Building artificial neural circuits for domain-general cognition: a primer on brain-inspired systems-level architecture

Jascha Achterberg,[1,2] **Danyal Akarca,**[1] **Moataz Assem,**[1] **Moritz Heimbach,**[3] **Duncan E. Astle,**[1,4] **John Duncan**[1]

[1] University of Cambridge, MRC Cognition and Brain Sciences Unit
[2] Intel Labs
[3] Julius-Maximilians-Universität Würzburg, Institute of Computer Science
[4] University of Cambridge, Department of Psychiatry
{ jascha.achterberg, danyal.akarca, moataz.assem, duncan.astle, john.duncan } @mrc-cbu.cam.ac.uk,
moritz.heimbach@uni-wuerzburg.de

Abstract

There is a concerted effort to build domain-general artificial intelligence in the form of universal neural network models with sufficient computational flexibility to solve a wide variety of cognitive tasks but without requiring fine-tuning on individual problem spaces and domains. To do this, models need appropriate priors and inductive biases, such that trained models can generalise to out-of-distribution examples and new problem sets. Here we provide an overview of the hallmarks endowing biological neural networks with the functionality needed for flexible cognition, in order to establish which features might also be important to achieve similar functionality in artificial systems. We specifically discuss the role of system-level distribution of network communication and recurrence, in addition to the role of short-term topological changes for efficient local computation. As machine learning models become more complex, these principles may provide valuable directions in an otherwise vast space of possible architectures. In addition, testing these inductive biases within artificial systems may help us to understand the biological principles underlying domain-general cognition.

Introduction

An aspiration of machine learning research is not just to create architectures capable of achieving increasingly high levels of task-specific performance, but the genesis of models able to achieve good performance across different domains simultaneously. Recent striking advances in network models have enabled them to solve many problems within a domain with just one architecture (Brown et al. 2020; Webb, Holyoak, and Lu 2022; Srivastava et al. 2022). Additionally, networks are increasingly acquiring multi-modal capabilities (Xu, Zhu, and Clifton 2022; Akkus et al. 2023) and learn in open-ended task environments (Fan et al. 2022; Adaptive Agent Team et al. 2023). These advances provide necessary building blocks for models capable of domain general cognition, as observed in intelligent human behaviour. Crucially, these new models may be able to go beyond simple generalisation to unseen data (Hardt and Recht 2022); they may be able to learn new abilities and directly abstract them, allowing for generalisation across entire input modalities and the reuse of skills learned in one domain to support learning in entirely new domains. Indeed, this parallels how children learn over the course of their own development (Kievit 2020). However, the extent to which current models can achieve this remains limited.

For decades, neuroscientists have been focused on identifying core features of the brain's structural and functional architecture. This allows us to connect our knowledge of human neural architectures that enable flexible domain-general cognition (Duncan, Assem, and Shashidhara 2020), with ideas on how we hope to achieve similar capabilities in artificial systems. Here we provide an overview of mechanisms underlying domain-general cognition in biological neural networks to derive which features of the systems-level architecture may be important to build flexible multimodal problem-solving capabilities into artificial systems. Previously published reviews have already outlined which cognitive ideas and modules might be essential (Russin, O'Reilly, and Bengio 2020; Goyal and Bengio 2022; VanRullen and Kanai 2021; LeCun 2022; Lake et al. 2017). We aim to expand these cognitive perspectives by providing a brief introduction to the system-level network structure underlying domain-general cognition in the brain, highlighting what structural optimisation processes we think could be used in machine learning models. In this, our goal is not to hard-code brain-like anatomy into a network model's architecture. Instead, we aim to identify computationally beneficial structural motifs which can be soft-coded into the network's learning process to serve as helpful inductive biases or priors. As we see increasingly complex machine learning models being built as a combination of functional submodules (Pfeiffer et al. 2023; Akkus et al. 2023), we believe that the system-level priors we outline may provide helpful guidance to coordinate information flow in the most complex artificial neural networks (Goyal et al. 2022).

A core domain-general network in the brain

The human brain, as with many complex physical systems, is economically organised to balance numerous competing objectives – including metabolic, computational, and geometric (Cajal et al. 1995; Bullmore and Sporns 2012).

These objectives have a strong influence on the topology of the brain's network; not only is it energetically expensive to fully build and sustain neural connections (Raichle and Gusnard 2002; Tomasi, Wang, and Volkow 2013) but it is highly costly to constantly communicate signals between neurons and assemblies of neurons, particularly over longer distances (Levy and Calvert 2021). Owing to its size, complexity, and these economic considerations, it is infeasible for each neural region to communicate directly with every other region equivalently (Horvát et al. 2016). To avoid this problem, evolutionary pressures have guided the brain towards a modularised network, with modules of very strong local connectivity and high-connection hub nodes connecting across these modules (Suarez et al. 2022; Luppi et al. 2022). Networks with this structure are described as having "small-world" characteristics, defined as having concurrently a highly clustered topology and short path lengths, meeting a balance between totally random versus regular networks (Bassett and Bullmore 2006, 2017). Small-world structures are commonly found in distributed systems under resource constraints, showing patterns of locally specialised computation alongside good propagation of signals within and between hubs. In brains, this locality of computations results in concentration of specific cognitive function within specific anatomical regions. Specialised regions act as foci for cognitive functions like sensory processing, semantic knowledge, and language abilities (Kaas and Collins 2001; Ralph et al. 2017; Skeide and Friederici 2016). These are likely semi-specialised, meaning that they mostly focus on unique local computation but also partially integrate meaningful information across areas and domains (Atilgan et al. 2018; Steinmetz et al. 2019).

