TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2017) 1–17

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12307

This article is part of the topic "Ubiquity of Surprise: Developments in Theory, Converging Evidence, and Implications for Cognition," Edward Munnich, Mark Keane, and Meadhbh Foster (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview.

The Role of the Anterior Cingulate Cortex in Prediction Error and Signaling Surprise

William H. Alexander,[a] Joshua W. Brown[b]

[a]*Department of Experimental Psychology, Ghent University*
[b]*Department of Psychological and Brain Sciences, Indiana University*

Received 27 June 2016; received in revised form 22 April 2017; accepted 16 August 2017

Abstract

In the past two decades, reinforcement learning (RL) has become a popular framework for understanding brain function. A key component of RL models, prediction error, has been associated with neural signals throughout the brain, including subcortical nuclei, primary sensory cortices, and prefrontal cortex. Depending on the location in which activity is observed, the functional interpretation of prediction error may change: Prediction errors may reflect a discrepancy in the anticipated and actual value of reward, a signal indicating the salience or novelty of a stimulus, and many other interpretations. Anterior cingulate cortex (ACC) has long been recognized as a region involved in processing behavioral error, and recent computational models of the region have expanded this interpretation to include a more general role for the region in predicting likely events, broadly construed, and signaling deviations between expected and observed events. Ongoing modeling work investigating the interaction between ACC and additional regions involved in cognitive control suggests an even broader role for cingulate in computing a hierarchically structured surprise signal critical for learning models of the environment. The result is a predictive coding model of the frontal lobes, suggesting that predictive coding may be a unifying computational principle across the neocortex.

Keywords: Surprise; Prediction error; Reinforcement learning; Anterior cingulate; Prefrontal cortex; Predictive coding

Correspondence should be sent to Joshua W. Brown, Department of Psychological and Brain Sciences, Indiana University, 1101 E. Tenth St, Bloomington, IN 47401. E-mail: jwmbrown@indiana.edu

1. Introduction

Although the quintessential example of a surprising event may be described as a "black swan" event, the philosophy of surprise actually began in earnest with sheep. The Dutch philosopher Baruch Spinoza (1632–1677) was the first modern philosopher to explore the nature and meaning of surprise, after Descartes and Aristotle who treated it only in passing. Spinoza's *Short Treatise on God, Man, and His Well-Being* (De Spinoza, 1910) devotes a whole section to surprise, from which a number of points remain relevant for the present. First, Spinoza considers what would surprise a person: "Since from a few particulars he draws a conclusion which is general, he stands surprised whenever he sees anything that goes against his conclusion; like one who, having never seen any sheep except with short tails, is surprised at the sheep from Morocco which have long ones." Thus, we have the first recognition that surprise can result from the occurrence of perceived low or zero probability events. Second, Spinoza recognized that surprise should lead to learning and adaptation, which in turn should minimize future surprise, as he stated (or perhaps overstated) succinctly: "Surprise is never felt by him who draws true inferences." Third, surprise is not exclusively good or bad: "... it [surprise] arises either from ignorance or prejudice, is an imperfection in the man who is subject to this perturbance. I say an imperfection, because, through itself, surprise does not lead to any evil." In Spinoza's view, surprise is not necessarily valenced, but it should lead to "inferences," which adjust expectations and thereby minimize future surprise. These three points aptly summarize the state of affairs of the neuroscience of surprise, as we discuss in the remainder of the paper.

1.1. Rescorla–Wagner model

In the present day, we have moved beyond philosophical discussions of sheep to probability theory and computational neural models to describe surprise, which may be described in a more focused way as *prediction error*. Errors generally entail prediction errors that are aversive. Perhaps the best known use of error in associative learning is the Rescorla–Wagner (RW) model of classical conditioning. Briefly, the RW model is formalized as follows (Miller, Barnet, & Grahame, 1995; Rescorla & Wagner, 1972):

$$\Delta V_X^{n+1} = \alpha_X \beta_1 (\lambda_1 - V_{\text{total}}^n) \tag{1}$$

$$V_X^{n+1} = \Delta V_X^n + \Delta V_X^{n+1} \tag{2}$$

The intuition behind the RW model is admirable in its simplicity: The association V between a CS and US changes by (ΔV), increasing if the level of the US is greater than predicted, and decreasing if the US level is less than predicted. In that sense, ΔV may be thought of as an error term, that is, the difference between the maximum possible strength λ_1 and the current total strength V_{total}^n. The association V changes at a rate

defined by the product of α and β. The RW model was among the first to argue for a biological basis of learning based on surprise, building on earlier theories such as the Widrow–Hoff delta rule (Widrow & Hoff, 1960).

