Rate-Coding Bundle Memory: A Unified Model of Memory and Control for Symbolic Computation in the Brain

Teun van Gils[1], Rowan Sommers[1], Markus Ostarek[1], and Peter Hagoort[1,2]
[1]Neurobiology of Language Department, Max Planck Institute for Psycholinguistics
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University

We propose a neurobiologically plausible model of cognition that combines the advantages of connectionist and symbolic systems, and that can solve a wide range of cognitive problems. This model, called Rate-Coding Bundle Memory (RCBM), is based on the Symbolic Subsystem Hypothesis, which posits that the brain implements a symbolic subsystem within its fundamentally connectionist nature. RCBM is a hybrid model that uses rate coding to represent symbols in a continuous space, and it uses a bundle memory system to store and retrieve these symbols. The model is capable of solving a wide range of cognitive problems, including one-shot learning, pattern separation, and the binding problem. We argue that RCBM provides a promising framework for understanding the nature of cognition, and that it can be used to develop more sophisticated models of cognition in the future.

Keywords:

Introduction

Few things defy a simple explanation as much as our mental model of the world: for any categorical, any box we create, there's an edge case, a nuance, a contradiction to be found. We may not usually pay much attention to the concepts we use, but when we do, we often find that they are unexpectedly tricky to pin down. Take a concept like "bird": it's easy to think of a robin or a sparrow, but what about a penguin, or an ostrich? What about a chicken, or its ancient ancestor, the T-Rex? Or what about a bat, or a dragonfly? Even though we can categorize these animals as birds or not, the boundaries of the category are fuzzy, and when asked to provide a rigid definition, we might struggle to find one that both conforms to our intuitions and covers all cases (Rosch, 1973; Wittgenstein et al., 2010). Is it being able to fly? Is it having feathers? Is it being warm-blooded? Is it having a beak? Is it being able to lay eggs? The answer is, perhaps, that it's all of these things and none of them. Our categorizations do not exist in a vacuum, but are always situated in a particular context and serve a particular goal, a position known as goal-directedness in embodied cognition (Barsalou, 2003, 2008a). Within this context, we can see that the concept of "bird" is not a single, fixed entity, but a complex, dynamic network of associations, with a grounding in the physical world and how we interact with it, and with meaning that is not fixed, but rather lies on a continuum of related concepts, each with their own fuzzy boundaries (Barsalou, 2008a; Varela et al., 2016). A connectionist system, one comprised of simple, interconnected units that can learn and adapt to their environment, is well-suited to model this kind of graded, context-dependent, and continuous meaning (Doerig et al., 2023; Hinton, 1984; Rogers & Mcclelland, 2014). It provides a way to reason about the world that is robust against noise and small differences, and that allows for associative reasoning and graded (i.e., less rigidly categorical) inference (Hinton, 1984; Rogers & Mcclelland, 2014). As such, one may pose that the brain is continuous, connectionist, and embodied (CCE): a fundamentally connectionist system, with its meaning grounded in the physical world and its reasoning associative and graded. Why, then, would we need anything else to explain cognition?

But there is discreteness, hidden in plain sight: this page or screen that you are looking at, it contains words. And no matter the kind of bird you are describing, the word "bird" remains the same. All of its nuances and contradictions are hidden behind a single, fixed symbol. And so the words that you are reading (and thinking, if your thoughts manifest as inner speech), are at least partially symbolic in nature. The meaning and associations they evoke may be continuous, but the words themselves are not. When you read the word "bird", you need not think of a robin or a sparrow, or any bird in particular, but you just think of the concept of "bird" itself, and despite the meaning of this concept refusing to be pinned down when examined closely, it still by and large *behaves* as if it were a single, fixed entity. Moreover, sentences are not just a collection of words with associations, but a structured and

Correspondence concerning this article should be addressed to Teun van Gils, Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands. E-mail: Teun.vanGils@mpi.nl

organized system of symbols that follow syntactic rules which dictate how meaning is composed (Fodor, 1975; Frankland & Greene, 2019). The sentence "wings allow an animal to fly", though strictly speaking not true, describes a logical relationship between the concepts of "wings" and "flying", not merely an association between the two. It involves the composition of simpler, "atomic" concepts — wings, flying — into a structured, "molecular" representation (Fodor & Pylyshyn, 1988). This structured representation has directionality, which is absent in bi-directional associations, i.e., the sentence does not mean that flying allows an animal to have wings. Moroever, there is quite some support for the idea that this kind of reasoning generalizes beyond language, and that a fundamental feature of human cognition is its "language of thought" (Fodor, 1975; Frankland & Greene, 2019; Quilty-Dunn et al., 2022). It is this kind of relational reasoning that is often difficult to capture in a purely connectionist system (Doumas et al., 2008; B. M. Lake & Baroni, 2018). In fact, a large body of literature has argued that there exist some problems that are very hard to solve with only a CCE system (B. M. Lake & Baroni, 2018; Loula et al., 2018; Marcus, 2001; van der Velde & de Kamps, 2006).

For example, many problems that are hard for a CCE system are easy for a symbolic system: reasoning using predicate-based logic, especially when using language, is a prime example of this. Variable assignment, a key feature of symbolic systems, trivially solves one-shot learning, the ability to learn a new concept after observing only a single example, where CCE systems often require many repetitions to do so (Hadley, 2009; B. Lake et al., 2014; Marcus, 2001). CCE systems also struggle with the problem of two (also known as the problem of interference or as pattern separation), where two distinct concepts with similar overlapping features are confused with each other (Fodor & Pylyshyn, 1988; Jackendoff, 2003; O'Reilly et al., 2014; van der Velde & de Kamps, 2006) — though see Sommers et al. (n.d.) for nuance. In symbolic systems, variable assignment also solves the problem of two, since it handily allows for the distinction between any number of concepts into separate variables, no matter how similar they are. Finally, compositionality is an inherent feature of many symbolic systems, allowing for the flexible generation and comprehension of an unbounded number of complex ideas (Frankland & Greene, 2019; B. M. Lake & Baroni, 2018). In fact, it is not uncommon for connectionist systems to be criticized for their lack of compositional freedom, as they struggle to allow the same degree of combinatorial freedom that symbolic systems do (Fodor & Pylyshyn, 1988; B. M. Lake & Baroni, 2018; van der Velde & de Kamps, 2006). More broadly then, it seems to be the case that symbolic systems have certain computational advantages over traditional connectionist systems — although the difference between a CCE system and a symbolic system is also not as clear-cut as assumed here (see Sommers et al. (n.d.) for a discussion).

All existing models of cognition solve a particular subset of problems — the problems they have targeted — while often struggling with others. Oftentimes, the problems that a model struggles with are partly inherited from which tradition it is from, e.g., symbolism or connectionism. We pose that any satisfactory model of cognition should be able to solve problems that are solved by the superset of both traditions, since a model that can solve a wide range of problems is less likely to be overoptimized for a particular subset of tasks at the expense of others. In order to address this, we propose a set of problems that we call Symbol Recombination, Retention, and Resolution (S3R). This set of problems is not exhaustive, nor is it static: our aim has been to keep collecting problems that are hard to solve by any (but not necessarily all) existing models, so as to ensure that our model is as comprehensive as possible. We divide this set of problems into broadly 5 categories: semantics, compositionality, retention, resolution, and control. Each of these categories detail what we believe to be important properties of cognition in some domain, and the problems that arise when these properties are not present in a model.

The brain stores and processes a vast network of semantic information. One key property of this semantic network is that it is context-dependent: the meaning of a concept is embedded in a semantic graph whose activation patterns are determined by the context in which the concept is used (Barsalou, 2008b). For instance, priming studies reveal that the activation of a concept can be influenced by the activation of related concepts, even when these concepts are not explicitly mentioned (Schacter & Buckner, 1998). Without such context-dependency, the evoked meaning of concepts would always be the same; rather absurdly, one might imagine an avid bird-watcher recalling all of their knowledge about the appearance and behaviour of ducks upon being presented with a plate of roasted duck. Graded inference is another property that is crucial for a semantic network: the ability to reason about the world in a way that is not categorical, but continuous (Rosch, 1973; Taylor, 2011). This starts with the concepts themselves, whose information is stored in a distributed manner across the network, and whose connections are not binary (Rosch, 1973). Priming studies further support this idea, as they show that the activation of a concept can be influenced in a subtle and graded manner (Schacter & Buckner, 1998). Though one may argue a certain level of black-and-white thinking is present in humans, it is evidently not the case that all of our reasoning is categorical, and a model that is capable of exhibiting graded and context-dependent reasoning is more likely to be able to capture the complexity of human cognition.

One large claim in cognition and, in particular, language, is that of compositionality: the idea that complex ideas can be built from simpler ones in a systematic and rule-governed manner (Fodor & Pylyshyn, 1988; Frankland & Greene, 2019). We will distinguish between two types of composition-

ality: flat compositionality, where concepts are combined in a single step, and recursive compositionality, where concepts are combined in a hierarchical manner. Flat compositionality is tightly connected to the binding problem in connectionist literature or (dynamic) variable binding in symbolic literature, and it is an ongoing question in neuroscience and AI (Feldman, 2013; Greff et al., 2020; Hummel, 2011). Feldman (2013) clearly relates the two, stating that dynamic variable binding is required to solve the (neural) binding problem, which he further distinguishes into four (related but) distinct subproblems, including visual feature-binding, and dynamic variable binding. A solution to the binding problem is required to allow arbitrary combinations of concepts and preventing interference between similar items. Without a solution to the binding problem, we would require brute force enumeration of all possible combinations, which is intractable due to the combinatorial explosion of options (Becke et al., 2015; van der Velde & de Kamps, 2015). When considering all possible combinations for N concepts, there are 2^N possible bindings. For example, for only 25 concepts, there are already more than 33 million possible combinations. Given that humans easily have more concepts than that, this approach would quickly exceed the number of neurons in the brain (or, for that matter, the number of atoms in the universe). In particular, this has been a hard problem for connectionist systems, which often struggle with the binding problem and the tractable implementation of compositionality (Fodor & Pylyshyn, 1988; van der Velde & de Kamps, 2006).

However, once we have a tractable mechanism for binding, we can effectively repeat the same process to combine the results of previous combinations, leading to recursive compositionality. This does mean that the selected solution to the binding problem must be able to handle multiple levels of binding. Recursive compositionality is a key feature of many symbolic systems, notably language, and it allows for the "flexible generation and comprehension of an unbounded number of complex ideas" (Chomsky, 2006; Frankland & Greene, 2019). One example of this, often used in language, is role-filler independence: it states that roles (e.g., one-place predicates such as "x walks", "x loves (active)", "x is loved (passive)") and fillers (e.g., "John", "Mary", "the dog") should be computationally kept separate such that they can be combined in a flexible and arbitrary manner (Hummel, 2011; Hummel et al., 2004). Without recursive compositionality, we would be limited to a single level of combination, e.g., being able to express that "John loves Mary", but not that "John loves the dog that Mary loves". Thus, in order to represent the complex capabilities of human cognition, as seen in language and other domains, we believe that a model should be able to account for both flat and recursive compositionality.

The ability to retain information over time is crucial for any cognitive system, as it is a prerequisite for the accumulation of knowledge and the building of complex ideas. Memory itself has been implemented in many ways in cognitive models, and so the ability to retain information itself is not generally a hard problem. However, the way in which information is acquired and retained can make this a difficult problem for some model classes. For example, this means that with their many parameters, connectionist systems often require many repetitions to learn a generalizable solution (Burgess et al., 2017; Fei-Fei et al., 2006; Vinyals et al., 2016). Humans show clear evidence of one-shot learning, also known as fast mapping in the case of word learning (Bion et al., 2013; Carey & Bartlett, 1978; Spiegel & Halberda, 2011) — if we would not, we would never be able to learn someone's name after only a single introduction. Another example is discourse incrementation, where the understanding of a discourse is built up incrementally, requiring both the retention and integration of information over time (Garrod et al., 1995; Lewis et al., 2006; Sanford & Garrod, 2005). If we were unable to perform discourse incrementation, we would not be able to create a mental image of a character described in a novel, or to gradually build up understanding about a complex concept. Working memory is a key component of this process, as it allows for the temporary storage and manipulation of information over short periods of time (Baddeley, 2010). Presumably, if all information were to be stored in memory, we would quickly accumulate so much information that we would no longer be able to see the forest for the trees. In that vein, forgetting is a crucial part of working memory (Lewis et al., 2006): if we were unable to forget information, our working memory would quickly become overloaded, and we would be unable to focus on the most relevant information. As such, to be able to learn quickly and flexibly, and to be able to build up complex ideas over time, we should expect both gradual and one-shot learning, both short- and long-term memory, and both retention and forgetting to be present in any cognitive model.

Of course, retention is only half of the story: the ability to retrieve information is just as important. We know that the brain can retrieve memories using content-addressable memory (CAM), where the activation of a partial memory can lead to the retrieval of the full memory (Lewis et al., 2006; McElree et al., 2003; Parker et al., 2017), effectively completing the stored pattern. This idea is very intuitive from a connectionist lens, as interconnected units can activate each other in a distributed manner. However, its implementation in more symbolic paradigms is less clear. Without the ability to retrieve information in this way, one would need to enumerate all memory traces and compare them to the partial memory, which would quickly become computationally costly as the number of memories grows. However, what happens in problem of two cases, where two similar but distinct concepts can cause interference (Jackendoff, 2003; van der Velde & de Kamps, 2006)? A classic example is the sentence "the little star is to the left of the big star". These are hard problems for

connectionist systems, as they struggle to distinguish between two similar but distinct concepts (van der Velde & de Kamps, 2006). One way in which connectionist systems can solve this problem is through hippocampal pattern separation, as argued by complementary learning system theory (Kumaran et al., 2016; McClelland & Goddard, 1996; McClelland et al., 1995; O'Reilly et al., 2014). By using a combination of expansion recoding, strong thresholding, lateral inhibition, and coincidence detection of memory, the hippocampus can effectively separate similar but distinct patterns (Albus, 1971; Cayco-Gajic & Silver, 2019; Marr, 1970). This effectively transforms a semantically continuous input space into an output space where semantic differences are maximally separated, leading to (more) discrete representations. The prevention of interference between memories also requires serializing the memory traces through cognitive control (Musslick & Cohen, 2021), so that only one memory can be activated at a single time. Besides activating singular memory traces, the brain can also store and retrieve sequences of memories, allowing for the retention of temporal information and the replay of past events composed of multiple time slices (Buhry et al., 2011; Buzsáki & Tingley, 2018; Pavlides & Winson, 1989). Without the presence of these retrieval mechanisms, our models would not be able to match the rich and complex capabilities of human memory.

Somewhat counterintuitively, an important aspect of memory retrieval is the ability to suppress this retrieval, often vital to prevent interference between some memory and other cognitive processes or memories (Herrmann et al., 2001; Levy & Anderson, 2002). This is especially important when forming a new memory that partially overlaps with an existing one but needs to be stored separately (Musslick & Cohen, 2021; van Kesteren et al., 2012). Without a mechanism to suppress retrieval, the new information would simply be merged with the existing memory, leading to a loss of information and a potential confusion between the two items. Suppression can also help solve problems like that of correlation violation, where learnt associations that are stored in long-term memory are violated by new information, but especially connectionist systems are unable to suppress the retrieval of these associations to allow for the learning of new ones (Puebla et al., 2021; van der Velde & de Kamps, 2006). All of these problems relate to the ability to control memory retrieval, and show that simple retrieval mechanisms are not sufficient to capture the complexity of human memory.

However, the ability to control the memory system extends beyond just retrieving memories. Different task contexts require different operations to be performed, and the memory system shoud not just be able to perform each of these operations, but also to switch between them in a controlled and selective manner. For instance, when a friend tells you that the person speaking to the bartender is a friend of theirs, you need to be able to identify the bartender, see who is speaking to them, and then assign to this person the role of being a friend of your friend. This role assignment goes beyond simple retrieval, and instead involves manipulating the information that is being stored. Moreover, it requires clear control: the system needs to prevent assigning the role to the bartender. This ability to manipulate memories distinguishes working memory from short-term memory, as working memory is not just a passive store of information, but an active system that can be used to perform computations on the stored information (Baddeley, 2010). In order to perform manipulations, the memory system should be able to distinguish between novel entities and existing ones (Bion et al., 2013), and in order to solve the problem two, it should also be able to detect when a memory is ambiguous. Detection of these different situations can then be used to selectively suppress parts of the memory system, so that only the relevant information is retrieved or manipulated, and that it is stored in the right place. Another operation that requires control is the retrieval of arbitrary memories, i.e., not through the content but by virtue of holding a particular position in the memory system. This is called addressable read-write memory, and it is a key feature of working memory that allows for the selective manipulation of memories (Atkinson & Shiffrin, 1968; Awh & Vogel, 2025; Gallistel & King, 2009). These operations are agnostic to the content of the memories, and are instead focused on the management of the location of memories and their arbitrary retrieval. Altogether, without control, the complex memory mechanisms outlined in the previous sections would be unable to function properly, making it especially complicated to construct a model that can solve all of these problems — yet vital in order to capture the full complexity of human cognition.

There are existing models that combine the advantages of both connectionist and symbolic systems, and that can solve a wide range of cognitive problems. One such model is the Adaptive Control of Thought-Rational (ACT-R) model, which combines a connectionist memory system with a symbolic control system (J. R. Anderson, 2009). However, these models are often criticized as duct-tape AI, as they are often not fully embedded in a connectionist system, nor neurobiologically plausible (Eliasmith, 2013). Other hybrid models like vector symbolic architectures (VSAs) and the Semantic Pointer Architecture (SPA) (Stewart et al., 2012) use either tensor products (Smolensky, 1990), circular convolution (Plate, 1995) or matrix multiplication (Gallant & Okaywe, 2013) to combine symbolic and connectionist representations. These models provide promising solutions to S3R problems (Gayler, 2004), but their binding operations are mathematical abstractions without clear biological implementations.

Although influenced by these approaches, we propose a more principled hybrid approach to solving the S3R problems, in which all components are fully embedded in a connectionist system and implemented in a neurobiologically plausible man-

ner. We call this proposal the Symbolic SubSystem Hypothesis. It makes the following three claims: 1. That at a lower level of abstraction, the brain is a connectionist system, where connections and activation patterns are used to represent and process information. 2. That at a somewhat higher level of abstraction, the brain is a CCE system, enabling it to accurately represent the continuous, semantic space of concepts. 3. That the brain has a symbolic subsystem—implemented as a connectionist system—that extracts discrete symbols from the continuous, semantic space, allowing the brain to perform symbolic operations such as (relational) composition. Consider the following analogy: a computer is fundamentally a machine that can perform arithmetic operations on numbers, fully symbolic by any definition. However, computers are often used to simulate connectionist systems, e.g., (spiking) neural networks, in a way that is clearly considered accurate enough for the fields of computational neuroscience and AI. Implementational connectionism proposes the reverse: that the brain is fundamentally a connectionist system, but that this connectionist system implements the operations required to perform symbolic computations (Pinker & Prince, 1988). The Symbolic SubSystem Hypothesis is highly similar to implementational connectionism, but more specific: it does not just propose that the brain is a connectionist system implementing a symbolic system, but that this symbolic system is a subsystem of the brain, which works in tandem with the larger CCE system. The algorithm of the brain is thus not purely symbolic, but is instead a symbolic-connectionist hybrid, fully implemented in neural/connectionist hardware.

