# Indirection and symbol-like processing in the prefrontal cortex and basal ganglia

Trenton Kriete[a,1], David C. Noelle[b], Jonathan D. Cohen[c], and Randall C. O'Reilly[a,1]

[a]Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO 80309; [b]Cognitive and Information Sciences, University of California, Merced, CA 95340; and [c]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544

Edited by John R. Anderson, Carnegie Mellon University, Pittsburgh, PA, and approved August 16, 2013 (received for review March 7, 2013)

**The ability to flexibly, rapidly, and accurately perform novel tasks is a hallmark of human behavior. In our everyday lives we are often faced with arbitrary instructions that we must understand and follow, and we are able to do so with remarkable ease. It has frequently been argued that this ability relies on symbol processing, which depends critically on the ability to represent variables and bind them to arbitrary values. Whereas symbol processing is a fundamental feature of all computer systems, it remains a mystery whether and how this ability is carried out by the brain. Here, we provide an example of how the structure and functioning of the prefrontal cortex/basal ganglia working memory system can support variable binding, through a form of indirection (akin to a pointer in computer science). We show how indirection enables the system to flexibly generalize its behavior substantially beyond its direct experience (i.e., systematicity). We argue that this provides a biologically plausible mechanism that approximates a key component of symbol processing, exhibiting both the flexibility, but also some of the limitations, that are associated with this ability in humans.**

generativity | generalization | computational model

One of the most impressive aspects of human cognition is also one of its most enduring mysteries: how it can respond in appropriate ways to novel circumstances. In our everyday lives, we are constantly confronted with the need to make sense of and respond appropriately to new situations. Almost always, the individual constituents of these situations (e.g., the people, places, and/or actions involved) are things with which we have had prior experience, and it is the particular combination that is new. A person may appear in a new context or carry out an action we have never before witnessed them perform, or a word may be used in a novel way within a sentence. Nevertheless, we are able to makes sense of and respond appropriately to such circumstances, drawn from a nearly infinite array of possible combinations, despite having had experience with only a limited number of them. It has frequently been argued that this flexibility, or systematicity, relies on symbol processing, that is, the ability to represent information in the form of abstract variables that can be bound to arbitrary values, as is possible in a symbol system. For example, in trying to understand a sentence, if the constituent parts can be represented as variables, then any possible word can be assigned, or "bound," to each (e.g., in the sentence "I want to desk you," "desk" can be understood as the verb). Such variable binding provides tremendous flexibility and is fundamental to the power of computer systems. However, whether and how this ability is implemented in the brain remains one of the great mysteries of neuroscience. Historically, this ability has been used as a key argument by those advocating for symbolic cognitive models, over "associationist" neural network models (1, 2). In response, some have argued that human symbol processing ability is limited at best and that many behaviors that might be construed as evidence of such an ability can be explained by general-purpose learning algorithms that can infer statistical regularities of the environment (3–5).

Here, we propose a set of neural mechanisms, involving the prefrontal cortex (PFC) and basal ganglia (BG), that support a form of variable binding through the use of indirection (corresponding to the use of a pointer, in computer science terms). We demonstrate that these mechanisms can exhibit the kind of systematicity in processing novel combinations of stimuli of which people are capable, typically attributed to a symbol processing system. However, it does so using standard neural network mechanisms for both learning and processing. As a consequence, its mechanisms for indirection and variable binding are limited. It can only assign pointers to memory locations with which it has had some previous experience, and those locations can only represent information that has been learned to be represented. Also, neural pointers cannot be nested at arbitrary levels of complexity or depth. In this respect, these neural representations fall short of qualifying as symbols in the most general sense. Accordingly, the processing capabilities of this system fall short of the more powerful capabilities of fully general symbol processing systems found in most computers. However, human systematicity has its limits (4, 5). These limits may be explained by the structure of the PFC/BG system, learning based on standard neural network mechanisms, and the distributed representation of information. In other words, although human behavior exhibits some of the characteristics of a classic symbol processing system, the mechanisms upon which this relies may not implement symbols in the usual sense. By understanding the limitations of the mechanisms our behavior is built upon we may shed a light on the limitations of human behavior itself.

In the following, we describe a model in which neurons in one part of the PFC (area A) encode and maintain a pattern of neural activity that represents the location (or address) of information maintained in another part of the PFC (area B). Furthermore, representations in area A can regulate the use of information in area B by way of the BG: Representations in area A project to, and are decoded by, a region of the BG associated with area B, which in turn can regulate efferent projections from area B to more posterior neocortical areas responsible for controlling behavior. Thus, area A of the PFC encodes a pointer to area B and, by way of the BG, can regulate the influence of information stored in area B on behavior. With reasonable assumptions about the connectivity between the PFC and BG, area A can point to a wide range of areas (i.e., not just B), providing considerable flexibility in processing.