1.2. The Temporal Difference model (TD)

Beginning in the 1980s, a series of papers investigating models of associative learning extended the RW model from a trial-level into real-time, ultimately producing Temporal Difference (TD) learning (Barto, Sutton, & Anderson, 1983). TD learning (and its precursor models) integrated previous ideas, such as eligibility traces (Klopf, 1972), regarding the influence of the temporal characteristic of CSs and USs on learning. Perhaps the key innovation of TD learning was the reformulation of the RW error computation (Eq. 2). While reward in the trial-level RW model is effectively a single time point event, in real-time models of learning the duration of stimuli matters. The challenge is not only to predict the total level of reward but to account for how associations between a CS and US change based on the duration of the US. In order to do so, Sutton and Barto reimagined associations V as reflecting a discounted sum of all future levels of a US:

$$V(t) = E[r(t) + \gamma V(t+1)] \tag{3}$$

Here E is the expected value, r is reward, and γ is a discount factor between 0 and 1, in which smaller values of γ reflect a greater discounting, that is, more weighting toward recent reward experience. In order to learn this quantity, TD learning extends the RW learning rule to include not only the level of a US at a given time, but also a term reflecting the future predicted rewards.

$$\delta(t) = r(t) + \gamma \hat{V}(t+1) - \hat{V}t \tag{4}$$

Here δ is then multiplied by a learning rate and then added to update the value term V in a manner analogous to Eq. (2) above. The RW model is a special case of TD learning model, as can be seen when the discount value γ is set to 0 in Eq. (4), and thus accounts for all effects captured by the RW model. TD learning additionally accounts for effects not captured by the RW, or trial-level models in general, related to the temporal relationships of CSs and USs, including second-order conditioning and the development of negative associations for CSs occurring near the end of a US presentation (Sutton & Barto, 1990).

While the TD learning is of interest from a psychological standpoint in terms of providing a framework for investigating temporal effects on associative learning, it became a foundational model of modern neuroscience during the 1990s ("The Decade of the Brain") when the relationship between the TD prediction error term and the activity of dopamine (DA) neurons in the primate midbrain was observed. In a series of influential papers, Schultz et al. documented this relationship and potential functional interpretations

of the DA signal in driving learning about reinforcement (Schultz, 1998; Waelti, Dickinson, & Schultz, 2001). The activity profile of DA neurons signals reinforcement surprise: In the substantia nigra pars compacta and ventral tegmental area, activity is briefly greater at the moment when information indicates that a reward will be better than expected, and activity is briefly reduced at the moment when new information indicates that a reward will be worse than expected (Schultz, Dayan, & Montague, 1997).

The influence of the TD model, especially the TD prediction error, on neuroscience cannot be overstated. Besides delivering an elegant account of a single neuromodulatory system, the interpretation of DA neurons as providing a global signal of reward has driven a vast research enterprise focused on the role of value in the brain, informing investigation in a diverse range of subfields, including cognitive control, affective and clinical neuroscience, judgment and decision-making, perception and action, and many others. Likewise, TD learning has driven subsequent theoretical work in attempting to construct a "Grand Unification Theory" of other major neuromodulatory systems using the framework of surprise and error, linking their function to, for example, prediction and prediction errors of aversive events by serotonin (Daw, Kakade, & Dayan, 2002), or uncertainty about the environment, that is, both expected (acetylcholine) and unexpected (noradrenaline) (Yu & Dayan, 2005), all aspects of the underlying concept of surprise. We will see below that the core computation of surprise in the TD model may generalize to simulate effects in the ACC as well, but first we discuss theories of predictive coding as a context for models of ACC and related frontal cortex areas.

