REVIEW

Biologically Based Computational Models of High-Level Cognition

Randall C. O'Reilly

Computer models based on the detailed biology of the brain can help us understand the myriad complexities of human cognition and intelligence. Here, we review models of the higher level aspects of human intelligence, which depend critically on the prefrontal cortex and associated subcortical areas. The picture emerging from a convergence of detailed mechanistic models and more abstract functional models represents a synthesis between analog and digital forms of computation. Specifically, the need for robust active maintenance and rapid updating of information in the prefrontal cortex appears to be satisfied by bistable activation states and dynamic gating mechanisms. These mechanisms are fundamental to digital computers and may be critical for the distinctive aspects of human intelligence.

Biologically based computational modeling has been an integral part of many basic areas of cognitive neuroscience (e.g., perception and memory). More recently, these mechanistic approaches have been encroaching on some of the most mysterious and challenging higher level areas of human cognition, including decision making, problem solving, and "executive" control of cognition and action. From a biological perspective, it is clear that the prefrontal cortex (PFC) and associated subcortical areas in the basal ganglia and midbrain play a disproportionately important role in these aspects of cognition (*1*, *2*). Furthermore, the PFC is the area of cortex most greatly expanded in humans relative to other mammals (*3*), suggesting that it is critical to human intellectual abilities.

Two examples are instructive. First, people with damage to PFC areas often exhibit environmental dependency syndrome (*4*), which is (provocatively) just a fancy name for a lack of free will: Behavior is driven more by the external environment than by internal plans or goals. For example, one person with PFC damage visiting a researcher's home saw a bed and proceeded to get undressed (including removing his toupee), got into bed, and prepared to sleep. The second example is something to which everyone can relate: the crazy world of dreams. One minute you are talking with a long-forgotten friend from high school and the next you are late for an airplane you cannot quite seem to find. The PFC is one of the primary brain areas deactivated during dream states (*5*), and its absence may have much to do with the lack of temporal contiguity and inability to stay on task that are characteristic of dreams. In short, the PFC is critical for maintaining current context, goals, and other information in an active state that guides ongoing behavior in a coherent, task-relevant manner (*2*, *6*).

It is not enough to state simply that the PFC houses our internal "executive" that decides what we want to do and keeps us focused on those goals in the face of various environmental distractions. That only labels and locates the mystery. The promise of biologically based computational models is that they can actually break open these mysteries by describing the underlying mechanisms in precise computational detail and showing that they are indeed capable of the functions attributed to the PFC. The success of this approach does not mean that we need to think of humans as robots. Instead, these models show that many subtle factors interact in complex ways to produce the emergent phenomenon of cognitive control, which cannot be simply reduced to its constituents. The modeling approach enables such complex and subtle interactions to be understood in a way that would be impossible using the comparatively blunt empirical methods available today. The risk run by these models is that they provide an elaborate fiction, instead of facts, about how the brain actually works. However, this risk can be mitigated by building models that integrate a wide range of empirical data spanning many different levels of analysis.

Here, we focus on converging models at multiple levels of analysis that together paint a remarkably coherent picture of PFC and associated systems. We first review some areas of fundamental agreement and provide an initial sketch of this emerging picture, followed by some specific recent developments.

The Standard Model of Prefrontal Cortex Function: Active Maintenance, Top-Down Control, and Rapid Updating

The PFC is important for actively maintaining information by sustained neural firing (*7*, *8*), which is robust in the face of potentially distracting information [i.e., working memory (*9*)]. This is in contrast with other cortical areas, which tend to be swayed by whatever stimulus is currently impinging on them (hence the environmental dependency syndrome in the absence of normal PFC function). From this basic mechanism of active maintenance, it is possible to explain a remarkable amount of what the PFC does. For example, the robust active maintenance of a goal or plan representation (e.g., "go to the grocery store before going home") can guide a sequence of behaviors (e.g., making the appropriate turns) simply by providing additional neural activation to these appropriate behaviors in the face of other possibly stronger competing actions (e.g., driving directly home). Because such goals can be maintained in the face of inevitable environmental distractors, they enable behavior to be consistent and coherent over time. Accordingly, when PFC is not functioning well, as in the dream state or in the prevalent attention deficit hyperactivity disorder (ADHD) (*10*, *11*), behavior becomes less consistent and coherent over time. Furthermore, because of this ability to focus on a task to the exclusion of other distracting information, the PFC is often characterized as inhibiting task-irrelevant information (*12*, *13*).

