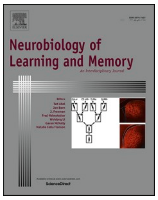
Neurobiology of Learning and Memory 186 (2021) 107552

Contents lists available at ScienceDirect

Neurobiology of Learning and Memory

journal homepage: www.elsevier.com/locate/ynlme



Neural circuits and symbolic processing

Quan Do, Michael E. Hasselmo

Center for Systems Neuroscience, Boston University, 610 Commonwealth Ave, Boston, MA 02215, United States

ARTICLE INFO

Keywords:
Conjunctive coding
Dynamic binding
Reinforcement learning
Bayesian inference
Program synthesis
Category theory

ABSTRACT

The ability to use symbols is a defining feature of human intelligence. However, neuroscience has yet to explain the fundamental neural circuit mechanisms for flexibly representing and manipulating abstract concepts. This article will review the research on neural models for symbolic processing. The review first focuses on the question of how symbols could possibly be represented in neural circuits. The review then addresses how neural symbolic representations could be flexibly combined to meet a wide range of reasoning demands. Finally, the review assesses the research on program synthesis and proposes that the most flexible neural representation of symbolic processing would involve the capacity to rapidly synthesize neural operations analogous to lambda calculus to solve complex cognitive tasks.

1. Introduction

The human brain has the capacity to guide behavior based on representations of the full scope of human knowledge, ranging from practical navigation through tasks in the daily world to the highest levels of abstract thought on topics of philosophical, social, scientific, or mathematical concepts. However, the fundamental neural circuit mechanisms for flexibly representing a broad range of concepts for a broad range of behaviors have not been elucidated. Within psychology, the flexible formation of new concepts for guiding a wide range of different behaviors is referred to as general intelligence. Most researchers seem to accept that we have not yet described the neural mechanisms for general intelligence. While much progress has been made on training neural models to perform specific tasks such as image recognition, neural models do not have the capacity to flexibly solve multiple new problems in the manner that humans can. This has been shown by relatively limited performance in tasks such as the Raven's progressive matrices task (Carpenter et al., 1990; Raven, 2003) and the Abstraction and Reasoning Challenge (Chollet, 2019) exhibited by existing neural models (Barrett et al., 2018; Kolev et al., 2020; Rasmussen & Eliasmith, 2011; Raudies & Hasselmo, 2017)

Most existing neural network models of brain function utilize a standard representation of neural activity as vectors of activity which spread through matrices of synaptic connections that influence activity in other vector representations. However, the flexible formation of new concepts and representations appears to require an intermediate level of representation that is not fully expressed in this neural network vector code. For instance, deep neural network models have been criticized for failing to represent the essential features of symbolic processing found in human behavior (Lake et al., 2017). These include the properties of productivity, compositionality and systematicity (Fodor and Pylyshyn, 1988). Productivity refers to the capacity for a set of representations to generate an infinite number of meaningful combinations. Compositionality and systematicity refer to the capacity to take component elements of representations and use them with the same meaning in other circumstances. Handwritten characters recognition, a task in which deep neural network and humans both excel in, can illustrate this point (Lake et al., 2017). People can learn to recognize a new handwritten character from a single example, whereas deep neural network require a lot more training data. Moreover, people learn a concept when they do pattern recognition. They can parse a character into its most important parts and relations (compositionality), apply them to different situations (systematicity), and generate new examples (productivity). Deep neural network models have yet to demonstrate these abilities (Marcus, 2018).

This article reviews research on neural models for symbolic processing. The review first focuses on the question of how symbols can be represented as role-filler interactions in neural circuits, which would satisfy the productivity, compositionality and systematicity requirements. The review then addresses how neural symbolic representations could systematically and flexibly be constructed and combined to solve a variety of behavioral tasks by forming hierarchical representations and planning actions. Finally, the review addresses the research on program synthesis and proposes that the most flexible neural representation of symbolic processing would involve the capacity for neural program synthesis, using a flexible set of neural operations analogous to lambda calculus.

E-mail addresses: qdo@gmail.com (Q. Do), hasselmo@bu.edu (M.E. Hasselmo).

https://doi.org/10.1016/j.nlm.2021.107552
Received 13 May 2021; Received in revised form 14 October 2021; Accepted 2 November 2021
Available online 8 November 2021

2. Historical background

For many decades, various fields of research in computer science, psychology and neuroscience have wrestled with the question of whether neural circuits in the brain have mechanisms for manipulations of symbols in a manner analogous to computers. One of the first researchers to emphasize this idea was Allen Newell. Inspired by the amazing feats accomplished by digital computers, Newell hypothesized that there is a physical symbol system that underlies humans' cognitive abilities, and that "these symbols are in fact the same symbols that we humans have and use every day of our lives" (Newell, 1980). Symbols are useful because they allow us to reason about relational roles in which objects or entities are engaged, rather than just viewing the literal features of the objects. This ability to think explicitly about relations is central to mathematics, science, engineering, art, or even simple tasks like planning a meal or making an analogy (Gentner, 1983; Goldstone et al., 1991; Holland et al., 1989; Holyoak & Thagard, 1995; Hummel, 2000; Palmer, 1978). It arguably underlies some of the most fascinating aspects of human experience.

2.1. The challenge of neural representations of symbols

If symbols are indeed stored and manipulated, then what is the data structure for the symbols in the brain, that is, how can symbols be represented? Classical artificial intelligence (AI) researchers, those who followed Newell's footsteps, essentially ignored this question and described symbols as the syntactic mechanisms developed in propositional logic (Minker, 2000), which deals with the ways statements relate to and interact with one another. In this view, intelligence behaviors can emerge from the "same symbols that we humans have and use every day of our lives", as Newell has hypothesized (Newell, 1980). One obvious problem with this approach is that syntax does not equate to semantics, and no matter how structured these statements or propositions are organized, the associative semantic links to the meaning behind each word, a fundamental aspect of cognitive processes, are missing. Indeed, many AI researchers today are stepping away from pure syntactic representations and emphasizing the importance of tying semantic meaning to symbol (Santoro et al., 2021).

Due to the initial failure of the classicists, an opposing camp proposed that information is stored non-symbolically within neurons and in the connections between networks of neurons (Churchland & Sejnowski, 1992). They argued that symbolic processing was an inadequate model for the cognitive flexibility exhibited by the mind (Cummins, 1991; Elman, 1990; McClelland et al., 1995; Rogers & McClelland, 2004; St. John & McClelland, 1990). Starting with the Parallel Distributed Processing framework published in the 80s (Rumelhart et al., 1986) and after going through several waves of variation in success, pure connectionist models have achieved some impressive results in recent years, especially in image and pattern recognition (Krizhevsky, Suthskever, & Hinton, 2017; LeCun, Bengio, & Hinton, 2015). Connectionist models have been developed to address performance of some specific aspects of intelligence tasks such as the Raven Progressive Matrices task (Barrett et al., 2018) partially addressing the long-standing criticism of its inability to represent symbolic structures and to model tasks requiring symbolic operations (Fodor & Pylyshyn, 1988). Connectionist models like Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have properties that allow them to encode and operate on the spatial and temporal relationships between stimuli (Battaglia et al., 2018). Canonical multilayered feed-forward neural networks are universal function approximators that can represent any arbitrary relations (Hornik et al., 1989). However, one of the major challenges that remains with this type of models is that the relationships are hard coded within these architectures as fixed connections with weight values that change slowly and are retained long term, diverging from the expressive power of the brain to rapidly represent and flexibly apply arbitrary relations to novel contexts with limited training (Diamond, 2013; Hasselmo, 2018; Miyake et al., 2000; Zelazo, 2015). Graph Neural Network (Battaglia et al., 2018), an approach that combines both symbolic representation and neural network, might be a promising remedy, but it is unclear how one can modify graph structures during the course of computation to adapt to the ever-evolving relations encountered in the real world.

2.2. Symbolic-Connectionist representations

Attempting to bridge the gap between connectionist and classical AI models, a view that reconciles the two suggests that networks of neurons implement a symbolic processor at a higher and more abstract level of description (Marcus, 2001), pertaining to Marr's levels of analysis (Marr, 1982). Marr's multi-level view ironically results in two competing camps whose difference reflects the fundamental conflict between the Connectionists and the Classicists. One side preferred a bottom-up approach focused on neural implementations, building up neural networks that can eventually manipulate symbols, while the other side favors a top-down approach, starting with algorithms and cognitive functions and using brain and behaviors as constraints.

One idea from classical AI that needs to be addressed by the bottom-up approaches to symbolic processing involves role-filler binding. This idea suggests that abstract ideas or concepts called roles can be occupied by or bound to fillers, arbitrary individuals, or instances. For example, in the sentence, Alice ordered a coffee from Bob, Alice and Bob are fillers, and the underlying roles that these fillers are bound to can be customer and barista. Role-filler can also be referred to as type-token or class-instance. A single role bound to different fillers can enable rapid generalization across new tasks and situations. If Bob and Alice occupied the same role of a barista, they should serve similar functions if placed in the same coffeehouse setting, despite being different individuals. Multiple role-filler bindings can also give rise to a structured and compositional representation that is characteristic of cognition (Fodor, 1975). Bob the barista can be bounded to another role, either in parallel such as Bob the graduate student, or hierarchically, such as Bob the barista, an employee at Starbucks. Thus, from the bottom-up approach perspective, role-filler binding presents an important test of the ability to link network of activity of neurons to more abstract symbolic representations. Unsurprisingly, there are disagreements within this approach as to how neurons might participate in the process.

The most prominent idea in this framework is conjunctive coding. The idea proposes that roles and fillers are represented by separate vectors of activity, and the binding is represented by a weight matrix that is a conjunction of the role and filler vectors (Fig. 1a). In other words, the encoding process of an arbitrary structure occurs through some forms of multiplication between its constituents. Multiplying the conjunctive weight matrix with a constituent should return the other component, hence simulating a retrieval process. Though present in the early days of connectionism (Rumelhart et al., 1986), conjunctive coding gained additional traction with work by Paul Smolensky (Smolensky, 1990) introducing tensor product as a potential binding mechanism. However, though this framework continues (Smolensky et al., 2016) and argues that the representations appear in recurrent neural networks (McCoy et al., 2019), this approach has not been broadly implemented within the field.

As an alternative, Tony Plate (Plate, 1991) proposed circular convolution as a mean to bind between role and filler to alleviate the exponentially increasing size of tensors required by tensor product representation. More recently, building on Plate's idea, Chris Eliasmith and colleagues created semantic pointer architecture using these circular convolutions to link role and filler and built a large-scale model (Eliasmith et al., 2012) capable of performing a variety of cognitive tasks.