To be clear, the Symbolic SubSystem Hypothesis is deliberately vague and does not make any claims about many of the specifics, such as whether the symbolic subsystem is distributed across the brain or localized in a particular area, or whether it is used for all cognitive tasks or only for some. It merely posits that current evidence suggests that the brain implements such a system in a connectionist manner, that such an implementation would allow for simple solutions to many of the S3R problems, and that the brain performs symbolic computations in some form. The model we will propose is a specific instantiation of this hypothesis, and while capable of interesting symbolic computations, does not yet provide the full story. Thus, the Symbolic SubSystem Hypothesis should be taken as somewhat distinct from the model we propose, and as a more general claim about the nature of cognition.

In the cognitive neuroscience literature, top-down control (of memory) is widely recognized as being an important aspect of (higher) cognition. It is used to guide attention, to suppress irrelevant information, and to select the most relevant information for the task at hand. Yet, the use of the word "control" is often used in a metaphorical sense, and it is not clear how this control is implemented in the brain, which invites the criticism that cognitive neuroscientists commit a homunculus fallacy, i.e., by delegating all hard problems to the homunculus supposed to control everything the brain does, they are just pushing the problem of cognition to a different level (Monsell & Driver, 2000). By providing a model of top-down memory control, we aim to show that the control system is not a separate entity but an integral part of the cognitive system, which can be implemented through specific neural circuitry.

Prior to this work, we have developed a functioning cognitive system for reference comprehension that solves a large part of the S3R problems (Sommers et al., n.d.). It solves the task of anaphoric reference (coreference) in linguistics, where a model is required to understand a sentence (in an artificial language) and to use this understanding to resolve what entities are being referred to. However, this system does not yet provide a detailed neurobiological implementation: there are several separate components, only one of which is connectionist in nature, and the control system is fully hard-coded through conditional logic. Although we do provide potential venues for the neurobiological implementation of the different components, this model still falls prey to many of the criticisms levelled at hybrid models, such as the homunculus problem and the duct-tape AI criticism. As such, we aim to provide a more specific, fully integrated neural model that can solve the S3R problems, and that can be used as a reference for future models of cognition.

Another model that has similar goals is the conjunction neuron model by Manohar et al. (2019), which implements a rate-coding model of symbol-like working memory slots or registers. This model uses a pool of neurons, called conjunctive neurons, that have facilitating synapses with all attributes, meaning they can bind to them. By using lateral inhibition, the model can perform a winner-take-all operation, effectively selecting a single memory trace from a pool of candidates. However, this model struggles with several S3R problems: it cannot reliably resolve cases with multiple candidates that share partially overlapping features (the problem of two), it cannot gradually build up memory traces over time (discourse incrementation), and there are several other semantic tasks that it cannot perform, such as transitive inference. We hypothesize that these limitations stem primarily from the lack of a control system that can distinguish between novel and existing memories, dynamically allocate new memory slots while protecting existing ones, and suppress the whole system during other activity. Despite this limitation, this model is a great example of a fully connectionist and neurobiologically plausible model capable of solving a broad set of S3R problems. Moreover, it explains a wide range of findings in the working memory literature and thus provides a good starting point.

We fine-tune and extend the conjunction neuron model with a control system to solve a large subset of S3R problems. This model, which we call the Rate-Coded Bundle Memory (RCBM), is a rate-coding model similar to that used

by Manohar et al. (2019) in their conjunctive neuron model. However, it contains several additional mechanisms that allow it to solve a broader set of S3R problems, the most important of which are required to implement a control system. In reference to our earlier work (Sommers et al., n.d.), RCBM functionally implements Bundle Memory (BM), which solves many S3R problems, but in contrast to our earlier work it is fully embedded in a connectionist system. Importantly, we argue that the combination of mechanisms available to RCBM enables the formation of a Symbolic Subsystem, and that it is exactly the presence of this subsystem that allows RCBM to solve such an extensive range of S3R problems.

Methods

In prior work (Sommers et al., n.d.), we developed a model of working memory called Bundle Memory (BM), named after the thesis by David Hume that objects are nothing but bundles of ideas (attributes) (Hume & Mossner, 2000). This model was designed to solve the S3R problems, and consists of three main components: a control system, a BM system, and a semantic system. Here, we implement this model using a rate-coding model, to embed all components in a single connectionist system. In order to achieve this, we provide neurobiologically plausible implementations of the control system and Bundle Memory modules. We call this model the Rate-Coded Bundle Memory (RCBM) model.

According to the framework constructed in our previous work (Sommers et al., n.d.), all three components of the model are crucial for the functioning of a system capable of solving the S3R problems. The semantic system is responsible for the representation of concepts and their relations, and it allows for graded inference, context-dependence, and groundedness. The BM system is responsible for the retention and manipulation of information, and it allows for one-shot learning and dynamic variable binding. It is implemented through monosynaptic binding, which allows for the creation of short- to medium-term memory traces. Finally, the control system is responsible for the dynamic allocation of memory slots, the distinction between novel and existing memories, and the suppression of the memory system during other activity. In RCBM, we implement all parts within the same rate-coding model, with specialized circuits and connection types to distinguish their functioning (see Figure 1).

In the RCBM model, we provide a more detailed implementation of all three systems. Here, the semantic system consists of a hierarchically organized set of attribute neurons, with connections to visual and auditory sensory neurons (Figure 1, left). Each of these attributes can be externally activated through either sensory input (left) or linguistic input (top). The BM system consists of two memory pools: a Winner-Take-All Memory (WTAM) pool and a Multiple Activation Memory (MAM) pool (Figure 1, bottom right). These pools have pair-wise connections between them, and each pair of memory neurons constitutes a single bundle (memory trace). Each attribute neuron in the semantic system is connected to all memory neurons in the WTAM and MAM pools through facilitating synapses that allow for monosynaptic binding. In this way, the model can store attributes in specific bundles by binding to their corresponding memory neurons. The control system consists of four control neurons. These neurons are responsible for identifying whether the current input is novel (N) or belongs to an existing memory (E), whether there are multiple candidate memories (A), or whether the memory system should be suppressed/cleared (S) (Figure 1, top right). These control neurons interact heavily with the memory system to allocate new memory bundles, to protect existing ones, and activate the appropriate bundle upon retrieval.

Together, these components allow the RCBM model to retain bundles of information over time, to retrieve them when needed, and to manipulate them in a controlled manner. Upon being presented with sensory or linguistic input, the corresponding semantic attribute neurons are activated, and the control system detects whether this is a novel memory or an existing one (see Figure 2 for example sequence). If the input is novel, the control system allocates a new memory slot in the WTAM pool. If there already exists a bundle that matches the input, that memory emerges from the WTAM pool through winner-take-all dynamics. In cases where there are multiple candidate memories, the control system detects this through the MAM pool and awaits disambiguating information before proceeding. Semantic information that is bound to the currently active bundle is reactivated through connections back to the semantic system from the WTAM pool. Any new semantic information is simultaneously bound to the respective WTAM and MAM neurons. In between sequences of words, to prevent interference between bundles, the memory system can be cleared through the suppression neuron. The interactions between these systems is what allows the RCBM model to turn sequences of sensory and linguistic input into bundles of information in a controlled and organized manner.

Rate Coding Model

Our motivation for using a rate coding model is that it balances computability with neurobiological plausibility. Rate coding models are directly informed by observations from neural recordings, and are based on the idea that the firing rate of a neuron can be used to encode information. At the same time, they significantly reduce computational requirements compared to spiking neural networks, which makes them more suitable for large-scale simulations. Our RCBM simulation is split into two main parts per time step: signal propagation and neural activation.

During signal propagation, all neurons propagate their signals to all neurons they are connected to, implemented in the form of a matrix multiplication.

We use leaky neurons, which means that the new state

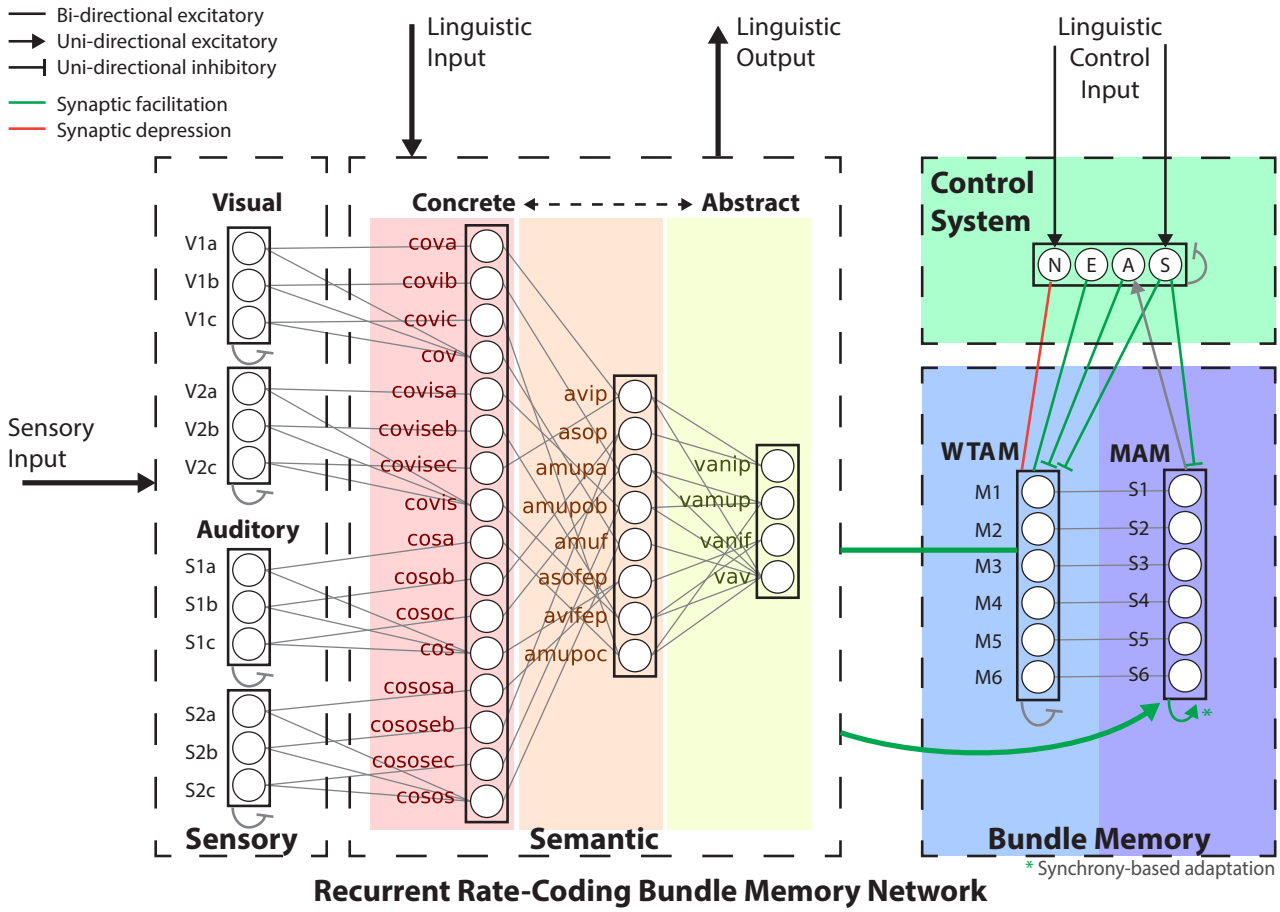

Figure 1

The RCBM model consists of three main components: a semantic system, a control system, and a memory system. The semantic system (left and center) consists of both visual and auditory sensory feature neurons (left) and conceptual feature neurons (center). Each neuron in the semantic system can directly receive external input through corresponding sensory or linguistic input. Linguistic input consists of multiple sentences in an artificial language, presented word-for-word. Each word is represented by a conceptual feature neuron, and these conceptual feature neurons are organized in increasingly abstract layers, with the rightmost layer representing the most abstract concepts. Each non-sensory semantic neuron can directly output corresponding artificial words for the model to produce a response. The control system (top right) consists of four control neurons: a novelty detection neuron (N), an existing memory detection neuron (E), an ambiguity detection neuron (A), and a memory suppression neuron (S). Special linguistic inputs can directly activate the novelty detection neuron (determiner, e.g., ***a*** *man v.s.* ***the*** *man) and the suppression neuron (end-of-sentence token). The memory system (bottom right) consists of two memory pools: a Winner-Take-All Memory pool (WTAM) and a Multiple Activation Memory pool (MAM). Each memory pool consists of a set of memory neurons, which are connected to the conceptual feature neurons and have pairwise connections between them. The control system interacts heavily with the memory system, as it is responsible for activating the appropriate memory neurons and suppressing the memory system when necessary.*

of a neuron is determined by the old state ($R(t)$), and some exponential decay through which the signal returns to baseline over time:

$$R(t+1) = f(\gamma \cdot WR(t) + (1-d_R)R(t) + b + \zeta(t)) - b$$

Where R is the neuronal firing rate, f is a neural activation function, γ is a scaling factor that determines the general signal strength, W is the weight matrix which defines the connections between neurons, d_R is the rate at which the firing rate decays to baseline, b is the baseline activation of the neurons, and $\zeta(t)$ is sampled from a zero-mean normal distribution with a standard deviation equal to noise parameter ζ. For most neurons, we set the noise to 0. Excitatory and inhibitory connections are determined by the sign of the scaling factor. The activation function we use is a linear function clipped between 0 and 1, with 0 representing no firing and 1 representing the maximum firing rate. We use a decay rate of

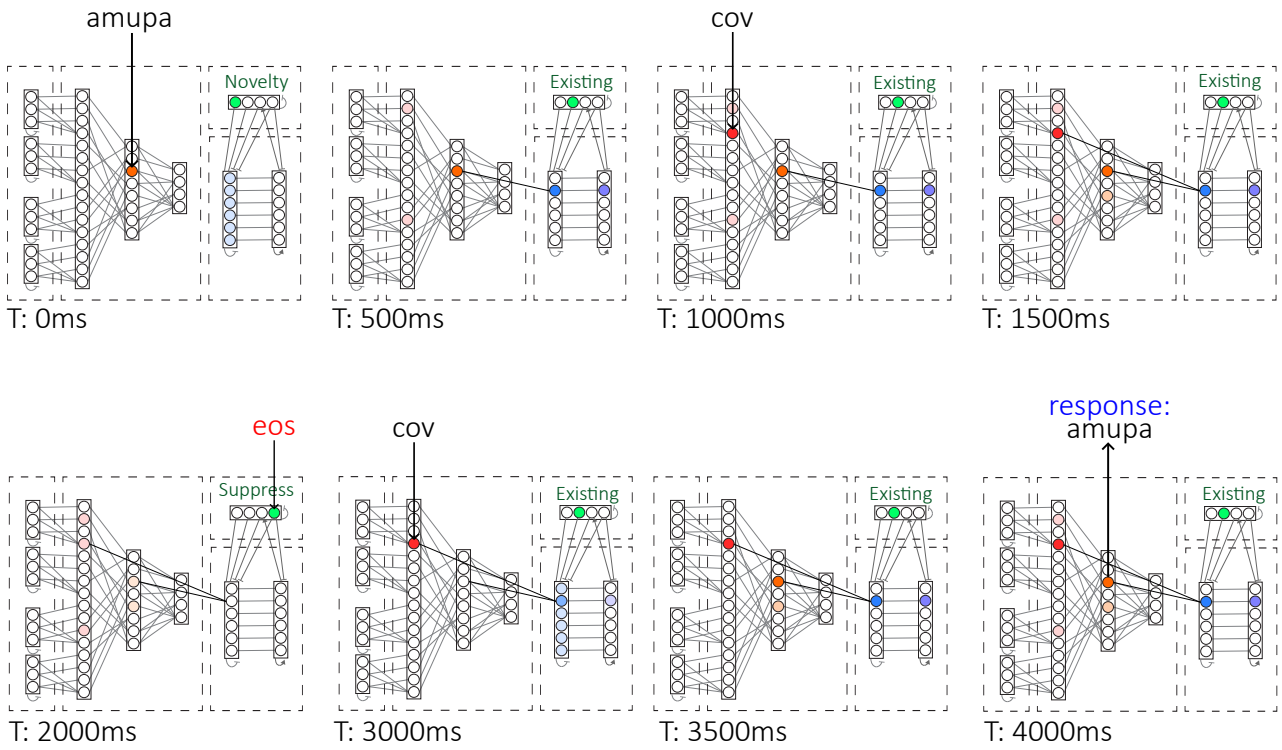

Figure 2

The RCBM model stores and retrieves information to and from a bundle in response to an input sequence. In response to the first word (amupa), the corresponding attribute neuron is activated, and the control system detects that this is a novel memory ($T = 0ms - 500ms$). The control system allocates a new memory slot in the Winner-Take-All Memory (WTAM) pool, which then activates the existing control neuron. The semantic information is then bound to the WTAM and MAM neurons belonging to the same bundle ($T = 500ms - 1000ms$). After the first word, the model receives a second word (cov), again activating the corresponding attribute neuron which is then also bound to the WTAM and MAM neurons of the current bundle ($T = 1000ms - 2000ms$). The model then receives an end-of-sentence signal (EOS), which causes the control system to suppress the memory system and to prepare for the next input sequence ($T = 2000ms - 3000ms$). Finally, the model receives the second word again (cov), which causes the control system to detect that this is an existing memory and to reactivate the corresponding bundle ($T = 3000ms - 3500ms$). This bundle will then lead to the reactivation of the other semantic information that was bound to it (amupa), and the model will be able to retrieve the information that was stored in the bundle ($T = 3500ms - 5000ms$).

0.05, which corresponds to 20 time steps for half the signal to decay, and a baseline of 0.2 for most neurons.

Conjunction Neurons as the Basis of Bundle Memory

Beyond the general rate coding model, which dictates the way in which neurons interact with each other, we also implement several specialized circuits and connection types to distinguish the functioning of the different components of the model. We use the conjunction neuron model by Manohar et al. (2019) as a basis for our model, as it provides a rate coding model of working memory that can already solve a large part of the S3R problems. Primarily, we use the idea of conjunction neurons, which have facilitating synapses with all attributes allowing them to bind to them, to form a Bundle Memory-like model of working memory. Their model also incorporates basic mechanisms for memory allocation and selection, which we extend with a more sophisticated control system to solve a broader set of S3R problems.