This use of indirection to implement variable binding extends a more traditional view of neural representation, in which a given population of neurons is assumed to encode a particular type of information, and different patterns of activity over that population correspond to different possible contents of each type. For example, in the case of sentence processing, there might be separate populations for encoding the various constituents, or

Author contributions: T.K., D.C.N., J.D.C., and R.C.O. designed research; T.K. and D.C.N. performed research; R.C.O. contributed new reagents/analytic tools; T.K. analyzed data; and T.K., D.C.N., J.D.C., and R.C.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: Trenton.Kriete@Colorado.edu or randy.oreilly@colorado.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1303547110/-/DCSupplemental.

roles, within a sentence (e.g., agent, verb, or patient). The pattern of activity within each of these populations would then represent the current value, or filler, of that role. With a sufficient number of different populations (e.g., to represent the different possible roles), each with a sufficient number of neurons (e.g., to represent the possible fillers of each role), rich combinations of information can be expressed (e.g., sentences can be formed). This traditional strategy can support the learning of compositional representations when the range of training experiences spans the space of possible combinations of constituents (6), but three fundamental problems arise as the space of possible representations gets very large. First, this approach risks extensive duplication of resources. A second related, but more important, problem is that this scheme cannot benefit from generalization resulting from similarity of representation without the inclusion of additional mechanisms. This is because the representation learned for fillers in one role are specific to the neural population used to represent that role and isolated from the representation of the same fillers in other roles. For example, any similarity that boys and girls share in the role of agent will not generalize to their roles as patient, because agents and patients are represented over separate populations of neurons. Finally, learning from a limited number of experiences can shape neural representations based on accidental correlations (or anticorrelations) in the sampled set of experiences (7, 8). For example, if "girl" had never been experienced in the role of agent, then the population of neurons for the agent role would have no opportunity to learn to represent the "girl" filler in that role. Because of these problems, some researchers have proposed that compositional representation schemes in the brain are not learned but arise from specialized neural binding circuits that combine role and filler representations using a fixed method, such as the tensor product (9) or circular convolution (10). The resulting representations can seem identical to traditional ones involving the association of roles to isolated populations of neurons, but many of the problems involving generalization are avoided by making the binding operation nonadaptive and insensitive to experience. Avoiding the learning of compositional representations, in this way, makes it difficult to account for ways in which human behavior departs from pure systematicity (4, 5), however. Perhaps more importantly, although fixed binding operations of this kind have been formally implemented in simulated neural circuits (11), there is little anatomical or physiological evidence for such specialized circuitry in the brain.

Introducing a mechanism for indirection can avert the problems described above while still allowing compositional representations to be learned from experience. For example, this can be used to separate the representation of roles and fillers. Role populations can be used to represent pointers (e.g., in the PFC area A of our example above), which in turn can be assigned to reference fillers represented elsewhere (e.g., area B). This allows each filler to be represented only once rather than separately for each role. Furthermore, any role that points to that filler can exploit the similarity that filler shares with other fillers represented in that population, allowing what is learned about the relationship between fillers in one role to generalize to others. This separation between form and content is central to the classic symbol processing model (1). Using the model described below, we demonstrate that under certain architectural assumptions—that are consistent with the known anatomy of the PFC and BG systems—the relationship between representations of pointers and their referents (e.g., between roles and fillers) can be learned. However, although this mechanism exhibits considerable flexibility, its performance is constrained by its learning experiences—a feature that seems consistent with human performance.

We begin by providing a brief review of the neurobiological mechanisms that we propose support variable binding in the brain, followed by a computational model that embodies these properties. We then use the model to simulate a simple sentence processing task that requires variable binding. We demonstrate that, through trial-and-error learning, the model can learn to bind words (fillers) to their appropriate roles in the sentence and, critically, can do so for words that it has never seen in a particular role. We evaluate this by testing the model on several benchmark tests of generalization and by comparing it to others, including neural network architectures that have been used in previous work to model variable binding.

### PFC, BG, and Indirection

Our model focuses on functional specializations within the BG that we have previously proposed implement a dynamic, adaptive gating mechanism for regulating the updating, maintenance, and output of information in the PFC (12–14) (Figs. 1 *A* and *B* and 2 *A* and *B*). Importantly, the model assumes that separate pathways through the BG can independently regulate different subregions of the PFC. This is consistent with the anatomy of both the PFC and BG. Within the PFC, relatively isolated stripe-like patches of neurons have been observed, each of which exhibits dense within-stripe interconnectivity and sparse (largely inhibitory) between-stripe connectivity (15, 16). Furthermore, these stripes project to distinct regions within the BG which, in turn, project back to different stripes within the PFC (17). In previous modeling work, we have shown that this anatomy and physiology can support a system of control in which gating signals from the BG regulate when a given stripe within the PFC will either (*i*) update to encode new information, (*ii*) continue to robustly maintain information in an active state (through sustained neural firing), or (*iii*) output the information it is currently encoding to drive information processing elsewhere in the brain (18). That is, it provides a separation of function that allows different signals to control when and where information is encoded or used (BG) compared with the signals that encode the information content itself (PFC) (14). Furthermore, we have shown that if the output of one PFC stripe controls the BG gating signal that regulates a different stripe, this architecture can support a system of hierarchically nested control (19). This, in turn, can be used to separate the representation of variables and their values, upon which a mechanism for indirection can be built.