The PFC system is also capable of rapidly updating what is being maintained, which is critical for behavioral flexibility—the ability to quickly adapt to the changing demands of the environment. People with PFC damage tend to perseverate in the face of changing task demands (*14*), as do young children with immature PFC function (*15*). Areas within PFC also play a key role in monitoring of behavior [necessary for applying appropriate levels of control, e.g., (*16*)] and in emotional and reward processing (*17*). These are beyond the scope of the present paper, but it may be possible to understand many aspects of these functions using the basic mechanisms elaborated below (*18*).

We return to our main question: How does the brain actually perform these active maintenance and rapid updating functions in terms of detailed biological mechanisms? Biologically based computational models have explored this question in depth. The emerging picture can be summarized in terms of analog versus digital computation; whereas the rest of cortex can be characterized as a fundamentally analog system operating on graded, distributed information, the prefrontal cortex has a more discrete, digital character. Robust active maintenance is supported by a form of bistability, which means that neurons switch between two stable states (off or on), much as bits in a computer. Rapid updating requires a mechanism for gating or switching between these bistable states—this gating/switching is the essential function of a transistor in a digital

Department of Psychology, University of Colorado Boulder, Boulder, CO 80309, USA. E-mail: oreilly@psych.colorado.edu

computer. Even though each of these mechanisms is strongly motivated from basic biological and computational factors having nothing to do with digital computation, the parallels are striking and might perhaps provide some critical insight into what makes humans distinctively intelligent.

Biological Mechanisms of Active Maintenance and Rapid Updating: Bistability and Gating

Perhaps the most obvious mechanism for active maintenance is recurrent excitatory connectivity, which amounts to a form of you-rub-my-back-and-I'll-rub-yours. Active neurons send excitation to other neurons that then send excitation back, creating a stable "attractor" state (*19*). Although appealingly simple, several problems with this model have arisen. For example, biologically detailed models have shown that it is difficult to sustain this positive feedback system because individual spikes of neural firing may not come frequently enough to keep it going (*20*). Furthermore, when using these attractor states to integrate information over time, it has become clear that noise (which is ubiquitous in the brain) quickly swamps any signal present in these systems (*21*). Intuitively, the system operates like the classic "telephone" game, where a message passed continuously along a chain (or repeatedly among an interconnected population of neurons) is rapidly distorted.

A solution to both of the above problems is to incorporate some form of intrinsic bistability into the neural systems (*20*–*22*). In these models, bistability comes from gated ion channels that require specific levels of neural depolarization to be activated, and once active they remain so for hundreds of milliseconds or more [e.g., the *N*-methyl-D-aspartate (NMDA) channel]. This imparts a critical degree of robustness to the active maintenance abilities of PFC neurons, enabling them to span the gaps between spikes and also not to get blown around by the winds of neural noise (*21*). To encode analog (graded) information with bistable neurons, the system must use distributed binary representations that work somewhat like a binary encoding of a floating-point number on a computer: Many neurons (bits) work together such that the combined pattern of activity represents different values. Although this is less efficient than a direct analog representation (which could be done with a single neuron), the improvement in robustness may be worth this cost. Certainly, this is the case with computers. We use digital computers because analog computers are quickly swamped by noise.

Rapid updating provides another important challenge for neural mechanisms to solve, because it directly conflicts with the need for robust active maintenance. Once a set of neurons is locked into a stable state, how can it subsequently be updated to hold on to new information? A number of mechanistic solutions to this problem have been proposed, all of which amount to a dynamic gating mechanism, which modulates the stability of PFC active maintenance. When the gate is open, PFC is rapidly updated with new information. When the gate is closed, it robustly maintains existing information.