The primary ingredient to enable the formation of dynamic memory traces in our model is fast Hebbian short-term plasticity (STP), which has been proposed for neurobiological computational models of working memory (Durstewitz et al., 2000; Fiebig & Lansner, 2017; Fiebig et al., 2020; Manohar et al., 2019; Sandberg et al., 2003). Hebbian learning is a mechanism by which the connection strength between two neurons increases when they fire at the same time (Hebb, 2005; Kempter et al., 1999). We implement STP in the memory and control modules of our model, allowing for dynamic changes in the connectivity between neurons:

$$\Delta W = \eta \cdot \mathcal{H}(R_{\text{source}}, R_{\text{target}})$$

Where η is the STP rate and:

$$\mathcal{H}(R_{\text{source}}, R_{\text{target}}) = g(R_{\text{source}}) \otimes g(R_{\text{target}})$$

Where g is some STP function, R_{source} and R_{target} are the firing rates of the source and target neurons. We disallow self-connections, as these would lead to neurons always reinforcing their own activation. By default, and in the original model by Manohar et al. (2019), the STP function is the identity function, meaning that the connection strength increases linearly with the activation of the neurons. However, we also use different STP functions in our model for specialized neurons, to allow for more complex learning dynamics. Such specialized neurons will be discussed in more detail in later sections. By implementing Hebbian learning, we create the conditions for a system that can dynamically adapt its connectivity to represent new information.

In Manohar et al. (2019), they implement the learning of new memory traces through a pool of neurons called conjunctive neurons, which have facilitating synapses with all attributes, allowing them to dynamically bind to them. These neurons very much resemble a certain type of neurons in our model, and as such we use a very similar mechanism to implement these here. We use an STP rate (η) of 0.5 for connections between the memory and attribute neurons, meaning that the connection strength increases by 50% when both neurons are fully active. For connections from attributes to memory neurons we use a scaling factor (γ) of 0.01, meaning that about 1% of the activation of the attribute neurons is transferred to the memory neurons on each time step, so that there is a temporal delay in the activation of the memory neurons. For connections back from memory to attribute neurons we use a scaling factor of 0.02, so that the memory neurons are faster to activate the attribute neurons when a memory trace is activated.

To enable competition between memory neurons, like Manohar et al. (2019), we implement lateral inhibition between the memory neurons, which leads to a winner-take-all paradigm. Because of this, we will hereafter refer to the primary memory neurons as Winner-Take-All Memory (WTAM) neurons (Figure 1, bottom right). This lateral inhibition is implemented as a negative scaling factor (γ) between the memory neurons, which ensures that when one memory neuron is activated, the other memory neurons are suppressed. As such, when multiple memory neurons are activated, the lateral inhibition ensures that, in the end, only one of them is selected. This allows for the selection of a single memory neuron when attributes are activated that are associated with multiple memory neurons, where the memory neuron most similar to the activated attributes is selected. Taken together, this is what allows for Content Addressable Memory (CAM) in our model, where the activation of a partial memory can lead to the retrieval of the full memory.

To allocate a new memory neuron when no other distinguishing attributes are present, i.e., when a novel memory trace is being formed, we introduce random noise to the system (Manohar et al., 2019). This noise is sampled from a zero-mean normal distribution with a standard deviation equal to some noise parameter, and is added to the activation of the memory neurons on each time step. For our WTAM neurons, we use a noise parameter of 0.005, meaning that there is a 0.5% standard deviation relative to the total possible signal strength (0 to 1). This noise allows for the selection of a random WTAM neuron when no other distinguishing attributes are present, effectively allowing for the assignment of random unassigned WTAM neurons to novel bundle memories.

Increasing Network Stability

The original model by Manohar et al. (2019) uses a relatively simple rate-coding model to implement their conjunction neurons. This model works well for the tasks they set out to solve, but it struggles with some of the more complex S3R problems, such as discourse incrementation and the problem of two. To solve these problems, we need to extend the model with more specialized neural mechanisms that allow for the assignment and retention of information more reliably, and that allow the model to act in a more discrete and symbolic fashion. We will discuss these specialized neural mechanisms in more detail in the following sections.

First of all, we implement decay to our Hebbian weights to enable gradual forgetting of memory traces, regardless of network activity. This is important to prevent the network from becoming saturated with memory traces, and to allow for the formation of new memory traces over time. The original model by Manohar et al. (2019) relied on active unlearning over time, where activations below baseline would lead to negative Hebbian learning and thus to the reduction of the connection strength. We implement a more passive form of decay, where the connection strength of Hebbian weights decreases over time to some specified baseline:

$$\Delta W = (W - w_0)d_W$$

Where w_0 is the baseline connection strength, and d_W is the weight decay rate. For our WTAM neurons, we used a baseline of 0.2 for connections from attributes to memory neurons, ensuring a baseline level of activation to allow and reinforce the assignment of new memory neurons when attributes are active. For connections back from memory to attribute neurons, we used a baseline of 0, as we do not want the memory neurons to activate any attributes that are not actively bound to them. The decay rate is determined through the half-time, which is calculated as follows:

$$d = 1 - 0.5^{1/t_{1/2}}$$

Where $t_{1/2}$ is the half-time, which is set to 30000 time steps for connections between memory and attribute neurons.

To further increase the robustness of our Bundle Memory model, we extend our Hebbian learning with sigmoidal synaptic activation thresholds. This means that binding only takes place under strong pre- and post-synaptic activation, which is desirable to support the binary nature of our WTAM neurons, where attributes are either part of a bundle or they are not. The sigmoidal activation function we use is defined as:

$$f(x) = \frac{1}{1 + e^{-k(x-\theta)}}$$

Where k is the slope of the sigmoid, and θ is the offset.

For our WTAM neurons, we use a sigmoid STP activation function with a slope of 50 and an offset of 0.7 for connections between our attribute and memory neurons. This means that the connection strength only increases when both the attribute and memory neurons are strongly activated, which is important to ensure that the memory neurons are only activated when the attributes are strongly associated with them.

Another mechanism to increase the discreteness of our memory neurons is the use of bistable neurons. These neurons have two stable states, an upper attractor state and a baseline state, and they can switch between these states depending on the input they receive. This is opposed to the leaky neurons we used before, which only have one stable state: their baseline state, to which they return over time without any external input. This further supports the idea that attributes are either part of a bundle or they are not, and that a bundle should either be activated or not. We use an extended version of our decay equation to account for the bistable nature of these neurons:

$$\Delta R = -d_R \cdot c \cdot (R-b) \cdot (R-l) \cdot (R-a)$$

Where b is the baseline state, l is the boundary state, and a is the attractor state. The boundary state is defined as the average of the baseline and attractor states, meaning that the neuron will switch between these states when it crosses the midpoint between them. The decay constant is modelled to ensure that the meaning of the half-life is comparable between normal leaky neurons and bistable neurons, although it is necessarily an approximation because the decay rate depends on how close the state is to the boundary. As such, the decay rate is determined by the maximum gradient of the decay function, which is calculated as follows:

$$c = \frac{r-b}{m}$$

Where r is the root, which can be found by solving for the quadratic equation and determining the discriminant, and m is the maximum gradient, i.e., the gradient of the decay function at this root:

$$m = (r-b) \cdot (r-l) \cdot (r-a)$$

By using bistable neurons, we ensure that the memory neurons are more stable and concrete, leading to a more pronounced winner-take-all dynamic, where only one memory neuron is activated at any given time.

To stabilize the lateral inhibition between competing WTAM neurons, we introduce STP in the inhibitory connections between these neurons. The need for this mechanism arises from the fact that high lateral inhibition makes the network more sensitive to noise, because only a small difference in one neuron can lead to the suppression of all others, despite the fact that they might be equally activated. On the other hand, low lateral inhibition makes the network slow to converge when selecting memory neurons through competition, leading to problems especially in cases with ambiguity or when a novel memory bundle is being formed. Adaptive lateral inhibition balances between these two extremes, where consistent co-activation between memory neurons leads to stronger lateral inhibition (and thus competition) between the two.

We use facilitating synapses with baseline weights of 0.1, a half-time of 10, an STP rate of 0.1, and a scaling factor of -0.2 for the inhibitory connections between the WTAM neurons.

This means that, when two memory neurons are consistently activated together, the lateral inhibition between them will increase, making it more likely that only one of them will be selected in the end.

Control System

While a winner-take-all mechanism is sufficient to assign novel combinations of attributes to Bundle memories, there are several other operations that are required to solve a bigger set of S3R problems. To this end, we add a control system (Figure 1, right top) to our model so that it can distinguish between novel and existing memories, dynamically allocate new memory slots while protecting existing ones, detecting when there are multiple ambiguous candidates, and suppress the whole system during other activity. We implement this control system using a set of specialized neurons that interact with the memory system, and that are responsible for activating the appropriate memory neurons and suppressing the memory system when necessary. We will discuss these control neurons in more detail in the following sections.

One of the key operations that the control system needs to perform is novelty detection, which allows the network to selectively boost unassigned memory neurons when attributes are activated that are not associated with any existing bundles.

The novelty detection neuron (Figure 1, control system, N) is a normal leaky neuron with a baseline activation of 0 and a half-time of 20. It is connected with the memory neurons through facilitating synapses with a baseline of 1 and an STP rate of -0.005, meaning that they are on by default and are gradually suppressed when both pre- and postsynaptic neurons are active. This has the effect that, once a memory neuron

is assigned and turns from a "novel" Bundle into an existing one, it will no longer be activated by (or itself activate) the novelty detection neuron. These facilitating synapses have a half-time of 40000 and a sigmoidal STP function with a slope of 50 and an offset of 0.7, meaning that the STP only occurs when both neurons are sufficiently strongly activated (i.e., when both the memory and novelty neurons are fully active). The scaling factor for connections from the novelty neuron to the memory neurons is 0.025, facilitating unassigned memory neurons when the novelty neuron is activated. The connections from the memory neurons to the novelty neuron have a scaling factor of 0.08, activating the novelty neuron when an unassigned memory neuron is activated.

In practice, this means that when a novel combination of attributes is activated, since no WTAM neuron is yet associated with these attributes, the novelty detection neuron will be activated, which in turn will boost the activation of the unassigned WTAM neurons, allowing one of them to be selected rather than a neuron that is already associated with another set of attributes.

The other side of the coin is existing memory detection, which allows the network to boost assigned memory neurons to promote competition between them, until one is eventually selected. The existing memory detection neuron (Figure 1, control system, E) is a normal leaky neuron with the same properties as the novelty neuron. The memory neurons are connected to the existing memory detection neuron through sigmoidal synapses, meaning that only activation above a certain threshold is propagated.

This sigmoidal connection has a slope of 50 and an offset of 0.7, and the connections generally have a scaling factor of 0.04. The existing memory detection neuron is connected back to the memory neurons through facilitating synapses with a baseline of 0, a half-time of 30000, an STP rate of 0.01, and a scaling factor of 0.02, meaning that memory neurons, once assigned, will also activate the existing memory detection neuron.

This has the effect that, as a WTAM neuron is bound to a set of attributes, it will also activate and bind to the existing memory detection neuron. When, subsequently, another set of attributes is activated that is associated with the same WTAM neuron, the existing memory detection neuron will be activated, which in turn will boost the activation of the WTAM neurons, keeping unassigned WTAM neurons from being activated.

Since the memory system heavily interacts with the semantic system, it may also interfere with other cognitive processes that require the activation of the semantic system without immediate binding or activation of associated bundles. To this end, we add a memory suppression neuron (Figure 1, control system, S) to our control system, which turns off all WTAM neurons, allowing for cortical activity without immediate binding or activation of associated bundles. This same mechanism can also be used to turn off currently active Bundle memory, for example when a new sentence starts or the currently active memory is no longer relevant.

The memory suppression neurons connect to the memory neurons using inhibitory facilitating synapses with a scaling of -0.2, a baseline of 0, a half-time of 10, and an STP rate of 0.1.

They are facilitating to still allow some activation to come through when the network is already silent, while retaining the ability to silence very strong network activity.

All control neurons have lateral inhibition with each other, meaning that only one state can be active at any given time.

The scaling factor for these connections is either 0.02 or 0.03, with suppression acting more strongly on the other control neurons, and novelty acting more strongly on the existing memory detection neuron to ensure a smooth transition from novel to existing memory.

This makes the control system function as a sort of Markov chain with certain transition pathways between the different states. In particular, novel activation should always automatically transition to existing, as the connections activating the novelty detection neuron get slowly suppressed while the connections activating the existing memory detection neuron get facilitated.

Ambiguity Detection

Given these control neurons, it turns out that the model can already solve a large part of the S3R problems. However, it struggles when faced with input with overlapping features, since this leads to ambiguity in the retrieval process. This retrieval ambiguity is difficult because the memory neurons are part of a winner-take-all circuit with noise, making them very eager to converge on a single memory neuron even when the evidence is fully ambiguous. Reducing the strength of lateral inhibition improved this, but directly reduced the network's ability to assign memory neurons to novel bundles of attributes, as this relies on the same noise-driven competition mechanism. To solve this problem, we implement a mechanism to detect ambiguity and suppress the winner-take-all mechanism selectively when the network is presented with ambiguous input, but not when presented with novel input. For instance, when a novel input like "big" is first introduced, the winner-take-all mechanism engages to store the relevant memory trace to a single bundle, but if a "big cat" and "big dog" are already in memory, the network must recognize the retrieval ambiguity caused by "big" and temporarily suppress the winner-take-all mechanism to allow both memory traces to be activated simultaneously.

To detect ambiguity, we add a secondary pool of Multiple Activation Memory (MAM) neurons (Figure 1, bottom right), which bind in parallel with the actual memory neurons and have a 1-on-1 mapping between them, but do not have lateral inhibition so that multiple memory neurons can be activated at

once. The mapping is created by sigmoidal synapses connecting each WTAM neuron with a corresponding MAM neuron in this secondary pool, while inhibiting all other neurons in this pool — in other words, when the WTAM neurons converge, this will also lead to convergence in the MAM neuron pool.

This dampening is 75% of the strength of the activation of the corresponding MAM neuron, which is connected with a scaling factor of 0.05 through a sigmoid with an offset of 0.70 and a slope of 50. Though they have no lateral inhibition between them, the MAM neurons are still suppressed by the memory suppression neuron, using the same parameters as those suppressing the WTAM neurons.

By implementing this secondary memory pool, each bundle memory now consists of two neurons: one in the primary WTAM pool and one in the secondary MAM pool. When the WTAM neurons converge on a single memory neuron, the corresponding MAM neuron will also be activated, but when the WTAM neurons are ambiguous, multiple MAM neurons can be activated.

To allow the MAM neurons to be activated directly and independently from the WTAM neurons, they have their own facilitating synapses coming from the attribute neurons.

These connections have the same parameters as those from the attribute neurons to the WTAM neurons, except with a scaling factor of 0.02 instead of 0.01.

This means that, when a set of attributes is activated that is associated with one or more memories, the corresponding MAM neurons will directly be activated. Note that there are no reciprocal connections from the MAM neurons to the attribute neurons, as there are from the WTAM neurons, meaning that the MAM neurons are not able to activate the attributes they are associated with. This is necessary to prevent mixing up attributes between different memory neurons: since multiple MAM neurons can be activated at once, this would lead to the activation of multiple sets of attributes, which would then become indistinguishable at the moment of retrieval and, in the worst case, lead to the overwriting of both bundles with a mix of their attributes. By having these asymmetric facilitating synapses between the MAM neurons and the attribute neurons, we allow for the identification of multiple (ambiguous) memory bundles without automatically retrieving them.

In order to detect ambiguity, we introduce an ambiguity detection neuron (Figure 1, control system, A) in the control system, which detects the simultaneous activation of two or more ambiguous bundles. Neurons in the MAM pool feed into this ambiguity detection neuron, whose threshold is set such that it requires multiple active MAM neurons to be activated.

The neuron itself has a sigmoidal activation function instead of a clipped linear one, with a slope of 50 and an offset of 0.9, meaning that it only remains on under the presence of enough input from the MAM pool, and otherwise transitions to the existing or novel control state when disambiguating evidence is found. The activation is done through a simple excitatory connection with a scaling of 0.5. The neuron again has lateral inhibition (scaling 0.02) with all other control neurons.

The ambiguity detection neuron inhibits the WTAM neuron pool, thus postponing the resolution of the winner-take-all mechanism until disambiguating information is provided.

This suppression mechanism uses facilitating synapses with the same properties as the connections between the memory suppression neuron and the WTAM neuron pool, but with a strength of only -0.02 to not suppress all activity, but only postpone convergence.

Taken together, this mechanism allows the model to detect ambiguity through the MAM pool and suppress the convergence of the WTAM neurons until disambiguating evidence is provided.

In order for this ambiguity detection to be stable, we need competing bundles to both fully activate their corresponding MAM neurons, rather than one suppressing the other like in the WTAM pool. It turns out that, in itself, the level of activation of the MAM neurons varies greatly depending on the number of activated attributes, the amount of overlap, and the time between memory encoding and retrieval. To ensure that the MAM neurons are activated in a consistent and comparable manner, we use Hebbian learning driven by synchronization-locking, where STP only takes place when the rates of both the pre- and post-synaptic neurons are similar. This mechanism is somewhat similar to auto-associative networks, e.g., Hopfield networks, although the neurons here are memory traces rather than individual attributes. Such rate-synchronization could be explained through spiking coincidence in lower-level neuron simulations, since synchronized spiking leads to coinciding postsynaptic potentials and thus to a higher likelihood of long-term potentiation. This provides a sort of oscillatory resonance that allows specifically for mutually boosting similar signals, so that ambiguous bundles fully activate their corresponding MAM neurons.

First, we calculate the pairwise differences between the activation states of all neurons in the MAM pool, which gives us a matrix of differences between all pairs of neurons (excluding self-connections):

$$\Delta_R = \left| R_{\text{source}} - R^T_{\text{target}} \right|$$

Next, we threshold these differences using a sigmoidal function, which when inverted gives us a matrix S of synchrony values between all pairs of neurons:

$$S = 1 - \frac{1}{1 + e^{-k_S(\Delta_R - \theta_S)}}$$

where $k_S = 50$ and $\theta_S = 0.1$ These values are then used to calculate the STP, which is the product of the synchrony and the normal STP mechanism:

$$\Delta W_{\text{MAM}} = S \cdot \eta \cdot \mathcal{H}(R_{\text{source}}, R_{\text{target}})$$

Where η is the STP rate, as before, and the decay works the same as for other facilitating synapses.

These synchrony detection synapses between neurons in the MAM pool use a baseline of 0, a half-time of 10, an STP rate of 0.1, a sigmoidal STP activation function with a slope of 50 and an offset of 0.3, and an overall scaling of 0.03.

Semantic Network

Of course, a memory system can only meaningfully store information if it is connected to a semantic network that encodes this information. Though the focus of this paper is on the memory system, we aimed to also provide a richer and more extensive semantic network than the one used by Manohar et al. (2019). An important reason for this is that we aim to not just solve a problem in visual working memory, but to solve a wider range of S3R problems which require a more complex conceptual space. To this end, we implement a semantic network based roughly on a Hub-and-Spoke-like architecture (Patterson & Lambon Ralph, 2016), meaning that we have different sensory areas (the spokes; see Figure 1, left) that converge on a central area (the hub; see Figure 1, center) where multimodal and abstract concepts are represented. In addition, we implement semantic information in increasing levels of abstraction, with the first layer connecting to the sensory areas, the second layer connecting to the first, and the third layer connecting to the second, similar to the convergence zones and Binder hierarchies of abstraction (Binder & Desai, 2011). Finally, we implement linguistic input and output layers that map to and from the semantic neurons (Figure 1, top, neurons not shown). All of these semantic neurons represent arbitrary concepts in an artificial conceptual space and language, though they were picked to provide a conceptual space that is rich enough to solve the S3R problems we defined.

