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

Jascha Achterberg, Danyal Akarca, Moataz Assem, Moritz Heimbach, Duncan E. Astle, John Duncan

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

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 (Attilgan 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 Krügelbach 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
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 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 (Löhwe, O’Connor, and Velling 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


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.


Neuron, 0(0). Publisher: Elsevier.


Power, J. D.; Schlaggar, B. L.; Lessov-Schlaggar, C. N.; and Petersen, S. E. 2013. Evidence for Hubs in Human Functional Brain Networks. 