One class of gating mechanisms depends on the neuromodulator dopamine (*22*–*25*), which is transmitted to the PFC by the midbrain ventral tegmental area (VTA). In all such models, dynamic changes in the level of dopamine in PFC, caused by phasic VTA firing above the normal tonic level, switch the system between rapid updating and robust maintenance (Fig. 1). Biologically detailed computational modeling has made sense of the dense and confusing thicket of studies on dopamine modulation of PFC circuits (*22*, *25*). In this model, dopamine D1 receptor activation produces a net overall effect of stabilizing working memory states in PFC through a complex combination of seemingly opposing effects, including increased NMDA current activation (*20*). In contrast, D2 receptor activation produces opposing destabilizing effects. Given that both receptors are activated by the same neuromodulator, how does the system alternate between maintenance and updating? D2 receptors are largely synaptic and respond only to high concentrations of dopamine, whereas D1 receptors are extrasynaptic and respond to lower concentrations. Thus, a phasic burst of dopamine will activate D2 receptors located in the synapses and trigger rapid updating, whereas lower tonic concentrations diffusing extrasynaptically produce a default level of robust maintenance (*25*) (Fig. 1). This biologically detailed model converges remarkably well with earlier, more abstract computational models hypothesizing that dopamine modulates the gain or signal-to-noise ratio of neurons (*24*, *26*).

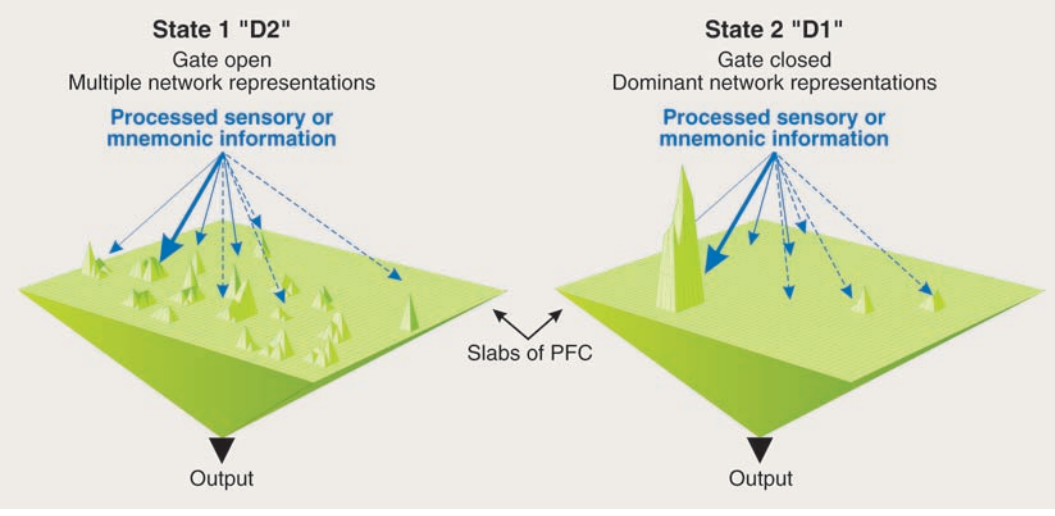

Fig. 1. Dopamine-based gating mechanism that emerges from the detailed biological model of Durstewitz, Seamans, and colleagues (*22*, *25*). The opening of the gate occurs in the dopamine D2-receptor–dominated state (State 1), in which any existing active maintenance is destabilized and the system is more responsive to inputs. The closing of the gate occurs in the D1-receptor–dominated state (State 2), which stabilizes the strongest activation pattern for robust active maintenance. D2 receptors are located synaptically and require high concentrations of dopamine and are therefore activated only during phasic dopamine bursts, which thus trigger rapid updating. D1 receptors are extrasynaptic and respond to lower concentrations, so robust maintenance is the default state of the system with normal tonic levels of dopamine firing.

