DEEP LEARNING NEEDS A PREFRONTAL CORTEX

Jacob Russin
Psychology Department
University of California Davis
Davis, California
jlrussin@ucdavis.edu

Randall C. O'Reilly
Psychology Department
Center for Neuroscience
Computer Science Department
University of California Davis
oreilly@ucdavis.edu

Yoshua Bengio
MILA
Université de Montréal
CIFAR Senior Fellow

ABSTRACT

Research seeking to build artificial systems capable of reproducing elements of human intelligence may benefit from a deeper consideration of the architecture and learning mechanisms of the human brain. In this brief review, we note a connection between many current challenges facing artificial intelligence and the functions of a particular brain area — the prefrontal cortex (PFC). This brain area is known to be involved in executive functions such as reasoning, rule-learning, deliberate or controlled processing, and abstract planning. Motivated by the hypothesis that these functions provide a form of out-of-distribution robustness currently not available in state-of-the-art AI systems, we elaborate on this connection and highlight some computational principles thought to be at work in PFC, with the goal of enhancing the synergy between neuroscience and machine learning.

1 INTRODUCTION

Deep learning, which has historically taken inspiration from the brain, has had unexpected and massive success in many applications. Though much progress has been made, new advances will be needed to meet the substantial challenges remaining on the path toward recreating the most powerful aspects of human intelligence. State-of-the-art methods remain inferior to human learners in their ability to transfer knowledge to new domains (Lake et al., 2017), to capture compositional or systematic structure (Lake & Baroni, 2018), to plan efficiently (Hamrick, 2019), and to reason abstractly (Bhagavatula et al., 2019; Xu et al., 2020). All of these abilities share similarities with the collection of human capacities known as executive functions, and are often associated with conscious processing, i.e., they can be reported verbally by human subjects. The human brain must embody principles lacking in current deep learning systems that allow it to perform these powerful functions. This has led some to consider the possibility of taking inspiration from the architecture of the human brain to build more flexible learning systems (Marblestone et al., 2016; Hassabis et al., 2017). Here, we observe an intriguing correspondence between some of the current open questions in deep learning research and the functions of the human prefrontal cortex (PFC), a brain area known to be involved in executive functions such as planning (Duncan, 1986), abstract reasoning (Donoso et al., 2014), rule-learning (Wallis et al., 2001), and controlled or deliberate processing (Miller & Cohen, 2001). We explore this connection, and the potential of translating what is known about this brain area into architectural assumptions or inductive biases in deep learning (Marblestone et al., 2016; Battaglia et al., 2018). First, we elaborate on some of the current challenges in deep learning research mentioned above, and then briefly survey some findings from neuroscience about the PFC, noting connections to these current challenges. We then discuss some theoretical ideas about PFC function from cognitive and computational neuroscience, with the aim of stimulating a fruitful synergy between neurocience and deep learning research.

2 THE NEED FOR NEURAL NETWORKS WITH EXECUTIVE FUNCTIONS

Current deep learning methods excel in perceptual tasks in which complicated patterns must be recognized in high-dimensional data. However, no one yet knows how to build learning machines which fare well on tasks that require deliberate, controlled processing over multiple steps or dealing with changes in distribution (Bengio, 2017; 2019; Lake et al., 2017; Marcus, 2018). In the following, we highlight some of the aspects of human cognition that have so far proven difficult for neural networks to reproduce, and have become active areas of research in deep learning.

Reasoning Bottou (2011) offers a helpful working definition of reasoning as "algebraically manipulating previously acquired knowledge in order to answer a new question." What this definition entails is the reuse of dynamically selected computational modules, with the results of recently produced computations feeding

the currently selected computation. Most of the tasks at the center of the rise of deep learning (e.g., object recognition, video-game playing, machine translation) generally do not require reasoning, i.e., algebraic manipulation of existing knowledge. Current neural nets involve composition of functions (e.g. layers) but in a fixed order. Recently, there has been growing interest and much progress on datasets and tasks that require reasoning over multiple steps (e.g. Bhagavatula et al., 2019; Johnson et al., 2017; Graves et al., 2016; Hudson & Manning, 2019; Weston et al., 2015; Xu et al., 2020; Barrett et al., 2018), with some cases of systems surpassing human performance (e.g. Hudson & Manning, 2018; Perez et al., 2018; Yi et al., 2018). However, as we discuss in the next section, most models trained on these tasks still fail to "answer new questions", in the sense of generalizing outside of the training distribution (Barrett et al., 2018; Bahdanau et al., 2019a).