This picture of a functionally modular system becomes more nuanced when we consider human domain-general cognition. As the tasks to be solved become more complicated, the brain increasingly abandons solely relying on its specialised modular structure. Instead, neural architectures must increasingly integrate signals across modules (Power et al. 2013) and rely on its Multiple Demand system (MD) (Duncan, Assem, and Shashidhara 2020; Assem et al. 2020). This is a core network in the brain (depicted in Figure 1A) which is highly active when a complex task of any nature is solved. It is thought that the MD system serves as a central processing unit, receiving information from more specialised input notes, to compress it into meaningful abstract representations on which it can run problem solving algorithms (schematic shown in Figure 1B). It also plays a central role in controlling information in other brain regions, using knowledge and complex analysis of the situation to control thought processes in specialised brain regions through top-down control processes (Duncan, Assem, and Shashidhara 2020; Deco, Vidaurre, and Kringelbach 2021; Dehaene, Kerszberg, and Changeux 1998; Miller and Cohen 2001; Norman and Shallice 1986). Ultimately it is this central processing circuit that likely gives the primate brain the ability to have abstract thoughts used to solve complex problems to reach long-term goals.

What are the key principles underpinning this system? In the following we will discuss three computational / structural motifs which vary across the hierarchy from specialised regions to the integrative MD system, which allow the network to show domain-general cognitive skills. These are: Recurrence, communicability, and short-term topological changes. We will review each of these in terms of their relevance in biological networks before then discussing possible directions for artificial implementations, in the context of related existing implementations.

Computational motifs supporting domain-general cognition

Global recurrence

Computations in functionally more specialised regions depend strongly on a feed-forward structure that extracts increasingly abstract features from sensory inputs (Grill-Spector and Malach 2004; Hackett 2011; Mashour et al. 2020). Much work shows how this process can be modelled using an artificial feed-forward network (Schrimpf et al. 2018; Lindsay 2021). While there is also recurrent processing in these specialised systems (Grill-Spector and Malach 2004; Hackett 2011; Kietzmann et al. 2019), the recurrent loops in these systems are likely very local and cover relatively short distances and timescales. This means that a signal sent from a node will only travel a short path before arriving back at its starting point. As we move towards more integrative and domain-general cognition, recurrent connections become a hallmark feature of the brain's systems-level design. The frontal cortex, where a large part of the MD system lies, is often thought of as implementing recurrent loops for abstract information processing (Mashour et al. 2020; Miller and Buschman 2007). Importantly, these loops not only process information locally but also broadcast information widely across the brain, influencing and controlling computations in specialised regions. It does so by not only having local recurrent connections within the circuit but also many loops spanning large distances in the brain, reaching out to nodes which lie far outside the core (Miller and Cohen 2001; Munakata et al. 2011; Mashour et al. 2020; Dehaene, Kerszberg, and Changeux 1998). With nodes widely distributed over the cortex, coupled to strong communication between these nodes, the MD system is well positioned for widespread integration and communication. A large set of recursive processing loops with varied scales in terms of time and spatial distance likely facilitate the MD system's abstract domain-general processing and deliver the ability to coordinate computation in a large distributed system (Duncan, Assem, and Shashidhara 2020).

The use of recurrent loops in artificial neural networks has a long history (Schmidhuber 2022). They proved to be useful tools for processing and predicting time series data but also suffered from problems of vanishing gradient and computational complexity when capturing long-range dependencies in the input (Hochreiter and Schmidhuber 1997; Vaswani et al. 2017; Fawaz et al. 2019). To avoid these issues, feed-forward based architectures can be used as substitutes (Fawaz et al. 2019) and various attention-based architectures have recently been very effective in capturing dependencies in language time courses and multiple other

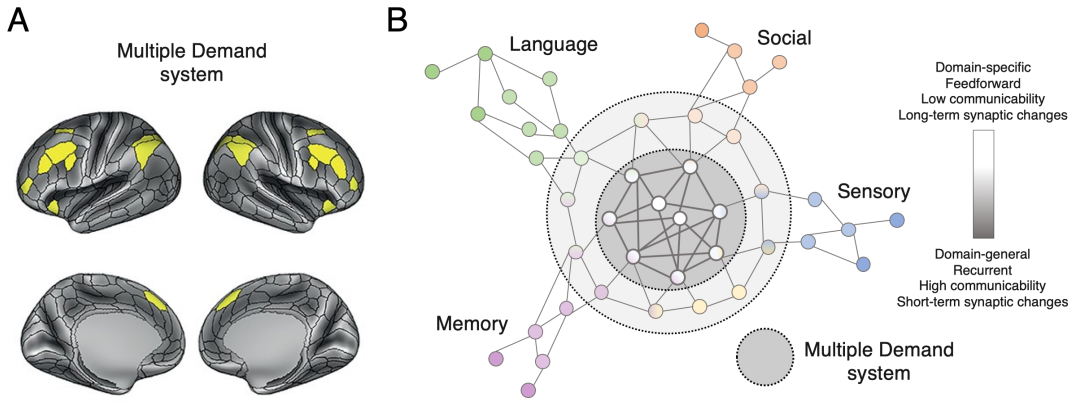

Figure 1: A - The cortical areas forming the core Multiple Demand system in the human brain, from (Assem et al. 2020). B – Schematic depiction of a systems-level view of the brain. The Multiple Demand system lies at the core of information processing in the brain, exchanging inputs with more specialised regions such as language, memory, sensory and social processing. Due to its central position, the MD core can influence computations in multiple specialised areas by broadcasting information it constructed from integrating across domains back to specialised regions, e.g., influencing perception by abstract understanding of the environment / situation at large. Refer to (Assem et al. 2020) for detailed anatomical perspective of the MD system's core and penumbra regions not discussed here.