Consider the sentence processing example introduced above (and illustrated in Fig. 1). Imagine there are different PFC stripes dedicated to representing the different roles of a sentence (e.g., agent, verb, and patient). In the simplest case, the pattern of activity in a given role-specific stripe would represent the current filler for that role. However, now consider that the pattern of activity within each role-specific stripe, instead of representing any particular filler, represents the address of another PFC stripe that represents that filler (Figs. 1 and 2*C*). There could then be a large number of different such filler stripes, organized in useful ways (e.g., according to semantic relationships among the fillers). The BG system can then be used to update the address information in the role-specific stripes as new sentences are presented and new fillers need to be assigned to a given role, whereas the role-specific stripe, when queried, can signal the BG to trigger the output of information from the filler stripe to which it currently points, permitting a read-out of the current filler for that role. Below, we show not only that such a system can self-organize through learning, but also that, with only a modicum of such learning, it can accurately process a wide range of role-filler combinations to which it has never been exposed.

## Methods

**Generalization Tests.** To test these ideas, we used a simple sentence encoding and decoding task. Each sentence was composed of three roles: an agent, verb, and patient. Each role could be filled with words drawn from a set of 10 words, and each word could be used in any role, resulting in 1,000 possible sentences. The network was trained on only a small subset (20%) of the 1,000 possible sentences and then tested on sentences it had not previously seen, to evaluate its ability to generalize. On each trial, the network was presented

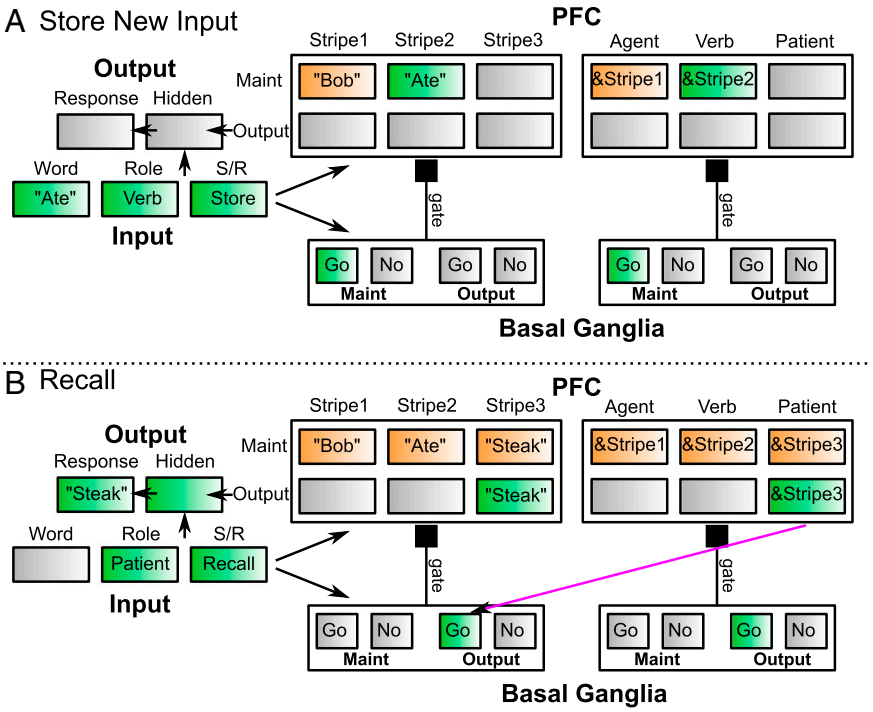


**Fig. 1.** Simple sentence encoding task demonstrating indirection in the PFC/BG working memory system. Three-word sentences are encoded one word at time along with a sentential role. After encoding, the network is probed for each word using the associated roles. Green indicates currently active inputs; orange represents actively maintained information. (*A*) One step of the encoding process for the sentence "Bob ate steak." "Ate" is presented along with its current role (verb) and the instruction to store, or encode, this information. In this example, the word "ate" is stored in Stripe2 of PFC filler stripes (*Left*). The identity/location of Stripe2 is subsequently stored in the verb stripe of PFC role stripes (*Right*). This process repeats for each of the other two words in the sentence. (*B*) One step of the recall process. A role (Patient in the example) and the instruction Recall are presented as input. This drives output gating of the address information stored that role stripe (highlighted by purple arrow), driving the BG units corresponding to that address to output gate the corresponding filler stripe, thus outputting the contents of that stripe (Steak).

with a sentence, word by word (Fig. 1*A*), requiring the maintenance of each word (filler) and its assigned role. The network was then tested by presenting each of the three roles as input, one at a time, to which it had to respond by generating the corresponding filler (Fig. 1*B*). The large majority (80%) of these were sentences it had not seen before, involving not only novel combinations of role-filler pairs, but also fillers in novel roles (i.e., novel role-filler pairs). The network's ability to generalize was indexed by its ability to process these and other types of novel sentences.