Our semantic model consists of two unimodal input layers (representing the visual and auditory modalities), each encoding two attribute classes with three mutually exclusive attributes each, very much like the semantic network used by Manohar et al. (2019). These are the spokes of our Hub-and-Spoke-like architecture, and they are connected to a concrete-semantic hidden layer, which forms the first layer of the semantic hub. This layer contains eight attribute neurons per unimodal area, with four neurons per attribute: one for the attribute itself and three for the properties. The second layer of the semantic hub connects to the first, and contains eight attribute neurons with combinations of concrete attributes, i.e., multimodal concepts like dog (e.g., barking plus visual features). The third layer of the semantic hub connects to the second, and contains four attribute neurons with combinations of multimodal attributes, i.e., abstract concepts like animal (e.g., dog plus cat). We use linguistic input and output layers of 28 neurons each, the same as the total number of attribute neurons, mapping to and from these attribute neurons.

All our semantic neurons have a baseline activation rate of 0.2 and a half-time of 20.

This makes for a relatively rich semantic network, with 2 modalities and 28 attributes of different kinds, for a total of 96 neurons in the semantic system.

Task Design

To test the model's ability to solve a wider range of S3R problems, we implement an artificial language and a set of tasks that require the model to solve (many of) these problems. In total, we define 14 tasks divided over 4 categories: controlled storage, memory retrieval, multiple memories, and semantic association. We evaluate the model's performance on these tasks relative to expected outputs, with the expectation that our full model should be able to solve these tasks with high accuracy.

Each test consists of 3 to 4 trials, representing different sentences that test the same phenomenon. These tests could test the same problem with different semantic content, or different aspects of the same problem (e.g., to test the model's ability to forget over time, we could test it on different time scales). Each trial consists of a sequence of inputs, with each input presented for 500 timesteps. Each input was either a word (one of the 28 conceptual input neurons), a sensory input (one of the 12 unimodal input neurons), a full stop (activating the suppression neuron), a novelty signal (A person vs. THE person, where the former indicates novelty; activating the novelty neuron), or no signal. This means that there are 42 different input possibilities, which are combined in different ways to create the tests. The absence of any input is used to measure responses at important moments in the sequence, as the model will maintain its current state without interference from input signals. Each trial is repeated 25 times, and performance is measured by checking whether these responses match the expected outputs. In total, we ran approximately 1200 trials for our primary model.

Our test suite was divided into four categories: controlled storage, memory retrieval, multiple memories, and semantic association. Controlled storage tests evaluate the model's ability to store information in a controlled manner, i.e., whether memory doesn't *just* store and retrieve any information, but always the *right* information. Memory retrieval tests evaluate the model's ability to retrieve information from memory, both now and over longer periods of time. Multiple memories tests evaluate the model's ability to store and retrieve multiple memories at once, and to distinguish between them. Semantic association tests evaluate the model's ability to use the semantic network in combination with the memory system to solve problems that require the association of different concepts.

First of all, **controlled storage** contains the following tests:

	Semantic	WTAM	MAM	Ambiguity	Novelty	Suppression	Existing
Semantic		$\gamma = .01$ $w_0 = .2$ $\eta = .5$ $k_\eta = 50$ $\theta_\eta = .7$	$\gamma = .02$ $w_0 = .2$ $\eta = .5$ $k_\eta = 50$ $\theta_\eta = .7$				
WTAM	$\gamma = .02$ $w_0 = 0$ $\eta = .5$	$\gamma = -.2$ $w_0 = .1$ $\eta = .1$ $t_{1/2,W} = 10$	$\gamma = \pm .05$ $k_\eta = 50$ $\theta_\eta = .7$		$\gamma = .08$ $w_0 = 1$ $\eta = -.005$ $t_{1/2,W} = 40000$ $k_\eta = 50$ $\theta_\eta = .7$		$\gamma = .04$ $k_\eta = 50$ $\theta_\eta = .7$
MAM			$\gamma = .03$ $w_0 = 0$ $\eta = .1$ $t_{1/2,W} = 10$ $k_\eta = 50$ $\theta_\eta = .3$	$\gamma = .5$			
Ambiguity		$\gamma = -.02$ $w_0 = 0$ $\eta = .1$ $t_{1/2,W} = 10$			$\gamma = -.02$	$\gamma = -.02$	$\gamma = -.02$
Novelty		$\gamma = .025$ $w_0 = 1$ $\eta = -.005$ $t_{1/2,W} = 40000$ $k_\eta = 50$ $\theta_\eta = .7$		$\gamma = -.02$		$\gamma = -.02$	$\gamma = -.03$
Suppression		$\gamma = -.2$ $w_0 = 0$ $\eta = .1$ $t_{1/2,W} = 10$	$\gamma = -.2$ $w_0 = 0$ $\eta = .1$ $t_{1/2,W} = 10$	$\gamma = -.03$	$\gamma = -.03$		$\gamma = -.03$
Existing		$\gamma = .02$ $w_0 = 0$ $\eta = .01$ $t_{1/2,W} = 30000$		$\gamma = -.02$	$\gamma = -.02$	$\gamma = -.02$	

Note: γ *= firing rate scaling factor,* w_0 *= baseline weight,* η *= STP rate,* $t_{1/2,W}$ *= weight decay half-time,* k_η *= STP sigmoid slope,* θ_η *= STP sigmoid offset.*

- Correlation violation: assesses the model's ability to store anti-correlated attributes in memory neurons, i.e., to combine attributes that do not normally occur together without interference.
- Inferred novelty: evaluates the model's ability to infer the novelty of an information sequence, i.e., to allocate a new memory neuron for a sequence that describes a new concept, rather than associating it with an existing memory neuron.
- Determined novelty: evaluates the model's ability to create a novel memory neuron, even in cases where the information might otherwise be inferred as part of an existing memory neuron, by explicitly marking the information as novel.

	Semantic Neurons	WTAM Neurons	MAM Neurons	Control Neurons
ζ	0	.005	0	0
b	.2	.2	0	0
$t_{1/2,R}$	20			20

Note: ζ *= noise parameter,* b *= baseline activation,* $t_{1/2,R}$ *= firing rate decay half-time.*

Secondly, **memory retrieval** contains the following tests:

- Gradual forgetting: assesses the model's ability to forget attributes of a memory over sufficiently long periods of time.
- Long-term remembering: assesses the model's ability to retain information for a certain amount of time.
- Memory maintenance: evaluates the model's basic capability to retain a coherent memory from uncorrelated attributes, i.e., to implement binding and one-shot learning.
- Pattern completion: measures the model's ability to complete a pattern based on partial input of a previously learned sequence, i.e., content-addressable memory.

Thirdly, **multiple memories** contains the following tests:

- Problem of two: assesses the model's ability to distinguish between two memory neurons with overlapping attributes, i.e., to handle competition and resolve to choose one memory neuron over another.
- Discourse incrementation: assesses the model's capability for incremental binding of new information to existing memory neurons over multiple sentences, i.e., to build up an integrated memory representation over time.
- Many memories: assesses the model's capability to create and retrieve up to four different memory bundles in various sequence patterns, without interference.
- Pattern completion with two referents: measures the model's ability to complete a pattern based on partial input of a previously learned sequence when presented with another distractor sequence in between.

Finally, **semantic association** contains the following tests:

- Deduced association: assesses the model's ability to infer that a novel attribute is associated with an existing memory neuron through an existing association with another attribute that is part of that bundle, i.e., to deduce reference to a memory through a concept that has not been mentioned before in combination with that memory.
- Correlation inference: assesses the model's ability to infer the presence of an attribute based on the presence of another attribute in the same memory bundle, based on an association between these two attributes that has been learned before and is stored in the memory system.
- Sensory binding: assesses the model's ability to indirectly bind sensory input to a memory representation, i.e., to use sensory input to activate a conceptual attribute that is then stored in memory, and vice versa to again activate the sensory representation from the memory representation.

Lesioning the Model

To investigate which connections and mechanisms are causally required for the model's functioning, we rerun each test for lesioned versions of the network. These lesions are introduced by not including certain connections in the network, while leaving the neurons themselves and other connections intact. This allows for a clean side-by-side comparison of how the model functions without some of its connections. We define lesions to different parts of the network:

- Existing: removes the connections to and from the existing memory detection control neuron.
- Ambiguity: removes the connections to and from the ambiguity detection control neuron.
- MAM: removes the connections between the WTAM neurons and the MAM neurons, as well as the connections from the MAM neurons to the attribute neurons.
- MAM synchrony: removes the lateral facilitatory connections between the MAM neurons that drive the synchrony detection mechanism.
- Sensory grounding: removes the connections between semantic layers, leading to loss of semantic information.
- WTAM completition: removes the lateral inhibitory connections between the WTAM neurons that drive the winner-take-all mechanism.
- Suppression: removes the connections to and from the memory system suppression control neuron.
- Novelty: removes the connections to and from the novelty detection control neuron.

- WTAM: removes the connections between the attribute neurons and the WTAM neurons, in both directions.

We rerun each test for each lesioned version of the network besides the full unlesioned model, for a total of around 12000 trials across all tests and lesions.

Results

To evaluate the capabilities of the Rate-Coding Bundle Memory (RCBM) model, we designed a comprehensive set of tasks aimed at addressing key challenges in cognitive modeling, particularly those related to Symbol Recombination, Retention, and Resolution (S3R). These tasks were carefully constructed to test the model's ability to store, retrieve, and manipulate information in a controlled manner in a way that is relevant to how humans construct memories based on linguistic input (Baggio & Hagoort, 2011; McElree et al., 2003; Seuren, 2009). Specifically, we assessed the model's performance across four categories: controlled storage, memory retrieval, tracking multiple memories, and semantic association. Each task was designed to probe specific cognitive phenomena, such as one-shot learning, ambiguity resolution, and incremental memory integration, which are critical for understanding higher-order cognition. Additionally, we conducted lesion studies to systematically investigate the causal necessity of individual components and mechanisms within the model, providing insights into how different subsystems contribute to its overall functionality. This approach allowed us to rigorously test the model's ability to solve a wide range of cognitive problems, and to identify the specific mechanisms that are required for performing these tasks effectively.

We show that the RCBM model is capable of solving a wide range of S3R problems, achieving >95% performance on all tasks (Figure 4, top row). In the category of controlled storage, the model successfully demonstrated that it could identify when a new entity was being introduced (inferred novelty), but also that it could be forced to create a new memory neuron even when the information could be inferred as part of an existing memory neuron (determined novelty). It also demonstrated that it could store anti-correlated attributes in memory neurons and retrieve them correctly (correlation violation). Under memory retrieval, the model was able to retain a coherent memory from uncorrelated attributes (memory maintenance) for prolonged periods of time (long-term remembering), and to complete a pattern based on partial input of a previously learned sequence (pattern completion). However, after sufficiently long periods of time, this information was forgotten (gradual forgetting), freeing up space for new information. All these processes worked for multiple memories as well: the model was able to distinguish between two memory neurons with overlapping attributes (problem of two; pattern completion two referents), to incrementally bind new information to these entities over time (discourse incrementation), and to create and retrieve up to four different memory bundles in various sequence patterns (many memories). Finally, the model was able to use the semantic network in combination with the memory system to solve problems, such as inferring that a novel attribute was associated with an existing memory neuron through an existing association (deduced association), inferring the presence of an attribute based on the presence of another attribute in the same memory bundle (correlation inference), and the binding of sensory input to a memory representation (sensory binding). Altogether, these results demonstrate that the RCBM model is a versatile model that succeeds at its task of reliably solving a wide range of S3R problems.

In constructing the RCBM model, we demonstrate that it is possible to solve many problems that are symbolic in nature in a connectionist and neurobiologically plausible manner. In particular, the control system of our model exemplifies this by using mechanisms based on known neural processes, such as Hebbian learning, lateral inhibition, and facilitating synapses. These mechanisms allow the control system to dynamically allocate memory, detect novelty and ambiguity, suppress irrelevant information, and resolve competition between overlapping memory traces. Importantly, all components of the control system are embedded within a single rate-coding framework, avoiding abstract or biologically implausible operations. This implementation provides a clear example of how cognitive processes can be modeled in a way that is consistent with our current understanding of neural function.

We redesigned the original model by Manohar et al. (2019) for use in a more linguistic task by both extending the semantic network and altering the input and output layers to be sequential and symbolic, and thus more language-like, while still retaining the original more graded sensory input layers. By presenting pseudowords to the network one by one, the network identifies entities that are being referred to and is able to e.g., distinguish between two similar referents (Figure 3). This requires the model to incrementally build up a coherent semantic representation of the presented information over time, which is a key aspect of language comprehension. In particular, here, tasks like novelty detection, the problem of two, and discourse incrementation show that the model is able to identify whether a new entity is being introduced or an existing one is being referred to and, if so, to which one. In addition, it shows it can incrementally bind new information to these entities over time.

We lesioned different parts of the model to investigate which connections and mechanisms are causally required for the model's functioning. In doing so, we found that the model's performance was measurably reduced on particular tasks when lesions were introduced to the model (Figure 4), indicating that the control and memory systems are crucial for the model's ability to perform these tasks. We identify

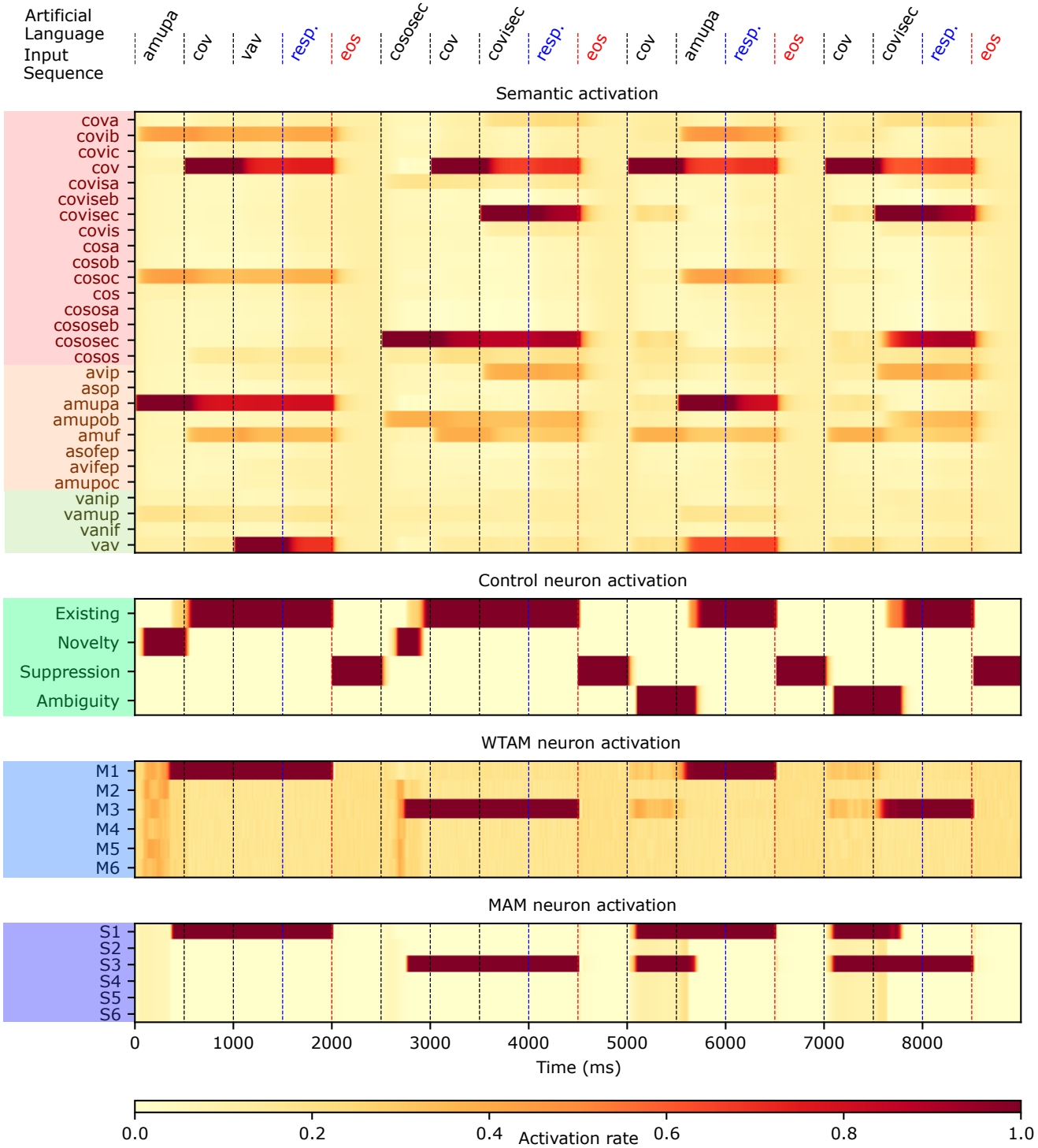

Figure 3

An example trial of the problem of two test, presented as a sequence of words in an artificial language (top). In the first two sentences, the model is introduced to two entities with an overlapping feature, e.g., a blue square and a purple square. The next two sentences each start with this overlapping feature (e.g., square) and then introduce a distinguishing feature (e.g., either blue or purple), requiring the model to temporarily withold its response until the distinguishing feature is presented. The model stores the two entities in M1/S1 (sentence 1) and M3/S3 (sentence 2) in the WTAM and MAM pools, respectively (bottom); when presented with the overlapping feature (sentence 3/4), the ambiguity detection neuron is activated (middle), supressing the winner-takes-all process until disambiguating information is presented.

4 types of impairment that we observed across the different lesions: impaired disambiguation, impaired semantic association, impaired memory capacity, and complete impairment. In impaired disambiguation, the model was unable to perform the problem of two task (Figure 4), meaning it was unable to distinguish between two memory neurons with overlapping attributes. This impairment occurred when the model was lesioned in the ambiguity detection control neuron, the MAM neurons, or the MAM synchrony connections (Figure 4); all of which are involved in the detection and resolution of ambiguity in the input. In impaired semantic association, the model was unable to perform the deduced association and correlation inference tasks, both of which rely on the retrieval of semantic information. Logically, this impairment occurred when the model was lesioned in the sensory grounding connections (Figure 4), which encode the semantic information in the model. In impaired memory capacity, the model was unable to remember more than a single memory at a time, conflating multiple memories into a single bundle and failing at a wide range of tasks. It occurred in two different ways: one when the model was lesioned in the WTAM competition connections (Figure 4), which made the model unable to select a neuron in the memory system (thus selecting them all); and one when the model was lesioned in the suppression control neuron (Figure 4), which made the model unable to clear its state to be able to switch between memories. Finally, in complete impairment, the model was unable to perform any of the tasks, indicating that the model was unable to store or retrieve any information at all. This occurred when the model was lesioned in the novelty control neuron or the WTAM neurons (Figure 4), which are both crucial for the operation of the memory system. Interestingly, the existing control neuron was not strictly necessary for the model to perform any of the tasks, and it seems to act like a neutral state that does not affect the model's performance when it is lesioned (Figure 4).