Another class of gating mechanism leverages the extensive connectivity between the PFC and the basal ganglia (*27*–*29*) (Fig. 2). Direct pathway "Go" neurons in the basal ganglia can trigger a phasic wave of activation into PFC through a modulatory disinhibition effect (Fig. 2), which results in rapid updating of PFC states. These Go neurons are opposed by a set of indirect pathway "NoGo" neurons that prevent this phasic wave of activation and enable the default state of robust active maintenance to continue. This basal ganglia mechanism is functionally very similar to the dopamine-based one. However, a key difference is that the basal ganglia mechanism enables selective gating of only some regions of PFC, whereas dopamine modulation is more broad and diffuse. Furthermore,

the dopaminergic effects in PFC are thought to operate on a slower time scale than the basal ganglia (*25*). Thus, it is likely that the basal ganglia gating supports more rapid, selective updating of specific regions in PFC, whereas the dopaminergic gating provides a longer time scale, broader gating signal that is modulated by overall performance levels.

The presence of a gating mechanism raises the question of how the gate knows when to open and close. Several models have shown that reinforcement learning mechanisms, which also involve the dopaminergic system (synergistically with its role in gating), can learn to control this gating mechanism (*23*, *29*). One of these models was shown to compare favorably with some of the most advanced but biologically implausible learning mechanisms on complex temporally extended working memory tasks (*29*). This model is now being tested on a wide range of benchmark higher level cognition tasks to determine whether the biologically based mechanisms converge with behavioral data (*30*). Thus, it is plausible, but not yet established, that such a system could learn to perform the many different higher level cognitive tasks that humans can perform.

Toward Higher Digital Intelligence

The "digital" picture emerging from these bistable, dynamically gated PFC neurons supports a certain amount of intelligent behavior, at least in the terms that were considered above: robust active maintenance of goals and other task-relevant information, and rapid updating of this information to keep pace with a changing environment or task. Is that the full story, or are there other important ingredients to our intelligence?

These digital PFC dynamics may also support more abstract forms of reasoning, which is another important aspect of human intelligence (*31*). The presence of a dynamic gating mechanism and robust active maintenance in the PFC led to the development of more abstract, rulelike representations in a simulated PFC. Although these representations do not have the arbitrary flexibility of the symbols present in digital computer programs (and symbolic models of human cognition), they are nevertheless closer than the graded, distributed representations typically associated with other areas of cortex. Specifically, these PFC representations in the model enabled the behavior of the overall system to be more regular (i.e., describable by an abstract rule), in that it could more consistently apply a previously learned rule to novel situations. This model is consistent with recent recordings from PFC and posterior cortex neurons in monkeys, which showed that PFC neurons exhibit more abstract rulelike encodings of categories and other task-relevant information (*32*–*34*).

The fact that monkeys also show some degree of abstract representation in PFC raises the perennial question of what exactly is different between us and them. The critical difference may be that people have a basic social instinct for sharing experiences and knowledge with each other that is largely absent in even our closest primate relatives (*35*). Thus, the qualitative difference comes not from the hardware [which is still quantitatively better (*3*)] but from the motivations that drive us to spend so much time learning and communicating what we have learned to others. This account dovetails nicely with the above modeling work (*31*), which found that the abstract PFC representations took a long time to develop and required integrating knowledge across multiple different but related tasks. Furthermore, the development of the PFC is the most protracted of any brain area, not fully maturing until adolescence (*15*, *36*, *37*). Thus, the full glory of human intelligence requires the extensive, culturally supported developmental learning process that takes place over the first decades of life.

Another potential implication of a dynamic gating mechanism is the ability to perform transistor-like dynamic switching, which can enable a form of variable binding (i.e., assigning arbitrary information to a given functional role, as in "let $X = 7$") that is not otherwise possible in statically connected neural circuits (*29*, *38*). Figure 3 provides a simple illustration of this form of variable binding, in which one input signal (a verb in a sentence) can dynamically control (through the basal ganglia) which of two different PFC representations encode the name of a person. If the verb form is active, then the name is encoded in the PFC neurons that represent the agent (actor) of the sentence; if the verb is passive, then the patient (object) representations are activated. This system can be more flexible than other more static neural circuits because the gating signal can be completely independent of the content that is being gated. However, unlike a memory buffer in a standard digital computer, the PFC areas must learn slowly over time to be able to represent all the things that they can maintain, and other areas of the brain must similarly learn to decode both the content and role meaning of these PFC representations. Thus, the dynamic variable binding operates in the context of the relatively more static learned