modalities (Vaswani et al. 2017; Tay et al. 2022). This works by inputting an entire time series in a single time step so that the attention mechanism learns the relationship between timesteps without needing to hold past time points in memory. While these architectures likely can be good substitutes for the local recurrent loops, we believe that ultimately, researchers are going to have to find a way to also introduce global recurrent loops to arrive at domain-general cognition in artificial systems. Approaches like weight-sharing in deep models paired with skip-connections may allow us to mimic a recurrent process in a regular forward pass but it seems likely that alternative ways will be needed to allow abstract multimodal knowledge to be broadcasted through the network to inform distributed computations. This seems even more timely now that models generate impressive responses to inputs such as images or language (Brown et al. 2020; Rombach et al. 2022), but struggle to be constrained by meaningful world models (e.g., intuitive physics, (Lake et al. 2017)). Instead, researchers rely on human feedback signals in the training pipeline (Ouyang et al. 2022). As such, machine learning models may need to be adapted to allow for the introduction of a global recurrent architecture similar to the MD system.

Communicability in large scale networks

For any complex network which is concerned with processing information, it is of central importance to optimise how signals are communicated between the nodes within the network (Estrada, Hatano, and Benzi 2012). This becomes an increasingly challenging problem as a network grows, leaving nodes to only be able to communicate with a smaller proportion of the network. This limited communication capacity naturally leads to variation in terms of how much information is exchanged between different pairs of nodes across the network. This results in a very real challenge for any large-scale network system to optimise its structure to integrate information most effectively and efficiently across its functional hubs. This is constrained, ultimately, by the topological arrangement of the network. The idea of how much information is exchanged between nodes is captured by the concept of communicability (Estrada, Hatano, and Benzi 2012; Crofts and Higham 2009; Srivastava et al. 2020) and is a highly effective framework to understand how the structure of the brain guides function (Goñi et al. 2014; Seguin, Razi, and Zalesky 2019; Betzel et al. 2022; Griffa et al. 2022; Avena-Koenigsberger, Misic, and Sporns 2018; Avena-Koenigsberger et al. 2019; Laughlin and Sejnowski 2003). Specifically, across the brain's complex network, regions vary in terms of how well they can communicate to other regions, and the macro-scale dynamics and capabilities of the brain will be determined by this interareal communication. This heterogeneous communicability becomes especially interesting when one considers how system-level communication link to domain-general cognition. In the previous section, we described how more specialised regions tend to have a mostly feed-forward structure with some local recurrence. As such, information tends to be communicated locally between adjacent and functionally related regions. This changes as information approaches the domain-general MD system with its wider communicative influence. In its central position, the MD system not only receives information from all over the brain but utilises its widespread connectivity as global recurrent loops to broadcast processed information to a distributed set of brain regions (Mashour et al. 2020; Duncan, Assem, and Shashidhara 2020; Dehaene, Kerszberg, and Changeux 1998). On

the systems-level perspective of the brain, a given region's communicative structure heavily depends on its functional role and hence its degree of specialisation.

The concept of heterogeneous communicability between regions and modules of the brain has not been particularly relevant in artificial neural network architectures which were state-of-the-art until very recently. Take convolutional neural networks (CNNs) as an example. In CNNs, which dominated processing of visual information for several years (Schmidhuber 2022), information is mostly passed along from layer to layer in a relatively even fashion. This means regions do not stick out has having a particular communicative ability (though see work like (Shrivastava et al. 2017) for interesting communicative extensions of CNNs). However, this is changing with new architectures which have been growing in scale (Kaplan et al. 2020). Especially for network models which utilise multiple modalities, architectures have increasingly been created by combining existing pre-trained models into more complex modular architectures (Rombach et al. 2022; Akkus et al. 2023). Once we build complex system like these, it becomes increasingly important to not only think about which models to combine, but also how to combine them. This means that the communicability between parts of the network can be optimised to achieve better information flow between components and hence improve performance. A first step in this direction was made by a multimodal transformer model which outperformed prior networks by introducing a set of special bridge layers to connect two modality specific models. These bridge layers allow the model to learn a communicative structure in which abstract semantic knowledge is gradually merged across modalities. This increased performance in several relevant benchmark tasks (Xu et al. 2023). In addition, other implementations have shown that bringing ideas from highly communicative small world graph structures into a Transformer's attention mechanism can help with processing longer sequences (Zaheer et al. 2021). In simple recurrent neural networks, we also have seen that system-level communicability can easily be used as a regularisation term to optimise the communicative structure of a sparsely connected network to arrive at a network with many brain-like structural and functional properties (Achterberg et al. 2022a). As network models grow in complexity and increasingly make use of composite structures which combine sub models into larger networks, it will be important to fine tune the communicative structure of a network. Having good priors and inductive biases for these linkages can help circumvent problems arising from adding the extensive set of connections it would require to fully connect multiple models which already have a complex structure themselves. Following this line of thinking, we believe that making use of work on communicability and how it can be optimised in complex networks will be of central importance to inform model building on a systems-level.