1.3. Predictive coding

While the functions of major neuromodulatory systems have been interpreted under the framework of surprise, prediction error computation is also frequently deployed to explain the role of a diverse array of cortical regions, from primary sensory cortices (Rao & Ballard, 1999) to language (Osterhout & Holcomb, 1992) to prefrontal regions involved in sophisticated cognitive processes (Gehring, Goss, Coles, Meyer, & Donchin, 1993; Gemba, Sasaki, & Brooks, 1986; Jessup, Busemeyer, & Brown, 2010). The apparent ubiquity of error signals throughout the brain has in turn led to additional theoretical work to develop yet another Grand Unification Theory, this time of neocortex. Proposed frameworks, including Predictive Coding (Rao & Ballard, 1999), Free Energy (Friston, 2010), and Hierarchical Bayesian Inference (Lee & Mumford, 2003), while differing in their details, use prediction and prediction error as the fundamental currency of communication among neurons and brain regions. Generally, these related approaches propose that the brain is organized hierarchically, with each hierarchical level attempting to predict the likely causes of input received by an inferior hierarchical level. Predictions that are not sufficient to explain input are used to compute error signals that are passed to superior hierarchical levels. This process of top-down prediction and bottom-up error signaling may be repeated an arbitrary number of times, producing a system composed of relatively simple learning mechanisms that nevertheless can engage in sophisticated processing of input.

Predictive coding accounts have primarily concerned themselves with the processing of sensory input, and they have been deployed to explain early (pre-cortical) visual processing, where the lateral and temporal antagonism of receptive fields in the retina and lateral geniculate nucleus (Huang & Rao, 2011), as well as biphasic responses of neurons in LGN (Jehee & Ballard, 2009), arise from predictive coding formulations. Predictive coding similarly provides an account of how receptive fields in early visual cortex (V1) may be learned (Jehee, Rothkopf, Beck, & Ballard, 2006), an explanation for extra-classical receptive fields throughout visual cortex (Rao & Ballard, 1999), binocular rivalry as the need to account for error due to multiple possible percepts (Hohwy, Roepstorff, & Friston, 2008), the processing of complex stimuli in high-level visual areas (Egner, Monti, & Summerfield, 2010), and the response of regions of frontal cortex in coding predictive information regarding prospective percepts (Summerfield et al., 2006). Although less well elaborated, predictive coding has likewise been offered as an explanatory framework for processing in auditory cortex (Huang & Rao, 2011) as well as the structure of primary motor cortex (Shipp, Adams, & Friston, 2013). More recent work has argued that predictive coding could be structured into hierarchical representations (Clark, 2013), a point to which we return below.

1.4. Frontal cortex and predictive coding

The success of predictive coding and related approaches in accounting for neural data is compelling, as well as suggestive of a general organizing principle of the brain. However, it is unclear how the motif of top-down prediction and bottom-up error signals might be extended to account for the function of frontal regions involved in higher level cognition. These approaches have been most successfully applied to explain results in sensory and motor cortices, regions of the brain where neural activity is most closely related to concrete and observable phenomena: perceptual input on the one end, and overt action on the other. Speculation that hierarchical predictive coding schemes might also be deployed to explain more abstract, cognitive behaviors rests largely on the premise of a repeating motif of prediction and error that can be extended indefinitely. The implicit assumption, then, is that by layering prediction error loops "enough" times, one will arrive at a system capable of complex cognition.

While it is unclear why or how a predictive coding framework with a sufficient number of hierarchical levels should eventually result in a cognitive system, there is substantial reason to believe that the frontal lobes—the region of the brain generally regarded as critical for cognition—may implement some version of predictive coding. Prefrontal cortex is generally thought to be organized along a rostrocaudal abstraction gradient, with activity in caudal regions associated with concrete stimulus-response associations, whereas rostral regions are more typically implicated in maintaining abstract information related to goals, rules, and task sets (Badre & D'Esposito, 2009; Koechlin, Ody, & Kouneiher, 2003). While the observed hierarchical organization of PFC, especially its lateral aspects, hints at the possibility that a predictive coding scheme may be utilized by

the frontal lobes, a critical component of such a framework is the calculation of prediction errors used to drive learning throughout the hierarchy.