Discussion

Summary of findings

In this study, we introduced the Rate-Coding Bundle Memory (RCBM) model, a neurobiologically plausible framework designed to address key challenges in cognitive modeling, particularly those related to Symbol Recombination, Retention, and Resolution (S3R). Through a series of carefully constructed tasks, we demonstrated that the RCBM model is capable of solving a wide range of cognitive problems, including one-shot learning, ambiguity resolution, incremental memory integration, and pattern separation. These tasks required the model to dynamically store, retrieve, and manipulate information in a controlled manner, mimicking the processes involved in human memory and cognition.

A key feature of the RCBM model is its control system, which enables the detection of novelty, ambiguity, and existing memory states, as well as the suppression of memory activity when necessary. This control system, implemented using mechanisms such as lateral inhibition and specialized facilitating synapses, allows the model to perform complex memory management tasks. By embedding all components within a single rate-coding framework, the model avoids abstract or biologically implausible operations, providing a concrete example of how symbolic-like computations can emerge from connectionist principles.

Our results show that the RCBM model achieves near-perfect performance across a diverse set of S3R tasks, demonstrating its versatility and robustness. Our lesion study further revealed the causal necessity of specific components and mechanisms within the model, highlighting the importance of the control system and memory subsystems for solving S3R problems. These findings underscore the potential of the RCBM model as a proof-of-concept of the Symbolic Subsystem Hypothesis, and provides a starting point for more comprehensive computational models of higher-order cognition.

One limitation that is visible from our lesion results, is that our RCBM model may be over-reliant on the bundle memory subsystem. When the novelty control neuron is lesioned, almost everything stops working (Figure 4) because no new information is stored in the bundle memory subsystem anymore. Subsequently, our model loses all information not in bundle memory through decay, leading to loss of all information. In contrast, the brain also has cortical short term memory in the semantic network, which could maintain information independently of the bundle memory system (Christophel et al., 2017; Sreenivasan et al., 2014). Extending our model with a more robust cortical short term memory system could be one fruitful direction to make the model less dependent on the bundle memory system.

Neurobiological grounding

We have mentioned the term "neurobiologically plausible" several times in this paper, but there is of course a difference between a mechanism being plausible and it being actually observed in the brain. It turns out that the RCBM model includes many mechanisms for which there is substantial empirical support in the neuroscientific literature. Here, we will discuss several ways in which the RCBM model is consistent with existing theories and findings in the field of neuroscience, and how properties of the RCBM model may lead to new predictions about the brain. We will primarily focus on the hippocampus and dorsolateral prefrontal cortex, as these are the brain areas that are most relevant for the S3R problems that our model solves, given their involvement in episodic- and working memory (Baddeley, 2003; Curtis & D'Esposito, 2003; D'Esposito & Postle, 2015; Moscovitch et al., 2016). It has recently been argued that the hippocampus and prefrontal cortex employ a comparable sequence memory algorithm

Figure 4

The RCBM model performs well on all tests, while lesions to the model lead to systematic reductions in performance. Each bar represents the proportion of trials (right y-axis) that the model successfully (blue) or unsuccessfully (red) completed for each test. The tests are displayed on the x-axis (top) and grouped into four categories. The different lesions are displayed on the y-axis (left), with each lesion representing a different part of the model that was turned off. Per test, there were up to 4 trial types that were each repeated 25 times, for a total of 12000 trials across all tests and lesions.

(Whittington et al., 2025), which supports considering these two brain regions simultaneously.

Binding mechanisms

The RCBM model can store and retrieve information through fast Hebbian plasticity. It is well-established that the hippocampus is capable of fast (and long-term) Hebbian plasticity, and that this can already occur after a single presentation (Kesner et al., 2008; E. Rolls, 2013). There is more debate on the mechanisms of synaptic plasticity in the prefrontal cortex, with the dominant hypothesis being that it stores information through persistent neural firing, i.e., in the activity of the neurons themselves rather than in the synapses (Curtis & D'Esposito, 2003; Fuster & Alexander, 1971). However, there is increasing evidence that working memory in the dorsolateral prefrontal cortex also requires fast synaptic plasticity mechanisms, not unlike those used in the hippocampus, in particular for dealing with multi-item tasks (Lansner et al., 2023; Lundqvist et al., 2018; Mongillo et al., 2008; Sreenivasan et al., 2014; Stokes, 2015). The RCBM model proposes circuitry that the hippocampus and dorsolateral prefrontal cortex could implement to utilize fast Hebbian plasticity to store and retrieve information in memory.

In addition, the RCBM model predicts that the attributes are bound to the bundle memory nodes in a conjunctive manner, meaning that multiple attributes can be bound to the same object. Several theories and models in the literature on the hippocampus (O'Reilly et al., 2014; Schapiro et al., 2017; Teyler & Rudy, 2007) and in (visual) working memory/PFC literatures (Fougnie & Alvarez, 2011; Luck & Vogel, 1997; Manohar et al., 2019; Wheeler & Treisman, 2002), have found evidence that attributes are indeed conjunctively bound to objects for both episodic as well as working memory. Our model is consistent with these findings and theories, and provides a more detailed account of how this binding process could be implemented in the brain.

To allow storage and retrieval of the same information, the RCBM model predicts that the connections between attributes and bundle neurons must be reciprocal: the strengthening of an incoming connection between an attribute and a bundle should lead to the strengthening of the relevant outgoing connection as well. Hippocampal replay and cortical consolidation provides some indirect evidence for this idea, since these phenomena can only be effective if the connections to and from the entorhinal cortex and the hippocampus are similar enough that memories encoded in the hippocampus can be successfully retrieved via the cortical-hippocampal-cortical loop (Maingret et al., 2016; Rothschild et al., 2017; Siapas & Wilson, 1998). Additional support for this idea is that there is topographical reciprocity: regions of the entorhinal cortex that project to particular hippocampal parts receive output from those same parts (Naber et al., 2001; Tamamaki & Nojyo, 1995; Witter & Amaral, 2021). The RCBM model more strongly predicts that such reciprocal connections are necessary for the storage and retrieval of information in the brain, and that this is a general mechanism that can be used in other brain areas as well.

The RCBM model predicts selective strengthening of some connections (the attributes of the same memory) and not others (the attributes of an interfering memory). This can be done in models with compartmentalized synapses, where the strengthening of one connection does not necessarily lead to the strengthening of all other connections of the same neuron: instead synaptic plasticity can be localized to individual dendritic spines (Sabatini et al., 2002; Yuste & Denk, 1995). Indeed, there is evidence that synapses can be compartmentalized in the hippocampus (Matsuzaki et al., 2004; Yuste & Denk, 1995), and that compartmentalized synapses are important for encoding memories there (Govindarajan et al., 2011; Yang et al., 2008). In fact, the simplifications that many computational neuron models make limit the range of computations that they are able to perform, not just in the hippocampus but throughout the brain (Cazé et al., 2013; Legenstein & Maass, 2011; Morita, 2008; Poirazi et al., 2003). Nevertheless, more evidence is required to confirm the prediction that the memory-related compartmentalized synapses in the hippocampus (and elsewhere) do indeed bind semantic attributes only to certain memory neurons and not others.

Finally, a major computational advantage of bundle memory is its ability to do dynamic variable binding. Dynamic variable binding is an important principle of symbolic computation (see Discussion Section on why Bundle Memory is symbolic). Like variables in a computer program, hippocampal place cells only have stable representations within particular contexts: When the context changes, the representations dynamically change or remap and place cells that were active in one environment, become silent in another, while different cells become active (Fenton, 2024; Jeffery, 2011). Moreover, similar to how a variable can represent a variety of different values, the hippocampus is capable of binding a variety of different (grid cell) features to individual (place) cells (Fenton, 2024; Jeffery, 2011). Place cells can even concurrently store multiple distinct locations in a context-dependent manner (Park et al., 2011), suggesting memories can be multiplexed, something that is not implemented in the RCBM model. At any rate, the flexible binding mechanism of the hippocampus makes it ideally suited for symbolic computation (Kazanina & Poeppel, 2023; Kurth-Nelson et al., 2023)

Distinction between the WTAM and MAM pools

We likewise see parallels between our model and the hippocampal circuit when it comes to the distinction between Winner-Takes-All Memory (WTAM) and Multiple Activation Memory (MAM) pools. More specifically, the functional distinction between WTAM (separating memories through lateral inhibition) and MAM (keeping multiple memories

active through auto-association) can be seen as similar to the distinction between neuronal properties of the dentate gyrus and CA3 in the hippocampal circuit.

In order to prevent interference between memories, which may have overlapping or contrasting attributes, the WTAM pool allows activation of only a single bundle. Similarly, the dentate gyrus has very sparce coding, meaning that only a small number of dentate gyrus granule cells are active during encoding or retrieval of a particular memory (Diamantaki et al., 2016; X. Liu et al., 2012; Neunuebel & Knierim, 2012). To achieve activation of only a single bundle, the RCBM model implements a winner-takes-all mechanism in the the WTAM pool, which uses lateral inhibition between the different bundles. The dentate gyrus, one of the main gateways between entorhinal cortex and hippocampus, is known to have exceptionally strong lateral inhibition (Cayco-Gajic & Silver, 2019; Espinoza et al., 2018). This same mechanism is used when encoding new information into memory, as the lateral inhibition between unbound WTAM neurons allows the model to allocate a single random new bundle when presented with a new item (Manohar et al., 2019; E. T. Rolls & Treves, 2024). This process of competition between memories also provides a decision mechanism when multiple different candidates are present, as it allows the model to accentuate the differences (=pattern separation) and select the most relevant candidate for the current context. Evidence from multiple sources have implicated the dentate gyrus in such pattern separation (see Schmidt et al. (2012) for a review).

However, the WTA mechanism runs into problems when competing candidates are so similar that selecting one over the other would be too hasty. In these cases, a WTA mechanism by itself would force the model to arbitrarily pick one candidate over the other, even when they are equally relevant. Instead, we would like the model to be able to keep both candidates active at once, and to suspend the WTA mechanism until the model has gained enough information to select one of them. In order to solve this problem, we implemented a second pool of memory neurons, the MAM pool, which is activated in parallel with the WTAM pool. This MAM pool is connected to the WTAM pool in a 1-on-1 mapping, meaning that each WTAM-MAM pair stores a single bundle memory. Instead of lateral inhibition, the MAM pool has facilitating synchrony-based synapses, which allow for the parallel activation of multiple similar memories. We view these synchrony-based facilitating synapses as a rate-based version of spike-coincidence boosting mechanisms that can be found in (hippocampal) neurons (Bi & Poo, 1998) and spiking neural networks (Izhikevich et al., 2004). With these lateral synchrony-based facilitating synapses, the MAM pool implements a form of auto-associative memory (Hopfield, 1982; Izhikevich et al., 2004), but rather than storing features of the same pattern as is typically done with auto-associative Hopfield networks, the MAM pool instead stores multiple previously activated bundle memories. The CA3, unlike the dentate gyrus (DG), has many excitatory recurrent connections that bind to one another, and is therefore (at the network level) thought to implement a form of auto-associative memory (McNaughton & Morris, 1987; E. T. Rolls & Treves, 2024).

By implementing both the WTAM and MAM pools, the RCBM model is able to keep multiple similar memories active at once without interference of the semantic content of the respective memories. Based on the above parallels, we predict that the hippocampal DG and CA3 exhibit a similar division of labour as our WTAM and MAM pools.

Controlled Sequential retrieval

As a result of the aforementioned winner-take-all mechanism, the RCBM model retrieves only one bundle at a time. Thus, the RCBM model predicts sequential retrieval of individual bundles. Previous theoretical and empirical work strongly supports the idea that the hippocampus encodes and replays episodic memories in a sequential manner (Buzsáki & Tingley, 2018; Ólafsdóttir et al., 2018). For instance, when navigating a maze, rats are known to sequentially retrieve the memories of the different paths they have taken (or will take) (Foster & Wilson, 2006; Takahashi, 2015). There is also evidence suggesting that working memory in the dlPFC similarly involves sequential memory operations (Lisman & Jensen, 2013; Miller et al., 2018).

In order to control the sequential retrieval, the model requires a mechanism to switch between active bundles. One way the RCBM model switches between bundles is through the suppression control neuron. This neuron is responsible for suppressing the activity of the current WTAM and MAM neurons, allowing the model to re-initialize competition and activate a different bundle. Currently, the RCBM model relies on new sensory input to trigger the retrieval of a different bundle after suppression of the previous one.

But the RCBM model requires more than just suppression to control retrieval, since merely suppressing the current bundle does not afford much control. Right now, the additional control signals are primarily the novelty and ambiguity signals. Through our lesion analysis we show that such control mechanisms are necessary for our model to perform particular symbolic computations. But human behaviour is complex enough that more control can be expected: For instance, the RCBM model does not implement a mechanism to directly activate a different bundle memory after suppression, which could be useful for actively searching through or iterating over bundles.

Though we expect our control system to be greatly simplified compared to the actual brain, we believe that the model's control states are used in the brain, and that they are necessary for the brain to solve accurate retrieval, especially in problem of two cases, which requires the detection of ambiguity. In

particular, we predict that the brain should be able to detect novelty and ambiguity in the input, and to use this information to control the memory system. Given this prediction, it should be possible to read out these control variables from the brain, using neuroimaging techniques such as fMRI or EEG.

Localization of control signals

There already exists a wealth of neuroimaging evidence reporting novelty and ambiguity signals in the brain. Event related potentials (ERPs) such as the N400 and the Late Positive Complex (LPC/P300) are sensitive to novelty (Barry et al., 2020; Friedman & Johnson, 2000; Kutas & Federmeier, 2009), and the Nref has been found to track (referential) ambiguity (van Berkum et al., 1999; Van Berkum et al., 2003). Furthermore, a sensitivity to novelty has been observed in the medial temporal lobe, which includes the hippocampus (HPC), the dorsolateral prefrontal cortex (dlPFC), the orbitofrontal cortex (OFC), the striatum (S), and the basal forebrain (BF) (Fredes & Shigemoto, 2021; Geiger et al., 2018; Kafkas & Montaldi, 2018; Kishiyama et al., 2009; Petrides, 2007; Zhang et al., 2022). Specific evidence for a sensitivity to memory retrieval ambiguity or interference is less clearcut, but brain areas like the anterior cingulate cortex (ACC), the medial prefrontal cortex (mPFC), the left inferior frontal gyrus (lIFG) and the hippocampus have all been associated to these processes (M. C. Anderson et al., 2016; Badre & Wagner, 2007; Jonides et al., 2008; van Kesteren et al., 2012). The above areas are logical candidates either for the control networks in our model (ACC, OFC, S, BF, mPFC, lIFG) or for the bundle memory module itself (HPC, dlPFC).

However, detection of control signals alone is clearly not sufficient. The RCBM model further predicts that these ambiguity and novelty are used in the control of memory. The strongest evidence for a relationship between novelty signals and memory control comes from the VTA-hippocampus loop: The hippocampus detects newly arrived information not yet stored in memory, sends this novelty signal (indirectly) to the ventral tegmental area (VTA), which in turn sends dopamine back to the hippocampus to boost memory encoding through the effect dopamine has on long-term potentation there (Duszkiewicz et al., 2019; Lisman & Grace, 2005). Similar mechanisms might be at play in the prefrontal cortex as well: For instance, the dorsolateral prefrontal cortex has been implicated in novelty-based benefits on memory (Geiger et al., 2018; Kishiyama et al., 2009); Similarly, the orbitofrontal cortex is associated with the detection of novelty, and becomes more active during memory encoding specifically (Petrides, 2007; E. T. Rolls & Treves, 2024). As for the relationship between ambiguity detection and memory control, the negative effects of similarity-based interference on memory encoding and retrieval are well-established (M. C. Anderson & Neely, 1996). For instance, McElree et al. (2003) found that memory for a target word was worse when it was preceded by a similar word, compared to when it was preceded by a dissimilar word. However, the brain can overcome similarity-based interference by applying cognitive control (Amer & Davachi, 2023; M. C. Anderson et al., 2016; Badre & Wagner, 2007; Jonides et al., 2008; van Kesteren et al., 2012), as in our model. This supports the idea that the brain is capable of actively mitigating negative effects of ambiguity to resolve interference between similar memories. Together, these findings suggest that the brain is capable of detecting novelty and ambiguity, and moreover that these signals are used to affect memory encoding.

Transmission of control signals

The release of neuromodulators such as dopamine and norepinephrine for novelty and acetylcholine for both novelty and ambiguity are good candidates for ways in which these control signals could cascade and causally influence the neurons in the memory system (Bazzari & Parri, 2019; Kafkas & Montaldi, 2018; Palacios-Filardo & Mellor, 2019). Nuclei in the midbrain are able to release these neuromodulators to many (frontal) areas at the same time, including to the hippocampus and the prefrontal cortex and they are implicated in cognitive control of behaviour (Aston-Jones & Cohen, 2005; Beeler & Dreyer, 2019; Cools & D'Esposito, 2011; Hasselmo & Giocomo, 2006; Záborszky et al., 2018). Release of neuromodulators can be triggered by prefrontal areas, such as the anterior cingulate cortex (ACC) and the orbitofrontal cortex (OPFC), two areas that are linked to cognitive control (Avery & Krichmar, 2017). As mentioned in the previous section, dopamine — released in response to novelty — is known to have an effect on memory encoding, in both the hippocampus and the dorsolateral prefrontal cortex, through the effect dopamine has on long-term potentation (Duszkiewicz et al., 2019; Lisman & Grace, 2005). In the hippocampus, acetylcholine is able to push certain neurons into high and sustained firing rates, which is thought to be a mechanism for memory encoding (Douchamps et al., 2013; Hasselmo et al., 2002). Moreover, when the cholinergic system is blocked, memory encoding is impaired (Maurer & Williams, 2017). In addition, the cholinergic system — known for its role in attention and cognitive control — is also important for pattern separation in the hippocampus (Myers & Scharfman, 2009; Raza et al., 2017). Acetylcholine from the medial septum to the dentate gyrus (DG) in the hippocampus can increase lateral inhibition in the DG, which is thought to be an important mechanism by which pattern separation is achieved (Myers & Scharfman, 2009; Raza et al., 2017). Acetylcholine could therefore provide a mechanism to exert novelty and/or ambiguity control in the brain.

Neuromodulators like dopamine and acetylcholine act diffusely: unspecific to particular neurons (C. Liu et al., 2021; Özçete et al., 2024). This is similar to how in our model the novelty node is connected to all memory nodes. One differ-

ence is that in our model the novelty signal only has an effect on memory nodes that are unbound, as determined by the connections from the novelty node to the memory nodes. If neuromodulators like dopamine and acetylcholine are indeed novelty signals and unspecific, then we would predict that bound memory nodes have an internal state that is different from unbound memory nodes, which determines whether they engage in new long-term potentiation and how they respond to the dopamine and acetylcholine signals. There is some empirical evidence for such internal states: for instance, neurons have homeostatic or metaplastic mechanisms in which recent long-term potentiation can lead to inhibition of further long-term potentiation (Abraham, 2008; Huang et al., 1992). This could provide a concrete neurobiological mechanism for the assignment of novel memories to unbound memory nodes used in our model.