Short-term topological changes

The discussion so far has focused on how the systems-level network structure of the brain and the unique communicative structure of its MD system play a key role in the domain-general cognition we see in humans. An important element of its flexible and multimodal information processing capabilities is how the MD system's network structure is not fully fixed but often rapidly changing. This means that while the MD system is running multimodal computations internally, the connections between its neurons are in continuous flux. As such, the general problem-solving ability of this network is assumed to be due to its inherent flexibility. In it, local computations are organised by rapid changes to the network structure (Stokes 2015; Tang et al. 2022; García-Cabezas et al. 2017), often called short-term plasticity. This allows the network to continuously reassign its neurons and modules new computational roles while solving complex sequential problems (Duncan 2001; Miller and Cohen 2001; Crowe, Averbeck, and Chafee 2010; Meyers et al. 2008; Achterberg et al. 2022b). Rapid topological changes are likely induced by local learning rules which supplement the more long-term optimisation of the global network structure. These mechanisms likely underly a multitude of complex abilities of the human brain (Assem et al. 2020; Duncan 2010) and some of them strongly overlap with timely discussions in machine learning. As one example, research points to the fact that the MD system uses its short-term dynamics for attentional control, to focus on information which is relevant for the current operation (Sakagami and Niki 1994; Rainer, Asaad, and Miller 1998; Buschman et al. 2012) and break complex pieces of information down into simple computable bits (Duncan, Assem, and Shashidhara 2020) – a function which has played a central role in machine learning discussion recently (Shrivastava et al. 2017; Lindsay 2020). Another example is the MD system's ability to construct abstract representations of problems (Wallis, Anderson, and Miller 2001) to then tie observed stimuli rapidly to their roles in this abstract problem representation (Duncan, Assem, and Shashidhara 2020; Achterberg et al. 2022b), a phenomenon going by the name of variable binding (Smolensky 1990) or meta-learning (Botvinick et al. 2019). These are very related to few-shot learning (Brown et al. 2020) and in-context learning abilities (von Oswald et al. 2022) observed in large Transformer models.

As we already see foundations of these skills emerging in currently existing architectures it is reasonable to believe that they will continue to improve purely by scaling existing architectures (Kaplan et al. 2020). In this case models would use their unit activations to implement rapid in-context learning and it has been shown that this can work well without any short-term synaptic changes (Wang et al. 2018). In fact, even in the brain many complex computations are likely facilitated due to dynamics of the network activations which do not necessarily have to rely on changes in the network structure (Vyas et al. 2020). But once computations reach the scale of using network-wide attention processes to controlling the flow of information across the entire brain network and flexibly combining task modules to solve the task at hand (Duncan, Assem, and Shashidhara 2020; Buschman and Miller 2014; MacDowell et al. 2023), rapid topological network changes might be necessary for domain-general computations (Stokes 2015; Duncan, Assem, and Shashidhara 2020). Reaching this level of flexible and multimodal

cognition might not be possible in current static architectures and hence might require us to allow models to modify some of their connections in the moment through local learning rules. Some work in smaller network models is highlighting how local learning mechanisms can complement network-wide optimisation processes (Whittington et al. 2020; Dekker, Otto, and Summerfield 2022) with relevant comparisons to Transformer implementations (Whittington, Warren, and Behrens 2022). Other examples point to how local learning rules and single neuron-based optimisation principles by themselves can be sufficient to solve meaningful cognitive tasks (Masse et al. 2019; Falandays et al. 2023). In addition, we have seen how standard network optimisers can be updated with certainty judgements to support rapid relational learning (Nelli et al. 2023). If we could scale these rapid learning dynamics to large Transformer models, this might allow models to flexibly combine abstract task structures with capabilities learned in the past, to flexibly apply skills across modalities in a truly domain-general way. One research direction which might support rapid learning processes is work on using local loss functions and learning mechanisms to substitute costly global optimisation processes (Löwe, O'Connor, and Veeling 2020; Ren et al. 2023; Hinton 2022). Combining these local optimisation processes with more wide-spread recurrent loops and an optimised communicative structure in large networks might bring us closer to observing flexible domain-general cognition in artificial neural networks.

Conclusion

We believe that in the pursuit of building artificial intelligence which is able to engage in domain-general problem solving, a systems-level view of the human brain will provide useful guidance (Hassabis et al. 2017; Zador et al. 2023). We believe this will become increasingly relevant as AI systems become more and more complex. The topics of recurrence, communication and rapid structural changes are particularly relevant at the current point due to their central role in theories of domain-general cognition in the brain and their links to existing works in neural network models. As such, they might be key drivers behind efficient and flexible information processing in large multimodal networks. But we do not believe that any of these features should be fully hard-coded – instead we should think of them as useful priors and inductive biases which can guide complex learning processes. Ultimately, bringing these features into machine learning models opens up the perspective of not only improving the performance of artificial neural networks but also for us to understand which core principles underly domain-general and multimodal computations in neural networks - may these be biological or artificial.