1.5. Anterior cingulate cortex and the PRO model

A wealth of findings have established anterior cingulate cortex (ACC) and surrounding medial prefrontal cortex (mPFC) as a region critically involved in prediction and processing error. Findings from neurophysiological studies in monkeys observed activity in single neurons in ACC associated with the anticipation of an imminent reward (Shidara & Richmond, 2002), as well as the occurrence of behavioral error or when a reward was surprisingly withheld (Gemba et al., 1986; Ito, Stuphorn, Brown, & Schall, 2003). Evidence from EEG and fMRI studies in humans has likewise implicated the region in signaling behavioral error (Gehring et al., 1993; Hohnsbein, Falkenstein, & Hoorman, 1989; Kiehl, Liddle, & Hopfinger, 2000), resolving response conflict (MacDonald, Cohen, Stenger, & Carter, 2000), predicting the likelihood of error and indicating the unexpectedness of an error (Brown & Braver, 2005), and several others. While the array of disparate findings regarding cingulate seemed to indicate a diversity of functions and operating modes subserved by ACC, the emphasis on interpreting cingulate activity in terms of, on one hand, signaling aversive events such as pain (Lieberman & Eisenberger, 2015; Wager et al., 2016), error, and behavioral conflict or, on the other hand, appetitive events such as reward prediction and reward detection (Ito et al., 2003), suggested that the primary function of ACC might not depend on the valence of events. In simpler terms, ACC doesn't care about "good" or "bad" events, but only if the event was expected, and whether or not the event occurred. This intuition, that ACC attempts to both predict and evaluate the surprisingness of all outcomes, positive and negative, led to the development of the Predicted Response-Outcome (PRO) model of ACC (Alexander & Brown, 2010, 2011).

1.6. The PRO model

The PRO model (Fig. 1) as originally published casts the ACC as learning to predict the likely outcomes of actions, regardless of their affective import, and signaling in particular the unexpected non-occurrences of predicted events. Initially, the model was aimed at accounting for results from the cognitive control literature, including effects of conflict, error likelihood, conflict, and the surprising occurrence of error. To account for these results, the model exploited formulations from RL theory, especially the TD Learning model described above. As in TD learning, the PRO model computes a temporal prediction error

$$\delta_{i,t} = r_{i,t} + \gamma V_{i,t+1} - V_{i,t} \tag{5}$$

in order to train predictions. The PRO model extends the TD model in two key ways. First, rather than predicting a scalar quantity such as value—which, in RL represents the

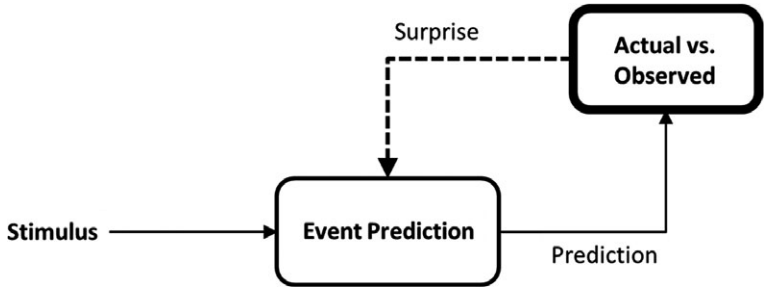

Fig. 1. The PRO model. In the PRO model, predictions regarding likely future events are associated with relevant stimuli. Deviations from predictions ("surprise") drives learning as in several models of associative learning.