One major criticism of conjunctive coding is that it fails to preserve role-filler independence. Role-filler independence, also known as type-token distinction, refers to the difference between two processes of recognizing a symbol as a certain type, and individuating that symbol as

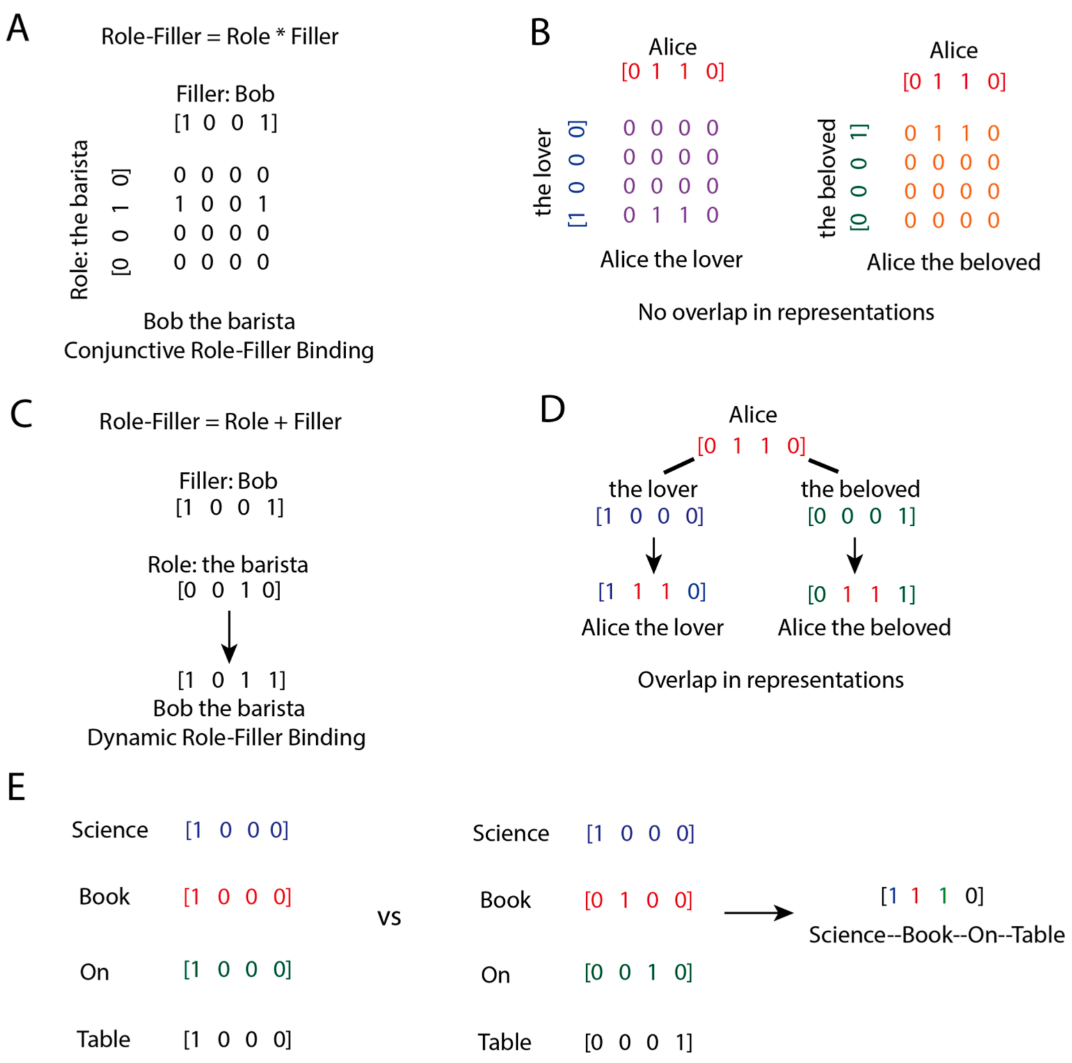


Fig. 1. Vector Representations of Role-Filler Interactions. (a) Conjunctive Coding returns a multiplicative interaction between a filler and a role. (b) The returned role-filler representation from conjunctive coding for the same filler with different roles are distinct. (c) Dynamic Binding via temporal synchrony involves summing and maintaining the individual role and filler. (d) The returned representations for the same filler bound to different roles are similar in dynamic binding, facilitating generalization. (e) Temporal synchrony does not perverse causality. Temporal asynchrony alleviates this issue by keeping track of the order of spike timing.

a particular token of that type. A viewer being asked to count the number of pears or pick out a red apple in a large bowl of fruits cannot rely solely on type information. He or she must be able to quickly individuate token of the same type (Kanwisher, 1987). This requirement has been helpful in explaining repetition blindness and illusory conjunctions (Chun, 1997; Kanwisher, 1991; Kanwisher, 1987). In repetition blindless, when a stimulus is repeated, it is sometimes identified only by type but not tokenized, and therefore is not identified correctly as another occurrence. In illusory conjunctions, subjects erroneously combine features of two objects into one object, essentially linking one token to two types. Conjunctive coding returns a stable representation where role and filler are always mixed (Fig. 1a) and therefore cannot explain the above phenomena. Another common example (Hummel et al., 2004) to illustrate the failure of conjunctive coding to preserve role-filler independence is the sentence Alice loves Bob and its inverse, Bob loves Alice. In both cases, even though the two sentences have very different meanings, the example of Alice should intuitively be independent from her role as a lover or a beloved. One can easily think of cases when Alice the lover is the same as Alice the beloved. In fact, it is precisely this capacity that allows us to rapidly generalize across novel contexts. Conjunctive coding, however, would always create two separate representations or units for Alice the lover and Alice the beloved that are dissimilar to each other (Fig. 1b). If Alice the lover discovered that Bob is having an affair, conjunctive coding would describe Alice the beloved as having no clue what is going on.

2.3. Role-filler interactions - dynamic binding with conjunctive coding

A potential solution to this issue is perhaps to create distinct links arising from a single node representing Alice, connecting her to the distinct roles of a lover and of a beloved. These links would have to be rapidly created or destroyed, since if Bob cheated on Alice, then Bob no longer loves Alice, thereby breaking the link between Alice and her role as the beloved. Alice will probably no longer love Bob at that point, so the remaining link between Alice and her role as the lover should be destroyed too. This solution, termed Dynamic Binding (Fig. 1c,d), will satisfy the role-filler independent requirement and can be implemented in the brain using the spike timing of neurons. The idea first took root when Von der Malsburg proposed temporal correlation in neural spiking activity (von der Malsburg, 1994) as a potential way to form complex representation. Basically, spikes that co-occur can be summed together by synaptic interactions on the dendritic tree. Dynamic Binding by the temporal synchrony of spiking activity (Hummel & Biederman, 1992; Hummel & Holyoak, 1993; Konen & von der Malsburg, 1993; Shastri & Ajjanagadde, 1993) was a natural next step. In this framework, the relationship between role and filler is represented by temporal synchrony of spikes that can be created and destroyed on the fly, and the same units can participate in different bindings either simultaneously or at different times. Role-filler independence is preserved in this framework. However, these models have not been scaled to large cognitive processing systems.

Temporal synchrony runs into issue when its unit must act simultaneously as a role and a filler. An example from Gary Marcus (Marcus, 2001) did a good job illustrating this point. Imagine a book about science sitting on a table. Temporal synchrony tells us how to bind science with book, and book with table. Thing gets a bit fuzzy when all these three units, science, book, and table fire together. Temporal synchrony does not preserve directionality in its binding, so it is not clear which unit is a predicate, and which is an argument. In other words, we can't tell from the example whether the science book is sitting on top of the table or whether the table is sitting on top of the science book (Fig. 1e). There would have to be additional mechanisms that marks what purpose book serves in each binding, since one can imagine how things get complicated quickly when we consider the more complex recursive structure that is found in human language or even bird songs (Gentner et al., 2006).

The potential problems with the previous dynamic binding approach were addressed by the use temporal asynchrony (Love, 1998) as a solution that enables more complex recursive binding. The intuition is simple. If unit A fired before unit B, which fired before C, ordering information or causality is preserved (Fig. 1e). Asynchronous binding can also simulate complex tree structures, the abstract representations of hierarchy and recursion. There does not appear to be further work implementing temporal asynchrony to scale.

Synchrony with spike timing models could also be energetically expensive, especially if used constantly to keep track of the causal interaction between the diverse contents in the environment. It is estimated that less than 1% of neurons can be activated simultaneously (Lennie, 2003). The biological cost of maintaining an active code through persistent firing in spike timing models led some researchers to propose dynamic binding through synaptic traces as an alternative (Mongillo et al., 2008). Formally referred to as synaptic or 'activity-silent' working memory theory (Mongillo et al., 2008; Stokes, 2015; Trübutschek et al., 2019), this framework suggests that information can be stored silently as pattern of synaptic weights. A rapid reconfiguration of the functional connectivity through short term synaptic plasticity (Zucker & Regehr, 2002) would modify the information encoded and allow for dynamic binding (Stokes, 2015). It is worth noting that neural oscillations can also reconfigure functional connectivity by phase aligning different periods of excitability to maximize or suppress the opportunity for information transfer (Fries, 2005). There is also evidence for rule-specific subnetworks formed by oscillatory synchronization of local field potentials (Buschman et al., 2012). It is therefore possible that there is some complementary interaction between synaptic plasticity and phase coding to enable dynamic binding (Lundqvist et al., 2011, 2018; Miller et al., 2018; Stokes, 2015), but more experimental findings are needed to verify and probe this interaction.

Another big problem with spike timing models is their inability to store binding in long term memory (Hummel & Holyoak, 1997) due to the transient nature of spikes. To overcome this capacity issue, researchers have looked for ways to integrate conjunctive coding to take advantage of long-term memory storage by the creation of persistent patterns of weight strengths between units, while maintaining dynamic binding to achieve rapid role-filler binding, role-filler independence, and simultaneous role-filler representation (Hummel et al., 2004). The exciting structural and dynamical constraints found in grid cells (Giocomo & Hasselmo, 2008; Hafting et al., 2005; Yoon et al., 2013) might hint at how the nervous system can implement a joint conjunctive and dynamic binding strategy. Cognitive variables can be stored in conjunctive grid cells (Constantinescu et al., 2016), but in attractor network models, modification of those variables would require slow circuit reconfiguration, and formation of higher dimensional state spaces would suffer from the curse of dimensionality (Klukas et al., 2020). Phase information from the periodic firing fields of grid cells can be rapidly and dynamically combined (Bush & Burgess, 2014) to allow for a combinatorial coding range, but can't efficiently represent high-dimensional space (Fiete et al., 2008). A combination of the two strategies, termed "mixed modular coding" (Klukas et al, 2020), employs spatial phase coding of conjunctive grid modules and enables flexible on-demand coding as well as efficient memory states for variables in high dimensional vector spaces. Conjunctive grid cells can exhibit temporal phase coding in 2D space (Climer et al., 2013). It is an open question whether grid modules can employ spatial phase coding in higher dimensional cognitive space. Recent data suggests that phase coding might be behaviorally important because manipulations of neuronal oscillations regulated by the medial septum cause impairments of behavior in tasks such as the radial arm maze (Chrobak et al., 1989) and spatial alternation (Zutshi et al., 2018), but driving medial septum at different frequencies does not prevent firing of place cells (Zutshi et al., 2018) or grid cells (Lepperød et al., 2021).

2.4. *Role-filler interactions - coding by neural sequences*

Research in episodic memory can offer additional clues as to how the brain might utilize spike timing and conjunctive coding to represent and store abstract variables. The main advantage of all these spike timing models over conjunctive coding can essentially be boiled down to its ability to preserve similarity between the encoded representations. This is a rephrasing of role-filler independence violation discussed previously, where the representation for Alice the lover radically differs from that of Alice the beloved. Our capacity for relational generalization and transitive inference would prefer a unified representation of fillers (Hummel, 2011; Piaget, 1928). In the hippocampus, there is evidence supporting the presence of overlapping representations (Brown et al., 2010; Eichenbaum et al., 1999; Hasselmo, 2012; Howard et al., 2005; Kraus et al., 2013, 2015; Shohamy & Wagner, 2008; Wood et al., 2000; Zeithamova & Preston, 2010) that capture and preserve similarity, and are arguably critical for relational memory (Eichenbaum et al., 1999). These overlapping 'sequences' of neural activity could bridge representations of distinct events, linking related episodes in a memory space (Eichenbaum et al., 1999). Sequential firing cells, or 'neural sequences' have been attributed to coding for place, time and other cognitive variables (Aronov et al., 2017; Dombeck et al., 2010; Hasselmo, 2012; Howard et al., 2014; Kinsky et al., 2020; Koay et al., 2021; Kraus et al., 2013, 2015). Neural sequences provide a common syntactic mechanism that the brain could employ for cognition and is highly reminiscent of the aforementioned temporal asynchrony model (Love, 1998), as well as the abstract syntax utilized in classical AI.

