama, Yisong Yue, et al. Neurosymbolic programming. *Foundations and Trends® in Programming Languages*, 7(3):158–243, 2021.

Guozhang Chen, Franz Scherr, and Wolfgang Maass. A data-based large-scale model for primary visual cortex enables brain-like robust and versatile visual processing. *science advances*, 8(44):eabq7592, 2022.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Paul Cisek. Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, 81:2265–2287, 2019.

Jeff Clune, Jean-Baptiste Mouret, and Hod Lipson. The evolutionary origins of modularity. *Proceedings of the Royal Society b: Biological sciences*, 280(1755):20122863, 2013.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537, 2011.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? inspecting functional modularity through differentiable weight masks. *arXiv preprint arXiv:2010.02066*, 2020.

Ian J Deary, Lars Penke, and Wendy Johnson. The neuroscience of human intelligence differences. *Nature reviews neuroscience*, 11(3):201–211, 2010.

Katharina Dobs, Julio Martinez, Alexander JE Kell, and Nancy Kanwisher. Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, 8(11):eabl8913, 2022.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pp. 647–655. PMLR, 2014.

Rodney J Douglas and Kevan AC Martin. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27(1): 419–451, 2004.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Kai Olav Ellefsen, Jean-Baptiste Mouret, and Jeff Clune. Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLoS Computational Biology*, 11(4):e1004128, 2015.

Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

Jerry A Fodor. *The modularity of mind*. MIT press, 1983.

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

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135, 1999.

Artur d'Avila Garcez and Luis C Lamb. Neurosymbolic ai: The 3rd wave. *Artificial Intelligence Review*, 56 (11):12387–12406, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Matthew F Glasser, Stephen M Smith, Daniel S Marcus, Jesper LR Andersson, Edward J Auerbach, Timothy EJ Behrens, Timothy S Coalson, Michael P Harms, Mark Jenkinson, Steen Moeller, et al. The human connectome project's neuroimaging approach. *Nature neuroscience*, 19(9):1175–1187, 2016.

Ramón Calvo González, Daniele Paliotta, Matteo Pagliardini, Martin Jaggi, and François Fleuret. Leveraging the true depth of llms. *arXiv preprint arXiv:2502.02790*, 2025.

Ian Goodfellow. Deep learning, 2016.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Stephen Grossberg. Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23(3):121–134, 1976.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking, 2025. URL `https://arxiv.org/abs/2501.04519`.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

Stefan Haeusler and Wolfgang Maass. A statistical analysis of information-processing properties of lamina-specific cortical microcircuit models. *Cerebral Cortex*, 17(1):149–162, 02 2006. ISSN 1047-3211. doi: 10.1093/cercor/bhj132. URL `https://doi.org/10.1093/cercor/bhj132`.

Thomas F Hansen. Is modularity necessary for evolvability?: Remarks on the relationship between pleiotropy and evolvability. *Biosystems*, 69(2-3):83–94, 2003.

Bart LM Happel and Jacob MJ Murre. Design and evolution of modular neural network architectures. *Neural Networks*, 7(6-7):985–1004, 1994.

Kenneth D Harris and Gordon MG Shepherd. The neocortical circuit: themes and variations. *Nature neuroscience*, 18(2):170–181, 2015.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Johan Håstad. *Computational limitations for small depth circuits*. PhD thesis, Massachusetts Institute of Technology, 1986.

Jeff Hawkins. *A thousand brains: A new theory of intelligence*. Basic Books, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

Xu Owen He. Mixture of a million experts. *arXiv preprint arXiv:2407.04153*, 2024.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*, 2024.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Takuya Ito, Tim Klinger, Doug Schultz, John Murray, Michael Cole, and Mattia Rigotti. Compositional generalization through abstract representations in human and artificial neural networks. *Advances in Neural Information Processing Systems*, 35:32225–32239, 2022.

Robert A Jacobs and Michael I Jordan. Computational consequences of a bias toward short connections. *Journal of cognitive neuroscience*, 4(4):323–336, 1992.

Robert A Jacobs, Michael I Jordan, and Andrew G Barto. Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, 15(2):219–250, 1991a.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991b.

Devon Jarvis, Richard Klein, Benjamin Rosman, and Andrew M Saxe. On the specialization of neural modules. *arXiv preprint arXiv:2409.14981*, 2024.