To further characterize generalization in the model, we used three variants of the basic learning and testing procedures described above. These tested for standard generalization, spurious correlations, and full combinatoric generalization, as described below. In addition to the full indirection model, we trained and tested two other variants as well as a standard neural network architecture that has been used extensively to model sequential processes, including language. A strict performance criterion of 95% correct was used to determine the amount of experience each network received during the training phase, requiring that every network performed nearly perfectly on the training sentences. Each network was trained and tested using each of the protocols listed below.

***Standard generalization.*** After training on a random selection of 200 out of the 1,000 different sentences, 100 novel sentences were selected randomly from the remaining 800 to be used as the testing set. These sentences had not been experienced by the network during training. However, the training set was constrained so as to ensure that every individual word was presented as a filler in each of the different roles during training. This generalization protocol tested the ability of the network to encode and decode arbitrary combinations of role-filler pairs, but not its ability to process novel role-filler pairs.

***Spurious anticorrelations.*** An anticorrelation occurs whenever one member of a pair of words is never seen in the same sentence as the other during training. The anticorrelation is spurious if there is no valid reason why the two words should not appear together. Such anticorrelations are pervasive in natural environments, in which there are large numbers of possible combinations of stimuli, all of which are plausible but few of which have been experienced. Learning such anticorrelations can be maladaptive, because the anticorrelated pairs could occur in the future. Many artificial neural network learning mechanisms show susceptibility to such spurious anticorrelations, which interferes with their ability to generalize to novel combinations of stimuli (7, 8). To evaluate the response to such anticorrelations, sentences in the training set were selected so that certain words never occurred together during learning (e.g., the words "knife" and "tire" were never part of the same sentence). Then, at test, networks were evaluated on their ability to process sentences that contained the anticorrelated words.

***Full combinatorial generalization.*** This tested the network's ability to generalize not only to novel combinations of role-filler pairs, but also to novel role-filler pairs themselves. To do so, 2 words out of the 10 possible were selected and never used in the role of patient in sentences during training. Networks were then evaluated on their ability to process sentences containing those words as patients in the test set. This is an extremely difficult task for systems that learn from experience about the structure of the world.

**Models.** Fig. 2 shows the networks that we simulated to determine which architectural and processing features were critical to performance on each of the three generalization tests (*Supporting Information*). This includes three progressively elaborated variants of the PFC/BG working memory (PBWM) model (13, 14, 20). The first variant (PBWM Maintenance Only, Fig. 2*A*) was the simplest version of the PBWM model, in which each of the three words in a sentence was encoded and maintained in a separate PFC stripe. The identity of the PFC stripes that stored each word was determined by a learning process in the gating system (BG). The filler structure was learned randomly but included an additional assumption that if a PFC stripe recently gated in new information (e.g., a new word), then it was unlikely to do so again immediately. The specific "role" inputs were provided to the BG, allowing the network to learn a policy that specialized different PFC stripes to the different sentential roles. A response was generated by projections from the PFC and a sentential role input layer to a hidden (association) layer that, in turn, activated a representation of the filler for the specified role over the output units. The sentential role input layer was used during encoding to tell the network what the current role of a word should be (agent, patient, etc.) and, during recall, to identity the role of the filler to be retrieved (e.g., "What was the agent in the sentence you just saw?"). A total of 15 PFC/BG stripes were used in the maintenance-only version. This number was determined by use of a grid search procedure, optimizing the performance of the network across tasks. The other PBWM networks also used 15 stripes to store the "filler" information based on the grid search performed in the maintenance-only version. In other words, we optimized the number of stripes based on the performance of the simplest of the PBWM networks—to give that network the greatest advantage—and used that number for the other PBWM variants.

In the second variant (PBWM Output Gating, Fig. 2*B*), an output gating mechanism was added to the network that selectively decided when information maintained in the PFC stripes should be output, to drive responding. This allowed multiple items to be actively maintained in working memory while allowing only one to be selected to influence the response. A total of 15 PFC/BG stripes were used in the output gating