Compositionality and systematicity It has been argued that one of the most powerful aspects of human cognition is its systematicity: concepts can be composed in novel ways, so that the number of expressible combinations grows exponentially in the number of primitive concepts learned (Fodor & Pylyshyn, 1988; Lake et al., 2017; Lake & Baroni, 2018). This topic is closely related to reasoning, because "algebraic manipulation" requires that existing knowledge be represented in a form that is systematically composable. Interest in compositionality among deep learning researchers has grown over the past few years, where experiments have revealed that standard approaches in deep learning show weak generalization for compositions of known elements which are unlikely under the training distribution (Lake & Baroni, 2018; Bahdanau et al., 2019b;a; Keysers et al., 2020, but see also Hill et al. 2020). These experiments show that standard architectures fail to capture the compositional structure or systematic rules governing the data-generation process.

Control in novel environments Just as standard deep networks have weaker generalization outside of their training distribution in the settings described above, they are less efficient than humans at transferring knowledge about learned environments to novel ones (Hagendorff & Wezel, 2019; Kansky et al., 2017; Lake et al., 2019b; 2017). For example, when trained on the Atari games, the generalization of standard methods in deep reinforcement learning is not robust to slight changes in the rules of the game or the layout of the inputs (Kansky et al., 2017). Generalization to novel environments has continued to be an important topic in deep learning research, where an increased focus on one-shot learning (e.g. Vinyals et al., 2017), transfer (Weiss et al., 2016), and meta-learning (e.g. Finn et al., 2017; Bengio et al., 2019) has emerged. Much progress has been made in these areas, but human-level transfer remains elusive (Lansdell & Kording, 2019; Griffiths et al., 2019).

Abstract planning It has long been recognized that the standard planning algorithms used in model-based reinforcement learning (RL) are too computationally expensive to be useful in many real-world domains (Barto & Mahadevan, 2003), and that humans and other animals seem to possess planning strategies that avoid much of this computational cost (Botvinick, 2008; Botvinick et al., 2009). In particular, it has been suggested that humans plan using temporally abstract representations, whereas model-based algorithms usually treat each time-step independently (Botvinick et al., 2009; Botvinick & Weinstein, 2014). The most successful algorithms in deep RL are model-free (Arulkumaran et al., 2017; Hamrick, 2019), and though model-based deep RL methods have had some recent success (Corneil et al., 2018; Finn & Levine, 2017; Nagabandi et al., 2018; Feinberg et al., 2018), most still plan each time step individually or lack the abstraction and compositionality displayed in human planning (Hamrick, 2019).

3 Some functions of the prefrontal cortex

All of the challenges described above have been noted by others, and are active areas of research. The first of our main contributions is to draw connections between them and the functioning of the human PFC. The PFC comprises a large swath of the most anterior portion of the cerebral cortex and appears to have undergone a disproportionate amount of development over the course of human evolution (Schoenemann et al., 2005; Rilling, 2006; Semendeferi et al., 2001; Falk, 2012). It receives highly processed, multimodal information from perceptual areas, and sits at the top of the decision-making hierarchy (Fuster, 2009; Hunt & Hayden, 2017; O'Reilly et al., 2012). Much remains unknown about the PFC, and in particular there is ongoing investigation into functional differentiation between different areas within it (e.g. Hunt et al., 2018). However, it has been argued that much of the PFC retains a canonical computational role, with functional differentiation among subareas emerging due to differences in connectivity (Miller & Cohen, 2001; O'Reilly, 2010; Thompson-Schill, 2004). Here we highlight some aspects of the general functionality of the PFC.

Reasoning One of the most well-established findings about the PFC is that it is specialized for working memory, or the ability to maintain and manipulate information over short periods of time (Fuster & Alexander, 1971; Kubota & Niki, 1971; Miller & Desimone, 1994; Goldman-Rakic, 1995; Sommer & Wurtz, 2000; Lara & Wallis, 2015). Working memory can be seen as an important aspect of the capacity to reason, as it allows for 1) computation on information that is not currently observable in the environment, and 2) the integration of intermediate results in a larger reasoning process (e.g., in a serial summation of a list of numbers; Menon, 2016). Indeed, evidence of prefrontal engagement has been found in many experiments investigating the

neural underpinnings of human reasoning (Donoso et al., 2014), including deductive (Goel, 2007), inductive (Crescentini et al., 2011), relational (Krawczyk et al., 2011), and analogical reasoning (Hampshire et al., 2011).