combined, discounted sum of future rewards obtainable from a current state—the PRO model learns a vector-valued prediction of future states that may be associated with actions given a current state. In effect, it predicts the whole distribution of likely outcomes as a vector of predictions, rather than collapsing to the scalar mean as in RL. Second, whereas positive and negative prediction errors in RL are associated with appetitive and aversive events, respectively, the PRO model jettisons valence and instead signals events that occurred unexpectedly ("positive surprise") and events that were predicted but failed to occur ("negative surprise"). These provide a signal akin to an unsigned prediction error in RL. Single-unit studies in monkey suggest that neurons in ACC signal both kinds of surprise—an unexpected reward or an unexpected punishment may cause separate neurons in ACC to fire phasically (Ito et al., 2003)—both examples of positive surprise. Negative surprise, on the other hand, requires neurons that exhibit increased activity when an event fails to occur as expected, as has been found previously in the medial PFC (Alexander & Brown, 2011; Amador, Schlag-Rey, & Schlag, 2000). In the absence of overt sensory cues, neurons showing increased activity following surprising omissions must have access to information that allows them to predict both the nature and timing of the event, that is, expected to happen—exactly the kind of information trained in TD learning. In ACC, single-unit activity has been observed in which firing rates "ramp up" in the periods prior to the occurrence of a trained (affectively positive or negative) outcome (Shidara & Richmond, 2002). According to the PRO model (Alexander & Brown, 2011), if the outcome occurs as expected, the firing of that unit is suppressed, while if the outcome is withheld, the unit's activity peaks slightly after the time when the outcome was expected and gradually decays. Positive and negative surprise tend to be positively correlated in most experimental paradigms—the surprising absence of "correct" feedback (negative surprise), for example, is often accompanied by the surprising occurrence of "error" feedback" (positive surprise). However, units related to negative surprise signals additionally have access to temporal information, allowing them to capture a broader range of effects related to anticipation and prediction in ACC.

The PRO model thus offers a compelling reconciliation of findings linking ACC primarily with reward processing (Hayden & Platt, 2010; Ito et al., 2003) with those finding

effects primarily related to error and conflict (Botvinick, Braver, Carter, Barch, & Cohen, 1998; Botvinick, Nystrom, Fissel, Carter, & Cohen, 1999), despite concerns that such a reconciliation may not have been possible (Cole, Yeung, Freiwald, & Botvinick, 2009). This reconciliation is achieved by assigning the calculation of surprise as the principal function of ACC. Rather than assuming that ACC is composed of a number of functional modules devoted to processing conflict, signaling error, signaling reward, registering pain, etc., to explain its involvement across a range of neuroimaging studies, the PRO model suggests that ACC activity is routinely observed by virtue of a single function, surprise calculation, that is, applied generally across modalities and paradigms (Alexander & Brown, 2014).

1.7. Unifying prefrontal cortex

The proposed role of ACC in signaling valence-neutral prediction error as its main function suggests one possible avenue by which predictive coding might be extended into PFC. Given the wealth of effects accounted for by the PRO model, any effort to interpret PFC function under the predictive coding framework is likely to map bottom-up error signaling to ACC/mPFC. However, to satisfy the requirements of predictive coding formulations, two conditions must be met. First, under predictive coding, multiple errors signals are generated corresponding to hierarchical levels—errors at lower levels are more closely related to concrete events, while higher levels signal abstract errors. Second, predictions at lower levels are contextualized by top-down predictions that explain away input that leads to prediction errors.

With regard to the first condition, recent evidence suggests that distinct regions within ACC and mPFC may support hierarchical error processing. Evidence from studies on humans and monkeys demonstrates that ACC follows a dual topography (Amiez & Petrides, 2012; Procyk et al., 2014) in which, within a particular ACC region, the activity of distinct subregions corresponds to an anatomical map of the body during feedback processing, and this anatomical map is repeated at distinct loci along the rostrocaudal axis, suggestive of a repeated, hierarchical organization. More direct evidence from neuroimaging studies in which the degree of abstraction of an error was manipulated demonstrates spatially distinct error activity within ACC, with more abstract errors processed more rostrally than concrete errors (Fig. 2) (Kim, Johnson, Cilles, & Gold, 2011; Zarr & Brown, 2016). These findings suggest that the essential function ACC proposed by the PRO model of surprise calculation may apply at multiple regions within cingulate.