If neural sequences provide syntactic structure, is there any mechanism that can provide symbols with the necessary semantics, or meanings that were lacking in classical AI? Disambiguation of overlapping experience is in fact critical to episodic memory, and modelling work has shown that contexts can provide the necessary guidance to distinguish between similar events (Hasselmo & Eichenbaum, 2005; Hasselmo & Stern, 2018; Howard et al., 2005; Katz et al., 2007). Experimental findings provide ample evidence of cells that could be coding for temporal and spatial contexts in the hippocampal formation (Bright et al., 2020; Dudchenko & Wood, 2014; Frank et al., 2000; Kinsky et al., 2020; Komorowski et al., 2009; Solstad et al., 2008; Tsao et al., 2018; Wood et al., 2000). It is likely that these 'context' cells are tightly coupled with cells coding for stimulus identity, perhaps through some form of associative outer product mediated by synaptic hebbian plasticity (Bliss & Collingridge, 1993; Hasselmo & Eichenbaum, 2005; Lisman et al., 2002; Tiganj et al., 2018), forming an item-place-time conjunction (Cruzado et al., 2020; Hasselmo, 2012; Hasselmo et al., 2010; Komorowski et al., 2009), providing the necessary information about what, where and when underlying abstract syntax. Neural sequences alone should be sufficient to code for spatial and temporal contexts, and to participate in conjunctive coding with cells representing stimulus features. However, the latest theoretical framework (Howard & Hasselmo, 2020) emphasizes the existence of a distinct population closely related to neural sequences but with different receptive fields that code for time and space in a more efficient manner, facillitating more complex computations in abstract cognitive space. There is some evidence for the existence of such populations (Bright et al., 2020; Kraus et al., 2013; Mau et al., 2018; Tsao et al., 2018), but it is unclear whether these cells are causally related to neural sequences, as well as whether they participate in conjunctive coding with stimulus cells.

Whatever the case, there is evidence to suggest that item-place-time conjunctions, once formed, can be stored in long term memory, therefore solving the capacity issue of spike timing models. Perhaps the most famous example in memory research, the selective sensitivity of recent episodic memories but not remote memories in patient H.M. (Scoville & Milner, 2000) led researchers to propose that during offline periods like sleep, hippocampal episodic memories become stored in long-term neocortical semantic memory over time through system consolidation (Hasselmo et al., 1996; McClelland et al., 1995). This process is thought to involve a loss of time and place, contextual information, and a transition to more fact-based semantic representation. There are debates as to whether this theory is valid (Nadel & Moscovitch, 1997; Yassa & Reagh, 2013; Yonelinas et al., 2019), but experimental findings of sequential reactivation of neurons encoding previous experience, or replay in hippocampal and cortical regions (Davidson et al., 2009; Euston et al., 2007; Foster & Wilson, 2006; Karlsson & Frank, 2009), as well as neocortical-hippocampal coupling during sleep (Logothetis et al., 2012; Siapas & Wilson, 1998; Sirota et al., 2003), do support the idea that item-place-time binding started in the hippocampus and propagated to the cortex, with additional processing there to extract the regularities in item information from the detailed contextual information, perhaps through matrix factorization (Bengio et al., 2012; Higgins et al., 2018; Koren et al., 2009; Liu et al., 2019; Morin et al., 2021; Zhu et al., 2020). This is a simplistic view, however, since information can also flow in the opposite direction during sleep, with neocortex activity preceding that of hippocampus (Hahn et al., 2006; Ji & Wilson, 2007; Karimi Abadchi et al., 2020; Liu et al., 2021; Sirota et al., 2003), hinting at a more complex bidirectional process of consolidation. More comprehensive modeling and experimental designs are needed to untangle this intricate process.

2.5. *Flexible planning for behavioral tasks - reinforcement learning*

Since the discussion thus far deals mainly with how symbols are represented in a bottom-up fashion, the next step in describing mechanisms of general intelligence is to discover the algorithms that select and organize operations acting on said representations, i.e, planning or searching. These issues were the major focus of researchers who follow the symbolic-connectionists perspective but favor the more abstract modeling of behavior rather than the neural mechanisms, which we will refer to as the top-down approach. Then there is also the question of identifying and mapping elements from the external world to internal representations that are suitable for relational reasoning, i.e detection and recognition. Both are difficult questions. This paper will only focus on the question of planning algorithms, since this is where the top-down approach really shines, but see (Hinton, 2021) for an interesting discussion on the mapping of external world to internal representations.

On the topic of planning, perhaps the most famous top-down approach is reinforcement learning (Sutton & Barto, 2018). Progress in optimal control theory (Eveleigh, 1967) and evidence from trial-and-error learning in animals (Pavlov, 1927) helped guide and constraint a model of behavior in which the learning agent aims to learn a policy that would maximize reward obtained while interacting with states in an environment to achieve a goal. Essentially, the bulk of reinforcement learning (RL) methods focus on estimating a value function that determines the total amount of reward one can expect to accumulate over time. The best course of actions is one that maximizes this value function. The main problem with classical RL is scalability, since the agent needs to encounter and store a large number of states and actions combinations in the environment to learn the best course of actions, or a policy, but it is not feasible to explore all possible combinations of states and actions. Deep RL promises a potential solution by introducing a function approximation that outputs values given a continuous range of actions and states, and one can represent and train this function with data using deep neural networks and gradient descent. This is beneficial for generalization but also problematic because the known sample inefficiency problem of deep learning will create an RL agent that requires massive amounts of training data; hence this is the reason most applications of Deep RL are in controlled simulated environments like video games where data generation is not an issue (Mnih et al., 2015). Another problem is transferability. Agents encountering a novel, yet similar environment will have to learn an entirely new policy and can’t rely on previous experience (Bhandari & Badre, 2018; Gamrian & Goldberg, 2018; Kansky et al., 2017). Recent approaches to life-long RL promises to minimize this issue of catastrophic forgetting (McCloskey & Cohen, 1989) by explicitly retaining learned knowledge, leveraging shared structure and learning to adapt and learn (Khetarpal et al., 2020).

The paper will focus mainly on leveraging shared structure since it is the most relevant to the discussion on planning with symbolic representations. RL agents navigating the world can find repeatable structure through state and action abstraction, and thereby reuse solutions from previous problems to novel situations. This involves reducing the number of states observed by the agents and aggregating primitive actions into higher level action in frameworks like Options (Bacon et al., 2016; Frank & Badre, 2012; Sutton et al., 1999), Feudal RL (Dayan & Hinton, 1992; Vezhnevets et al., 2017), Hierarchical Abstract Machines (Parr & Russell, 1997) and MAXQ (Dietterich, 2000). The general idea is that state abstraction and action or temporal abstraction allow for better knowledge representations that significantly reduce search space and are easily transferable across environments once learned. For example, instead of telling an experienced chef to choose the bread, stack the meat and cheese followed by some condiments, we can just ask this person to make us a burger. The chef through experience has already abstracted away the trivial steps in between. This is action or temporal abstraction. To illustrate state abstraction, imagine working from home when it is sunny versus working from home when it is raining out. The weather indicates different states of the environment but working from home long enough and each day would eventually feel the same. We have clustered all these different states together, abstracted away the differences, and every day of working from home feels the same as any other day.

Though promising, these frameworks however have not been widely adopted since hierarchy cannot yet be defined automatically, and sampling inefficiency is still a big issue for RL in general. In fact, on simple tasks, it has been shown that a simple random search can outperform reinforcement learning algorithms (Mania et al., 2018). However, progress in RL is rapidly guiding and informing neurophysiological findings in memory, navigation (Banino et al., 2018; Stachenfeld et al., 2017), and decision making (Wang et al., 2018). Early neural implementations of RL sought to find mechanisms for selection of specific pathways through the environment. In recent work, grid cells can be thought of as the eigenvectors of the graph Laplacian or equivalently the successor representation in 2D navigation (Dayan, 1993; Stachenfeld et al., 2017), and is therefore useful for clustering connected components in a graph. This is a well-known algorithm in machine learning called spectral clustering (Ng et al., 2001). Assuming an environment can be represented as a graph, these grid-like eigenvectors can help discover a hierarchical decomposition of the environment, finding connected states, essentially performing state abstraction in a hierarchical RL framework. These theory-guided findings suggest that we will soon gain a better understanding of what neural representations underly states and actions in the RL framework, and perhaps use that knowledge to better inform algorithm design.

2.6. Flexible planning for behavioral tasks – Bayesian inference

Tackling the issue of sample inefficiency, a success story might be that of Bayesian Program Learning (Lake et al., 2015). This is a top-down approach that can capture one-shot learning, the ability to learn concepts after a single example. This approach follows a long line of researchers who attempted to explain human learning and inductive reasoning in terms of Bayesian Inference (Griffiths et al., 2010). Specifically, in this framework, when given an inductive problem, one specifies the hypotheses under consideration, the relation between these hypotheses and observable data, as well as the prior probability of each hypothesis. To rephrase in term of learning, a learner considers a set of hypotheses H that might explain observed data D and assigns a probability p(H) before even observing the data. This is a prior probability that depends on previously acquired knowledge. Then according to Bayes' rule, the chosen hypothesis after observing the data will be determined by how well the hypotheses cohere with prior knowledge, as well as how well they explain the data (the likelihood or the probability of observing data D if hypothesis H is true). One major strength of this framework is that it is representation-agnostic. Hypotheses can take any form as long as they specify a probability distribution over the observed data. Furthermore, inductive biases can be controlled by changing the prior probability, allowing one to model how previously learned knowledge can aid or interfere with new learning. In fact, this ability to arbitrarily modify inductive biases is what is missing in most Connectionist models (Battaglia et al., 2018). However, this incredible flexibility of the Bayesian framework might also be its biggest weakness in modelling brain and behaviors, since the models can become immune to falsification (Bowers & Davis, 2012). It is further highly unlikely that neurons employ an optimal method for inference like the classical Bayes Rule since most biological features are evolved to be good enough rather than optimal, and non-Bayesian approaches sometimes provide a better account of human performance than Bayesian inference (Bowers & Davis, 2012). A framework that aims to connect naturalistic behaviors to neural implementations would therefore need to capture these biological constraints.

Major theoretical attempts to describe how neurons can represent probability distribution and perform probabilistic computation are the Neural Sampling Hypothesis (Fiser et al., 2010; Hoyer & Hyvärinen, 2002; Kutschireiter et al., 2017) and Probabilistic Population Code (PPC) (Beck et al., 2008; Ma et al., 2006). Essentially, these frameworks model neural activity as either representing samples from a probability distribution or coding for the natural parameters of said distribution. According to the Neural Sampling Hypothesis, each neuron can represent a particular feature of external stimuli, with higher spike counts correspond to higher confidence that a feature is present, and the high variability in spike counts can correspond to high uncertainty. This framework was inspired by a powerful machine learning algorithm called Markov chain Monte Carlo sampling (Brooks et al., 2011), which is particularly useful for computing large hierarchical models and representing high-dimensional probability distributions. On the other hand, in the PPC framework, each neuron has its own tuning curve with some parameters that encode some distribution, and the combined population activity would code for the probability distribution over stimuli. In the case of Poisson-like variability of single cell recording observed in cortex (Ma et al., 2006), Bayesian inference is simple since the log-probability distribution over stimuli is a linear combination of tuning curves. The question of which framework is a better candidate to describe how neural circuits represent probability is currently a topic of substantial debate and controversy that goes beyond the scope of this paper.