Arthur R Jensen. The factor. *Westport, CT: Prager*, 1998.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Ilenna Simone Jones and Konrad Paul Kording. Can single neurons solve mnist? the computational power of biological dendritic trees. *arXiv preprint arXiv:2009.01269*, 2020.

Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences*, 102(39):13773–13778, 2005.

Nadav Kashtan, Elad Noor, and Uri Alon. Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences*, 104(34):13711–13716, 2007.

Alan S Kaufman. *Contemporary intellectual assessment: Theories, tests, and issues*. Guilford Publications, 2018.

Henry Kautz. The third ai summer: Aaai robert s. engelmore memorial lecture. *AI Magazine*, 43(1):105–125, 2022.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*, 2019.

Mikail Khona, Sarthak Chandra, Joy J Ma, and Ila R Fiete. Winning the lottery with neural connectivity constraints: Faster learning across cognitive tasks with spatially constrained sparse rnns. *Neural Computation*, 35(11):1850–1869, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Louis Kirsch, Julius Kunze, and David Barber. Modular networks: Learning to decompose neural computation. *Advances in neural information processing systems*, 31, 2018.

Vedang Lad, Wes Gurnee, and Max Tegmark. The remarkable robustness of llms: Stages of inference? *arXiv preprint arXiv:2406.19384*, 2024.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pp. 2873–2882. PMLR, 2018.

Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. *Advances in neural information processing systems*, 32, 2019.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Large memory layers with product keys. *Advances in Neural Information Processing Systems*, 32, 2019.

Richard D Lange, David S Rolnick, and Konrad P Kording. Clustering units in neural networks: upstream vs downstream information. *arXiv preprint arXiv:2203.11815*, 2022.

Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 62(1):1–62, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Michael Lepori, Thomas Serre, and Ellie Pavlick. Break it down: Evidence for structural compositionality in neural networks. *Advances in Neural Information Processing Systems*, 36:42623–42660, 2023a.

Michael A Lepori, Thomas Serre, and Ellie Pavlick. Uncovering intermediate variables in transformers using circuit probing. *arXiv preprint arXiv:2311.04354*, 2023b.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

Hod Lipson, Jordan B Pollack, and Nam P Suh. On the origin of modular variation. *Evolution*, 56(8): 1549–1556, 2002.

Hod Lipson et al. Principles of modularity, regularity, and hierarchy for scalable systems. *Journal of Biological Physics and Chemistry*, 7(4):125, 2007.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024a.

Ziming Liu, Eric Gan, and Max Tegmark. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. *Entropy*, 26(1):41, 2023.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024b.

Michael London and Michael Häusser. Dendritic computation. *Annu. Rev. Neurosci.*, 28(1):503–532, 2005.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:215943, 2016.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

Prasanna Mayilvahanan, Roland S Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila Juhos, Matthias Bethge, and Wieland Brendel. In search of forgotten domain generalization. *arXiv preprint arXiv:2410.08258*, 2024.

John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, 27(4): 12, Dec. 2006. doi: 10.1609/aimag.v27i4.1904. URL `https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904`.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

Josh Merel, Matthew Botvinick, and Greg Wayne. Hierarchical motor control in mammals and machines. *Nature Communications*, 10(1):1–12, 2019.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. *arXiv preprint arXiv:2310.08744*, 2023.

David Meunier, Renaud Lambiotte, Alex Fornito, Karen Ersche, and Edward T Bullmore. Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3:571, 2009.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

Sarthak Mittal, Yoshua Bengio, and Guillaume Lajoie. Is a modular architecture enough? *Advances in Neural Information Processing Systems*, 35:28747–28760, 2022.

Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Steven E Petersen, Benjamin A Seitzman, Steven M Nelson, Gagan S Wig, and Evan M Gordon. Principles of cortical areas and their implications for neuroimaging. *Neuron*, 2024.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020a.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*, 2020b.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. Modular deep learning. *arXiv preprint arXiv:2302.11529*, 2023.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3593–3602, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.

Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Marcos Román-González, Juan-Carlos Pérez-González, and Carmen Jiménez-Fernández. Which cognitive abilities underlie computational thinking? criterion validity of the computational thinking test. *Computers in human behavior*, 72:678–691, 2017.

Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*, 2017.

Jay G Rueckl, Kyle R Cave, and Stephen M Kosslyn. Why are "what" and "where" processed by separate cortical visual systems? a computational investigation. *Journal of Cognitive Neuroscience*, 1(2):171–186, 1989.