Representing abstract rules One domain in which humans excel at generalizing outside of the training distribution is the ability to apply known rules to novel elements (Lake et al., 2019a). The PFC has been found to be important for success on tasks that require the induction, maintenance, updating, or application of rules (Mian et al., 2014; Milner, 1963; Wallis et al., 2001; Shallice & Burgess, 1991). For example, patients with damage to PFC struggle to sort cards according to a changing rule (e.g., color or shape) (Milner, 1963; Buchsbaum et al., 2005; Berg, 1948). In a seminal electrophysiology study on rule application (Wallis et al., 2001), monkeys were trained to either select the picture that was a 'match' to the previously presented one, or select the 'nonmatch'. Single neurons in PFC were found to respond when invoking such abstract rules, regardless of the particular pictures presented on a given trial (Wallis et al., 2001). Some computational models (Rougier et al., 2005) have attempted to capture this important property of the PFC, showing, e.g., how indirection might be implemented in a canonical PFC circuit (Kriete et al., 2013, see also Hayworth & Marblestone 2018 and Müller et al. 2016).

Control in novel environments Overwhelming evidence implicates the PFC in decision-making and control processes (Domenech & Koechlin, 2015; Miller & Cohen, 2001). However, it is in general not crucial for the execution of habitual responses that have been trained extensively (as would be the case, e.g., in a model that had played an Atari game for hundreds of hours) — rather, it is required for *overriding* these habitual responses in novel situations, with new rules or in the pursuit of a novel goal (Miller & Cohen, 2001; Botvinick & Cohen, 2014). This function, generally termed "cognitive control," is illustrated well in studies using the classic Stroop task (Stroop, 1935). In this task, participants are presented with color words (e.g., 'red', 'blue') written in colored ink, which may or may not match the words. Patients with damage to PFC perform reliably poorly on this task, which requires them to override habitual responses (reading text) according to the color-naming rule (Perret, 1974; Vendrell et al., 1995). In general, it is thought that the functioning of the PFC is crucially important when a novel goal is being pursued in a familiar environment where habits have become entrenched, or in novel environments when no such habits yet exist (Miller & Cohen, 2001).

Abstract planning Humans and other mammals demonstrate evidence of both model-free and model-based RL (Momennejad et al., 2017; Daw et al., 2011; 2005), but the PFC has been implicated in model-based RL in particular (Daw et al., 2005; Smittenaar et al., 2013). Humans with damage to the PFC can exhibit deficits in routine behaviors that require planning and coordinating sequences of actions like cooking or making coffee (Miller & Cohen, 2001; Levine et al., 1998; Duncan, 1986; Shallice, 1982). Some have theorized that the planning processes in PFC are temporally abstract or hierarchical, as in, e.g., the options framework (Sutton et al., 1999; Botvinick, 2008; Botvinick et al., 2009; Botvinick & Weinstein, 2014; Frank & Badre, 2012). This idea accords well with experiments indicating that PFC represents actions at multiple timescales simultaneously (Hunt & Hayden, 2017; Botvinick et al., 2009; Sarafyazd & Jazayeri, 2019).

4 Computational principles and learning mechanisms in PFC

The section above describing some of the functions of the PFC was structured to draw out their connection to current challenges facing deep learning. However, the structure of this section is somewhat arbitrary, as all of these functions are related to one another. Here, we cover some theoretical ideas about the underlying computational mechanisms of PFC that can unify these various functions, with an eye toward principles that may be transferable to deep learning.

Top-down attention and modulation In an influential framework, Miller & Cohen (2001) argue that many of the cognitive capacities associated with the PFC, including reasoning, rule-learning, planning, and cognitive control, can be explained by its role in top-down attentional modulation of other brain areas. The PFC sends projections to much of neocortex, allowing it to modulate activity in other areas, possibly according to a current goal or in agreement with currently conscious contents. In the Stroop task, e.g., the PFC represents the instruction to name the colors rather than read the words, and modulates the activity of color features in higher-order visual areas of the brain to bias behavior toward naming them (Miller & Cohen, 2001).

Top-down attentional modulation has some analogues in deep learning research. The use of attention has become an increasingly popular approach in many tasks (e.g. Bahdanau et al., 2014; Xu et al., 2016; Hudson & Manning, 2018). One major difference with these mechanisms may be that PFC is thought to modulate activity through multiple brain areas at once, conditioned on the current goal. This kind of conditioning may be more similar to HyperNetworks (Ha et al., 2016), FiLM (Perez et al., 2018), where the mapping learned by a single feedforward network can be modulated with transformations at each layer, or RIM, which tries explicitly to model a top-down attentional modulation mechanism (Goyal et al., 2019).