Going back to Bayesian program learning (Lake et al., 2015), to enable one-shot learning, one can impose a strong inductive bias, or a strong prior to restrict the number of possible solutions or hypotheses induced. A learner taking advantage of this strong prior only requires a small number of examples to converge to a solution. Combining this simple principle with the compositionality found in Hierarchical Bayesian modeling (Gelman & Hill, 2007), one can learn complex concepts from basic primitives or building blocks given a single training sample. One classic example to illustrate the power of hierarchical Bayesian modeling is the eight-school problem (Gelman & Hill, 2007; Rubin, 1981) which considers the effectiveness of SAT coaching programs conducted at eight parallel schools. A naïve observer viewing the SAT score might calculate the average score from each school to compare the difference or look at the standard error (uncertainty) of the score to judge whether all coaching programs are similar. In the Bayesian language, this is essentially forming a hypothesis by looking only at the immediate data, for instance the average score, presumably with some bias on whether the observer thinks the effectiveness of different coaching programs should be the same or different. A meticulous researcher, however, would reason that there are other factors at-play here, since even though each school has different teachers teaching different students, the SAT curriculum is standardized across schools, so there must be some mixed effects on the score. Thus, this researcher would form another hypothesis that the observed average score for each school comes from another distribution whose parameters depend on teachers' variability and the curriculum's similarity. What is powerful about this chained hypothesis-forming framework is that the researcher can now reliably predict how effective an SAT coaching program is for new incoming students given only some prior information about those eight schools. Applying this hierarchical framework to Bayesian problem learning, researchers have simulated inventing a new handwritten letter from basic strokes, or designing a new mode of transportation from simple building blocks (Lake et al., 2015). Furthermore, Hierarchical Bayesian program learning can be used to construct complex role-filler interactions from basic primitives (Fig. 2, adapted from Lake et al., 2015), capturing the productivity, compositionality and systematicity requirements of symbolic processing (Lake et al., 2017). However, from the SAT example, it is easy to recognize that the major challenges with Bayesian program learning, and essentially Hierarchical Bayesian Modelling, are hierarchy design and choosing a good hierarchical prior.

The idea that complex concepts are probabilistic programs compositionally built from simpler primitives can go a long way, as shown with the recent progress made by the DreamCoder model (Ellis et al., 2020), which tackles both aforementioned challenges to hierarchical Bayesian modelling. This system can learn to solve problems by writing computer programs given tasks in many different domains like text editing, graphics generation, symbolic regression, or even physics. Taking inspiration from the field of program induction (Solomonoff, 1964), DreamCoder treats learning a new task as searching for, or synthesizing a new program that solves it. Viewed in term of a probabilistic inference problem in the Bayesian framework, DreamCoder observes a training set of task X, and infers both a program p for solving each task x in X as well as a prior distribution over programs, encoded as a library L. Basically, it is a system that can generate hypotheses and update its prior. This is made possible due to the incorporation of separate encoding and consolidation phases of the Wake-Sleep cycle (Hinton et al., 1995), in which samples during encoding are replayed during consolidation to explore the problem space (Fig. 3). This allows training of a neural network to link specific examples from the problem space to bias the selection of specific program elements from the library. The wake-sleep algorithm was inspired by the changes in functional connectivity regulated by cholinergic modulation between waking and sleep (Hasselmo & Bower, 1993; Hasselmo, 1999). The inferred program is a hierarchy automatically formed by chaining together basic building blocks. This is related to the idea of planning discussed in the aforementioned reinforcement learning framework, and program synthesis can be implemented in an reinforcement learning fashion (Simmons-Edler et al., 2018). During the sleep phase, programs are syntactically compressed to speed up search. This is reminiscent of the action abstraction in Hierarchical Reinforcement Learning. The elemental 'actions' or basic

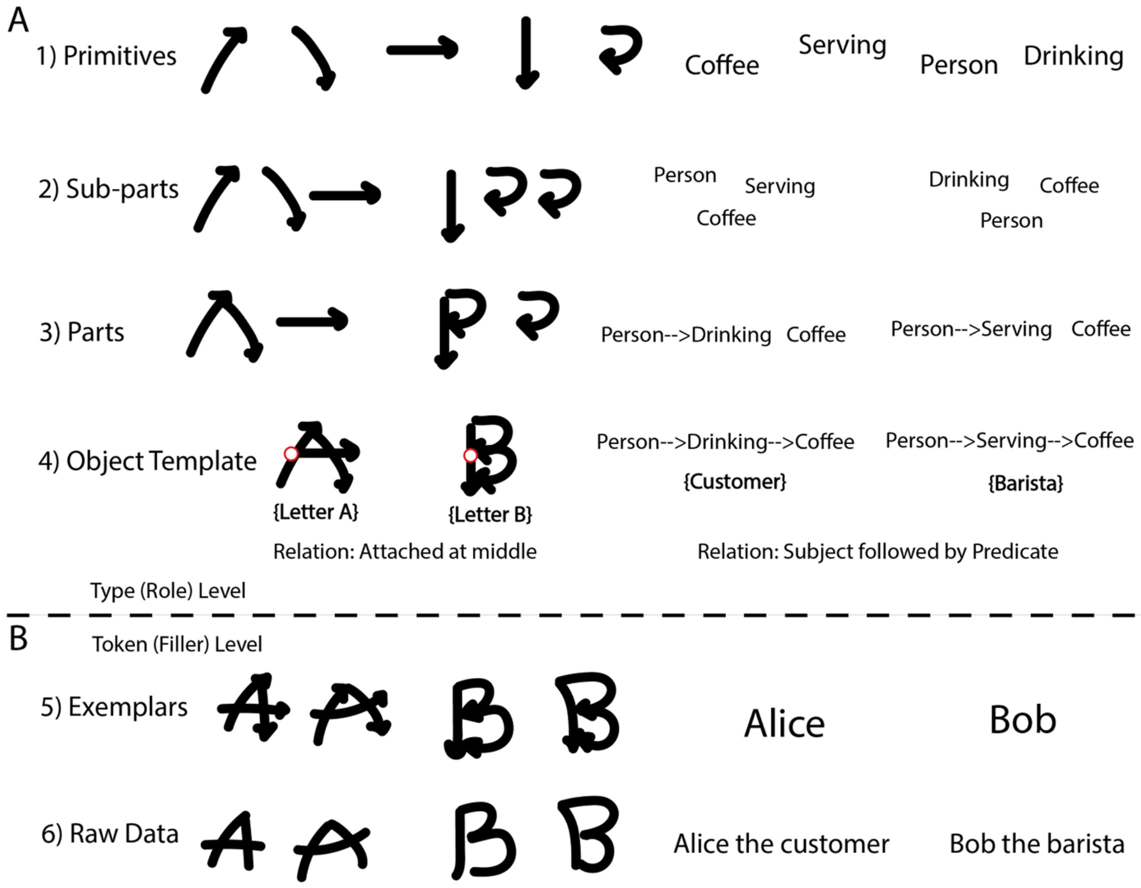


Fig. 2. Building Concepts with Bayesian Program Learning. (a) In Bayesian Program Learning, Type or Role level involves constructing template from basic primitives, sub-parts and parts, forming complex relations in the process. (b) Token or Filler level involves adding variation to the different features of the preexisting object template, for instance applying an affine transform to each letter, or changing the name of a person. The results are symbols or concepts.

building blocks to represent programs in DreamCoder are lambda calculus expressions (Fig. 4), which are Turing-complete and can therefore perform any computation a programmable computer can. Fig. 4 provides an overview of the use of lambda calculus to provide elements for constructing programs. If there is a neural analogy to lambda calculus, it could provide the elements for building neural programs. This framework is indeed well in line with the symbolic-connectionists' perspective (Piantadosi, 2021). This raises the important question of whether neural circuits could implement primitive operations similar to lambda calculus or combinatory logic (Piantadosi, 2021) and could allow a process of program synthesis in neural circuits.

2.7. *Flexible planning for behavioral tasks - operations*

This paper has so far discussed how variables and complex data structures can be represented in neural circuits, as well as how operations acting on said representations can be combined in a biologically plausible way. The discussion on lambda calculus operations provides a segue to the next section, since the next piece of the puzzle is on what elemental operations the brain utilizes to build programs. The field has not converged on a standard representation of operations within multitudes of different binding and non-binding strategies that have been proposed. On the pure connectionist front, the latest theoretical work (Domingos, 2020) suggested that the deep learning models trained on gradient descent are approximately equivalent to kernel machines, and their main operation is to superimpose training data for storage in the kernel spaces, enabling efficient matching with future query. This is reminiscent of the operations performed by Willshaw's associative network model (Willshaw et al., 1969). Biological neural network models on the other hand have been used to model various forms of attractor dynamics (Ben-Yishai et al., 1995; Brody et al., 2003; Chaudhuri & Fiete, 2016; Hopfield, 1982; Redish et al., 1996; Seung, 1996; Wang, 2001), supported by evidence from a growing number of large-scale neurophysiological recording and manipulation studies (Bassett et al., 2018; Inagaki et al., 2019; Knierim & Zhang, 2012; Yoon et al., 2013). Essentially, these networks are dynamical systems that over time settle to a stable pattern termed 'attractor'. That pattern might be stationary, cyclic, or chaotic. The networks' state at stability could then be described as residing on some low-dimensional manifold (point, line, circle, plane, toroid, etc.), which enables various robust and reliable information processing capabilities like noise reduction (Pouget et al., 1998), categorization (Wong et al., 2007), integration (Seung, 1996), or memorization (Hopfield, 1982). Interestingly, these observations are consistent with the Manifold Hypothesis (Fefferman et al., 2013) in machine learning, which states that the embeddings of high-dimensional real-world data tend to lie in the vicinity of a low dimensional manifold. The challenge for the connectionists then is to establish the neural operations that can manipulate manifolds by controlling and constructing attractors, perhaps by introducing translation to move the network's state to another location inside or outside of an existing attractor, or by transforming or changing the kind of attractor the network is implementing on-the-fly (Eliasmith, 2005). Such operations are likely to exist since they are theoretically realizable through short-term synaptic potentiation (Igarashi et al., 2012; Itskov et al., 2011; Katori et al., 2011; Mongillo et al., 2008; Seeholzer et al., 2019; Torres et al., 2007).