David E. Rumelhart, James L. McClelland, and PDP Research Group. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press, 07 1986. ISBN 9780262291408. doi: 10.7551/mitpress/5236.001.0001. URL `https://doi.org/10.7551/mitpress/5236.001.0001`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

W Joel Schneider and Kevin S McGrew. The cattell-horn-carroll model of intelligence. 2012.

Erwin Schrödinger and Roger Penrose. What is life? *(No Title)*, 1992.

Sebastian Seung. *Connectome: How the brain's wiring makes us who we are*. Houghton Mifflin Harcourt, 2012.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Gordon MG Shepherd and Naoki Yamawaki. Untangling the cortico-thalamo-cortical loop: cellular pieces of a knotty circuit puzzle. *Nature Reviews Neuroscience*, 22(7):389–406, 2021.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Herbert A Simon. The architecture of complexity. In *The Roots of Logistics*, pp. 335–361. Springer, 1962.

Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2505–2515, 2024.

Fabian H Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S Tolias. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979, 2019.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Louis Sokoloff. The metabolism of the central nervous system in vivo. *Handbook of physiology, section I, neurophysiology*, 3:1843–1864, 1960.

Binyang Song, Rui Zhou, and Faez Ahmed. Multi-modal machine learning in engineering design: A review and future directions. *Journal of Computing and Information Science in Engineering*, 24(1):010801, 2024.

Larry R Squire. The legacy of patient hm for neuroscience. *Neuron*, 61(1):6–9, 2009.

N.P. Suh. *The Principles of Design*. Oxford series on advanced manufacturing. Oxford University Press, 1990. URL `https://books.google.it/books?id=DEUk0AEACAAJ`.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. Adashare: Learning what to share for efficient deep multi-task learning. *Advances in Neural Information Processing Systems*, 33:8728–8740, 2020.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Lucina Q Uddin, BT Yeo, and R Nathan Spreng. Towards a universal taxonomy of macro-scale functional human brain networks. *Brain topography*, 32(6):926–942, 2019.

David C Van Essen and Matthew F Glasser. Parcellating cerebral cortex: how invasive animal studies inform noninvasive mapmaking in humans. *Neuron*, 99(4):640–663, 2018.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Günter P Wagner, Jason Mezey, and Raffaele Calabretta. *Natural selection and the origin of modules*. na, 2001.

Günter P Wagner, Mihaela Pavlicev, and James M Cheverud. The road to modularity. *Nature Reviews Genetics*, 8(12):921–931, 2007.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Chihiro Watanabe. Interpreting layered neural networks via hierarchical modular representation. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part V 26*, pp. 376–388. Springer, 2019.

Chihiro Watanabe, Kaoru Hiramatsu, and Kunio Kashino. Modular representation of layered neural networks. *Neural Networks*, 97:62–73, 2018.

Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

David H Wolpert, William G Macready, et al. No free lunch theorems for search. *CiteSeer*, 1995.

Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. Do large language models have compositional ability? an investigation into limitations and scalability. *arXiv preprint arXiv:2407.15720*, 2024.

Prateek Yadav, Colin Raffel, Mohammed Muqeeth, Lucas Caccia, Haokun Liu, Tianlong Chen, Mohit Bansal, Leshem Choshen, and Alessandro Sordoni. A survey on model moerging: Recycling and routing among specialized experts for collaborative learning. *arXiv preprint arXiv:2408.07057*, 2024.

Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, 2019.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

Lorijn Zaadnoordijk, Tarek R Besold, and Rhodri Cusack. Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, 4(6):510–520, 2022.

Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature Communications*, 14(1):1597, 2023.

Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1):3770, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

Ruiyi Zhang, Xaq Pitkow, and Dora E Angelaki. Inductive biases of neural network modularity in spatial navigation. *Science Advances*, 10(29):eadk1256, 2024.

A Approaches to Brain-Inspired Artificial Intelligence

Artificial Neural Networks (ANNs) can be traced back to efforts by McCulloch and Pitts aimed at understanding how networks of biological neurons could compute logical functions (McCulloch & Pitts, 1943). Similarly, the birth of Artificial Intelligence as a research field — typically traced back to the famous 1956 Dartmouth Workshop organized by McCarthy, Minsky, Rochester, and Shannon (McCarthy et al., 2006) — was driven by initiatives intended to prove that a thorough, computational description of human learning and intelligence could lead to the creation of machines that can simulate such natural processes. Thus, clearly, a desire to understand and model the brain and its unique capabilities was a major goal underlying the creation of the first ANNs and the AI research field.