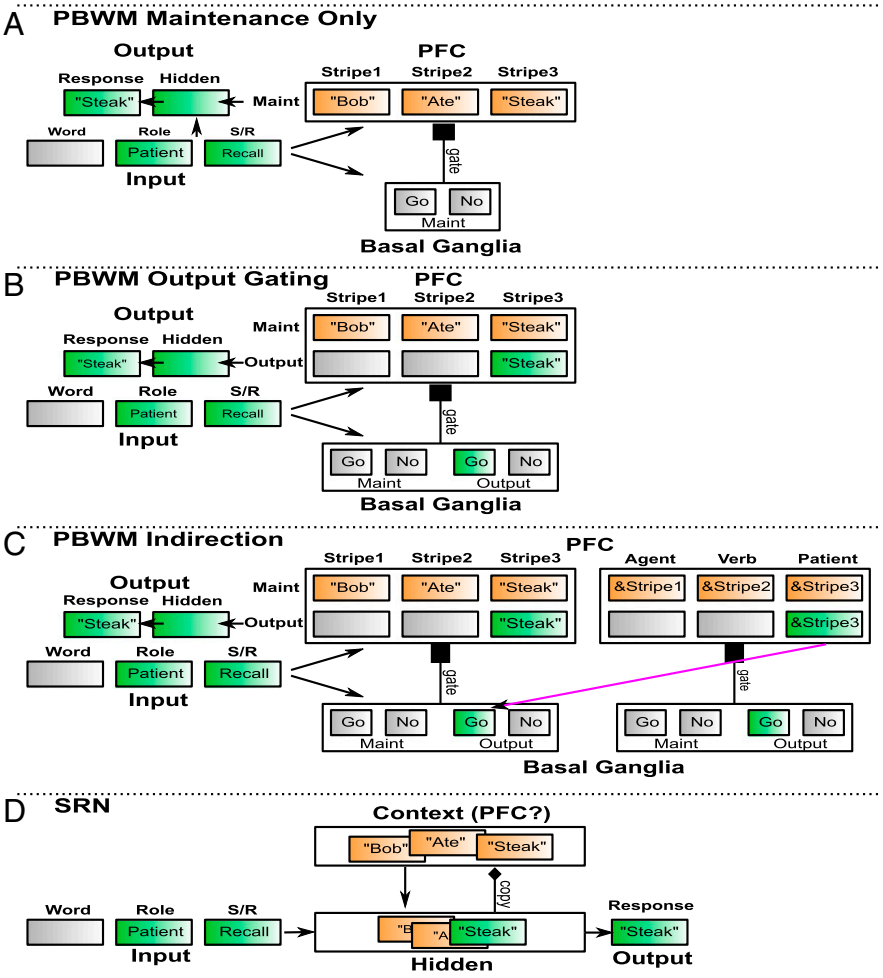


**Fig. 2.** Model architectures. Green indicates currently active inputs; orange represents actively maintained information. The inputs to all networks include the sentential role, the content word (filler), and a signal indicating whether this is an encoding (Store) or retrieval (Recall) trial. Please note that the actual number of PFC stripes used is greater than what is presented here for reasons of clarity. (*A*) PBWM maintenance-only network that implements the key components of the PBWM architecture. (*B*) PBWM output gating network, which separates active maintenance of representations in PFC from the driving a response. (*C*) Full PBWM indirection network. A word is presented as input and is gated into the filler-specific network on the left; its stripe address is propagated to gating units for the role-specific network on the right, which then gate that information into a role-specific PFC stripe. At recall, presentation of the role, together with the recall instruction, activates the output gating unit for that role-specific stripe. That stripe is storing the address of the corresponding filler-specific stripe, which activates the output gating unit for that stripe in the filler-specific network. (*D*) The simple recurrent network.

network (based on the same optimization procedure used for the maintenance-only version of the model).

The third variant (PBWM Indirection, Fig. 2*C*) implemented the full model. This possessed two complete and distinct PBWM networks as well as output gating. One PBWM network (filler-specific) learned to store each word in a set of different filler stripes. Each filler stripe learned to represent multiple fillers, and each filler was represented across multiple stripes. In the results discussed below, the PBWM "role-specific" network contained three sets of three PFC/BG stripes (nine in total), each set dedicated to a different role. Nine stripes were sufficient for satisfactory performance in the indirection network and, unlike the filler-specific network, this parameter was not optimized across models, because it is unique to the indirection network. The "pointer" that specified the location of the appropriate content in the filler-specific portion of the network was maintained until an output gating signal allowed a particular stripe's contents to influence the response. Output gating of the filler-specific network was tightly coupled with the role-specific PBWM network. For each word, the role-specific network learned to store the location of the stripe in the filler-specific network that represented that word. At test, presentation of a given role as input to the network generated an ouptut gating signal in the BG for the stripe corresponding to that role in the role-specific network. The address information stored in the role-specific stripe then generated, in the BG, an output gating signal for the stripe in the filler-specific network corresponding to that address. Output gating of the filler stripe then drove a response, via the hidden layer, corresponding to the word in the probed role. Note that the sentential role information was always explicitly provided in the input, instead of requiring the network to learn to recognize the role based on syntactic or other cues.

Finally, we tested a simple recurrent network (SRN) (Fig. 2*D*). SRNs have been used successfully to model a diverse set of complex behavioral phenomena associated with sequence processing, including early language acquisition, patterns of implicit learning, and routine task performance (21–23). Importantly, SRNs have been shown to exhibit behavior that seems componential and structured, using distributed representations without any explicit componential or hierarchical structure (23). This makes SRNs a useful comparison with our networks, which include structural elements to support componential and hierarchical relationships.