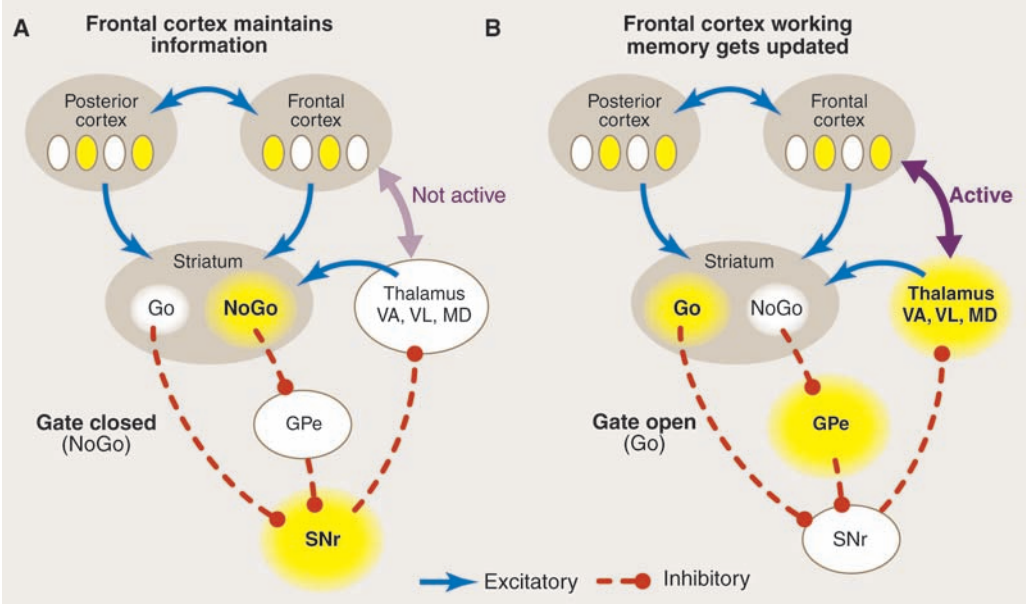

Fig. 2. Dynamic gating produced by disinhibitory circuits through the basal ganglia and frontal cortex/PFC (one of multiple parallel circuits shown). (**A**) In the base state (no striatum activity) and when NoGo (indirect pathway) striatum neurons are firing more than Go, the SNr (substantia nigra pars reticulata) is tonically active and inhibits excitatory loops through the basal ganglia and PFC through the thalamus. This corresponds to the gate being closed, and PFC continues to robustly maintain ongoing activity (which does not match the activity pattern in the posterior cortex, as indicated). (**B**) When direct pathway Go neurons in striatum fire, they inhibit the SNr and thus disinhibit the excitatory loops through the thalamus and the frontal cortex, producing a gating-like modulation that triggers the update of working memory representations in prefrontal cortex. This corresponds to the gate being open.