Recurrence, gating, and seriality Recurrence is ubiquitous in the brain, but the PFC has a special role in maintaining information in working memory over longer timescales (Lara & Wallis, 2015). Work

in computational neuroscience examining the detailed biological mechanisms that would allow PFC to accomplish this has emphasized its interaction with the basal ganglia and the importance of LSTM-like gating operations (O'Reilly & Frank, 2006). Although computations in the brain are massively parallelized, the amount of information that can be maintained in working memory at a given time is notoriously small (Petri et al., 2017; Feng et al., 2014; Oberauer & Kliegl, 2006). This means that seriality is also an important aspect of how the PFC operates: top-down attention must be applied serially over the course of a planning or reasoning process, and intermediate results must be integrated over time. However, the serial processing of a few elements at a time can also be an advantage, as it enables arbitrary sequences of complex computational processing at each of these steps to be combined to obtain more powerful and compositional computation. This may be an important factor for supporting Turing-machine like universal computation (Graves et al., 2014; Newell, 1990), and for generalizing outside of the training distribution (Bengio, 2017; 2019).

Learning: Dopamine and reinforcement Recent proposals from deep learning researchers have encouraged neuroscientists to focus on the architectures, learning algorithms, and cost functions in the brain, as opposed to the more traditional approach of characterizing the low-level biological mechanisms or tuning properties of neurons or neuronal populations (Marblestone et al., 2016; Richards et al., 2019). This approach has been emphasized in research in connectionism and parallel distributed processing for decades (Rumelhart et al., 1986), but much remains unknown about the learning mechanisms and cost functions that might be at work in biological neurons (Richards et al., 2019). Reinforcement learning is thought to be especially important for learning in the PFC, which receives ample dopamine signals conveying reward prediction errors (O'Reilly & Frank, 2006), and is heavily involved in decision-making and planning (Rushworth & Behrens, 2008). A recent proposal shows that a number of empirical findings can be explained by a model in which the PFC implements a meta-reinforcement learning system, trained by dopamine to instantiate an RL procedure within the dynamics of its neural activity (Wang et al., 2018).

A PFC module for deep learning? Much remains unknown, but an overall picture of the PFC that has emerged in cognitive and computational neuroscience is one where it selects, maintains, and manipulates learned representations in other areas of the brain through a serial process of top-down attentional modulation (Miller & Cohen, 2001; O'Reilly & Frank, 2006; Hazy et al., 2007). This serial processing may be tuned through reinforcement or meta-reinforcement learning and dopamine signals to optimize performance on tasks that require reasoning, rule-like representations, sequential and dynamic recombination of computations, cognitive control, or temporally abstract planning. This kind of system may be critical to ensuring flexibility in familiar environments and controlled decision-making in novel ones, and may allow for efficient planning on multiple timescales.

Many of the current major challenges facing deep learning research involve tasks that require an extended notion of generalization, not just to examples from the same distribution as the past observations, but also to out-of-distribution inputs (Lake & Baroni, 2018; Bahdanau et al., 2019b; Bengio, 2019). The ability to handle such non-stationarities would naturally evolve because learning agents (who change their policy and thus end up visiting different states of the environment) naturally face them, and even more so in a social multi-agent context where the environment itself changes. Some of the paradigmatic cases in which humans are able to do this involve the application of known rules to novel elements (Lake et al., 2019a) — a cognitive function that has been associated with the PFC (Miller & Cohen, 2001; Wallis et al., 2001). This systematicity is natural in symbolic systems typical of classical approaches to AI, but these lack many of the powerful advantages brought by deep learning (such as the ability to learn efficiently on a large scale, to handle uncertainty, to generalize well across symbols through distributed representations, and to ground these symbols in a complex perceptual reality). These symbolic systems utilize the notions of indirection or of variables — arrays of memory that can be manipulated by computations *that do not depend on the specific content stored there* — ensuring the kind of abstraction necessary for this kind of systematic generalization to emerge. An analogous independence may exist between the PFC and posterior sensory and association areas: the PFC may be able to select and manipulate representational content in these areas according to learned rules that can be applied to many different elements (Russin et al., 2019; Kriete et al., 2013). This may provide the kind of abstraction and compositionality currently missing from standard architectures in deep learning (Bengio, 2017; Bengio et al., 2019).