## Results

The generalization results for each network are shown in Fig. 3. The SRN struggled to generalize successfully on all three tasks, indicating that it could not represent each of the sentence elements in a way that supported generalization to novel sentences. Based on prior models, it may require training on a larger portion of all possible sentences (23–25).

The PBWM maintenance-only network performed well on the standard generalization task. This is because it was able to store each word as the filler of the appropriate role in dedicated, independently updatable PFC stripes for each role. As long as it experienced each word in each role, it could process sentences involving novel combinations of these pairs. This was confirmed by examination of the network, which revealed that the PFC stripes learned to encode words for a specific sentential role (e.g., all of the verbs). This replicates findings from both other PBWM models and other models of representation learning in the PFC (26). However, this network failed to generalize in both the anticorrelation and full combinatorial generalization tasks, for reasons that are clarified by examination of the other two PBWM networks.

The PBWM output gating network did well on both the standard and anticorrelation generalization tasks, but not on the full combinatorial task. Its performance on the anticorrelation task sheds light on why the basic PBWM model performed poorly

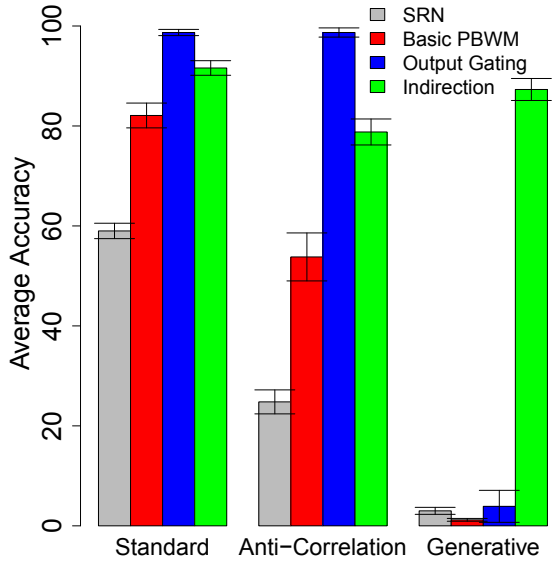


**Fig. 3.** Results for the four networks tested in the three generalization tasks (error bars represent SEM). The results are grouped by task: standard, anticorrelation, and generative.

on this task. Including an output gating mechanism restricted the influence of the items maintained across the various stripes of the PFC, encouraging a componential item-by-item influence on downstream cortical processing (e.g., in the hidden to response output pathway). Thus, the hidden layer only had to process a single, relevant, active representation from the output PFC stripes, instead of the entire set of active information across all of the PFC stripes, as in both the SRN and basic PBWM. This protected it from learning anticorrelations. For example, it would selectively allow the verb item to be processed by the response pathway, without any "contamination" from the maintained agent and patient role items. This is consistent with and replicates previous findings concerning the generalization benefits of output gating within the PBWM framework (27). However, the output gating network failed on the full combinatorial task because it had not been exposed to (many of) the role-filler pairings in the test set, and therefore did not learn filler representations for those roles.

Finally, the PBWM indirection network was able to generalize effectively in all of the tasks. This was due to its ability to encode fillers in roles as pointers to filler-specific stripes, rather than directly within the role-specific stripes themselves. So long as the network had experienced encoding a given stripe address where a filler was being maintained, it could then process anything stored at that location, allowing it to generalize to novel role-filler pairs. The network learned these joint content and location representations across the two sets of interconnected PFC/BG networks.

## Discussion

We have described a unique neural network architecture that implements arbitrary variable binding using biologically plausible mechanisms and demonstrated its ability to generalize, from limited experience, to a rich combinatorial domain. Using a simple sentence processing task, we demonstrated the model's ability to generalize not only to novel combinations of role-filler pairs, but also to novel assignments of fillers to roles. This level of generalization was accomplished only by the full PBWM indirection mode, and could not be achieved by models without a mechanism for indirection.

The design of the architecture of the PBWM indirection network was motivated in part by the computational demands of full combinatorial generalization that require a mechanism for variable binding. However, it was also inspired by, and is consistent with, a growing body of neuroscientific evidence concerning the anatomy and physiology of the PFC and BG. There is long-standing evidence of an asymmetric "spiral" pattern of projections between the PFC and BG, in which the PFC projects to areas of the BG that project to a nearby but distinct area of the PFC (17). Such evidence continues to accrue, suggesting that projections from the PFC to striatum exhibit a rostral-to-caudal bias (28, 29); similar spiraling patterns of connectivity are seen within the striatum proper (30). Gradients of function have also been proposed within the PFC. For example, some neuroimaging studies have suggested a dorsal–ventral organization according to content (31–35). This has also been suggested by previous modeling work using the PBWM, in which more dorsal areas have been proposed to represent higher-level information (e.g., dimensional or category information and task specifications), whereas more ventral areas represent more specific information (e.g., featural) (18, 35). Another proposal suggests that there is an anterior–posterior functional hierarchy in the frontal cortex (36, 37), in which anterior areas represent more abstract information (e.g., high-level goals), whereas more posterior areas represent more concrete information (e.g., motor plans). The separation of subnetworks within the PFC and the pattern of projections between the PFC and BG in the PBWM indirection network are broadly consistent with these proposals. Importantly, it provides a more specific characterization of the function of distinct PFC regions that it should be possible to test empirically.