Acknowledgments

J.A., Da.A., M.A., Du.A., and J.D. are supported by UKRI MRC funding and as a result the authors have applied a Creative Commons Attribution (CC BY) license to this manuscript for the purpose of open access. J.A. receives a Gates Cambridge Scholarship. Da.A. receives a Cambridge Trust Vice Chancellor's Scholarship. Da.A. and Du.A. are both supported by the James S. McDonnell Foundation Opportunity Award. J.A. was a research intern at Intel Labs at the time of writing this manuscript.

References

Achterberg, J.; Akarca, D.; Strouse, D. J.; Duncan, J.; and Astle, D. E. 2022a. Spatially-embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. BioRxiv: 2022.11.17.516914 Section: New Results.

Achterberg, J.; Kadohisa, M.; Watanabe, K.; Kusunoki, M.; Buckley, M. J.; and Duncan, J. 2022b. A One-Shot Shift from Explore to Exploit in Monkey Prefrontal Cortex. *Journal of Neuroscience*, 42(2): 276–287. Publisher: Society for Neuroscience Section: Research Articles.

Adaptive Agent Team; Bauer, J.; Baumli, K.; Baveja, S.; Behbahani, F.; Bhoopchand, A.; Bradley-Schmieg, N.; Chang, M.; Clay, N.; Collister, A.; Dasagi, V.; Gonzalez, L.; Gregor, K.; Hughes, E.; Kashem, S.; Loks-Thompson, M.; Openshaw, H.; Parker-Holder, J.; Pathak, S.; Perez-Nieves, N.; Rakicevic, N.; Rocktäschel, T.; Schroecker, Y.; Sygnowski, J.; Tuyls, K.; York, S.; Zacherl, A.; and Zhang, L. 2023. Human-Timescale Adaptation in an Open-Ended Task Space. ArXiv:2301.07608 [cs].

Akkus, C.; Chu, L.; Djakovic, V.; Jauch-Walser, S.; Koch, P.; Loss, G.; Marquardt, C.; Moldovan, M.; Sauter, N.; Schneider, M.; Schulte, R.; Urbanczyk, K.; Goschenhofer, J.; Heumann, C.; Hvingelby, R.; Schalk, D.; and Aßenmacher, M. 2023. Multimodal Deep Learning. ArXiv:2301.04856 [cs, stat].

Assem, M.; Glasser, M. F.; Van Essen, D. C.; and Duncan, J. 2020. A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. *Cerebral Cortex*, 30(8): 4361–4380.

Atilgan, H.; Town, S. M.; Wood, K. C.; Jones, G. P.; Maddox, R. K.; Lee, A. K. C.; and Bizley, J. K. 2018. Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. *Neuron*, 97(3): 640–655.e4.

Avena-Koenigsberger, A.; Misic, B.; and Sporns, O. 2018. Communication dynamics in complex brain networks. *Nature Reviews Neuroscience*, 19(1): 17–33. Number: 1 Publisher: Nature Publishing Group.

Avena-Koenigsberger, A.; Yan, X.; Kolchinsky, A.; Heuvel, M. P. v. d.; Hagmann, P.; and Sporns, O. 2019. A spectrum of routing strategies for brain networks. *PLOS Computational Biology*, 15(3): e1006833. Publisher: Public Library of Science.

Bassett, D. S.; and Bullmore, E. 2006. Small-World Brain Networks. *The Neuroscientist*, 12(6): 512–523. Publisher: SAGE Publications Inc STM.

Bassett, D. S.; and Bullmore, E. T. 2017. Small-World Brain Networks Revisited. *The Neuroscientist*, 23(5): 499–516. Publisher: SAGE Publications Inc STM.

Betzel, R. F.; Faskowitz, J.; Mišić, B.; Sporns, O.; and Seguin, C. 2022. Multi-policy models of interregional

communication in the human connectome. BioRxiv: 2022.05.08.490752 Section: New Results.

Botvinick, M.; Ritter, S.; Wang, J. X.; Kurth-Nelson, Z.; Blundell, C.; and Hassabis, D. 2019. Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, 23(5): 408–422. Publisher: Elsevier.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Bullmore, E.; and Sporns, O. 2012. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5): 336–349. Number: 5 Publisher: Nature Publishing Group.

Buschman, T. J.; Denovellis, E. L.; Diogo, C.; Bullock, D.; and Miller, E. K. 2012. Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. *Neuron*, 76(4): 838–846.

Buschman, T. J.; and Miller, E. K. 2014. Goal-direction and top-down control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655): 20130471. Publisher: Royal Society.

Cajal, S. R. y.; Swanson, N.; Swanson, L. W.; Cajal, S. R. y.; Swanson, N.; and Swanson, L. W. 1995. *Cajal's Histology of the Nervous System of Man and Vertebrates*. History of Neuroscience. Oxford, New York: Oxford University Press. ISBN 978-0-19-507401-7.

Crofts, J. J.; and Higham, D. J. 2009. A weighted communicability measure applied to complex brain networks. *Journal of The Royal Society Interface*, 6(33): 411–414. Publisher: Royal Society.

Crowe, D. A.; Averbeck, B. B.; and Chafee, M. V. 2010. Rapid Sequences of Population Activity Patterns Dynamically Encode Task-Critical Spatial Information in Parietal Cortex. *Journal of Neuroscience*, 30(35): 11640–11653. Publisher: Society for Neuroscience Section: Articles.