5 Conclusion

We have argued that there is a striking correspondence between the tasks on which humans outperform current AI systems and the executive functions associated with the PFC. We believe that a greater focus on the principles and inductive biases at work in the PFC may inspire novel architectures that can accomplish similar functions. Much remains to be learned in making these principles more concrete and in implementing them in working systems, but we hope that we have taken a step in this direction and that this work will facilitate greater synergy between neuroscience and AI in the future.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their thorough comments and useful suggestions and references. We would also like to thank Jason Jo, and all of the members of the Computational Cognitive Neuroscience Lab at UC Davis for helpful ongoing discussion on these topics. This work was supported by ONR N00014-19-1-2684 / N00014-18-1-2116, ONR N00014-14-1-0670 / N00014-16-1-2128, and ONR N00014-18-C-2067.

REFERENCES

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, 34(6):26–38, November 2017. ISSN 1558-0792. doi: 10.1109/MSP.2017.2743240.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*, May 2014.

Dzmitry Bahdanau, Harm de Vries, Timothy J. O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. CLOSURE: Assessing Systematic Generalization of CLEVR Models. *arXiv:1912.05783 [cs]*, December 2019a.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron C. Courville. Systematic Generalization: What Is Required and Can It Be Learned? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b.

David G. T. Barrett, Felix Hill, Adam Santoro, Ari S. Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. *arXiv:1807.04225 [cs, stat]*, July 2018.

Andrew G. Barto and Sridhar Mahadevan. Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems*, 13(4):341–379, October 2003. ISSN 0924-6703, 1573-7594. doi: 10.1023/A:1025696116075.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [cs, stat]*, June 2018.

Yoshua Bengio. The Consciousness Prior. *arXiv:1709.08568 [cs, stat]*, December 2017.

Yoshua Bengio. From System 1 Deep Learning to System 2 Deep Learning. November 2019.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *ICLR'2020*, 2019.

E. A. Berg. A simple objective technique for measuring flexibility in thinking. *The Journal of General Psychology*, 39:15–22, July 1948. ISSN 0022-1309. doi: 10.1080/00221309.1948.9918159.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive Commonsense Reasoning. *arXiv:1908.05739 [cs]*, August 2019.

Leon Bottou. From Machine Learning to Machine Reasoning. *arXiv:1102.1808 [cs]*, February 2011.

Matthew Botvinick and Ari Weinstein. Model-based hierarchical reinforcement learning and human action control. *Phil. Trans. R. Soc. B*, 369(1655):20130480, November 2014. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2013.0480.

Matthew M. Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, 12(5):201–208, May 2008. ISSN 1364-6613. doi: 10.1016/j.tics.2008.02.009.

Matthew M. Botvinick and Jonathan D. Cohen. The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, 38(6):1249–1285, August 2014. ISSN 1551-6709. doi: 10.1111/cogs.12126.

M.M. Botvinick, Y. Niv, and A C. Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009. ISSN 1873-7838. doi: 10.1016/j.cognition.2008.08.011.

Bradley R. Buchsbaum, Stephanie Greer, Wei-Li Chang, and Karen Faith Berman. Meta-analysis of neuroimaging studies of the Wisconsin Card-Sorting task and component processes. *Human Brain Mapping*, 25(1):35–45, 2005. ISSN 1097-0193. doi: 10.1002/hbm.20128.

Dane Corneil, Wulfram Gerstner, and Johanni Brea. Efficient Model-Based Deep Reinforcement Learning with Variational State Tabulation. *arXiv:1802.04325 [cs, stat]*, June 2018.

Cristiano Crescentini, Shima Seyed-Allaei, Nicola De Pisapia, Jorge Jovicich, Daniele Amati, and Tim Shallice. Mechanisms of Rule Acquisition and Rule Following in Inductive Reasoning. *Journal of Neuroscience*, 31(21):7763–7774, May 2011. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4579-10.2011.

Nathaniel D. Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, December 2005. ISSN 1097-6256. doi: 10.1038/nn1560.

Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, March 2011. ISSN 1097-4199. doi: 10.1016/j.neuron.2011.02.027.

Philippe Domenech and Etienne Koechlin. Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, 1:101–106, February 2015. ISSN 2352-1546. doi: 10.1016/j.cobeha.2014.10.007.

Maël Donoso, Anne G. E. Collins, and Etienne Koechlin. Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191):1481–1486, June 2014. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1252254.

John Duncan. Disorganisation of behaviour after frontal lobe damage. *Cognitive Neuropsychology*, 3(3): 271–290, August 1986. ISSN 0264-3294. doi: 10.1080/02643298608253360.