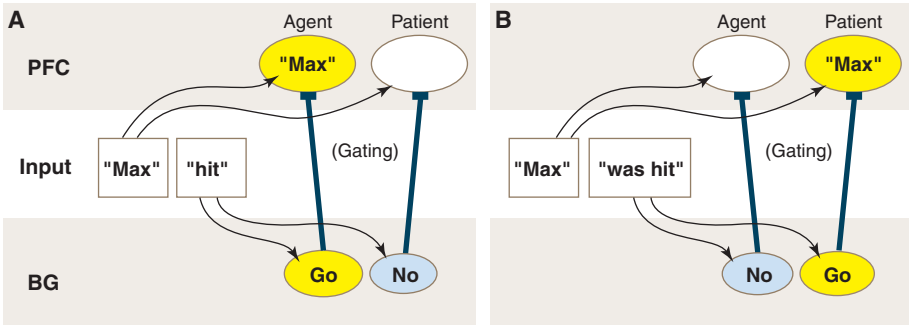

Fig. 3. Dynamic gating can achieve a form of dynamic variable binding, illustrated here for assigning the semantic role of a person based on the grammatical structure of incoming text. The basal ganglia (BG) provides dynamic gating signals to different PFC areas that have learned to encode either the Agent (actor) or Patient (object) semantic role information. If a given set of BG neurons fire a Go (update) signal, then current sensory information is updated into corresponding PFC neurons; if the corresponding BG neurons have a NoGo (do not update) signal, the PFC area is not updated. In (**A**), the active form of the verb ("hit") causes the BG gating units to fire a "Go" (update) signal for the Agent role representations in PFC, which then represent the incoming name ("Max"). The Patient role is not updated because of a NoGo (do not update) signal. In (**B**), the passive form ("was hit") activates the opposite pattern of BG gating, resulting in "Max" being encoded in the Patient role. This ability for one signal (the verb in this case) to modulate where another piece of information ("Max") is encoded provides a basic form of variable binding.

structures typical of other cortical areas and therefore does not achieve the completely arbitrary character of a digital computer. This constraint has benefits, however, because the PFC neurons automatically have meaning through their learned connections with other neurons, and this "grounds" what would otherwise be arbitrary, meaningless symbols.

Having a biologically based mechanism for limited variable binding opens up new opportunities to develop links between these models and more abstract cognitive architectures such as ACT-R (adaptive control of thought–rational) (*39*) that can actually perform complex problem solving and other higher level cognitive tasks that are beyond the reach of existing biologically detailed models. A highly constrained form of variable binding is critical for most cognitive operations in ACT-R, and there is some recent indication that a somewhat more flexible form of binding ("dynamic pattern matching") is necessary for distinctively human cognitive abilities (*40*). Establishing a clear neural basis for these properties would almost certainly provide important insights into what makes us so smart.

Conclusions

Scientists are always concerned about strongly differentiating theoretical positions: the long dominance and current disfavor of the computer metaphor for understanding the mind has led the new generation of biological neural network theorists to emphasize the graded, analog, distributed character of the brain. It is clear that the brain is much more like a social network than a digital computer, with learning, memory and processing all being performed locally through graded communication between interconnected neurons. These neurons build up strong, complex "relationships" over a long period of time; a neuron buried deep in the brain can only function by learning which of the other neurons it can trust to convey useful information. In contrast, a digital computer functions like the post office, routing arbitrary symbolic packages between passive memory structures through a centralized processing unit, without consideration for the contents of these packages. This affords arbitrary flexibility (any symbol is as good as any other), but at some cost: When everything is arbitrary, then it is difficult to encode the subtle and complex relationships present in our commonsense knowledge of the real world. In contrast, the highly social neural networks of the brain are great at keeping track of "who's who and what's what," but they lack flexibility, treating a new symbol like a stranger crashing the party.

The digital features of the PFC and associated areas help to broaden the horizons of naturally parochial neural networks. The dynamic gating mechanisms work more like a post-office, with the basal ganglia reading the zip code of which PFC stripe to update, whereas the PFC cares more about the contents of the package. Furthermore, the binary rulelike representations in the PFC are more symbol-like. Thus, perhaps a fuller understanding of this synthesis of analog and digital computation will finally unlock the mysteries of human intelligence.