A critical finding from the current simulations is that a neural network architecture can learn to implement indirection and thus achieve the capacity for variable binding required for full combinatorial generalization. This required a particular predetermined macroarchitecture (e.g., the pattern of projections between the PFC and BG) and set of physiological functions (e.g., the learning algorithm, input and output gating). However, the microarchitecture of the model (e.g., the representations within the filler subnetwork) and its information-processing functions (e.g., the conditions under which individual stripes were gated) were not prespecified. These were learned through experience. It should be noted, however, that the pointer representations were specified as a localist map of the possible filler stripe locations. Although there is evidence of topographic connections throughout the cortex, it seems plausible that these could be learned through an extended developmental process. Thus, the model can be characterized as a hybrid, in which general characteristics of anatomy and physiology are predetermined (presumably by genetics) but the functional characteristics are learned by exposure to the environment. This included the capacity for indirection, and thereby variable binding. This was made possible by the separate, but anatomically related, sets of PFC/BG subnetworks, allowing the system to segregate function (role) from content (filler). As a consequence, the role-specific stripes needed only to concern themselves with learning, at a high level, where to look for content when the time was appropriate. This greatly reduced the representational burden of these stripes, relieving them of the burden of encoding all possible fillers that may ever be needed in a particular role.

Although the model we described focused on a simple sentence processing task, the principles of function can be readily generalized. The role-specific subnetwork can be thought of as representing elements of task context and the filler-specific network as representing stimuli and/or actions that are appropriate in that particular context. Thus, the model could also be used to simulate not only the interpretation of novel sequences of inputs, but also the production of novel, appropriately structured sequences of actions. In this regard, it can be considered as providing the functionality necessary not only for generalization but also generativity (the ability to generate meaningful novel behavior). In this respect, it offers a middle ground in the long-standing

debate between symbolic and subsymbolic models of cognition (1, 38). Advocates of symbolic models have long pointed to the kind of combinatorial generalization we have tested here and, critically, capacity for generativity as support for the claim that human cognitive function relies on symbol processing and that this cannot be accomplished by associationist (neural network) architectures. In contrast, advocates of subsymbolic models have argued that the evident limitations on human symbolic reasoning (39, 40) suggest that actual everyday human cognition is better characterized by subsymbolic processing. Our model suggests that both approaches may be partially correct: The architecture of the brain supports a form of indirection and thereby variable binding, but it is limited. In particular, it relies on extensive learning, and the representations it develops are embedded, and distributed—the representations our indirection model learned were distributed across multiple stripes, which worked together to represent the information. This characterization also differs significantly from other biologically based attempts to account for symbolic processing abilities. These accounts build in automatic low-level variable binding mechanisms (11, 41–45), which currently lack an explanation for how these systems can be learned from experience and at the extreme may suggest that this symbolic capacity is present even at the lowest levels of processing, without any learning necessary. Furthermore, we expect that the subsymbolic foundation of our indirection model, and its grounding in neural learning mechanisms, will render it more robust and powerful than purely symbolic models, which are often brittle.

The computational architecture we have described could be used to explore the development of specific systems of content (filler) and pointer (role) representations in a range of different cognitive processing domains. This holds promise for generating predictions about the stages of cognitive flexibility and systematicity as a function of learning experience that can be compared with available human developmental data to further test and inform the model.

**ACKNOWLEDGMENTS.** This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) Contract D10PC20021. The US government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the US government.