For the symbolic-connectionists, the focus is on the algebraic operations acting on high dimensional vectors (Kanerva, 2009). For example, these operations can be dot product, component-wise addition,

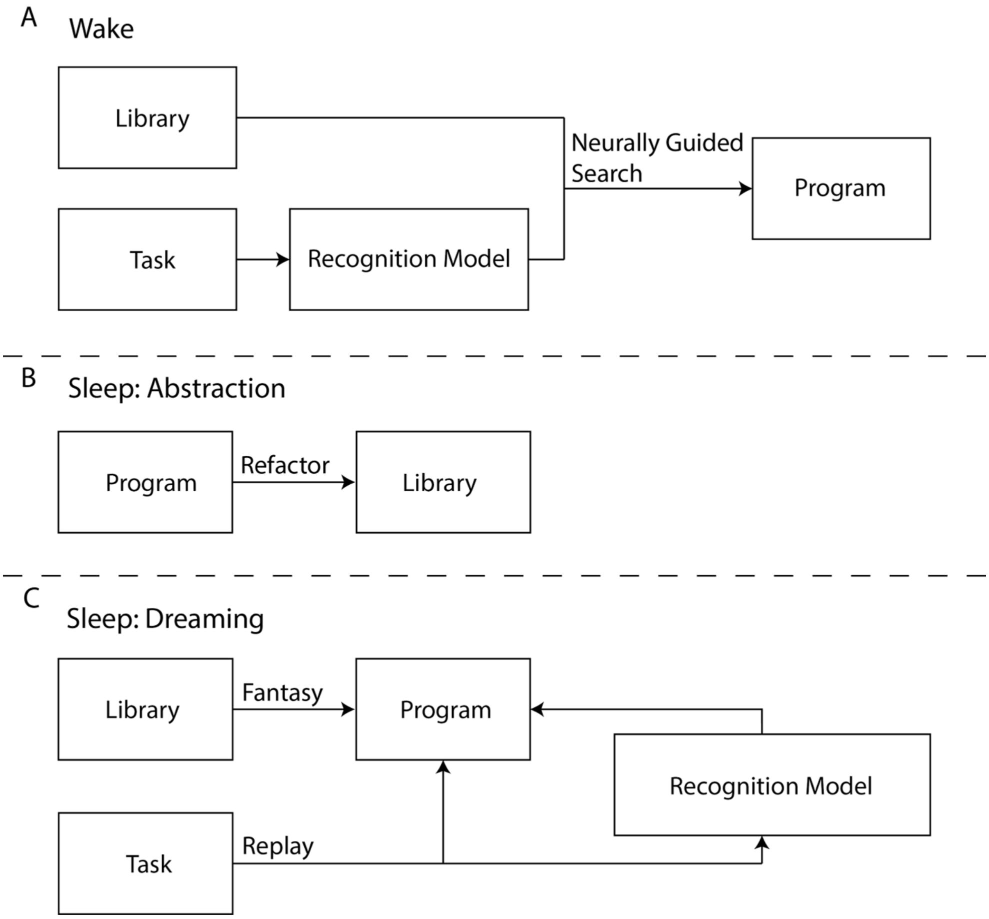


Fig. 3. DreamCoder Diagram. (a) In the wake stage, a recognition model (i.e., convolution neural network) takes the task as input and proposed a set of programs from the existing library that might be able to provide the solution. (b) In the sleep abstraction stage, proposed programs are compressed and refactored to reusable modules in the library. (c) In the sleep dreaming stage, task encountered during the wake stage as well as random combinations of these tasks are replayed, and the recognition model are then trained to map these tasks to the programs found during waking, as well as novel programs generated from the library.

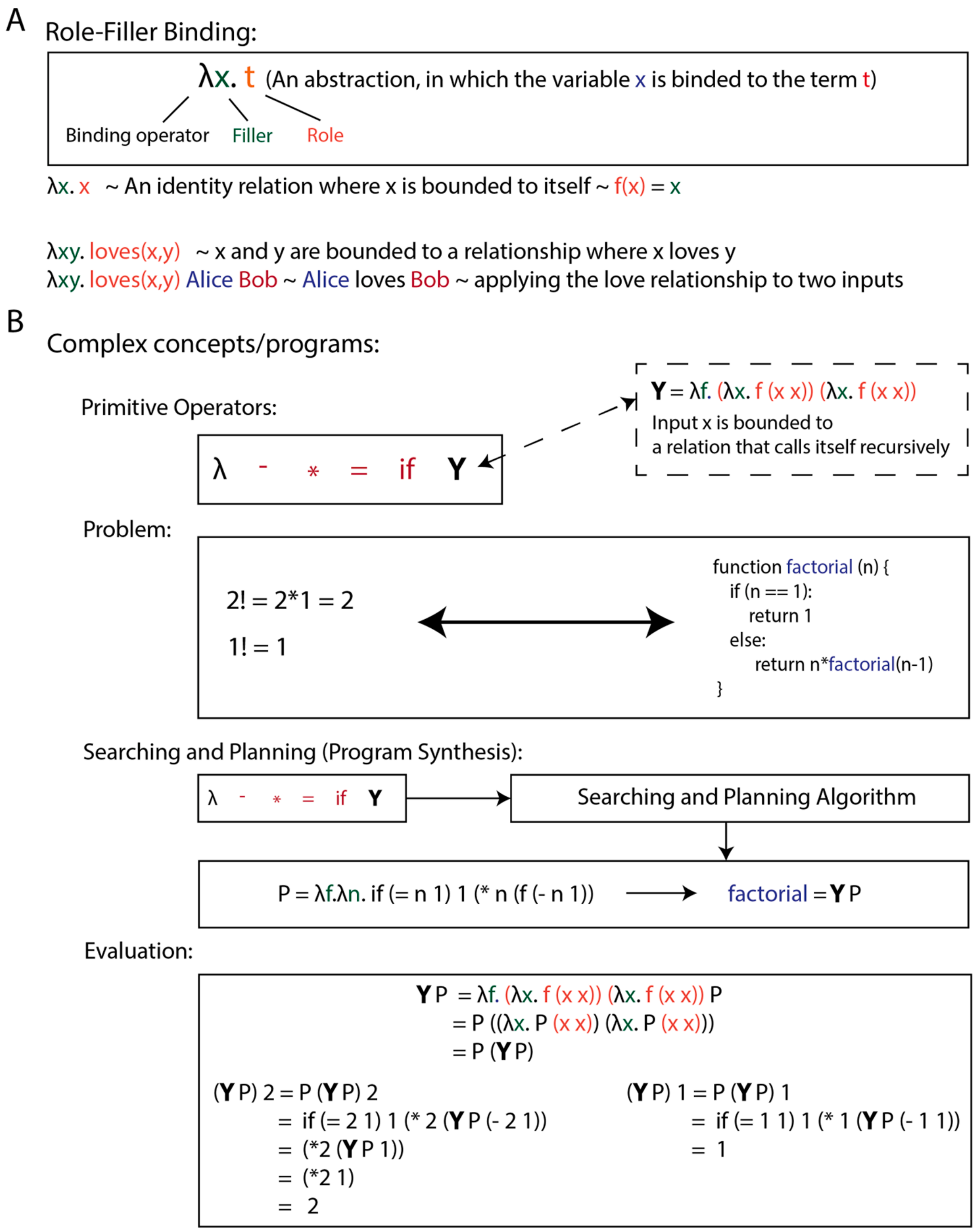
component-wise multiplication, or permutation. The closest analogy is the built-in Arithmetic Logic Unit (ALU) found in a Von-Neumann computer, and it is reasonable to believe that the brain has a diverse set of innate operations that are utilized. Indeed, several researchers (Dyer & Dickinson, 1996; Gallistel, 1998) have argued for an innate arithmetic calculator in honeybees and desert ants that allow them to calculate the sun's position to stay on course during foraging. The latest theory on pre-existing neural operations is a framework called Assembly Calculus (Papadimitriou et al., 2020), which proposed a set of operations that could result from the activity of neurons and synapses, and diverge from the vector algebraic operations like add, multiply and permute. Some examples of operations in the Assembly Calculus framework are projection, associate, merge, disinhibit, inhibit, and fire.

Symbolic-connectionist operations integrate quite nicely into the program synthesis framework, where problem-solving is defined as the process of searching for and utilizing different combinations of operations (Fig. 4b). Program synthesis can use the lambda calculus as the basic building blocks for initial construction of programs (Ellis et al., 2020). The question of whether neural circuits can implement lambda calculus operations can be addressed with Category Theory (Eilenberg & MacLane, 1945; Leinster, 2014; Spivak, 2014), a bird's-eye view theory about the common patterns, trends, structures between different mathematical realms like set theory, group theory, topology, linear algebra, etc. Each mathematical area has objects in it that can be related to each other in some ways, termed morphisms (set theory have sets that relate via functions, linear algebra has vector spaces that relate via linear transformation, topology have topological spaces that relate via continuous functions). A category is defined as a collection of objects that relate to each other via morphisms, such that the morphisms (functions, linear transformations, etc.) can be composed associatively, and there exists an identity morphism for each object. Different categories like Set (Set Theory) or Top (Topology) or Vect (Linear Algebra) can be related or mapped through functors. Simply typed lambda calculus is equivalent to a category called Cartesian closed category (CCC). Thus, finding a neural implementation of lambda calculus can be thought of as finding a mapping (functor) between cartesian closed category and the category that neural operations belong to. This is in fact the same process that compiler can use to map Haskell (a lambda calculus-based programming language) to digital hardware (Elliott, 2017). Hypothetically, if neural operations utilized high dimensional vector algebras, then they would belong to the FinVect category (the category of finite dimensional vector spaces). The question then becomes, what is the functor that maps CCC to FinVect, and indeed there are several existing attempts that try to derive a vector interpretation of lambda calculus using this framework (Elliott, 2017; Valiron & Zdancewic, 2014).

3. Discussion

It is particularly challenging to study how the brain can manipulate symbols, since this question is spread out across so many fields, with no agreement upon a common theory or framework. This review paper outlines the current debate and present some opposing viewpoints, discusses the progress made and emphasizes the remaining challenges to coming up with a falsifiable unifying theory detailing the circuit mechanisms of symbolic processing.

Juggling evidence and perspectives across fields, it seems that a good representation for symbols that satisfy the productivity, compositionality, systematicity requirements (Fodor & Pylyshyn, 1988) is the conjunctive binding between neural sequences and neural populations coding for stimulus information. Sequences provide a natural way to order information, hence providing the syntax for thoughts. Stimulus


A
Role-Filler Binding:
λx. t (An abstraction, in which the variable x is binded to the term t)
Binding operator
Filler
Role
λx. x ~ An identity relation where x is bounded to itself ~ f(x) = x
λxy. loves(x,y) ~ x and y are bounded to a relationship where x loves y
λxy. loves(x,y) Alice Bob ~ Alice loves Bob ~ applying the love relationship to two inputs
B
Complex concepts/programs:
Primitive Operators:
λ - * = if Y
Y = λf. (λx. f (x x)) (λx. f (x x))
Input x is bounded to
a relation that calls itself recursively
Problem:
2! = 2*1 = 2
1! = 1
Searching and Planning (Program Synthesis):
λ - * = if Y
Searching and Planning Algorithm
P = λf.λn. if (= n 1) 1 (* n (f (- n 1))
factorial = Y P
Evaluation:
Y P = λf. (λx. f (x x)) (λx. f (x x)) P
= P ((λx. P (x x)) (λx. P (x x))
= P (Y P)
(Y P) 2 = P (Y P) 2
= if (= 2 1) 1 (* 2 (Y P (- 2 1))
= (*2 (Y P 1))
= (*2 1)
= 2
(Y P) 1 = P (Y P) 1
= if (= 1 1) 1 (* 1 (Y P (- 1 1))
= 1

```
function factorial (n) {
  if (n == 1):
     return 1
  else:
      return n*factorial(n-1)
 }
```

Fig. 4. Lambda Calculus. (a) The most basic operation in lambda calculus is role-filler binding (also known as abstraction), specified by the λ operator. Multiple fillers can be bound to different relations using relatively simple notations. Placing different inputs to the right of the role-filler terms (also called lambda terms) signifies an application of the relations to the different fillers. (b) All operators in lambda calculus can be further reduced to the different combinations of the λ operator, for instance the Y-combinator. Operators can be combined to form complex concepts. In this example, the mathematical concept of a factorial can be broken down into a simple program that calls itself recursively until the input integer as well as all the positive integers smaller in magnitude are multiplied and accounted for. From the primitive operators, using some searching or planning algorithm, a program is proposed and evaluated that would eventually satisfy the requirement of the computation of the mathematical concept of factorial.

information provide meanings or semantics that are bound to sequences and can be retrieved effortlessly. The efficient and flexible arrangement and organization of these symbols could then depend on the fascinating nature of grid patterns. It is likely that these grid populations facilitate efficient planning by clustering and compressing similar and frequently used syntactic structures, while phase coding enables rapid combination and manipulation of those structures.