The idea of using the brain as a source of inspiration for designing more robust, reliable, and performant Deep Neural Networks (DNNs) has also been influential (though not central) in AI in the more recent past. Over the last decade, a few review papers have discussed this perspective from different angles. These papers identified the most evident limitations of DNNs that emerge from comparisons with humans and elaborated on different ways the brain can be harnessed as a guide to bridge the gap.

Hassabis et al. (2017) highlighted how the history of deep learning is strongly intertwined with that of neuroscience and how different neural features have profoundly impacted deep learning research or are poised to do so. For example, visual cortical neurons' tuning and normalization properties directly inspired CNNs. Similarly, efforts to model animal conditioning and experience replay profoundly impacted RL, while synaptic plasticity phenomena inspired several influential continual learning algorithms. Conversely, other active research areas, such as efficient learning, transfer learning, and long-term planning, stand to gain from yet unexplored neural features. Zador (2019), on the other hand, argued that to understand the brain and improve deep learning models, we should focus on identifying the fundamental *wiring rules* that are encoded in the 1 GB human genome; such rules are critical as they must provide a highly compressed representation of the entire 200 TB[10] human connectome (Seung, 2012), distilling knowledge acquired over evolutionary timescales into brain networks that support critical innate behaviors and fast learning. Taking a more pragmatic approach, Richards et al. (2019) recommended identifying the three main computational building blocks of the brain: its *cost functions* — which reflect the networks' learning goals — its *optimization algorithms* — which guide synaptic plasticity — and its *backbone architectures* — which constrain how the information can flow across the network. Sinz et al. (2019) went one step further and suggested a purely data-driven approach to boosting the generalization performance of DNNs based on aligning the latent features DNNs learn with the recorded brain activity patterns.

While the prevailing view in AI is that we should aim to directly try to reproduce the high-level cognitive abilities unique to human adults to advance the development of AI systems, more recent work (Zaadnoordijk et al., 2022; Zador et al., 2023) has emphasized a different perspective. For instance, Zaadnoordijk et al. (2022) highlighted the importance of examining infants' learning. Studies have shown that the infant brain already possesses adult-like structural and functional connectivity patterns that allow them to perform efficient multimodal, unsupervised learning by exploiting attentional, processing, and cognitive biases as well as curriculum and active learning strategies. Interestingly, this view is consistent with previous observations (Lake et al., 2017) stressing the importance of infants' *start-up software* consisting of causal world models and intuitive theories of psychology and physics that boost compositional meta-learning. In a similar vein, Zador et al. (2023) advocated a focus on reproducing animal-level intelligence first, as animals already possess a vast amount of developmentally inherited knowledge that allows them to thrive in their constantly changing environment through state-dependent decision-making and detailed world models.

The revised works offer interesting perspectives on selected brain properties that can guide the development of new models to move beyond narrow AI systems. However, none of them attempted to decompose brains into their fundamental building blocks, or modules, underlying intelligence. A notable exception is provided by recent work (Marblestone et al., 2016; Goyal & Bengio, 2022; Mahowald et al., 2024) that represents a significant step in this direction. Marblestone et al. (2016) took an optimization-centric approach and tried to organize the brain in terms of cost functions it appears to optimize. Specifically, they surveyed different AI-relevant functions the brain is known to perform effectively — including high-level planning, hierarchical

[10] Assuming 10^{14} synaptic weights stored in half-precision floating-point format (FP16)

predictive control, short- and long-term memory, selective attention, and information routing — attempted to map these functions to specific brain networks, and hypothesized how these networks might be coordinated to learn over different timescales. Goyal & Bengio (2022) attempted to identify the core cognitive principles of human intelligence and to suggest potential ways to translate them into inductive biases for deep learning architectures, building on influential cognitive neuroscience theories, such as the Global Workspace Theory (GWT) (Baars, 1997) and the Dual-Process Theory of Cognition (Kahneman, 2011). Mahowald et al. (2024) focused on LLMs and explained their inconsistent performance across different tasks as arising from a clear-cut separation of strictly linguistic, formal *abilities* — which they excel in — from higher-level, *functional abilities* — which they often struggle with. Neuroscientific evidence shows that these abilities are supported by separate brain networks, suggesting that architectural and emergent modularity approaches that mirror the specialization observed in the brain are essential for enhancing the capabilities of LLMs.