Deco, G.; Vidaurre, D.; and Kringelbach, M. L. 2021. Revisiting the global workspace orchestrating the hierarchical organization of the human brain. *Nature Human Behaviour*, 5(4): 497–511. Number: 4 Publisher: Nature Publishing Group.

Dehaene, S.; Kerszberg, M.; and Changeux, J.-P. 1998. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, 95(24): 14529–14534. Publisher: Proceedings of the National Academy of Sciences.

Dekker, R. B.; Otto, F.; and Summerfield, C. 2022. Determinants of human compositional generalization.

Duncan, J. 2001. An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, 2(11): 820–829. Number: 11 Publisher: Nature Publishing Group.

Duncan, J. 2010. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14(4): 172–179.

Duncan, J.; Assem, M.; and Shashidhara, S. 2020. Integrated Intelligence from Distributed Brain Activity. *Trends in Cognitive Sciences*, 24(10): 838–852.

Estrada, E.; Hatano, N.; and Benzi, M. 2012. The physics of communicability in complex networks. *Physics Reports*, 514(3): 89–119.

Falandays, J. B.; Yoshimi, J.; Warren, W.; and Spivey, M. 2023. A Potential Mechanism for Gibsonian Resonance: Behavioral Entrainment Emerges from Local Homeostasis in an Unsupervised Reservoir Network.

Fan, L.; Wang, G.; Jiang, Y.; Mandlekar, A.; Yang, Y.; Zhu, H.; Tang, A.; Huang, D.-A.; Zhu, Y.; and Anandkumar, A. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. ArXiv:2206.08853 [cs].

Fawaz, H. I.; Forestier, G.; Weber, J.; Idoumghar, L.; and Muller, P.-A. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4): 917–963. ArXiv:1809.04356 [cs, stat].

García-Cabezas, M. A.; Joyce, M. K. P.; John, Y. J.; Zikopoulos, B.; and Barbas, H. 2017. Mirror trends of plasticity and stability indicators in primate prefrontal cortex. *European Journal of Neuroscience*, 46(8): 2392–2405. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.13706.

Goyal, A.; and Bengio, Y. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478(2266): 20210068. Publisher: Royal Society.

Goyal, A.; Didolkar, A.; Lamb, A.; Badola, K.; Ke, N. R.; Rahaman, N.; Binas, J.; Blundell, C.; Mozer, M.; and Bengio, Y. 2022. Coordination Among Neural Modules Through a Shared Global Workspace. ArXiv:2103.01197 [cs, stat].

Goñi, J.; van den Heuvel, M. P.; Avena-Koenigsberger, A.; Velez de Mendizabal, N.; Betzel, R. F.; Griffa, A.; Hagmann, P.; Corominas-Murtra, B.; Thiran, J.-P.; and Sporns, O. 2014. Resting-brain functional connectivity predicted by analytic measures of network communication. *Proceedings of the National Academy of Sciences*, 111(2): 833–838. Publisher: Proceedings of the National Academy of Sciences.

Griffa, A.; Mach, M.; Dedelley, J.; Gutierrez-Barragan, D.; Gozzi, A.; Allali, G.; Grandjean, J.; Ville, D. V. D.; and Amico, E. 2022. The evolution of information transmission in mammalian brain networks. BioRxiv: 2022.05.09.491115 Section: New Results.

Grill-Spector, K.; and Malach, R. 2004. The Human Visual Cortex. *Annual Review of Neuroscience*, 27(1): 649–677. _eprint: https://doi.org/10.1146/annurev.neuro.27.070203.144220.

Hackett, T. A. 2011. Information flow in the auditory cortical network. *Hearing Research*, 271(1): 133–146.

Hardt, M.; and Recht, B. 2022. *Patterns, predictions, and actions: a story about machine learning*. Princeton: Princeton University Press. ISBN 978-0-691-23373-4.

Hassabis, D.; Kumaran, D.; Summerfield, C.; and Botvinick, M. 2017. Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2): 245–258.

Hinton, G. 2022. The Forward-Forward Algorithm: Some Preliminary Investigations. ArXiv:2212.13345 [cs].

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780. Conference Name: Neural Computation.

Horvát, S.; Gămănuț, R.; Ercsey-Ravasz, M.; Magrou, L.; Gămănuț, B.; Essen, D. C. V.; Burkhalter, A.; Knoblauch, K.; Toroczkai, Z.; and Kennedy, H. 2016. Spatial Embedding and Wiring Cost Constrain the Functional Layout of the Cortical Network of Rodents and Primates. *PLOS Biology*, 14(7): e1002512. Publisher: Public Library of Science.

Kaas, J. H.; and Collins, C. E. 2001. The organization of sensory cortex. *Current Opinion in Neurobiology*, 11(4): 498–504.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. ArXiv:2001.08361 [cs, stat].

Kietzmann, T. C.; Spoerer, C. J.; Sörensen, L. K. A.; Cichy, R. M.; Hauk, O.; and Kriegeskorte, N. 2019. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43): 21854–21863. Publisher: Proceedings of the National Academy of Sciences.

Kievit, R. A. 2020. Sensitive periods in cognitive development: a mutualistic perspective. *Current Opinion in Behavioral Sciences*, 36: 144–149.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40: e253. Publisher: Cambridge University Press.