Reinforcement learning is a good biologically plausible framework to guide the searching and planning process, but trial and error learning has to be coupled with the constructing and updating of prior and hypothesis, essentially forming an internal model as described in the Bayesian inference literature (Craik, 1943) to alleviate the sample inefficiency issue. In fact, Bayesian inference and the aforementioned optimal control theory in RL frameworks are mathematically equivalent, and goal-directed planning can be thought of as doing Bayesian inference to estimate a sequence of steps given the observation of a successful task completion (McNamee & Wolpert, 2019).

It is an important question how priors and hypotheses are represented in neural circuits. Neural sequence can have tuning curves that parametrically code a probability distribution as Probabilistic Population Code (Ma et al., 2006) has suggested, and its spike count can also sample said distribution in accordance with the Neural Sampling Hypothesis (Hoyer & Hyvärinen, 2002). Furthermore, the item-place-time conjunctive representation can be thought of as the joint probability of observing an item in a particular place at a particular time (Tiganj et al., 2018). These ideas could be useful for building a neural framework of program synthesis, since a program is represented as a hierarchy of priors and hypotheses in DreamCoder (Ellis et al., 2020), which can therefore flexibly address a broad range of tasks, building up to the capacity for general intelligence.

Mapping program synthesis to neural circuits could then be a promising approach to developing neural implementations of flexible symbolic processing but would also require development of neural operations analogous to the lambda calculus, since DreamCoder relies on such operations to perform computation. Category Theory is a powerful tool that can relate neural operations to existing mathematical frameworks, enabling us to utilize the rich theoretical foundations developed across multiple mathematical realms to build our intuition and strengthen our understanding of brain computations. It is an open question what Category neural operations fall into.

If the manifold hypothesis is correct, neural representations would lie on some low dimensional manifold. Short-term plasticity would likely play an essentially role in manipulating manifold by constructing and controlling attractors. It is therefore beneficial to study what neural operations are theoretically realizable given this biologically constraint and determine whether one can map those operations to lambda calculus using Category Theory. Such finding would bridge the gap between connectionism and classical AI and let us derive the circuit mechanism that would enable the brain to manipulate symbols.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Institutes of Health, grant numbers R01 MH120073, R01 MH60013, R01 MH052090 and by the Office of Naval Research MURI N00014-16-1-2832 and MURI N00014-19-1-2571 and DURIP N00014-17-1-2304. The authors have no conflict of interest.

References

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal/entorhinal circuit. *Nature, 543*(7647), 719–722. https://doi.org/10.1038/nature21692

Bacon, P.-L., Harb, J., & Precup, D. (2016). The Option-Critic Architecture. *ArXiv:1609.05140 [Cs]*. http://arxiv.org/abs/1609.05140.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., … Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature, 557*(7705), 429–433. https://doi.org/10.1038/s41586-018-0102-6

Barrett, D. G. T., Hill, F., Santoro, A., Morcos, A. S., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. *ArXiv:1807.04225 [Cs, Stat]*. http://arxiv.org/abs/1807.04225.

Bassett, J. P., Wills, T. J., & Cacucci, F. (2018). Self-Organized Attractor Dynamics in the Developing Head Direction Circuit. *Current Biology, 28*(4), 609–615.e3. https://doi.org/10.1016/j.cub.2018.01.010

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *ArXiv:1806.01261 [Cs, Stat]*. http://arxiv.org/abs/1806.01261.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron, 60*(6), 1142–1152. https://doi.org/10.1016/j.neuron.2008.09.021

Bengio, Y., Mesnil, G., Dauphin, Y., & Rifai, S. (2012). Better Mixing via Deep Representations. *ArXiv:1207.4404 [Cs]*. http://arxiv.org/abs/1207.4404.

Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences, 92*(9), 3844–3848. https://doi.org/10.1073/pnas.92.9.3844

Bhandari, A., & Badre, D. (2018). Learning and transfer of working memory gating policies. *Cognition, 172*, 89–100. https://doi.org/10.1016/j.cognition.2017.12.001

Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature, 361*(6407), 31–39. https://doi.org/10.1038/361031a0

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389–414. https://doi.org/10.1037/a0026450

Bright, I. M., Meister, M. L. R., Cruzado, N. A., Tiganj, Z., Buffalo, E. A., & Howard, M. W. (2020). A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proceedings of the National Academy of Sciences, 117*(33), 20274–20283. https://doi.org/10.1073/pnas.1917197117

Brody, C. D., Romo, R., & Kepecs, A. (2003). Basic mechanisms for graded persistent activity: Discrete attractors, continuous attractors, and dynamic representations. *Current Opinion in Neurobiology, 13*(2), 204–211. https://doi.org/10.1016/s0959-4388(03)00050-3

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Brown, T. I., Ross, R. S., Keller, J. B., Hasselmo, M. E., & Stern, C. E. (2010). Which way was I going? Contextual retrieval supports the disambiguation of well learned overlapping navigational routes. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(21), 7414–7422. https://doi.org/10.1523/JNEUROSCI.6021-09.2010

Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. *Neuron, 76*(4), 838–846. https://doi.org/10.1016/j.neuron.2012.09.029

Bush, D., & Burgess, N. (2014). A Hybrid Oscillatory Interference/Continuous Attractor Network Model of Grid Cell Firing. *Journal of Neuroscience, 34*(14), 5065–5079. https://doi.org/10.1523/JNEUROSCI.4017-13.2014

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431.

Chaudhuri, R., & Fiete, I. (2016). Computational principles of memory. *Nature Neuroscience, 19*(3), 394–403. https://doi.org/10.1038/nn.4237

Chollet, F. (2019). On the Measure of Intelligence. *ArXiv:1911.01547 [Cs]*. http://arxiv.org/abs/1911.01547.

Chrobak, J. J., Stackman, R. W., & Walsh, T. J. (1989). Intraseptal administration of muscimol produces dose-dependent memory impairments in the rat. *Behavioral and Neural Biology, 52*(3), 357–369. https://doi.org/10.1016/S0163-1047(89)90472-X

Chun, M. M. (1997). Types and tokens in visual processing: A double dissociation between the attentional blink and repetition blindness. *Journal of Experimental Psychology. Human Perception and Performance, 23*(3), 738–755. https://doi.org/10.1037//0096-1523.23.3.738

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain* (pp. xi, 544). The MIT Press.

Climer, J. R., Newman, E. L., & Hasselmo, M. E. (2013). Phase coding by grid cells in unconstrained environments: Two-dimensional phase precession. *The European Journal of Neuroscience, 38*(4), 2526–2541. https://doi.org/10.1111/ejn.2013.38.issue-410.1111/ejn.12256

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science (New York, N.Y.), 352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Craik, K. J. W. (1943). *The nature of explanation* (pp. viii, 123). University Press, Macmillan.

Cruzado, N. A., Tiganj, Z., Brincat, S. L., Miller, E. K., & Howard, M. W. (2020). Conjunctive representation of what and when in monkey hippocampus and lateral prefrontal cortex during an associative memory task. *Hippocampus, 30*(12), 1332–1346. https://doi.org/10.1002/hipo.v30.1210.1002/hipo.23282

Cummins, R. C. (1991). The Role of Representation in Connectionist Explanation of Cognitive Capacities. In W. Ramsey, S. P. Stich, & D. Rumelhart (Eds.), *Philosophy and Connectionist Theory* (pp. 91–114). Lawrence Erlbaum.

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal replay of extended experience. *Neuron, 63*(4), 497–507. https://doi.org/10.1016/j.neuron.2009.07.027

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation, 5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Dayan, P., & Hinton, G. E. (1992). Feudal Reinforcement Learning. *Advances in Neural Information Processing Systems, 5*. https://proceedings.neurips.cc/paper/1992/hash/d14220ee66aeec73c49038385428ec4c-Abstract.html.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dietterich, T. G. (2000). Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition. *Journal of Artificial Intelligence Research, 13*, 227–303.

Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L., & Tank, D. W. (2010). Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nature Neuroscience, 13*(11), 1433–1440. https://doi.org/10.1038/nn.2648

Domingos, P. (2020). Every Model Learned by Gradient Descent Is Approximately a Kernel Machine. *ArXiv:2012.00152 [Cs, Stat]*. http://arxiv.org/abs/2012.00152.

Dudchenko, P. A., & Wood, E. R. (2014). Splitter Cells: Hippocampal Place Cells Whose Firing Is Modulated by Where the Animal Is Going or Where It Has Been. In D. Derdikman & J. J. Knierim (Eds.), *Space, Time and Memory in the Hippocampal Formation* (pp. 253–272). Springer. https://doi.org/10.1007/978-3-7091-1292-2_10.

Dyer, F. C., & Dickinson, J. A. (1996). Sun-compass learning in insects: Representation in a simple mind. *Current Directions in Psychological Science, 5*(3), 67–72. https://doi.org/10.1111/1467-8721.ep10772759

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., & Tanila, H. (1999). The hippocampus, memory, and place cells: Is it spatial memory or a memory space? *Neuron, 23*(2), 209–226. https://doi.org/10.1016/s0896-6273(00)80773-4

Eilenberg, S., & MacLane, S. (1945). General Theory of Natural Equivalences. *Transactions of the American Mathematical Society, 58*(2), 231–294. https://doi.org/10.2307/1990284

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural Computation, 17*(6), 1276–1314. https://doi.org/10.1162/0899766053630332

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science, 338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266

Elliott, C. (2017). Compiling to categories. *Proceedings of the ACM on Programming Languages, 1*(ICFP), 1–27. https://doi.org/10.1145/3110271

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (2020). DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning. *ArXiv:2006.08381 [Cs]*. http://arxiv.org/abs/2006.08381.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Euston, D. R., Tatsuno, M., & McNaughton, B. L. (2007). Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science, 318*(5853), 1147–1150. https://doi.org/10.1126/science.1148979

Eveleigh, V. W. (1967). *Adaptive control and optimization techniques*. McGraw-Hill.

Fefferman, C., Mitter, S., & Narayanan, H. (2013). Testing the Manifold Hypothesis. *ArXiv:1310.0425 [Math, Stat]*. http://arxiv.org/abs/1310.0425.

Fiete, I. R., Burak, Y., & Brookings, T. (2008). What grid cells convey about rat location. *Journal of Neuroscience, 28*(27), 6858–6871. https://doi.org/10.1523/JNEUROSCI.5684-07.2008

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences, 14*(3), 119–130. https://doi.org/10.1016/j.tics.2010.01.003

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440*(7084), 680–683. https://doi.org/10.1038/nature04587

Frank, L. M., Brown, E. N., & Wilson, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron, 27*(1), 169–178. https://doi.org/10.1016/s0896-6273(00)00018-0

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex (New York, N.Y.: 1991), 22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences, 9*(10), 474–480. https://doi.org/10.1016/j.tics.2005.08.011

Gallistel, C. R. (1998). Insect navigation: Brains as symbol- processing organs. *Conceptual and Methodological Foundations, 4*, 61.

Gamrian, S., & Goldberg, Y. (2018). *Transfer Learning for Related Reinforcement Learning Tasks via Image-to-Image Translation*. https://openreview.net/forum?id=rkxjnjA5KQ.

Gelman, A., & Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://nyuscholars.nyu.edu/en/publications/data-analysis-using-regression-and-multilevelhierarchical-models.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive Science, 7*(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature, 440*(7088), 1204–1207. https://doi.org/10.1038/nature04675

Giocomo, L. M., & Hasselmo, M. E. (2008). Computation by oscillations: Implications of experimental data for theoretical models of grid cells. *Hippocampus, 18*(12), 1186–1199. https://doi.org/10.1002/hipo.20501

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology, 23*(2), 222–262. https://doi.org/10.1016/0010-0285(91)90010-L

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature, 436*(7052), 801–806. https://doi.org/10.1038/nature03721

Hahn, T. T. G., Sakmann, B., & Mehta, M. R. (2006). Phase-locking of hippocampal interneurons' membrane potential to neocortical up-down states. *Nature Neuroscience, 9*(11), 1359–1361. https://doi.org/10.1038/nn1788

Hasselmo, M. E. (1999). Neuromodulation: Acetylcholine and memory consolidation. *Trends in Cognitive Sciences, 3*(9), 351–359. https://doi.org/10.1016/s1364-6613(99)01365-0

Hasselmo, M. E. (2012). *How We Remember: Brain Mechanisms of Episodic Memory*. MIT Press.