Dean Falk. Hominin paleoneurology. In *Progress in Brain Research*, volume 195, pp. 255–272. Elsevier, 2012. ISBN 978-0-444-53860-4. doi: 10.1016/B978-0-444-53860-4.00012-X.

Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine. Model-Based Value Estimation for Efficient Model-Free Reinforcement Learning. *arXiv:1803.00101 [cs, stat]*, February 2018.

S. F. Feng, M. Schwemmer, S. J. Gershman, and J. D. Cohen. Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective & Behavioral Neuroscience*, 14(1):129–146, March 2014. ISSN 1531-135X. doi: 10.3758/s13415-013-0236-9.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2786–2793, May 2017. doi: 10.1109/ICRA.2017.7989324.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv:1703.03400 [cs]*, July 2017.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, March 1988. ISSN 00100277. doi: 10.1016/0010-0277(88)90031-5.

Michael J. Frank and David Badre. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex (New York, N.Y.: 1991)*, 22(3):509–526, March 2012. ISSN 1460-2199. doi: 10.1093/cercor/bhr114.

J. M. Fuster. Prefrontal Cortex. In Larry R. Squire (ed.), *Encyclopedia of Neuroscience*, pp. 905–908. Academic Press, Oxford, January 2009. ISBN 978-0-08-045046-9. doi: 10.1016/B978-008045046-9.01118-9.

J. M. Fuster and G. E. Alexander. Neuron activity related to short-term memory. *Science*, 173:652–654, January 1971.

Vinod Goel. Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, 11(10):435–441, October 2007. ISSN 1364-6613. doi: 10.1016/j.tics.2007.09.003.

P. S. Goldman-Rakic. Cellular basis of working memory. *Neuron*, 14(3):477–485, March 1995.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent Independent Mechanisms. *arXiv:1909.10893 [cs, stat]*, September 2019.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing Machines. *arXiv:1410.5401 [cs]*, December 2014.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, October 2016. ISSN 1476-4687. doi: 10.1038/nature20101.

Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30, October 2019. ISSN 23521546. doi: 10.1016/j.cobeha.2019.01.005.

David Ha, Andrew Dai, and Quoc V. Le. HyperNetworks. *arXiv:1609.09106 [cs]*, December 2016.

Thilo Hagendorff and Katharina Wezel. 15 challenges for AI: Or what AI (currently) can't do. *AI & SOCIETY*, March 2019. ISSN 1435-5655. doi: 10.1007/s00146-019-00886-y.

Adam Hampshire, Russell Thompson, John Duncan, and Adrian M. Owen. Lateral prefrontal cortex subregions make dissociable contributions during fluid reasoning. *Cerebral Cortex (New York, N.Y.: 1991)*, 21(1):1–10, January 2011. ISSN 1460-2199. doi: 10.1093/cercor/bhq085.

Jessica B Hamrick. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29:8–16, October 2019. ISSN 2352-1546. doi: 10.1016/j.cobeha.2018.12.011.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, July 2017. ISSN 0896-6273. doi: 10.1016/j.neuron.2017.06.011.

Kenneth J. Hayworth and Adam H. Marblestone. How thalamic relays might orchestrate supervised deep training and symbolic computation in the brain. *bioRxiv*, pp. 304980, April 2018. doi: 10.1101/304980.

Thomas E. Hazy, Michael J. Frank, and R. C. O'Reilly. Towards an executive without a homunculus: Computational models of the prefrontal Cortex/Basal ganglia system. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1485):1601–1613, August 2007.

Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Environmental drivers of systematicity and generalization in a situated agent. *arXiv:1910.00571 [cs]*, February 2020.

Drew A. Hudson and Christopher D. Manning. Compositional Attention Networks for Machine Reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. *arXiv:1902.09506 [cs]*, May 2019.

Laurence T. Hunt and Benjamin Y. Hayden. A distributed, hierarchical and recurrent framework for reward-based choice. *Nature Reviews Neuroscience*, 18(3):172–182, March 2017. ISSN 1471-0048. doi: 10.1038/nrn.2017.7.

Laurence T. Hunt, W. M. Nishantha Malalasekera, Archy O. de Berker, Bruno Miranda, Simon F. Farmer, Timothy E. J. Behrens, and Steven W. Kennerley. Triple dissociation of attention and decision computations across prefrontal cortex. *Nature Neuroscience*, 21(10):1471–1481, October 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0239-5.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, July 2017. doi: 10.1109/CVPR.2017.215.