1. Fodor JA, Pylyshyn ZW (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1–2):3–71.
2. Marcus GF (1998) Rethinking eliminative connectionism. *Cognit Psychol* 37(3): 243–282.
3. van Gelder T (1990) Compositionality: A connectionist variation on a classical theme. *Cogn Sci* 14:355–384.
4. Johnson K (2004) On the systematicity of language and thought. *J Philos* 101:111–139.
5. van Gelder T, Niklasson LF (1994) *Classicism and Cognitive Architecture* (Lawrence Erlbaum, Hillsdale, NJ), pp 905–909.
6. Plaut DC, McClelland JL, Seidenberg MS, Patterson K (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol Rev* 103(1):56–115.
7. Noelle DC, Cottrell GW (1996) *In Search of Articulated Attractors*, ed Cottrell GW (Lawrence Erlbaum, Mahwah, NJ), pp 329–334.
8. Noelle DC, Zimdars AL (1999) *Methods for Learning Articulated Attractors over Internal Representations*, eds Hahn M, Stoness SC (Lawrence Erlbaum, Mahwah, NJ), pp 480–485.
9. Smolensky P (1990) Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artif Intell* 46:159–216.
10. Plate TA (2008) Holographic reduced representations. *IEEE Trans Neural Networks* 6:623–641.
11. Stewart TC, Bekolay T, Eliasmith C (2011) Neural representations of compositional structures: Representing and manipulating vector spaces with spiking neurons. *Connect Sci* 23:145–153.
12. Frank MJ, Loughry B, O'Reilly RC (2001) Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cogn Affect Behav Neurosci* 1(2):137–160.
13. O'Reilly RC, Frank MJ (2006) Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput* 18(2):283–328.
14. O'Reilly RC (2006) Biologically based computational models of high-level cognition. *Science* 314(5796):91–94.
15. Levitt JB, Lewis DA, Yoshioka T, Lund JS (1993) Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 and 46). *J Comp Neurol* 338(3):360–376.
16. Pucak ML, Levitt JB, Lund JS, Lewis DA (1996) Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *J Comp Neurol* 376(4):614–630.
17. Alexander GE, DeLong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci* 9:357–381.
18. O'Reilly RC, Noelle DC, Braver TS, Cohen JD (2002) Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex* 12:246–257.
19. Reynolds JR, O'Reilly RC (2009) Developing PFC representations using reinforcement learning. *Cognition* 113(3):281–292.
20. Hazy TE, Frank MJ, O'Reilly RC (2006) Banishing the homunculus: Making working memory work. *Neuroscience* 139(1):105–118.
21. Hare M, Elman JL (1995) Learning and morphological change. *Cognition* 56(1):61–98.
22. Cleeremans A, McClelland JL (1991) Learning the structure of event sequences. *J Exp Psychol Gen* 120(3):235–253.
23. Botvinick M, Plaut DC (2004) Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychol Rev* 111(2):395–429.
24. O'Reilly RC (2001) Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Comput* 13(6):1199–1241.
25. Brousse O (1993) Generativity and systematicity in neural network combinatorial learning (Department of Computer Science, Univ of Colorado, Boulder, CO), Technical Report CU-CS-676-93.
26. Rougier NP, Noelle DC, Braver TS, Cohen JD, O'Reilly RC (2005) Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proc Natl Acad Sci USA* 102(20): 7338–7343.
27. Kriete T, Noelle DC (2011) Generalisation benefits of output gating in a model of prefrontal cortex. *Connect Sci* 23:119–129.
28. Haber SN, Kim KS, Mailly P, Calzavara R (2006) Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *J Neurosci* 26(32):8368–8376.
29. Verstynen TD, Badre D, Jarbo K, Schneider W (2012) Microstructural organizational patterns in the human corticostriatal system. *J Neurophysiol* 107(11):2984–2995.
30. Haber SN, Fudge JL, McFarland NR (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J Neurosci* 20(6): 2369–2382.
31. Funahashi S, Chafee MV, Goldman-Rakic PS (1993) Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature* 365(6448):753–756.
32. Smith EE, Jonides J (1999) Storage and executive processes in the frontal lobes. *Science* 283(5408):1657–1661.
33. Romanski LM (2004) Domain specificity in the primate prefrontal cortex. *Cogn Affect Behav Neurosci* 4(4):421–429.
34. Petrides M (2005) Lateral prefrontal cortex: architectonic and functional organization. *Philos Trans R Soc Lond B Biol Sci* 360(1456):781–795.
35. O'Reilly RC (2010) The *What* and *How* of prefrontal cortical organization. *Trends Neurosci* 33(8):355–361.
36. Koechlin E, Ody C, Kouneiher F (2003) The architecture of cognitive control in the human prefrontal cortex. *Science* 302(5648):1181–1185.
37. Badre D, D'Esposito M (2007) Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *J Cogn Neurosci* 19(12): 2082–2099.
38. McClelland JL, et al. (2010) Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends Cogn Sci* 14(8):348–356.
39. Johnson-Laird PN (2001) Mental models and deduction. *Trends Cogn Sci* 5(10): 434–442.
40. Todd PM, Gigerenzer G (2000) Précis of simple heuristics that make us smart. *Behav Brain Sci* 23(5):727–741, discussion 742–780.
41. Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. *Psychol Rev* 99(3):480–517.
42. David S, Touretzky GEH (1988) A distributed connectionist production system. *Cogn Sci* 12:423–466.
43. Lebiere C, Anderson JR (1993) *A Connectionist Implementation of the ACT-R Production System* (Lawrence Earlbaum, Hillsdale, NJ).
44. Stocco A, Lebiere C, Anderson JR (2010) Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination. *Psychol Rev* 117(2):541–574.
45. Hayworth KJ (2012) Dynamically partitionable autoassociative networks as a solution to the neural binding problem. *Front Comput Neurosci* 6:73, 10.3389/fncom.2012.00073.