Hasselmo, M. E. (2018). A model of cortical cognitive function using hierarchical interactions of gating matrices in internal agents coding relational representations. *ArXiv:1809.08203 [q-Bio]*. http://arxiv.org/abs/1809.08203.

Hasselmo, M. E., & Bower, J. M. (1993). Acetylcholine and memory. *Trends in Neurosciences, 16*(6), 218–222. https://doi.org/10.1016/0166-2236(93)90159-j

Hasselmo, M. E., & Eichenbaum, H. (2005). Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural Networks: The Official Journal of the International Neural Network Society, 18*(9), 1172–1190. https://doi.org/10.1016/j.neunet.2005.08.007

Hasselmo, M. E., Giocomo, L. M., Brandon, M. P., & Yoshida, M. (2010). Cellular dynamical mechanisms for encoding the time and place of events along spatiotemporal trajectories in episodic memory. *Behavioural Brain Research, 215*(2), 261–274. https://doi.org/10.1016/j.bbr.2009.12.010

Hasselmo, M. E., & Stern, C. E. (2018). A network model of behavioural performance in a rule learning task. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1744), 20170275. https://doi.org/10.1098/rstb.2017.0275

Hasselmo, M. E., Wyble, B. P., & Wallenstein, G. V. (1996). Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus, 6*(6), 693–708. https://doi.org/10.1002/(SICI)1098-1063(1996)6:6<693::AID-HIPO12>3.0.CO;2-W

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A. (2018). Towards a Definition of Disentangled Representations. *ArXiv:1812.02230 [Cs, Stat]*. http://arxiv.org/abs/1812.02230.

Hinton, G. (2021). How to represent part-whole hierarchies in a neural network. *ArXiv:2102.12627 [Cs]*. http://arxiv.org/abs/2102.12627.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268*(5214), 1158–1161. https://doi.org/10.1126/science.7761831

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of Inference, Learning, and Discovery*. MIT Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought* (pp. xiii, 320). The MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America, 79*(8), 2554–2558.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review, 112*(1), 75–116. https://doi.org/10.1037/0033-295X.112.1.75

Howard, M. W., & Hasselmo, M. E. (2020). Cognitive computation using neural representations of time and space in the Laplace domain. *ArXiv:2003.11668 [q-Bio]*. http://arxiv.org/abs/2003.11668.

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., & Eichenbaum, H. (2014). A Unified Mathematical Framework for Coding Time, Space, and Sequences in the Hippocampal Region. *The Journal of Neuroscience, 34*(13), 4692–4707. https://doi.org/10.1523/JNEUROSCI.5808-12.2014

Hoyer, P., & Hyvärinen, A. (2002). Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior. *Advances in Neural Information Processing Systems, 15*. https://proceedings.neurips.cc/paper/2002/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in human shape perception. In *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 157–185). Lawrence Erlbaum Associates Publishers.

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science, 23*(2), 109–118. https://doi.org/10.1080/09540091.2011.569880

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480–517. https://doi.org/10.1037/0033-295X.99.3.480

Hummel, J. E., & Holyoak, K. J. (1993). Distributing structure over time. *Behavioral and Brain Sciences, 16*(3), 464. https://doi.org/10.1017/S0140525X00031083

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*(3), 427–466. https://doi.org/10.1037/0033-295X.104.3.427

Hummel, J. E., Holyoak, K. J., Green, C., Doumas, L. A. A., Devnich, D., Kittur, A., & Kalar, D. J. (2004, December 1). *A solution to the binding problem for compositional connectionism*. 2004 AAAI Fall Symposium. https://experts.illinois.edu/en/publications/a-solution-to-the-binding-problem-for-compositional-connectionism.

Igarashi, Y., Oizumi, M., & Okada, M. (2012). Theory of correlation in a network with synaptic depression. *Physical Review E, 85*(1), Article 016108. https://doi.org/10.1103/PhysRevE.85.016108

Inagaki, H. K., Fontolan, L., Romani, S., & Svoboda, K. (2019). Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature, 566*(7743), 212–217. https://doi.org/10.1038/s41586-019-0919-7

Itskov, V., Hansel, D., & Tsodyks, M. (2011). Short-Term Facilitation may Stabilize Parametric Working Memory Trace. *Frontiers in Computational Neuroscience, 5*. https://doi.org/10.3389/fncom.2011.00040

Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience, 10*(1), 100–107. https://doi.org/10.1038/nn1825

John, & Raven, J. (2003). Raven Progressive Matrices. In R. S. McCallum (Ed.), *Handbook of Nonverbal Assessment* (pp. 223–237). Springer US. 10.1007/978-1-4615-0153-4_11.

Kanerva, P. (2009). Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors. *Cognitive Computation, 1*(2), 139–159. https://doi.org/10.1007/s12559-009-9009-8

Kansky, K., Silver, T., Mély, D. A., Eldawy, M., Lázaro-Gredilla, M., Lou, X., Dorfman, N., Sidor, S., Phoenix, S., & George, D. (2017). Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics. *ArXiv:1706.04317 [Cs]*. http://arxiv.org/abs/1706.04317.

Kanwisher, N. (1991). Repetition blindness and illusory conjunctions: Errors in binding visual types with visual tokens. *Journal of Experimental Psychology. Human Perception and Performance, 17*(2), 404–421. https://doi.org/10.1037//0096-1523.17.2.404

Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition, 27*(2), 117–143. https://doi.org/10.1016/0010-0277(87)90016-3

Karimi Abadchi, J., Nazari-Ahangarkolaee, M., Gattas, S., Bermudez-Contreras, E., Luczak, A., McNaughton, B. L., & Mohajerani, M. H. (2020). Spatiotemporal patterns of neocortical activity around hippocampal sharp-wave ripples. *ELife, 9*, Article e51972. https://doi.org/10.7554/eLife.51972

Karlsson, M. P., & Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nature Neuroscience, 12*(7), 913–918. https://doi.org/10.1038/nn.2344

Katori, Y., Sakamoto, K., Saito, N., Tanji, J., Mushiake, H., Aihara, K., & Sporns, O. (2011). Representational Switching by Dynamical Reorganization of Attractor Structure in a Network Model of the Prefrontal Cortex. *PLOS Computational Biology, 7*(11), e1002266. https://doi.org/10.1371/journal.pcbi.1002266

Katz, Y., Kath, W. L., Spruston, N., Hasselmo, M. E., & Friston, K. J. (2007). Coincidence detection of place and temporal context in a network model of spiking hippocampal neurons. *PLoS Computational Biology, 3*(12), e234. https://doi.org/10.1371/journal.pcbi.0030234

Khetarpal, K., Riemer, M., Rish, I., & Precup, D. (2020). Towards Continual Reinforcement Learning: A Review and Perspectives. *ArXiv:2012.13490 [Cs]*. http://arxiv.org/abs/2012.13490.

Kinsky, N. R., Mau, W., Sullivan, D. W., Levy, S. J., Ruesch, E. A., & Hasselmo, M. E. (2020). Trajectory-modulated hippocampal neurons persist throughout memory-guided navigation. *Nature Communications, 11*(1), 2443. https://doi.org/10.1038/s41467-020-16226-4

Klukas, M., Lewis, M., Fiete, I., & Bush, D. (2020). Efficient and flexible representation of higher-dimensional cognitive variables with grid cells. *PLOS Computational Biology, 16*(4), e1007796. https://doi.org/10.1371/journal.pcbi.1007796

Knierim, J. J., & Zhang, K. (2012). Attractor dynamics of spatially correlated neural activity in the limbic system. *Annual Review of Neuroscience, 35*(1), 267–285. https://doi.org/10.1146/annurev-neuro-062111-150351

Koay, S. A., Charles, A. S., Thiberge, S. Y., Brody, C. D., & Tank, D. W. (2021). Sequential and efficient neural-population coding of complex task information. *BioRxiv, 801654*. https://doi.org/10.1101/801654

Kolev, V., Georgiev, B., & Penkov, S. (2020). Neural Abstract Reasoner. *ArXiv:2011.09860 [Cs]*. http://arxiv.org/abs/2011.09860.

Komorowski, R. W., Manns, J. R., & Eichenbaum, H. (2009). Robust conjunctive item-place coding by hippocampal neurons parallels learning what happens where. *Journal of Neuroscience, 29*(31), 9918–9929. https://doi.org/10.1523/JNEUROSCI.1378-09.2009

Konen, W., & von der Malsburg, C. (1993). Learning to generalize from single examples in the dynamic link architecture. *Neural Computation, 5*(5), 719–735. https://doi.org/10.1162/neco.1993.5.5.719

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*(8), 30–37. https://doi.org/10.1109/MC.2009.263

Kraus, B. J., Brandon, M. P., Robinson, R. J., Connerney, M. A., Hasselmo, M. E., & Eichenbaum, H. (2015). During running in place, grid cells integrate elapsed time and distance run. *Neuron, 88*(3), 578–589. https://doi.org/10.1016/j.neuron.2015.09.031

Kraus, B. J., Robinson, R. J., White, J. A., Eichenbaum, H., & Hasselmo, M. E. (2013). Hippocampal "time cells": Time versus path integration. *Neuron, 78*(6), 1090–1101. https://doi.org/10.1016/j.neuron.2013.04.015

Krizhevsky, A., Suthskever, I., & Hinton, G. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6). https://doi.org/10.1145/3065386 (in press).

Kutschireiter, A., Surace, S. C., Sprekeler, H., & Pfister, J.-P. (2017). Nonlinear Bayesian filtering and learning: A neuronal dynamics for perception. *Scientific Reports, 7*(1), 8722. https://doi.org/10.1038/s41598-017-06519-y

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*. https://doi.org/10.1017/S0140525X16001837

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553). https://doi.org/10.1038/nature14539 (in press).

Leinster, T. (2014). *Basic category theory*. Cambridge University Press, 10.1017/CBO9781107360068.

Lennie, P. (2003). The cost of cortical computation. *Current Biology: CB, 13*(6), 493–497. https://doi.org/10.1016/s0960-9822(03)00135-0

Lepperød, M. E., Christensen, A. C., Lensjø, K. K., Buccino, A. P., Yu, J., Fyhn, M., & Hafting, T. (2021). Optogenetic pacing of medial septum parvalbumin-positive cells disrupts temporal but not spatial firing in grid cells. *Science Advances, 7*(19), eabd5684. https://doi.org/10.1126/sciadv.abd5684

Lisman, J., Schulman, H., & Cline, H. (2002). The molecular basis of CaMKII function in synaptic and behavioural memory. *Nature Reviews. Neuroscience, 3*(3), 175–190. https://doi.org/10.1038/nrn753

Liu, X., Ren, C., Lu, Y., Liu, Y., Kim, J.-H., Leutgeb, S., Komiyama, T., & Kuzum, D. (2021). Multimodal neural recordings with Neuro-FITM uncover diverse patterns of cortical–hippocampal interactions. *Nature Neuroscience*, 1–11. https://doi.org/10.1038/s41593-021-00841-5

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human replay spontaneously reorganizes experience. *Cell, 178*(3), 640–652.e14. https://doi.org/10.1016/j.cell.2019.06.012

Logothetis, N. K., Eschenko, O., Murayama, Y., Augath, M., Steudel, T., Evrard, H. C., Besserve, M., & Oeltermann, A. (2012). Hippocampal–cortical interaction during periods of subcortical silence. *Nature, 491*(7425), 547–553. https://doi.org/10.1038/nature11618

Love, B. C. (1998). Utilizing time: Asynchronous binding. In *Proceedings of the 11th International Conference on Neural Information Processing Systems* (pp. 38–44).