Laughlin, S. B.; and Sejnowski, T. J. 2003. Communication in Neuronal Networks. *Science*, 301(5641): 1870–1874. Publisher: American Association for the Advancement of Science.

LeCun, Y. 2022. A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27.

Levy, W. B.; and Calvert, V. G. 2021. Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number. *Proceedings of the National Academy of Sciences*, 118(18): e2008173118. Publisher: Proceedings of the National Academy of Sciences.

Lindsay, G. W. 2020. Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, 14.

Lindsay, G. W. 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10): 2017–2031.

Luppi, A. I.; Mediano, P. A. M.; Rosas, F. E.; Holland, N.; Fryer, T. D.; O'Brien, J. T.; Rowe, J. B.; Menon, D. K.; Bor, D.; and Stamatakis, E. A. 2022. A synergistic core for human brain evolution and cognition. *Nature Neuroscience*, 25(6): 771–782. Number: 6 Publisher: Nature Publishing Group.

Löwe, S.; O'Connor, P.; and Veeling, B. S. 2020. Putting An End to End-to-End: Gradient-Isolated Learning of Representations. ArXiv:1905.11786 [cs, stat].

MacDowell, C. J.; Libby, A.; Jahn, C. I.; Tafazoli, S.; and Buschman, T. J. 2023. Multiplexed Subspaces Route Neural Activity Across Brain-wide Networks. BioRxiv: 2023.02.08.527772 Section: New Results.

Mashour, G. A.; Roelfsema, P.; Changeux, J.-P.; and Dehaene, S. 2020. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, 105(5): 776–798.

Masse, N. Y.; Yang, G. R.; Song, H. F.; Wang, X.-J.; and Freedman, D. J. 2019. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, 22(7): 1159–1167. Number: 7 Publisher: Nature Publishing Group.

Meyers, E. M.; Freedman, D. J.; Kreiman, G.; Miller, E. K.; and Poggio, T. 2008. Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. *Journal of Neurophysiology*, 100(3): 1407–1419. Publisher: American Physiological Society.

Miller, E. K.; and Buschman, T. J. 2007. Rules through Recursion: How Interactions between the Frontal Cortex and Basal Ganglia May Build Abstract, Complex Rules from Concrete, Simple Ones. In Bunge, S. A.; and Wallis, J. D., eds., *Neuroscience of Rule-Guided Behavior*, 0. Oxford University Press. ISBN 978-0-19-531427-4.

Miller, E. K.; and Cohen, J. D. 2001. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24(1): 167–202. _eprint: https://doi.org/10.1146/annurev.neuro.24.1.167.

Munakata, Y.; Herd, S. A.; Chatham, C. H.; Depue, B. E.; Banich, M. T.; and O'Reilly, R. C. 2011. A unified framework for inhibitory control. *Trends in Cognitive Sciences*, 15(10): 453–459.

Nelli, S.; Braun, L.; Dumbalska, T.; Saxe, A.; and Summerfield, C. 2023. Neural knowledge assembly in humans and neural networks. *Neuron*, 0(0). Publisher: Elsevier.

Norman, D. A.; and Shallice, T. 1986. Attention to Action. In Davidson, R. J.; Schwartz, G. E.; and Shapiro, D., eds., *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4*, 1–18. Boston, MA: Springer US. ISBN 978-1-4757-0629-1.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

Pfeiffer, J.; Ruder, S.; Vulić, I.; and Ponti, E. M. 2023. Modular Deep Learning. ArXiv:2302.11529 [cs].

Power, J. D.; Schlaggar, B. L.; Lessov-Schlaggar, C. N.; and Petersen, S. E. 2013. Evidence for Hubs in Human Functional Brain Networks. *Neuron*, 79(4): 798–813.

Raichle, M. E.; and Gusnard, D. A. 2002. Appraising the brain's energy budget. *Proceedings of the National Academy of Sciences*, 99(16): 10237–10239. Publisher: Proceedings of the National Academy of Sciences.

Rainer, G.; Asaad, W. F.; and Miller, E. K. 1998. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, 393(6685): 577–579. Number: 6685 Publisher: Nature Publishing Group.

Ralph, M. A. L.; Jefferies, E.; Patterson, K.; and Rogers, T. T. 2017. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1): 42–55. Number: 1 Publisher: Nature Publishing Group.

Ren, M.; Kornblith, S.; Liao, R.; and Hinton, G. 2023. Scaling Forward Gradient With Local Losses. ArXiv:2210.03310 [cs].

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. ArXiv:2112.10752 [cs].

Russin, J.; O'Reilly, R. C.; and Bengio, Y. 2020. DEEP LEARNING NEEDS A PREFRONTAL CORTEX. *"Bridging AI and Cognitive Science" (ICLR 2020)*.

Sakagami, M.; and Niki, H. 1994. Encoding of behavioral significance of visual stimuli by primate prefrontal neurons: relation to relevant task conditions. *Experimental Brain Research*, 97(3): 423–436.

Schmidhuber, J. 2022. Annotated History of Modern AI and Deep Learning. ArXiv:2212.11279 [cs].

Schrimpf, M.; Kubilius, J.; Hong, H.; Majaj, N. J.; Rajalingham, R.; Issa, E. B.; Kar, K.; Bashivan, P.; Prescott-Roy, J.; Schmidt, K.; Yamins, D. L. K.; and DiCarlo, J. J. 2018. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? BioRxiv: 407007 Section: New Results.