Acetylcholine is also linked to sequential memory retrieval through its role in theta oscillations: Cholinergic (as well as GABAergic and glutamatergic) projections from the medial septum regulate the theta rhythm in the hippocampus (Nuñez & Buño, 2021). One prominent hypothesis is that memory encoding takes place on the peaks of this theta rhythm, while retrieval takes place on the troughs, and that this is mediated by the cholinergic signal (Douchamps et al., 2013; Hasselmo et al., 2002). On this view, acetylcholine also schedules serial memory encoding and retrieval operations, similar to how our model's novelty and existing control signals decide between designating either a new unbound memory node (≈encoding) or one of the bound memory nodes (≈retrieval). Thus, the separation of encoding and retrieval by the theta cycle could be the brain's way of implementing a form of serial scheduling of memory operations, similar to RCBM.

To summarize, the findings above corroborate that some of the control states in the RCBM model are measurable in the brain and that they are connected to memory processes. The brain may propagate these control states by controlled release of certain neuromodulators known to influence memory encoding and retrieval. This suggests that the brain has a memory control system that uses similar control signals as those in our model to regulate memory encoding and retrieval. However, it should also be clear from this discussion that the control system in the RCBM model is a great simplification of the complexity of the brain's memory control system and should thus not be taken literally. In that same vein, our RCBM model is ultimately a functional model, not a biologically realistic model of specific brain circuits, e.g., the cortico-hippocampal circuit. Thus, while there are many interesting parallels, there are also clear differences between the RCBM model and the brain. The RCBM model merely serves as a proof of concept of how a memory control system could be implemented through neurobiologically plausible mechanisms and clarifies how such mechanisms can be exploited in the service of symbolic computations.

What makes Bundle Memory symbolic?

We believe that a digital storage medium is a prerequisite for symbolic computation. It enables the capacity of storing discrete representations of information — variables or symbols. Since these representations are stored in a discrete way, they can be manipulated and later retrieved, enabling a form of read-write addressable memory (Atkinson & Shiffrin, 1968; Gallistel & King, 2009; Kumaran et al., 2016). To create a digital storage medium, we implemented mechanisms that discretize the continuous state space of neural networks, e.g., bistable neurons, winner-takes-all lateral inhibition, and sigmoidal activation functions. In our model, symbols are represented by WTAM-MAM neuron pairs, which form the discrete units that we refer to as bundles. This bundle can then be manipulated as a single unit, i.e., as an individual object (or token) distinct from its bound attributes (or types), yet simultaneously instantiating them (Fodor & Pylyshyn, 1988). The distinguishing of types from tokens, in turn, enables a form of indirection or abstraction where syntactic operations can be performed on the created symbols — instead of operating on the semantic content directly (Fodor & Pylyshyn, 1988). Since the model can perform syntactic operations on these symbols without needing to know the semantic content of the attributes, these operations are independent and generalizable to newly learned attributes, which is what affords systematicity (Fodor & Pylyshyn, 1988). Moreover, this means that the contents of the bundles can be completely arbitrary, much like the contents of a variable in a computer program. In essence, this arbitrary binding to discrete symbols, also known as dynamic variable binding, is what provides flexibility. It enables symbolic operations like fully compositional processing of information (Hummel, 2011; Hummel & Holyoak, 2003; Marcus, 2001), since it makes possible the binding of and control over never before seen combinations of attributes. The properties of discretization, read-write addressable memory, type-token distinction, syntactic operations, systematicity, dynamic variable binding and compositionality are all hallmarks of symbolic computation.

The symbolic operations implemented in RCBM allow it to regulate its own read-write operations to perform proper memory management, which is necessary for one-shot representation of more complex sequences of information, e.g., by preventing interference coming from multiple similar items. While the RCBM model right now only has a limited set of symbolic operations it can perform, it is easy to imagine new types of symbolic operations that can be implemented at a later time. For instance, it should be relatively straightforward to implement comparisons or logical conjunctions or disjunctions between different bundles. Likewise, it is easy to imagine some kind of serial search or other forms of iteration over multiple active bundles. Therefore, by taking this first step of implementing discrete symbols in a neural network alongside some symbolic operations, we have opened the way

to extend the RCBM model with other symbolic operations, including ones that are naturally performed by humans.

Demystifying the control homunculus

Executive control has long been a central concept in explaining a wide range of cognitive phenomena. However, the concept of executive control has been criticized for delegating the solving of all hard cognitive problems to an underspecified, centralized, top-down control system, which is sometimes derogatorily labeled a "homunculus" (Monsell & Driver, 2000). We will briefly address the two distinct concerns, proposed by Monsell and Driver (2000), raised in the debate about the control homunculus.

The first concern put forward by the control homunculus critique is that it often remains unclear how exactly this executive control is implemented in neural tissue. The control homunculus argument was levied against theories that posited the existence of a control system that could solve the hard problems for which cognitive scientists had no mechanism or concrete idea for how to implement them. In contrast, our model provides clear neurobiological mechanisms and circuitry for some control operations that the brain might be carrying out. Though we also propose potential future control mechanisms, all of our primary claims are supported by specific mechanisms that are currently implemented in the model, and it should thus be entirely clear how our current version of executive control is implemented.

The second concern is that it is often questioned whether the idea of a centralized, top-down control system can be reconciled with the evidence that the brain is a highly distributed system. On the one hand it seems the controlled decisions we make are discrete and unified, e.g., about what motor action to perform or what to pay attention to, and thus seemingly require a single central arbiter. On the other hand, the brain is a distributed system where there does not seem to be a unitary central command: multiple brain areas have been implicated in control processes, such as the prefrontal cortex, the anterior cingulate cortex, the basal ganglia, the thalamus, and the posterior parietal cortex (Menon & D'Esposito, 2022; Uddin et al., 2019). These areas do not act in a vacuum and instead interact with and are co-dependent on yet more brain areas, such as the sensory and motor cortices, the hippocampus, the midbrain, and the amygdala (Pessoa, 2018; Yavas et al., 2019). With our model we show that there is a way out of this paradox: control in our model has a dedicated top-down control structure that is distinct from the semantic network, abstracting away the bottom-up semantic details and enabling efficient repurposing of "syntactic" control operations. However, the control system is clearly not solving the problems by itself, since it is recurrently interacting with the semantic network. This recurrent interaction entails that while there is a division of labour between representing semantic information (bottom-up) and making more abstract decisions about the control of (cognitive) behaviour (top-down), this process itself is tightly coupled. In the bottom-up direction, the type of control exerted is strongly dependent on bottom-up input coming from the semantic network. For instance, if the bottom-up input is novel or ambiguous, the control system will decide to allocate a new unbound memory neuron in the bundle memory system. In the top-down direction, the control system can decide to inhibit the bundle retrieval process, thereby guiding if and when bottom-up information is processed, stored and retrieved. Thus, the control system does to some extent have centralized top-down control over the semantic network, yet their recurrent interactions imply that the two subsystems ultimately form a single cohesive system.

Moreover, the control system itself should not be seen as purely centralized either. It may appear from our depictions of the RCBM model that the control system is implemented in only four single neurons. To be clear, we do not believe these control states to literally be represented by single neurons in the brain, but rather that they are states represented by a distributed network of neurons. But more importantly, the RCBM control system is less centralized than it appears because the four neurons cannot exert control by themselves: the detection and resolution of control states is instead largely implemented in (synaptic connections to/from) the WTAM and MAM pools instead. The WTAM pool automatically assigns novel bundles to unused neurons through a combination of background noise and lateral inhibition, in tandem with the novelty and existing control neurons. The MAM pool detects ambiguity by boosting the activity of similar memories through synchrony-based facilitating synapses, which together with the ambiguity control neuron allows the model to suppress memory retrieval until disambiguating information is provided. In conclusion, the control system in our model is not a homunculus that solves all hard cognitive problems by itself. Instead, to be able to exert control, it depends on and interacts extensively with both the semantic network as well as the WTAM and MAM pools.

RCBM exemplifies how dynamical systems can be symbolic

In part due to the critique of the control homunculus, there has been a shift in cognitive neuroscience towards dynamical systems models of the brain. These models assume that the brain can be understood as a distributed dynamical system that operates over state spaces, and that the computations the brain performs can be seen as trajectories over these state spaces—rather than as a symbolic system with a centralized top-down control system (Buonomano & Maass, 2009; A. Clark, 1996; van Gelder, 1995, 1998). On this view, attractors represent stable states in the state space that the brain can settle into, and the transitions between these states are governed by the dynamics of the system. Bifurcation theory

can then be used to understand how the system can transition between different attractors, thereby controlling the behaviour of the system. These dynamical system models have been used to explain several cognitive phenomena, such as decision-making, semantics, and (verbal) working memory (Fitz et al., 2020; Spivey & Dale, 2006; Wang, 2002). Hopfield networks are perhaps the most famous example of a dynamical system model that can store memories as distinct attractors and retrieve them by moving back into the attractor state of the to-be-retrieved memory (Hopfield, 1982). Yet, so far it has remained difficult to reconcile dynamical system models with the symbolic nature of cognition.

Still, in contrast to what may be believed (van Gelder, 1995), a more abstract, symbolic interpretation of the brain is not incompatible with a dynamical systems view. Our model—though we can abstractly describe it as a discrete, symbolic computer that performs controlled read and write operations over states in (bundle) memory—at the same time just *is* a continuous dynamical system, as we have implemented it as a system of coupled differential equations. If we view our model as a single dynamical system, invariant to different input sequences (which we would not advise), then we may calculate a lower bound of the number of attractor states as follows: first of all, each possible combination of attributes (2^N) can be bound to each of the available bundles (M), for a total of $2^{(N\times M)}$ possible memory states. Additionally, for each of these memory states, one of the M memories can be active, or the entire system can be inactive (i.e., all neurons are silent). Under these assumptions, the total number of possible attractor states in our model is given by $(M+1)\times 2^{M\times N}$. For $M=6$ memory neurons and $N=28$ semantic attributes, this yields approximately 10^{51} possible attractor states. [1] In addition, one may expect many more attractor states still if we also consider the control system and, in particular, the MAM pool. This exemplifies the compositional properties of the symbolic subsystem of bundle memory and shows its effectiveness as a memory system (Sommers et al., n.d.). However, the more interesting point is not that there *are* many attractor states, but rather that transitions between these states can systematically and usefully be controlled. For instance, when the system is presented with an attribute of an existing bundle, the activation of the correct bundle memory is part of the attractor state: settling back into that attractor state then causes the retrieval of other bound features. When instead the system is presented with a novel set of features, the system detects an absence of memory attractor states (=novelty detection), which in turn causes the system to form a new attractor state that stores the new features. While viewing the RCBM model as a dynamical system can be enlightening, we believe that an abstract description helps to provide a more understandable explanation of the transitions between all of these different states, as it allows us to describe the model in terms of cognitive operations such as controlled memory storage and retrieval.

Hierarchy

While RCBM is capable of flat compositionality, it still lacks the capacity for hierarchical compositionality. This means that stereotypical symbolic operations like relational binding are still difficult for the conjunctive binding mechanism of the current model. Relational binding is the process of combining two or more elements into a single structured representation (Fodor & Pylyshyn, 1988; Marcus, 2001). This is especially important for sentences like "A man bit his dog", where the relation is not commutative, meaning that the order of words matters: in this case it signals who does what to whom. Previous models like LISA/DORA deal with relational binding by implementing two-place predicates as combinations of one-place predicates or roles, e.g., `Bites()` and `IsBitten()`. They then dynamically bind the fillers to their roles, e.g., `Bites(a man)` and `IsBitten(his dog)` (Hummel & Holyoak, 2003; Hummel & Holyoak, 1997; Seuren, 2009; Doumas et al., 2022) We can borrow this idea of using one-place predicates to "flatten" the relational binding problem into multiple bundles to allow storing structured representations. However, this would require binding between the bundles, which is not possible in the current model. We could potentially allow facilitating binding between WTAM neurons and MAM neurons, which would allow us to bind one bundle as a attribute to another bundle. This hierarchy of bundles would then be able to resemble the hierarchical structure found in linguistic tree structures or the graphs used to represent visual scenes. Without this or a similar capacity for hierarchical compositionality, our RCBM model will struggle to understand the many sentences and visual scenes containing hierarchical, non-commutative relations. The main future challenge will be to extend the control network so that it can orchestrate these hierarchical bindings.

Previous work

Existing theories and models already employ aspects of Bundle Memory (Green & Quilty-Dunn, 2021; Kriete et al., 2013; Lades et al., 1993; Meeter & Murre, 2005; Sanger et al., 2020). However, bundle memory-like models are spread across the vast literature and often do not solve the same behavioural problems. For instance, we build on work by Manohar et al. (2019) who has similarly used fast hebbian learning between attributes and conjunctive (=bundle) neurons to model working memory. The Dynamic Link Architecture is another example of a Bundle Memory-like model that implements dynamic on-the-fly representations based on short

[1] For more realistic values of M and N, the number of possible states quickly becomes astronomical. For instance, if we assume $M=100$ and $N=10000$, the number of possible states is on the order of 10^{300000}.

term synaptic plasticity (Lades et al., 1993; von der Malsburg, 1994). We build upon these existing bundle memory-like models by adding a memory control system, which enables more complex cognitive operations on these symbols than previously possible.

In addition to bundle memory-like theories and models, there are other connectionist/symbolic-hybrid models that attempt to solve the same problems through different mechanisms than Bundle Memory. Vector addition can potentially accomplish dynamic binding (Hummel, 2011), even if it requires a very specific and computationally inefficient representational format that inherently struggles with hierarchical compositionality (Fodor & Pylyshyn, 1988; Marcus, 2001). Vector symbolic architectures are another way to bind the vector activations of different objects together in various ways (Gayler, 2004; Padilla & Mcdonnell, 2014; Plate, 1995; Schlegel et al., 2022; Smolensky, 1990). These architectures are currently the state of the art in relational binding. One such model, Spaun, has used the idea of semantic pointers to solve the binding problem (Eliasmith, 2013; Stewart et al., 2012). Semantic pointers are reduced representations with little to no semantic content themselves that point to the detailed semantic content in the semantic network (Hinton, 1990). Another type of model, DORA/LISA, uses temporal synchrony, rather than synaptic binding, to dynamically bind roles to fillers (Doumas et al., 2008, 2022). At a coarse-grain level, these models are similar to Bundle Memory, though differ in the details of how they implement the binding operation — sometimes yielding different behavioural predictions. Not every such hybrid model has a neurobiologically plausible implementation, however: for instance, the binding operations of vector-symbolic architectures are typically not considered biologically plausible. In some sense, one contribution of bundle memory is that it provides a neurobiologically plausible implementation of these binding mechanisms.

In addition to these computational models, there are also several high-level theories that have proposed similar ideas. Object files are an existing bundle memory-like idea from the visual search literature (Green & Quilty-Dunn, 2021; Kahneman et al., 1992; Pylyshyn, 1989). Similarly, in the hippocampus literature, indexing theory speaks of binding between contentless indexes that point to information in semantic memory as a popular theory for episodic memory in the hippocampus (Teyler & DiScenna, 1986; Teyler & Rudy, 2007). Indexing theory has recently also been extended to model working memory in PFC (Fiebig et al., 2020). Although the ideas of object files, indirection, pointers or indexes have a long tradition in cognitive science (Kahneman et al., 1992; Pylyshyn, 1989), for a long time it has been largely forgotten or ignored in cognitive (computational) neuroscience. More recent literature has re-emphasised the need for these ideas in the domains of vision, language and working memory (Awh & Vogel, 2025; Green & Quilty-Dunn, 2021; Quilty-Dunn et al., 2022). We view the fact that this same conclusion is being drawn in a broad range of different functional behaviours as evidence that bundle-like memory models are a convergence point for many different cognitive functions, and thus that the brain is likely to implement a hybrid symbolic/connectionist system of this kind.

Emerging symbolic computation and LLMs

Connectionists have long argued that symbolic behaviour can emerge without being put in place by hand (McClelland et al., 1986). Transformer-based LLMs have finally demonstrated this: research on LLMs has shown that they are able to perform reasonably well on a wide range of cognitive tasks, including reasoning, problem solving, and even some forms of creativity (Brown et al., 2020; Wei et al., 2022). Moreover, they exhibit some aspects of symbolic computation like variable binding by using the residual stream as addressable memory (Feng & Steinhardt, 2024; Wu et al., 2025). Emergence, here, may refer to the fact that a model is able to perform these tasks without being explicitly programmed to do so, but rather by learning, e.g., from large amounts of data. Alternatively, "emergence" may mean that a higher-level property of a model arises from the interactions between its lower-level components, which may not exhibit these properties on their own. LLMs seem to exhibit both of these types of emergence. Our model only has the latter: the symbolic operations emerge entirely from the connectionist wiring of the neural network, but are not the result of a learning process. However, our model is not intended to be capable of the former type of emergence: it is a model of how these operations can be performed in a fully developed neural network, rather than a model of how these operations can be learned. As such, we do not make any (strong) claims about the learning process of our model, and leave the question of whether symbolic computation emerges through evolution or through developmental processes open for future work.

It should also be noted that the symbolic operations in connectionist models like LLMs do not emerge out of nowhere, but are at least in part a result of the input data used to train the model. Since language consists of discrete tokens/symbols, and has syntactic, compositional structure, these models must learn symbolic operations in order to be able to correctly process language. In that sense, all models that correctly process language must, by definition, be able to perform symbolic operations of some kind. This implies that at least some symbolic operations are multiple realizable. However, architecture does matter, and there is a reason why transformers are currently better at language processing than previous language model architectures like LSTMs or RNNs (Sorodoc et al., 2020; Vaswani et al., 2017). Unlike these other architectures, LLMs have discrete attention heads that can each attend to different parts of the input, which may allow them to function like bundles. For example, words that refer to some

earlier entity in the text often end up in the same attention head as the entity they refer to (K. Clark et al., 2019; Pandit & Hou, 2021). It could be that the ability to bundle these words together through attention is what allows LLMs to perform (certain) symbolic operations.

Another potential advantage of transformer-based LLMs over previous language models is that they can process the entire input at once, rather than sequentially. Consequently, they require no short-term memory, as all words in the context are simultaneously available to them as input. This computational advantage, however, translates into a disadvantage when viewed as a model of the brain, as we know that the brain lacks this advantage: we cannot read entire book chapters at once, but rather have to read them (more or less) word for word. This means that we require a way of incrementally parsing and building up representations of the input to then store these representations in memory and later retrieve them when needed Gernsbacher, 1990; Johnson-Laird, 1983; McElree et al., 2003, which is exactly the type of problem that Bundle Memory is designed to solve. Aside from the input, the learning trajectory of LLMs is also unlike that of humans: we do not require repeated exposure to all the information on the internet in order to learn language. As such, though they may provide an alternative way of performing symbolic operations, LLMs are unlikely to be a good model of how the brain performs these operations.