References and Notes

1. J. M. Fuster, *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe* (Lippincott-Raven, New York, ed. 3, 1997).
2. E. K. Miller, J. D. Cohen, *Annu. Rev. Neurosci.* **24**, 167 (2001).
3. P. T. Schoenemann, M. J. Sheehan, L. D. Glotzer, *Nat. Neurosci.* **8**, 242 (2005).
4. F. Lhermitte, *Ann. Neurol.* **19**, 335 (1986).
5. J. A. Hobson, E. F. Pace-Schott, R. Stickgold, *Behav. Brain Sci.* **23**, 793 (2000).
6. R. C. O'Reilly, T. S. Braver, J. D. Cohen, *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, A. Miyake, P. Shah, Eds. (Cambridge Univ. Press, New York, 1999), pp. 375–411.
7. J. M. Fuster, G. E. Alexander, *Science* **173**, 652 (1971).
8. P. S. Goldman-Rakic, *Handb. Physiol. Nervous Syst.* **5**, 373 (1987).
9. A. Miyake, P. Shah, Eds., *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (Cambridge Univ. Press, New York, 1999).
10. A. L. Krain, F. X. Castellanos, *Clin. Psychol. Rev.* (2006).
11. E. G. Willcutt, A. E. Doyle, J. T. Nigg, S. V. Faraone, B. F. Pennington, *Biol. Psychiatry* **57**, 1336 (2005).
12. A. Diamond, *Dev. Sci.* **1**, 185 (1998).
13. J. M. Stedron, S. D. Sahni, Y. Munakata, *J. Cognit. Neurosci.* **17**, 623 (2005).
14. D. T. Stuss *et al.*, *Neuropsychologia* **38**, 388 (2000).
15. A. Diamond, P. S. Goldman-Rakic, *Exp. Brain Res.* **74**, 24 (1989).
16. M. Botvinick, L. E. Nystrom, K. Fissel, C. S. Carter, J. D. Cohen, *Nature* **402**, 179 (1999).
17. E. T. Rolls, *Cereb. Cortex* **10**, 284 (2000).
18. M. J. Frank, E. D. Claus, *Psychol. Rev.* **113**, 300 (2006).
19. D. Zipser, *Neural Comput.* **3**, 179 (1991).
20. X.-J. Wang, *J. Neurosci.* **19**, 9587 (1999).
21. A. A. Koulakov, S. Raghavachari, A. Kepecs, J. E. Lisman, *Nat. Neurosci.* **5**, 775 (2002).
22. D. Durstewitz, J. K. Seamans, T. J. Sejnowski, *J. Neurophysiol.* **83**, 1733 (2000).
23. T. S. Braver, J. D. Cohen, *Control of Cognitive Processes: Attention and Performance XVIII*, S. Monsell, J. Driver, Eds. (MIT Press, Cambridge, MA, 2000), pp. 713–737.
24. J. D. Cohen, T. S. Braver, J. W. Brown, *Curr. Opin. Neurobiol.* **12**, 223 (2002).
25. J. K. Seamans, C. R. Yang, *Prog. Neurobiol.* **74**, 1 (2004).
26. D. Servan-Schreiber, H. Printz, J. D. Cohen, *Science* **249**, 892 (1990).
27. M. J. Frank, B. Loughry, R. C. O'Reilly, *Cognit. Affect. Behav. Neurosci.* **1**, 137 (2001).
28. J. W. Mink, *Prog. Neurobiol.* **50**, 381 (1996).
29. R. C. O'Reilly, M. J. Frank, *Neural Comput.* **18**, 283 (2006).
30. T. E. Hazy, M. J. Frank, R. C. O'Reilly, *Neuroscience* **139**, 105 (2006).
31. N. P. Rougier, D. Noelle, T. S. Braver, J. D. Cohen, R. C. O'Reilly, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7338 (2005).
32. I. M. White, S. P. Wise, *Exp. Brain Res.* **126**, 315 (1999).
33. J. D. Wallis, K. C. Anderson, E. K. Miller, *Nature* **411**, 953 (2001).
34. K. Sakai, R. E. Passingham, *Nat. Neurosci.* **6**, 75 (2003).
35. M. Tomasello, *The Cultural Origins of Human Cognition* (Harvard Univ. Press, Cambridge, MA, 2001).
36. P. R. Huttenlocher, *Neuropsychologia* **28**, 517 (1990).
37. J. B. Morton, Y. Munakata, *Dev. Sci.* **5**, 435 (2002).
38. F. van der Velde, M. de Kamps, *Behav. Brain Sci.* **29**, 37 (2006).
39. J. R. Anderson *et al.*, *Psychol. Rev.* **111**, 1036 (2004).
40. J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* (Oxford Univ. Press, New York, 2006).
41. Supported by Defense Advanced Research Projects Agency/Office of Naval Research grants N00014-05-1-0880 and N00014-03-1-0428 and NIH grants MH069597 and MH64445. I thank my long-time collaborators in developing these ideas: J. Cohen, T. Braver, D. Noelle, M. Frank, Y. Munakata, and the CCN Lab at the University of Colorado.

10.1126/science.1127242