Lundqvist, M., Herman, P., & Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *Journal of Cognitive Neuroscience, 23*(10), 3008–3020. https://doi.org/10.1162/jocn_a_00029

Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., & Miller, E. K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature Communications, 9*(1), 394. https://doi.org/10.1038/s41467-017-02791-8

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Mania, H., Guy, A., & Recht, B. (2018). Simple random search provides a competitive approach to reinforcement learning. *ArXiv:1803.07055 [Cs, Math, Stat]*. http://arxiv.org/abs/1803.07055.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *ArXiv:1801.00631 [Cs, Stat]*. http://arxiv.org/abs/1801.00631.

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science* (pp. xiii, 224). The MIT Press.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. In *Vision*. The MIT Press. https://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262514620.001.0001/upso-9780262514620.

Mau, W., Sullivan, D. W., Kinsky, N. R., Hasselmo, M. E., Howard, M. W., & Eichenbaum, H. (2018). The Same Hippocampal CA1 Population Simultaneously Codes Temporal Information over Multiple Timescales. *Current Biology: CB, 28*(10), 1499–1508.e4. https://doi.org/10.1016/j.cub.2018.03.051

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McCloskey, M., & Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 24, pp. 109–165). Academic Press. 10.1016/S0079-7421(08)60536-8.

McCoy, R. T., Linzen, T., Dunbar, E., & Smolensky, P. (2019). RNNs Implicitly Implement Tensor Product Representations. *ArXiv:1812.08718 [Cs]*. http://arxiv.org/abs/1812.08718.

McNamee, D., & Wolpert, D. M. (2019). Internal Models in Biological Control. *Annual Review of Control, Robotics, and Autonomous Systems, 2*(1), 339–364. https://doi.org/10.1146/annurev-control-060117-105206

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron, 100*(2), 463–475. https://doi.org/10.1016/j.neuron.2018.09.023

Minker, J. (2000). Introduction to Logic-Based Artificial Intelligence. In J. Minker (Ed.), *Logic-Based Artificial Intelligence* (pp. 3–33). Springer US. 10.1007/978-1-4615-1567-8_1.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science, 319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Morin, T. M., Chang, A. E., Ma, W., McGuire, J. T., & Stern, C. E. (2021). Dynamic network analysis demonstrates the formation of stable functional networks during rule learning. *Cerebral Cortex (New York, N.Y.: 1991)*, bhab175. https://doi.org/10.1093/cercor/bhab175

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7*(2), 217–227. https://doi.org/10.1016/s0959-4388(97)80010-4

Newell, A. (1980). Physical Symbol Systems*. *Cognitive Science, 4*(2), 135–183. https://doi.org/10.1207/s15516709cog0402_2

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (pp. 849–856).

Palmer, S. E. (1978). Structural aspects of visual similarity. *Memory & Cognition, 6*(2), 91–97. https://doi.org/10.3758/BF03197433

Papadimitriou, C. H., Vempala, S. S., Mitropolsky, D., Collins, M., & Maass, W. (2020). Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences, 117*(25), 14464–14472. https://doi.org/10.1073/pnas.2001893117

Parr, R., & Russell, S. (1997). Reinforcement Learning with Hierarchies of Machines. *Advances in Neural Information Processing Systems*, *10*. https://proceedings.neurips.cc/paper/1997/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* (pp. xv, 430). Oxford Univ. Press.

Piaget, J. (1928). *Judgment and reasoning in the child* (pp. viii, 260). Harcourt, Brace. 10.4324/9780203207260.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines, 31*(1), 1–58. https://doi.org/10.1007/s11023-020-09540-9

Plate, T. (1991). Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1* (pp. 30–35).

Pouget, A., Zhang, K., Deneve, S., & Latham, P. E. (1998). Statistically efficient estimation using population coding. *Neural Computation, 10*(2), 373–401. https://doi.org/10.1162/089976698300017809

Rasmussen, D., & Eliasmith, C. (2011). A Neural Model of Rule Generation in Inductive Reasoning. *Topics in Cognitive Science, 3*(1), 140–153. https://doi.org/10.1111/j.1756-8765.2010.01127.x

Raudies, F., & Hasselmo, M. E. (2017). A model of symbolic processing in Raven's progressive matrices. *Biologically Inspired Cognitive Architectures, 21*, 47–58. https://doi.org/10.1016/j.bica.2017.07.003

Redish, A. D., Elga, A. N., & Touretzky, D. S. (1996). A coupled attractor model of the rodent head direction system. *Network: Computation in Neural Systems, 7*(4), 671–685. https://doi.org/10.1088/0954-898X_7_4_004

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach* (pp. xiv, 425). MIT Press.

Rubin, D. B. (1981). Estimation in Parallel Randomized Experiments. *Journal of Educational Statistics, 6*(4), 377–401. https://doi.org/10.2307/1164617

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations* (pp. 45–76). MIT Press.

Santoro, A., Lampinen, A., Mathewson, K., Lillicrap, T., & Raposo, D. (2021). Symbolic Behaviour in Artificial Intelligence. *ArXiv:2102.03406 [Cs]*. http://arxiv.org/abs/2102.03406.

Scoville, W. B., & Milner, B. (2000). Loss of recent memory after bilateral hippocampal lesions 1957. *The Journal of Neuropsychiatry and Clinical Neurosciences, 12*(1), 103–113. https://doi.org/10.1176/jnp.12.1.103

Seeholzer, A., Deger, M., & Gerstner, W. (2019). Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLOS Computational Biology, 15*(4), e1006928. https://doi.org/10.1371/journal.pcbi.1006928

Seung, H. S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences, 93*(23), 13339–13344. https://doi.org/10.1073/pnas.93.23.13339

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences, 16*(3), 417–451. https://doi.org/10.1017/S0140525X00030910

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron, 60*(2), 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Siapas, A. G., & Wilson, M. A. (1998). Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron, 21*(5), 1123–1128. https://doi.org/10.1016/S0896-6273(00)80629-7

Simmons-Edler, R., Miltner, A., & Seung, S. (2018). *Program Synthesis Through Reinforcement Learning Guided Tree Search*. https://arxiv.org/abs/1806.02932v1.

Sirota, A., Csicsvari, J., Buhl, D., & Buzsáki, G. (2003). Communication between neocortex and hippocampus during sleep in rodents. *Proceedings of the National Academy of Sciences, 100*(4), 2065–2069. https://doi.org/10.1073/pnas.0437938100

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*(1–2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Smolensky, P., Lee, M., He, X., Yih, W., Gao, J., & Deng, L. (2016). Basic Reasoning with Tensor Product Representations. *ArXiv:1601.02745 [Cs]*. http://arxiv.org/abs/1601.02745.

Solomonoff, R. J. (1964). A formal theory of inductive inference Part I. *Information and Control, 7*(1), 1–22. https://doi.org/10.1016/S0019-9958(64)90223-2

Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B., & Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science, 322*(5909), 1865–1868. https://doi.org/10.1126/science.1166466

Spivak, D. I. (2014). *Category Theory for the Sciences*. MIT Press.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*(1-2), 217–257. https://doi.org/10.1016/0004-3702(90)90008-N

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience, 20*(11), 1643–1653. https://doi.org/10.1038/nn.4650

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences, 19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence, 112*(1), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

Tiganj, Z., Gershman, S. J., Sederberg, P. B., & Howard, M. W. (2018). Estimating scale-invariant future in continuous time. *ArXiv:1802.06426 [Cs, q-Bio]*. http://arxiv.org/abs/1802.06426.

Torres, J. J., Cortes, J. M., Marro, J., & Kappen, H. J. (2007). Competition between synaptic depression and facilitation in attractor neural networks. *Neural Computation, 19*(10), 2739–2755. https://doi.org/10.1162/neco.2007.19.10.2739

Trübutschek, D., Marti, S., Ueberschär, H., & Dehaene, S. (2019). Probing the limits of activity-silent non-conscious working memory. *Proceedings of the National Academy of Sciences, 116*(28), 14358–14367. https://doi.org/10.1073/pnas.1820730116

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., & Moser, E. I. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature, 561*(7721), 57–62. https://doi.org/10.1038/s41586-018-0459-6

Valiron, B., & Zdancewic, S. (2014). Modeling simply-typed lambda calculi in the category of finite vector spaces. *Scientific Annals of Computer Science, 24*(2), 325–368. https://doi.org/10.7561/SACS10.7561/SACS.2014.210.7561/SACS.2014.2.325

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., & Kavukcuoglu, K. (2017). FeUdal Networks for Hierarchical Reinforcement Learning. *ArXiv:1703.01161 [Cs]*. http://arxiv.org/abs/1703.01161.

von der Malsburg, C. (1994). The Correlation Theory of Brain Function. In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of Neural Networks: Temporal Aspects of Coding and Information Processing in Biological Systems* (pp. 95–119). Springer. 10.1007/978-1-4612-4320-5_2.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience, 21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Wang, X. J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences, 24*(8), 455–463. https://doi.org/10.1016/S0166-2236(00)01868-3

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature, 222*(5197), 960–962. https://doi.org/10.1038/222960a0

Wong, K.-F., Huk, A. C., Shadlen, M. N., & Wang, X.-J. (2007). Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making. *Frontiers in Computational Neuroscience, 1*. https://doi.org/10.3389/neuro.10.006.2007

Wood, E. R., Dudchenko, P. A., Robitsek, R. J., & Eichenbaum, H. (2000). Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron, 27*(3), 623–633. https://doi.org/10.1016/S0896-6273(00)00071-4

Yassa, M. A., & Reagh, Z. M. (2013). Competitive trace theory: A role for the hippocampus in contextual interference during retrieval. *Frontiers in Behavioral Neuroscience, 7*. https://doi.org/10.3389/fnbeh.2013.00107

Yonelinas, A., Ranganath, C., Ekstrom, A., & Wiltgen, B. (2019). A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews. Neuroscience, 20*(6), 364–375. https://doi.org/10.1038/s41583-019-0150-4

Yoon, K., Buice, M. A., Barry, C., Hayman, R., Burgess, N., & Fiete, I. R. (2013). Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nature Neuroscience, 16*(8), 1077–1084. https://doi.org/10.1038/nn.3450

Zeithamova, D., & Preston, A. R. (2010). Flexible memories: Differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *Journal of Neuroscience, 30*(44), 14676–14684. https://doi.org/10.1523/JNEUROSCI.3250-10.2010

Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. *Developmental Review, 38*, 55–68. https://doi.org/10.1016/j.dr.2015.07.001

Zhu, H., Paschalidis, I. C., Chang, A., Stern, C. E., & Hasselmo, M. E. (2020). A neural circuit model for a contextual association task inspired by recommender systems. *Hippocampus, 30*(4), 384–395. https://doi.org/10.1002/hipo.v30.410.1002/hipo.23194

Zucker, R. S., & Regehr, W. G. (2002). Short-term synaptic plasticity. *Annual Review of Physiology, 64*(1), 355–405. https://doi.org/10.1146/annurev.physiol.64.092501.114547

Zutshi, I., Brandon, M. P., Fu, M. L., Donegan, M. L., Leutgeb, J. K., & Leutgeb, S. (2018). Hippocampal neural circuits respond to optogenetic pacing of theta frequencies by generating accelerated oscillation frequencies. *Current Biology: CB, 28*(8), 1179–1188.e3. https://doi.org/10.1016/j.cub.2018.02.061