Seguin, C.; Razi, A.; and Zalesky, A. 2019. Inferring neural signalling directionality from undirected structural connectomes. *Nature Communications*, 10(1): 4289. Number: 1 Publisher: Nature Publishing Group.

Shrivastava, A.; Sukthankar, R.; Malik, J.; and Gupta, A. 2017. Beyond Skip Connections: Top-Down Modulation for Object Detection. ArXiv:1612.06851 [cs].

Skeide, M. A.; and Friederici, A. D. 2016. The ontogeny of the cortical language network. *Nature Reviews Neuroscience*, 17(5): 323–332. Number: 5 Publisher: Nature Publishing Group.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1): 159–216.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; Kluska, A.; Lewkowycz, A.; Agarwal, A.; Power, A.; et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs, stat].

Srivastava, P.; Nozari, E.; Kim, J. Z.; Ju, H.; Zhou, D.; Becker, C.; Pasqualetti, F.; Pappas, G. J.; and Bassett, D. S. 2020. Models of communication and control for brain networks: distinctions, convergence, and future outlook. *Network Neuroscience*, 4(4): 1122–1159.

Steinmetz, N. A.; Zatka-Haas, P.; Carandini, M.; and Harris, K. D. 2019. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786): 266–273. Number: 7786 Publisher: Nature Publishing Group.

Stokes, M. G. 2015. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, 19(7): 394–405.

Suarez, L. E.; Yovel, Y.; van den Heuvel, M. P.; Sporns, O.; Assaf, Y.; Lajoie, G.; and Misic, B. 2022. A connectomics-based taxonomy of mammals. *eLife*, 11: e78635. Publisher: eLife Sciences Publications, Ltd.

Tang, H.; Riley, M. R.; Singh, B.; Qi, X.-L.; Blake, D. T.; and Constantinidis, C. 2022. Prefrontal cortical plasticity during learning of cognitive tasks. *Nature Communications*, 13(1): 90. Number: 1 Publisher: Nature Publishing Group.

Tay, Y.; Dehghani, M.; Bahri, D.; and Metzler, D. 2022. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6): 109:1–109:28.

Tomasi, D.; Wang, G.-J.; and Volkow, N. D. 2013. Energetic cost of brain functional connectivity. *Proceedings of the National Academy of Sciences*, 110(33): 13642–13647. Publisher: Proceedings of the National Academy of Sciences.

VanRullen, R.; and Kanai, R. 2021. Deep learning and the Global Workspace Theory. *Trends in Neurosciences*, 44(9): 692–704.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

von Oswald, J.; Niklasson, E.; Randazzo, E.; Sacramento, J.; Mordvintsev, A.; Zhmoginov, A.; and Vladymyrov, M. 2022. Transformers learn in-context by gradient descent. ArXiv:2212.07677 [cs].

Vyas, S.; Golub, M. D.; Sussillo, D.; and Shenoy, K. V. 2020. Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, 43(1): 249–275. _eprint: https://doi.org/10.1146/annurev-neuro-092619-094115.

Wallis, J. D.; Anderson, K. C.; and Miller, E. K. 2001. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411(6840): 953–956. Number: 6840 Publisher: Nature Publishing Group.

Wang, J. X.; Kurth-Nelson, Z.; Kumaran, D.; Tirumala, D.; Soyer, H.; Leibo, J. Z.; Hassabis, D.; and Botvinick, M. 2018. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6): 860–868. Number: 6 Publisher: Nature Publishing Group.

Webb, T.; Holyoak, K. J.; and Lu, H. 2022. Emergent Analogical Reasoning in Large Language Models. ArXiv:2212.09196 [cs].

Whittington, J. C. R.; Muller, T. H.; Mark, S.; Chen, G.; Barry, C.; Burgess, N.; and Behrens, T. E. J. 2020. The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, 183(5): 1249–1263.e23.

Whittington, J. C. R.; Warren, J.; and Behrens, T. E. J. 2022. Relating transformers to models and neural representations of the hippocampal formation. ArXiv:2112.04035 [cs, q-bio].

Xu, P.; Zhu, X.; and Clifton, D. A. 2022. Multimodal Learning with Transformers: A Survey. ArXiv:2206.06488 [cs].

Xu, X.; Wu, C.; Rosenman, S.; Lal, V.; Che, W.; and Duan, N. 2023. BridgeTower: Building Bridges Between Encoders in Vision-Language Representation Learning. ArXiv:2206.08657 [cs].

Zador, A.; Escola, S.; Richards, B.; Ölveczky, B.; Bengio, Y.; Boahen, K.; Botvinick, M.; Chklovskii, D.; Churchland, A.; Clopath, C.; DiCarlo, J.; Ganguli, S.; Hawkins, J.; Koerding, K.; Koulakov, A.; LeCun, Y.; Lillicrap, T.; Marblestone, A.; Olshausen, B.; Pouget, A.; Savin, C.; Sejnowski, T.; Simoncelli, E.; Solla, S.; Sussillo, D.; Tolias, A. S.; and Tsao, D. 2023. Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution. ArXiv:2210.08340 [cs, q-bio].

Zaheer, M.; Guruganesh, G.; Dubey, A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; and Ahmed, A. 2021. Big Bird: Transformers for Longer Sequences. ArXiv:2007.14062 [cs, stat].