Ken Kansky, Tom Silver, David A. Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. Schema Networks: Zero-Shot Transfer with a Generative Causal Model of Intuitive Physics. *arXiv:1706.04317 [cs]*, August 2017.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. pp. 38, 2020.

Daniel C. Krawczyk, M. Michelle McClelland, and Colin M. Donovan. A hierarchy for relational reasoning in the prefrontal cortex. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 47(5):588–597, May 2011. ISSN 1973-8102. doi: 10.1016/j.cortex.2010.04.008.

T. Kriete, D. C. Noelle, J. D. Cohen, and R. C. O'Reilly. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, 110(41): 16390–16395, October 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1303547110.

K. Kubota and H. Niki. Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, 34(3):337–347, September 1971.

Brenden M. Lake and Marco Baroni. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2879–2888. PMLR, 2018.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *The Behavioral and Brain Sciences*, 40:e253, January 2017. ISSN 1469-1825. doi: 10.1017/S0140525X16001837.

Brenden M. Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. In Ashok K. Goel, Colleen M. Seifert, and Christian Freksa (eds.), *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pp. 611–617. cognitivesciencesociety.org, 2019a. ISBN 978-0-9911967-7-7.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. The Omniglot challenge: A 3-Year progress report. *arXiv:1902.03477 [cs]*, May 2019b.

Benjamin James Lansdell and Konrad Paul Kording. Towards learning-to-learn. *Current Opinion in Behavioral Sciences*, 29:45–50, October 2019. ISSN 2352-1546. doi: 10.1016/j.cobeha.2019.04.005.

Antonio H. Lara and Jonathan D. Wallis. The Role of Prefrontal Cortex in Working Memory: A Mini Review. *Frontiers in Systems Neuroscience*, 9, December 2015. ISSN 1662-5137. doi: 10.3389/fnsys.2015.00173.

B. Levine, D. T. Stuss, W. P. Milberg, M. P. Alexander, M. Schwartz, and R. Macdonald. The effects of focal and diffuse brain damage on strategy application: Evidence from focal lesions, traumatic brain injury and normal aging. *Journal of the International Neuropsychological Society: JINS*, 4(3):247–264, May 1998. ISSN 1355-6177.

Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, 10, 2016. ISSN 1662-5188. doi: 10.3389/fncom.2016.00094.

Gary Marcus. Deep learning: A critical appraisal. January 2018.

V. Menon. Memory and cognitive control circuits in mathematical cognition and learning. *Progress in brain research*, 227:159–186, 2016. ISSN 0079-6123. doi: 10.1016/bs.pbr.2016.04.026.

Matthew K. Mian, Sameer A. Sheth, Shaun R. Patel, Konstantinos Spiliopoulos, Emad N. Eskandar, and Ziv M. Williams. Encoding of Rules by Neurons in the Human Dorsolateral Prefrontal Cortex. *Cerebral Cortex (New York, NY)*, 24(3):807–816, March 2014. ISSN 1047-3211. doi: 10.1093/cercor/bhs361.

E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001.

E. K. Miller and R. Desimone. Parallel neuronal mechanisms for short-term memory. *Science (New York, N.Y.)*, 263:520–522, February 1994.

Brenda Milner. Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes. *Archives of Neurology*, 9(1):90–100, July 1963. ISSN 0003-9942. doi: 10.1001/archneur.1963.00460070100010.

I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692, September 2017. ISSN 2397-3374. doi: 10.1038/s41562-017-0180-8.

Michael G. Müller, Christos H. Papadimitriou, Wolfgang Maass, and Robert Legenstein. A model for structured information representation in neural networks. *arXiv:1611.03698 [q-bio]*, November 2016.

Anusha Nagabandi, Gregory Kahn, Ronald S. Fearing, and Sergey Levine. Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7559–7566, May 2018. doi: 10.1109/ICRA.2018.8463189.

A. Newell. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA, January 1990.

Klaus Oberauer and Reinhold Kliegl. A Formal Model of Capacity Limits in Working Memory. *Journal of Memory and Language*, 55(4):601–626, November 2006. ISSN 0749-596X. doi: 10.1016/j.jml.2006.08.009.

R. C. O'Reilly. The What and How of prefrontal cortical organization. *Trends in Neurosciences*, 33(8): 355–361, August 2010. ISSN 0166-2236. doi: 10.1016/j.tins.2010.05.002.