Conclusion

In this work, we have presented the Rate-Coding Bundle Memory (RCBM) model as a neurobiologically plausible instantiation of the Symbolic Subsystem Hypothesis. Our results demonstrate that a hybrid connectionist-symbolic architecture, grounded in neural mechanisms, can solve a wide range of cognitive problems that have traditionally been challenging for purely connectionist or symbolic models alone. By embedding a symbolic subsystem within a fundamentally connectionist framework, RCBM bridges the gap between the brain's ability to represent continuous, graded information and its capacity for discrete, symbolic operations.

Specifically, we showed that the RCBM model can perform one-shot learning, resolve ambiguity, incrementally integrate information, and manage multiple memory traces—key aspects of the S3R problem set. Lesion studies further revealed that these capabilities depend critically on the interplay between memory and control subsystems, highlighting the necessity of specialized mechanisms for novelty, ambiguity, and memory management. Importantly, all these operations emerge from neurobiologically plausible processes such as Hebbian learning, lateral inhibition, and facilitating synapses, without recourse to abstract or biologically implausible computations.

These findings support the central claim of the Symbolic Subsystem Hypothesis: that the brain implements a symbolic system within its connectionist substrate, enabling the emergence of discrete, manipulable symbols from continuous neural dynamics. This symbolic subsystem allows the brain to flexibly recombine, retain, and resolve information, supporting the structured, compositional reasoning that characterizes human cognition. At the same time, the underlying connectionist architecture ensures that meaning remains grounded, and robust to noise—reflecting the graded, embodied nature of our mental representations. Thus, the scientific debate between proponents of discrete symbols and proponents of continuous, graded representational spaces need not be decided in favour of a single camp: the brain most likely uses both.

References

Abraham, W. C. (2008). Metaplasticity: Tuning synapses and networks for plasticity [Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *9*(5), 387–387. https://doi.org/10.1038/nrn2356

Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences*, *10*(1), 25–61. https://doi.org/10.1016/0025-5564(71)90051-4

Amer, T., & Davachi, L. (2023). Extra-hippocampal contributions to pattern separation. *eLife*, *12*, e82250. https://doi.org/10.7554/eLife.82250

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, M. C., Bunce, J. G., & Barbas, H. (2016). Prefrontal–hippocampal pathways underlying inhibitory control over memory. *Neurobiology of Learning and Memory*, *134*, 145–161. https://doi.org/10.1016/j.nlm.2015.11.008

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In *Memory* (pp. 237–313). Academic Press. https://doi.org/10.1016/B978-012102570-0/50010-0

Aston-Jones, G., & Cohen, J. D. (2005). AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance [Publisher: Annual Reviews]. *Annual Review of Neuroscience*, *28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, *2*(4), 89–195.

Avery, M. C., & Krichmar, J. L. (2017). Neuromodulatory Systems and Their Interactions: A Review of Models, Theories, and Experiments. *Frontiers in Neural Circuits*, *11*, 108. https://doi.org/10.3389/fncir.2017.00108

Awh, E., & Vogel, E. K. (2025). Working memory needs pointers. *Trends in Cognitive Sciences*, *29*(3), 230–241. https://doi.org/10.1016/j.tics.2024.12.006

Baddeley, A. (2003). Working memory: Looking back and looking forward [Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136–R140. https://doi.org/10.1016/j.cub.2009.12.014

Badre, D., & Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia*, *45*(13), 2883–2901. https://doi.org/10.1016/j.neuropsychologia.2007.06.015

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400 [ISBN: 0169-0965]. *Language and Cognitive Processes*, *26*(9), 1338–1367. https://doi.org/10/bx8w4g

Barry, R. J., Steiner, G. Z., De Blasio, F. M., Fogarty, J. S., Karamacoska, D., & MacDonald, B. (2020). Components in the P300: Don't forget the Novelty P3! *Psychophysiology*, *57*(7), e13371. https://doi.org/10.1111/psyp.13371

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1435), 1177–1187. https://doi.org/10.1098/rstb.2003.1319

Barsalou, L. W. (2008a). Grounded Cognition [Publisher: Annual Reviews]. *Annual Review of Psychology*, *59*(Volume 59, 2008), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2008b). Situating concepts. *Cambridge handbook of situated cognition, ed. P. Robbins & M. Aydede*, 236–63.

Bazzari, A. H., & Parri, H. R. (2019). Neuromodulators and Long-Term Synaptic Plasticity in Learning and Memory: A Steered-Glutamatergic Perspective [Number: 11 Publisher: Multidisciplinary Digital Publishing Institute]. *Brain Sciences*, *9*(11), 300. https://doi.org/10.3390/brainsci9110300

Becke, A., Notger Müller, Vellage, A., Schoenfeld, M. A., & Hopf, J.-M. (2015). Neural sources of visual working memory maintenance in human parietal and ventral extrastriate visual cortex. *NeuroImage*, *110*, 78–86. https://doi.org/10.1016/j.neuroimage.2015.01.059

Beeler, J. A., & Dreyer, J. K. (2019). Synchronicity: The Role of Midbrain Dopamine in Whole-Brain Coordination [Publisher: Society for Neuroscience Section: Commentary]. *eNeuro*, *6*(2). https://doi.org/10.1523/ENEURO.0345-18.2019

Bi, G.-q., & Poo, M.-m. (1998). Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type [Publisher: Society for Neuroscience Section: ARTICLE]. *Journal of Neuroscience*, *18*(24), 10464–10472. https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory [arXiv: NIHMS150003 ISBN: 2122633255]. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10/djtc3r

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30months [ISBN: 0010-0277]. *Cognition*. https://doi.org/10.1016/j.cognition.2012.08.008

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020, July). Language Models are Few-Shot Learners [arXiv:2005.14165 [cs]]. https://doi.org/10.48550/arXiv.2005.14165

Buhry, L., Azizi, A. H., & Cheng, S. (2011). Reactivation, Replay, and Preplay: How It Might All Fit Together. *Neural Plasticity*, *2011*, 203462. https://doi.org/10.1155/2011/203462

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks [Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *10*(2), 113–125. https://doi.org/10.1038/nrn2558

Burgess, J., Lloyd, J. R., & Ghahramani, Z. (2017, July). One-Shot Learning in Discriminative Neural Networks [arXiv:1707.05562 [cs, stat]]. https://doi.org/10.48550/arXiv.1707.05562

Buzsáki, G., & Tingley, D. (2018). Space and Time: The Hippocampus as a Sequence Generator. *Trends in Cognitive Sciences*, *22*(10), 853–869. https://doi.org/10.1016/j.tics.2018.07.006

Carey, S., & Bartlett, E. (1978, August). *Acquiring a Single New Word* (tech. rep.) (ERIC Number: ED198703).

Cayco-Gajic, N. A., & Silver, R. A. (2019). Re-evaluating Circuit Mechanisms Underlying Pattern Separation. *Neuron*, *101*(4), 584–602. https://doi.org/10.1016/j.neuron.2019.01.044

Cazé, R. D., Humphries, M., & Gutkin, B. (2013). Passive Dendrites Enable Single Neurons to Compute Linearly Non-separable Functions [Publisher: Public Library of Science]. *PLOS Computational Biology*, *9*(2), e1002867. https://doi.org/10.1371/journal.pcbi.1002867

Chomsky, N. (2006). *Language and mind*. Cambridge University Press.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The distributed nature of working memory [Publisher: Elsevier]. *Trends in cognitive sciences*, *21*(2), 111–124.

Clark, A. (1996, November). *Being There: Putting Brain, Body, and World Together Again*. The MIT Press. https://doi.org/10.7551/mitpress/1552.001.0001

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019, August). What Does BERT Look at? An Analysis of BERT's Attention. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 276–286). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4828

Cools, R., & D'Esposito, M. (2011). Inverted-U–Shaped Dopamine Actions on Human Working Memory and Cognitive Control. *Biological Psychiatry*, *69*(12), e113–e125. https://doi.org/10.1016/j.biopsych.2011.03.028

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/S1364-6613(03)00197-9

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Diamantaki, M., Frey, M., Berens, P., Preston-Ferrer, P., & Burgalossi, A. (2016). Sparse activity of identified dentate granule cells during spatial exploration (K. Svoboda, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *5*, e20252. https://doi.org/10.7554/eLife.20252

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme [Number: 7 Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *24*(7), 431–450. https://doi.org/10/gr98fb

Douchamps, V., Jeewajee, A., Blundell, P., Burgess, N., & Lever, C. (2013). Evidence for Encoding versus Retrieval Scheduling in the Hippocampus by Theta Phase and Acetylcholine [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *33*(20), 8689–8704. https://doi.org/10.1523/JNEUROSCI.4483-12.2013

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

Doumas, L. A. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). *Relation learning in a neurocomputational architecture supports cross-domain transfer* (tech. rep.). https://arxiv.org/abs/1806.01709

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory [Number: 11 Publisher: Nature Publishing Group]. *Nature Neuroscience*, *3*(11), 1184–1191. https://doi.org/10.1038/81460

Duszkiewicz, A. J., McNamara, C. G., Takeuchi, T., & Genzel, L. (2019). Novelty and Dopaminergic Modulation of Memory Persistence: A Tale of Two Systems [Publisher: Trends Neurosci]. *Trends in Neurosciences*, *42*(2), 102–114. https://doi.org/10.1016/j.tins.2018.10.002

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Espinoza, C., Guzman, S. J., Zhang, X., & Jonas, P. (2018). Parvalbumin+ interneurons obey unique connectivity rules and establish a powerful lateral-inhibition microcircuit in dentate gyrus [Publisher: Nature Publishing Group]. *Nature Communications*, *9*(1), 4605. https://doi.org/10.1038/s41467-018-06899-3

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories [Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611. https://doi.org/10.1109/TPAMI.2006.79

Feldman, J. (2013). The neural binding problem(s). *Cognitive Neurodynamics*, *7*(1), 1–11. https://doi.org/10.1007/s11571-012-9219-8

Feng, J., & Steinhardt, J. (2024, May). How do Language Models Bind Entities in Context? [arXiv:2310.17191 [cs]]. https://doi.org/10.48550/arXiv.2310.17191

Fenton, A. A. (2024). Remapping revisited: How the hippocampus represents different spaces [Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *25*(6), 428–448. https://doi.org/10.1038/s41583-024-00817-x

Fiebig, F., Herman, P., & Lansner, A. (2020). An Indexing Theory for Working Memory Based on Fast Hebbian Plasticity. *eneuro*, *7*(2), ENEURO.0374–19.2020. https://doi.org/10.1523/ENEURO.0374-19.2020

Fiebig, F., & Lansner, A. (2017). A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *The Journal of Neuroscience*, *37*(1), 83–96. https://doi.org/10.1523/JNEUROSCI.1989-16.2016

Fitz, H., Uhlmann, M., van den Broek, D., Duarte, R., Hagoort, P., & Petersson, K. M. (2020). Neuronal spike-rate

adaptation supports working memory in language processing. *Proceedings of the National Academy of Sciences*, *117*(34), 20881–20889. https://doi.org/10.1073/pnas.2000222117

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.

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

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12), 3. https://doi.org/10.1167/11.12.3

Frankland, S. M., & Greene, J. D. (2019). Language of Thought. https://doi.org/10.1146/annurev-psych-122216

Fredes, F., & Shigemoto, R. (2021). The role of hippocampal mossy cells in novelty detection. *Neurobiology of Learning and Memory*, *183*, 107486. https://doi.org/10.1016/j.nlm.2021.107486

Friedman, D., & Johnson, R. (2000). Event-related potential (ERP) studies of memory encoding and retrieval: A selective review, 23.

Fuster, J. M., & Alexander, G. E. (1971). Neuron Activity Related to Short-Term Memory [Publisher: American Association for the Advancement of Science]. *Science*, *173*(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

Gallant, S. I., & Okaywe, T. W. (2013). Representing Objects, Relations, and Sequences. *Neural Computation*, *25*(8), 2038–2078. https://doi.org/10.1162/NECO_a_00467

Gallistel, C. R., & King, A. P. (2009, April). *Memory and the Computational Brain*. Wiley-Blackwell. https://doi.org/10.1002/9781444310498

Garrod, S., Sanford, A., Milward, D., & Sturt, P. (1995). Incrementality in discourse understanding [Publisher: Lawrence Erlbaum Associates Mahwah]. *Incremental Interpretation*, *11*, 99–122.

Gayler, R. W. (2004). *Vector Symbolic Architectures Answer Jackendoff's Challenges for Cognitive Neuroscience* (tech. rep.).

Geiger, L. S., Moessnang, C., Schäfer, A., Zang, Z., Zangl, M., Cao, H., van Raalten, T. R., Meyer-Lindenberg, A., & Tost, H. (2018). Novelty modulates human striatal activation and prefrontal–striatal effective connectivity during working memory encoding. *Brain Structure and Function*, *223*(7), 3121–3132. https://doi.org/10.1007/s00429-018-1679-0

Gernsbacher, M. A. (1990). THE PROCESS OF LAYING A FOUNDATION [Publisher: Earlbaum].

Govindarajan, A., Israely, I., Huang, S.-Y., & Tonegawa, S. (2011). The dendritic branch is the preferred integrative unit for protein synthesis-dependent LTP. *Neuron*, *69*(1), 132–146. https://doi.org/10.1016/j.neuron.2010.12.008

Green, E. J., & Quilty-Dunn, J. (2021). What Is an Object File? *The British Journal for the Philosophy of Science*, *72*(3), 665–699. https://doi.org/10.1093/bjps/axx055

Greff, K., van Steenkiste, S., & Schmidhuber, J. (2020, December). On the Binding Problem in Artificial Neural Networks [arXiv:2012.05208 [cs]]. Retrieved September 13, 2022, from http://arxiv.org/abs/2012.05208

Hadley, R. F. (2009). The Problem of Rapid Variable Creation. *Neural Computation*, *21*(2), 510–532. https://doi.org/10.1162/neco.2008.07-07-572

Hasselmo, M. E., & Giocomo, L. M. (2006). Cholinergic modulation of cortical function. *Journal of molecular neuroscience: MN*, *30*(1-2), 133–135. https://doi.org/10.1385/JMN:30:1:133

Hasselmo, M. E., Bodelón, C., & Wyble, B. P. (2002). A Proposed Function for Hippocampal Theta Rhythm: Separate Phases of Encoding and Retrieval Enhance Reversal of Prior Learning. *Neural Computation*, *14*(4), 793–817. https://doi.org/10/d7p8kw

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology press.

Herrmann, M., Rotte, M., Grubich, C., Ebert, A. D., Schiltz, K., Münte, T. F., & Heinze, H. J. (2001). Control of semantic interference in episodic memory retrieval is associated with an anterior cingulate-prefrontal activation pattern. *Human Brain Mapping*, *13*(2), 94–103. https://doi.org/10.1002/hbm.1027

Hinton, G. E. (1984). Distributed representations [Publisher: Carnegie Mellon University].

Hinton, G. E. (1990). Mapping Part-Whole Hierarchies into Connectionist Networks, 29.

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. https://doi.org/10.1073/pnas.79.8.2554

Huang, Y.-Y., Colino, A., Selig, D. K., & Malenka, R. C. (1992). The Influence of Prior Synaptic Activity on the Induction of Long-Term Potentiation [Publisher: American Association for the Advancement of Science]. *Science*, *255*(5045), 730–733. https://doi.org/10.1126/science.1346729

Hume, D., & Mossner, E. C. (2000). *A treatise of human nature* (Vol. 26) [Publication Title: A Treatise of Human Nature ISSN: 03197336]. Oxford University Press. https://doi.org/10.2307/2216614

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., & Holyoak, K. J. (2003). A Symbolic-Connectionist Theory of Relational Inference and Generalization. *Psychological Review*, *110*(2), 220–264. https://doi.org/10.1037/0033-295X.110.2.220

Hummel, J. E., Holyoak, K. J., Green, C., Doumas, L. A. A., Devnich, D., & Kalar, D. J. (2004). A Solution to the Binding Problem for Compositional Connectionism, 4.

Izhikevich, E. M., Gally, J. A., & Edelman, G. M. (2004). Spike-timing Dynamics of Neuronal Groups. *Cerebral Cortex*, *14*(8), 933–944. https://doi.org/10.1093/cercor/bhh053

Jackendoff, R. (2003, September). *Foundations of Language* (Reprint) [OCLC: 705605237]. Oxford University Press, USA.

Jeffery, K. J. (2011). Place Cells, Grid Cells, Attractors, and Remapping [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2011/182602]. *Neural Plasticity*, *2011*(1), 182602. https://doi.org/10.1155/2011/182602

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The Mind and Brain of Short-Term Memory. *Annual Review of Psychology*, *59*(1), 193–224. https://doi.org/10.1146/annurev.psych.59.103006.093615

Kafkas, A., & Montaldi, D. (2018). How do memory systems detect and respond to novelty? *Neuroscience Letters*, *680*, 60–68. https://doi.org/10.1016/j.neulet.2018.01.053

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219. https://doi.org/10.1016/0010-0285(92)90007-O

Kazanina, N., & Poeppel, D. (2023). The neural ingredients for a language of thought are available [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *27*(11), 996–1007. https://doi.org/10.1016/j.tics.2023.07.012

Kempter, R., Gerstner, W., & van Hemmen, J. L. (1999). Hebbian learning and spiking neurons [Publisher: American Physical Society]. *Physical Review E*, *59*(4), 4498–4514. https://doi.org/10.1103/PhysRevE.59.4498

Kesner, R. P., Hunsaker, M. R., & Warthen, M. W. (2008). The CA3 subregion of the hippocampus is critical for episodic memory processing by means of relational encoding in rats [Place: US Publisher: American Psychological Association]. *Behavioral Neuroscience*, *122*(6), 1217–1225. https://doi.org/10.1037/a0013592

Kishiyama, M. M., Yonelinas, A. P., & Knight, R. T. (2009). Novelty Enhancements in Memory Are Dependent on Lateral Prefrontal Cortex [Publisher: Society for Neuroscience Section: Brief Communications]. *Journal of Neuroscience*, *29*(25), 8114–8118. https://doi.org/10.1523/JNEUROSCI.5507-08.2009

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *110*(41), 16390–16395. https://doi.org/10.1073/pnas.1303547110

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated [arXiv: 1505.03711 ISBN: 1879-307X (Electronic)\r1364-6613 (Linking)]. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2016.05.004

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., Liu, Y., & Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*(4), 454–469. https://doi.org/10.1016/j.neuron.2022.12.028

Kutas, M., & Federmeier, K. D. (2009). N400. *Scholarpedia*, *4*(10), 7790. https://doi.org/10.4249/scholarpedia.7790

Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42*(3), 300–311. https://doi.org/10.1109/12.210173

Lake, B., Lee, C.-y., Glass, J., & Tenenbaum, J. (2014). One-shot learning of generative speech concepts [Issue: 36]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*. Retrieved November 12, 2023, from https://escholarship.org/content/qt3xf2n3vc/qt3xf2n3vc.pdf

Lake, B. M., & Baroni, M. (2018, June). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks [arXiv:1711.00350 [cs]]. https://doi.org/10.48550/arXiv.1711.00350

Lansner, A., Fiebig, F., & Herman, P. (2023). Fast Hebbian plasticity and working memory. *Current Opinion in*

Neurobiology, *83*, 102809. https://doi.org/10.1016/j.conb.2023.102809

Legenstein, R., & Maass, W. (2011). Branch-specific plasticity enables self-organization of nonlinear computation in single neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(30), 10787–10802. https://doi.org/10.1523/JNEUROSCI.5684-10.2011

Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *6*(7), 299–305. https://doi.org/10.1016/S1364-6613(02)01923-X

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, *10*(10), 447–54. https://doi.org/10.1016/j.tics.2006.08.007

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron*, *46*(5), 703–713. https://doi.org/10.1016/j.neuron.2005.05.002

Lisman, J. E., & Jensen, O. (2013). The theta-gamma neural code [Publisher: Elsevier]. *Neuron*, *77*(6), 1002–1016.