R. C. O'Reilly and Michael J. Frank. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2006.

R. C. O'Reilly, Yuko Munakata, Michael J. Frank, Thomas E. Hazy, and Contributors. *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu, 2012.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3942–3951. AAAI Press, 2018.

E. Perret. The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour. *Neuropsychologia*, 12(3):323–330, July 1974. ISSN 0028-3932. doi: 10.1016/0028-3932(74)90047-5.

Giovanni Petri, Sebastian Musslick, Biswadip Dey, Kayhan Ozcimder, Nesreen K. Ahmed, Theodore Willke, and Jonathan D. Cohen. Universal limits to parallel processing capability of network architectures. *arXiv:1708.03263 [q-bio]*, August 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, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, November 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0520-2.

James K. Rilling. Human and nonhuman primate brains: Are they allometrically scaled versions of the same design? *Evolutionary Anthropology: Issues, News, and Reviews*, 15(2):65–77, 2006. ISSN 1520-6505. doi: 10.1002/evan.20095.

N. P. Rougier, D. Noelle, T. S. Braver, J. D. Cohen, and R. C. O'Reilly. Prefrontal Cortex and the Flexibility of Cognitive Control: Rules Without Symbols. *Proceedings of the National Academy of Sciences*, 102 (20):7338–7343, January 2005.

David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group (eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*. MIT Press, Cambridge, MA, USA, 1986. ISBN 978-0-262-13218-3.

Matthew F. S. Rushworth and Timothy E. J. Behrens. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, 11(4):389–397, April 2008. ISSN 1546-1726. doi: 10.1038/nn2066.

Jake Russin, Jason Jo, Randall C. O'Reilly, and Yoshua Bengio. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv:1904.09708 [cs, stat]*, May 2019.

Morteza Sarafyazd and Mehrdad Jazayeri. Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441), May 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aav8911.

P. Thomas Schoenemann, Michael J. Sheehan, and L. Daniel Glotzer. Prefrontal white matter volume is disproportionately larger in humans than in other primates. *Nature Neuroscience*, 8(2):242–252, February 2005. ISSN 1546-1726. doi: 10.1038/nn1394.

Katerina Semendeferi, Este Armstrong, Axel Schleicher, Karl Zilles, and Gary W. Van Hoesen. Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Physical Anthropology*, 114(3):224–241, 2001. ISSN 1096-8644. doi: 10.1002/1096-8644(200103)114:3⟨224::AID-AJPA1022⟩3.0.CO;2-I.

T. Shallice. Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 298(1089):199–209, June 1982. ISSN 0962-8436. doi: 10.1098/rstb.1982.0082.

T. Shallice and P. W. Burgess. Deficits in strategy application following frontal lobe damage in man. *Brain: A Journal of Neurology*, 114 (Pt 2):727–741, April 1991. ISSN 0006-8950. doi: 10.1093/brain/114.2.727.

Peter Smittenaar, Thomas H.B. FitzGerald, Vincenzo Romei, Nicholas D. Wright, and Raymond J. Dolan. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-Free Control in Humans. *Neuron*, 80(4):914–919, November 2013. ISSN 0896-6273. doi: 10.1016/j.neuron.2013.08.009.

M. A. Sommer and R. H. Wurtz. Composition and topographic organization of signals sent from the frontal eye field to the superior colliculus. *Journal of Neurophysiology*, 83(4):1979–2001, April 2000.

J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18 (6):643–662, 1935. ISSN 0022-1015(Print). doi: 10.1037/h0054651.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, August 1999. ISSN 00043702. doi: 10.1016/S0004-3702(99)00052-1.

Sharon L. Thompson-Schill. Dissecting the language organ : A new look at the role of Broca ' s area in language processing. 2004.

P. Vendrell, C. Junqué, J. Pujol, M. A. Jurado, J. Molet, and J. Grafman. The role of prefrontal regions in the Stroop task. *Neuropsychologia*, 33(3):341–352, March 1995. ISSN 0028-3932. doi: 10.1016/0028-3932(94)00116-7.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. *arXiv:1606.04080 [cs, stat]*, December 2017.

J. D. Wallis, K. C. Anderson, and E. K. Miller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411:953–956, June 2001.

Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860–868, June 2018. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-018-0147-8.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016. ISSN 2196-1115. doi: 10.1186/s40537-016-0043-6.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv:1502.05698 [cs, stat]*, December 2015.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv:1502.03044 [cs]*, April 2016.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. WHAT CAN NEURAL NETWORKS REASON ABOUT? pp. 18, 2020.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1031–1042. Curran Associates, Inc., 2018.