Liu, C., Goel, P., & Kaeser, P. S. (2021). Spatial and temporal scales of dopamine transmission [Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *22*(6), 345–358. https://doi.org/10.1038/s41583-021-00455-7

Liu, X., Ramirez, S., Pang, P. T., Puryear, C. B., Govindarajan, A., Deisseroth, K., & Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall [Publisher: Nature Publishing Group]. *Nature*, *484*(7394), 381–385. https://doi.org/10.1038/nature11028

Loula, J., Baroni, M., & Lake, B. M. (2018, July). Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks [arXiv:1807.07545 [cs]]. Retrieved September 13, 2023, from http://arxiv.org/abs/1807.07545

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions [Publisher: Nature Publishing Group UK London]. *Nature*, *390*(6657), 279–281.

Lundqvist, M., Herman, P., & Miller, E. K. (2018). Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *The Journal of Neuroscience*, *38*(32), 7013–7019. https://doi.org/10.1523/JNEUROSCI.2485-17.2018

Maingret, N., Girardeau, G., Todorova, R., Goutierre, M., & Zugaro, M. (2016). Hippocampo-cortical coupling mediates memory consolidation during sleep [Publisher: Nature Publishing Group]. *Nature Neuroscience*, *19*(7), 959–964. https://doi.org/10.1038/nn.4304

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P., & Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews*, *101*, 1–12. https://doi.org/10.1016/j.neubiorev.2019.03.017

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Marr, D. C. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *176*(1043), 161–234. https://doi.org/10.1098/rspb.1970.0040

Matsuzaki, M., Honkura, N., Ellis-Davies, G. C. R., & Kasai, H. (2004). Structural basis of long-term potentiation in single dendritic spines [Publisher: Nature Publishing Group]. *Nature*, *429*(6993), 761–766. https://doi.org/10.1038/nature02617

Maurer, S. V., & Williams, C. L. (2017). The Cholinergic System Modulates Memory and Hippocampal Plasticity via Its Interactions with Non-Neuronal Cells [Publisher: Frontiers]. *Frontiers in Immunology*, *8*. https://doi.org/10.3389/fimmu.2017.01489

McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291098-1063%281996%296%3A6%3C654%3A%3AAID-HIPO8%3E3.0.CO%3B2-G]. *Hippocampus*, *6*(6), 654–665. https://doi.org/10.1002/(SICI)1098-1063(1996)6:6<654::AID-HIPO8>3.0.CO;2-G

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 [Place: US Publisher: American Psychological Association]. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McClelland, J. L., Rumelhart, D. E., & Group, P. R. (1986, July). *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*. The MIT Press. https://doi.org/10.7551/mitpress/5237.001.0001

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91. https://doi.org/10.1016/S0749-596X(02)00515-6

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, *10*(10), 408–415. https://doi.org/10.1016/0166-2236(87)90011-7

Meeter, M., & Murre, J. (2005). TraceLink: A model of amnesia and consolidation. *Cognitive Neuropsychology*, *22*(5), 559–587. https://doi.org/10.1080/02643290442000194

Menon, V., & D'Esposito, M. (2022). The role of PFC networks in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 90–103. https://doi.org/10.1038/s41386-021-01152-w

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0 [Publisher: Elsevier]. *Neuron*, *100*(2), 463–475.

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

Monsell, S., & Driver, J. (2000). *Control of Cognitive Processes: Attention and Performance XVIII* [Google-Books-ID: kO_baYlSVbwC]. MIT Press.

Morita, K. (2008). Possible Role of Dendritic Compartmentalization in the Spatial Working Memory Circuit [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *28*(30), 7699–7724. https://doi.org/10.1523/JNEUROSCI.0059-08.2008

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic Memory and Beyond: The Hippocampus and Neocortex in Transformation. *Annual Review of Psychology*, *67*(1), 105–134. https://doi.org/10/gdz7rv

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *25*(9), 757–775. https://doi.org/10.1016/j.tics.2021.06.001

Myers, C. E., & Scharfman, H. E. (2009). A Role for Hilar Cells in Pattern Separation in the Dentate Gyrus: A Computational Approach. *Hippocampus*, *19*(4), 321–337. https://doi.org/10.1002/hipo.20516

Naber, P. A., Lopes da Silva, F. H., & Witter, M. P. (2001). Reciprocal connections between the entorhinal cortex and hippocampal fields CA1 and the subiculum are in register with the projections from CA1 to the subiculum [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hipo.1028]. *Hippocampus*, *11*(2), 99–104. https://doi.org/10.1002/hipo.1028

Neunuebel, J. P., & Knierim, J. J. (2012). Spatial Firing Correlates of Physiologically Distinct Cell Types of the Rat Dentate Gyrus [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *32*(11), 3848–3858. https://doi.org/10.1523/JNEUROSCI.6038-11.2012

Nuñez, A., & Buño, W. (2021). The Theta Rhythm of the Hippocampus: From Neuronal and Circuit Mechanisms to Behavior [Publisher: Frontiers]. *Frontiers in Cellular Neuroscience*, *15*. https://doi.org/10.3389/fncel.2021.649262

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. *Current Biology*, *28*(1), R37–R50. https://doi.org/10.1016/j.cub.2017.10.073

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary Learning Systems [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2011.01214.x]. *Cognitive Science*, *38*(6), 1229–1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

Özçete, Ö. D., Banerjee, A., & Kaeser, P. S. (2024). Mechanisms of neuromodulatory volume transmission [Publisher: Nature Publishing Group]. *Molecular Psychiatry*, *29*(11), 3680–3693. https://doi.org/10.1038/s41380-024-02608-3

Padilla, D., & Mcdonnell, M. (2014). A Neurobiologically Plausible Vector Symbolic Architecture. *Proceedings - 2014 IEEE International Conference on Semantic Computing, ICSC 2014*, 242–245. https://doi.org/10.1109/ICSC.2014.40

Palacios-Filardo, J., & Mellor, J. R. (2019). Neuromodulation of hippocampal long-term synaptic plasticity. *Current Opinion in Neurobiology*, *54*, 37–43. https://doi.org/10.1016/j.conb.2018.08.009

Pandit, O., & Hou, Y. (2021, June). Probing for Bridging Inference in Transformer Language Models. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4153–4163). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.327

Park, E., Dvorak, D., & Fenton, A. A. (2011). Ensemble Place Codes in Hippocampus: CA1, CA3, and Dentate Gyrus Place Cells Have Multiple Place Fields in Large Environments [Publisher: Public Library of Science]. *PLOS ONE*, *6*(7), e22349. https://doi.org/10.1371/journal.pone.0022349

Parker, D., Shvartsman, M., & Van Dyke, J. A. (2017). *THE CUE-BASED RETRIEVAL THEORY OF SENTENCE COMPREHENSION: NEW FINDINGS AND NEW CHALLENGES* (tech. rep.).

Patterson, K., & Lambon Ralph, M. A. (2016, January). The Hub-and-Spoke Hypothesis of Semantic Memory. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 765–775). Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00061-4

Pavlides, C., & Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *9*(8), 2907–2918. https://doi.org/10.1523/JNEUROSCI.09-08-02907.1989

Pessoa, L. (2018). Emotion and the Interactive Brain: Insights From Comparative Neuroanatomy and Complex Systems. *Emotion review : journal of the International Society for Research on Emotion*, *10*(3), 204–216. https://doi.org/10.1177/1754073918765675

Petrides, M. (2007). The Orbitofrontal Cortex: Novelty, Deviation from Expectation, and Memory [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1196/annals.1401.035]. *Annals of the New York Academy of Sciences*, *1121*(1), 33–53. https://doi.org/10.1196/annals.1401.035

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Plate, T. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641. https://doi.org/10.1109/72.377968

Poirazi, P., Brannon, T., & Mel, B. W. (2003). Pyramidal Neuron as Two-Layer Neural Network [Publisher: Elsevier]. *Neuron*, *37*(6), 989–999. https://doi.org/10.1016/S0896-6273(03)00149-1

Puebla, G., Martin, A. E., & Doumas, L. A. A. (2021). The relational processing limits of classic and contemporary neural network models of language processing. *Language, Cognition and Neuroscience*, *36*(2), 240–254. https://doi.org/10.1080/23273798.2020.1821906

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97. https://doi.org/10.1016/0010-0277(89)90014-0

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences. *Behavioral and Brain Sciences*, 1–55. https://doi.org/10.1017/S0140525X22002849

Raza, S. A., Albrecht, A., Çalışkan, G., Müller, B., Demiray, Y. E., Ludewig, S., Meis, S., Faber, N., Hartig, R., Schraven, B., Lessmann, V., Schwegler, H., & Stork, O. (2017). HIPP neurons in the dentate gyrus mediate the cholinergic modulation of background context memory salience [Publisher: Nature Publishing Group]. *Nature Communications*, *8*(1), 189. https://doi.org/10.1038/s41467-017-00205-3

Rogers, T. T., & Mcclelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition [Publisher: Wiley-Blackwell Publishing]. *Cognitive Science*, *38*(6), 1024–1077. https://doi.org/10.1111/cogs.12148

Rolls, E. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus [Publisher: Frontiers]. *Frontiers in Systems Neuroscience*, *7*. https://doi.org/10.3389/fnsys.2013.00074

Rolls, E. T., & Treves, A. (2024). A theory of hippocampal function: New developments. *Progress in Neurobiology*, *238*, 102636. https://doi.org/10.1016/j.pneurobio.2024.102636

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Rothschild, G., Eban, E., & Frank, L. M. (2017). A cortical–hippocampal–cortical loop of information processing during memory consolidation [Publisher: Nature Publishing Group]. *Nature Neuroscience*, *20*(2), 251–259. https://doi.org/10.1038/nn.4457

Sabatini, B. L., Oertner, T. G., & Svoboda, K. (2002). The life cycle of Ca(2+) ions in dendritic spines. *Neuron*, *33*(3), 439–452. https://doi.org/10.1016/s0896-6273(02)00573-1

Sandberg, A., Tegnér, J., & Lansner, A. (2003). A working memory model based on fast Hebbian learning. *Network: Computation in Neural Systems*, *14*(4), 789–802. https://doi.org/10.1088/0954-898X_14_4_309

Sanford, A. J., & Garrod, S. C. (2005). Memory-based approaches and beyond [ISBN: 0163-853X]. *Discourse Processes*. https://doi.org/10.1207/s15326950dp3902&3_6

Sanger, T. D., Yamashita, O., & Kawato, M. (2020). Expansion coding and computation in the cerebellum: 50 years after the Marr–Albus codon theory [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1113/JP278745]. *The Journal of Physiology*, *598*(5), 913–928. https://doi.org/10.1113/JP278745

Schacter, D. L., & Buckner, R. L. (1998). Priming and the brain [Publisher: Elsevier]. *Neuron*, *20*(2), 185–195.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schlegel, K., Neubert, P., & Protzel, P. (2022). A comparison of vector symbolic architectures. *Artificial Intelligence Review*, *55*(6), 4523–4555. https://doi.org/10.1007/s10462-021-10110-3

Schmidt, B., Marrone, D. F., & Markus, E. J. (2012). Disambiguating the similar: The dentate gyrus and pattern

separation. *Behavioural Brain Research*, *226*(1), 56–65. https://doi.org/10.1016/j.bbr.2011.08.039

Seuren, P. A. M. (2009, October). *The Logic of Language: Language From Within Volume II*. Oxford University Press.

Siapas, A. G., & Wilson, M. A. (1998). Coordinated Interactions between Hippocampal Ripples and Cortical Spindles during Slow-Wave Sleep [Publisher: Elsevier]. *Neuron*, *21*(5), 1123–1128. https://doi.org/10.1016/S0896-6273(00)80629-7

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

Sommers, R. P., Gils, T. v., Hagoort, P., & Nieuwland, M. S. (n.d.). Bundle Memory: A Computational Model of Reference Comprehension.

Sorodoc, I.-T., Gulordava, K., & Boleda, G. (2020). Probing for Referential Information in Language Models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4177–4189. https://doi.org/10.18653/v1/2020.acl-main.384

Spiegel, C., & Halberda, J. (2011). Rapid fast-mapping abilities in 2-year-olds. *Journal of Experimental Child Psychology*, *109*(1), 132–140. https://doi.org/10.1016/j.jecp.2010.10.013

Spivey, M. J., & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. *Current Directions in Psychological Science*, *15*(5), 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Stewart, T., Choo, F.-X., & Eliasmith, C. (2012). Spaun: A Perception-Cognition-Action Model Using Spiking Neurons. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34). Retrieved November 22, 2023, from https://escholarship.org/uc/item/168466tf

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

Takahashi, S. (2015). Episodic-like memory trace in awake replay of hippocampal place cell activity sequences (U. S. Bhalla, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *4*, e08105. https://doi.org/10.7554/eLife.08105

Tamamaki, N., & Nojyo, Y. (1995). Preservation of topography in the connections between the subiculum, field CA1, and the entorhinal cortex in rats [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.903530306]. *Journal of Comparative Neurology*, *353*(3), 379–390. https://doi.org/10.1002/cne.903530306

Taylor, J. R. (2011). Prototype theory.

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory [Place: US Publisher: American Psychological Association]. *Behavioral Neuroscience*, *100*(2), 147–154. https://doi.org/10.1037/0735-7044.100.2.147

Teyler, T. J., & Rudy, J. W. (2007). The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus*, *17*(12), 1158–1169. https://doi.org/10.1002/hipo.20350

Uddin, L. Q., Yeo, B. T. T., & Spreng, R. N. (2019). Towards a Universal Taxonomy of Macro-scale Functional Human Brain Networks. *Brain Topography*, *32*(6), 926–942. https://doi.org/10.1007/s10548-019-00744-6

van Berkum, J. J., Brown, C. M., & Hagoort, P. (1999). Early Referential Context Effects in Sentence Processing: Evidence from Event-Related Brain Potentials. *Journal of Memory and Language*, *41*(2), 147–182. https://doi.org/10/d2m8cg

van Gelder, T. (1995). What Might Cognition Be, If Not Computation? [ISBN: 9780521831048]. *The Journal of Philosophy*, *92*(7), 345–381. https://doi.org/10/bt9md6

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–628. https://doi.org/10.1017/S0140525X98001733

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Van Berkum, J. J. A., Brown, C. M., Hagoort, P., & Zwitserlood, P. (2003). *Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension* (tech. rep.).

van der Velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–70. https://doi.org/10.1017/S0140525X06009022

van der Velde, F., & de Kamps, M. (2015). The necessity of connection structures in neural models of variable binding. *Cognitive Neurodynamics*, *9*(4), 359–370. https://doi.org/10.1007/s11571-015-9331-7

Varela, F. J., Thompson, E., & Rosch. (2016). *The Embodied Mind*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. Re-

trieved January 2, 2024, from https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu koray, k., & Wierstra, D. (2016). Matching Networks for One Shot Learning. *Advances in Neural Information Processing Systems*, *29*. Retrieved January 2, 2024, from https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html

von der Malsburg, C. (1994). The Correlation Theory of Brain Function [Series Title: Physics of Neural Networks]. In E. Domany, J. L. van Hemmen, K. Schulten, E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of Neural Networks* (pp. 95–119). Springer New York. https://doi.org/10.1007/978-1-4612-4320-5_2

Wang, X.-J. (2002). Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron*, *36*(5), 955–968. https://doi.org/10.1016/S0896-6273(02)01092-9

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: General*, *131*(1), 48–64. https://doi.org/10.1037/0096-3445.131.1.48

Whittington, J. C. R., Dorrell, W., Behrens, T. E. J., Ganguli, S., & El-Gaby, M. (2025). A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps [Publisher: Elsevier]. *Neuron*, *113*(2), 321–333.e6. https://doi.org/10.1016/j.neuron.2024.10.017

Witter, M. P., & Amaral, D. G. (2021). The entorhinal cortex of the monkey: VI. Organization of projections from the hippocampus, subiculum, presubiculum, and parasubiculum [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.24983]. *Journal of Comparative Neurology*, *529*(4), 828–852. https://doi.org/10.1002/cne.24983

Wittgenstein, L., Hacker, P. M. S., & Schulte, J. (2010). *Philosophical Investigations* [Google-Books-ID: XN9yyyhYMDoC]. Wiley & Sons, Incorporated, John.

Wu, Y., Geiger, A., & Millière, R. (2025, May). How Do Transformers Learn Variable Binding in Symbolic Programs? [arXiv:2505.20896 [cs]]. https://doi.org/10.48550/arXiv.2505.20896

Yang, Y., Wang, X.-b., Frerking, M., & Zhou, Q. (2008). Spine Expansion and Stabilization Associated with Long-Term Potentiation. *The Journal of Neuroscience*, *28*(22), 5740–5751. https://doi.org/10.1523/JNEUROSCI.3998-07.2008

Yavas, E., Gonzalez, S., & Fanselow, M. S. (2019). Interactions between the hippocampus, prefrontal cortex, and amygdala support complex learning and memory. *F1000Research*, *8*, F1000 Faculty Rev–1292. https://doi.org/10.12688/f1000research.19317.1

Yuste, R., & Denk, W. (1995). Dendritic spines as basic functional units of neuronal integration. *Nature*, *375*(6533), 682–684. https://doi.org/10.1038/375682a0

Záborszky, L., Gombkoto, P., Varsanyi, P., Gielow, M. R., Poe, G., Role, L. W., Ananth, M., Rajebhosale, P., Talmage, D. A., Hasselmo, M. E., Dannenberg, H., Minces, V. H., & Chiba, A. A. (2018). Specific Basal Forebrain–Cortical Cholinergic Circuits Coordinate Cognitive Operations [Publisher: Society for Neuroscience Section: Symposium and Mini-Symposium]. *Journal of Neuroscience*, *38*(44), 9446–9458. https://doi.org/10.1523/JNEUROSCI.1676-18.2018

Zhang, K., Bromberg-Martin, E. S., Sogukpinar, F., Kocher, K., & Monosov, I. E. (2022). Surprise and recency in novelty detection in the primate brain [Publisher: Elsevier]. *Current Biology*, *32*(10), 2160–2173.e6. https://doi.org/10.1016/j.cub.2022.03.064