With regard to the second condition, ACC may provide surprise signals to a number of other brain regions, which may in turn provide information critical to explaining the causes of prediction error. In this respect, dlPFC is a likely modeling target. DLPFC is generally recognized as a region concerned with representing rules and task sets beyond simple associations, and it is heavily implicated in maintaining items in working memory over protracted periods of time. Furthermore, dlPFC is known to be densely and reciprocally connected with ACC (Barbas & Pandya, 1989), suggesting that the interaction between these two regions is vital for supporting behavior and cognition. Indeed,

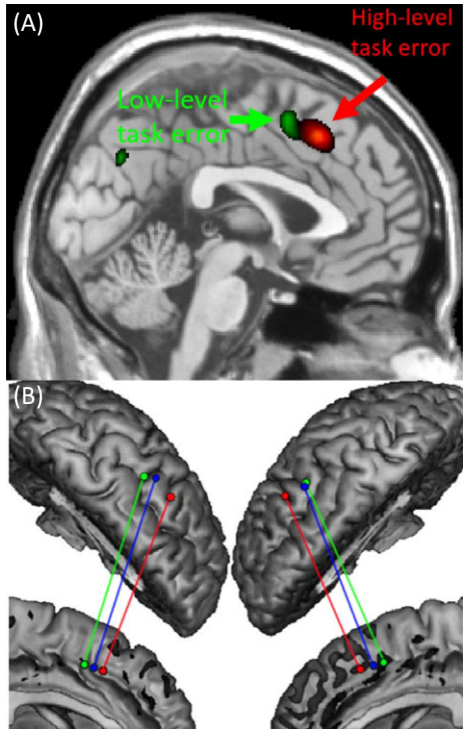

Fig. 2. Hierarchies of prediction error in the mPFC. (A) In a hierarchical task, higher level task errors led to more anterior medial PFC activation, but lower level task errors led to more posterior medial PFC activation. (B) The mPFC areas showing hierarchical error. Adapted with permission from Zarr & Brown (2016).

coactivation of ACC and dlPFC is routinely observed in neuroimaging studies, and the two regions are core components of the cognitive control network (Dosenbach et al., 2007). The DLPFC shows sustained activity ostensibly related to working memory (WM) (Niki & Watanabe, 1979), perhaps for strategies more than for stimuli (Nee & Brown, 2012; Riggall & Postle, 2012). Neuroimaging studies of subjects performing psychological tasks known to involve significant WM demands reliably elicit activity over broad regions of lPFC (Nee et al., 2013). Rather than acting as a simple buffer that retains information that may be needed at a later time, however, dlPFC activity is additionally sensitive to task demands. Distinct regions within dlPFC are differentially activated depending on how "abstract" an item that must be stored in WM is, with rostral regions showing increased activation for high-level context variables, and caudal regions responding to more concrete stimuli (Badre & D'Esposito, 2007; Koechlin et al., 2003; Nee & Brown, 2012; Nee, Jahn, & Brown, 2014).

1.8. The HER model

The above observations suggest a possible synthesis consistent with the predictive coding framework: The *error representation hypothesis* states that multi-dimensional error

signals generated by ACC are used to train prediction error representations in dlPFC (i.e., representations that predict the prediction errors), which are then trained to be activated by task-relevant stimuli. Subsequent presentations of a stimulus elicit activity representing the expected prediction error generated by ACC, and this activity is used to modulate predictions in order to support adaptive behavior. The error representation hypothesis was formalized in a recent computational model of mPFC and dlPFC, the *Hierarchical Error Representation* (HER) model (Alexander & Brown, 2015), which elaborates the prediction and error representations functions proposed by the error representation hypothesis.

1.9. Basic HER model principles

The HER model (Fig. 3) is an extension of the PRO model of ACC/mPFC, using the architecture of the PRO model as a computational motif that is repeated to form hierarchical levels. Indeed, at the core level, the HER model and the PRO model are identical in their function and account for the same effects described above, using the same principles of prediction of likely events and discrepancies between predictions and observations. Each additional hierarchical level engages in exactly the same processes of prediction and prediction error computation. However, while the predictions and prediction errors at the base level (and in the PRO model) are computed based on observed sensory events, additional hierarchical levels use the error signal generated by the immediately lower level as a kind of proxy outcome—rather than learning to predict

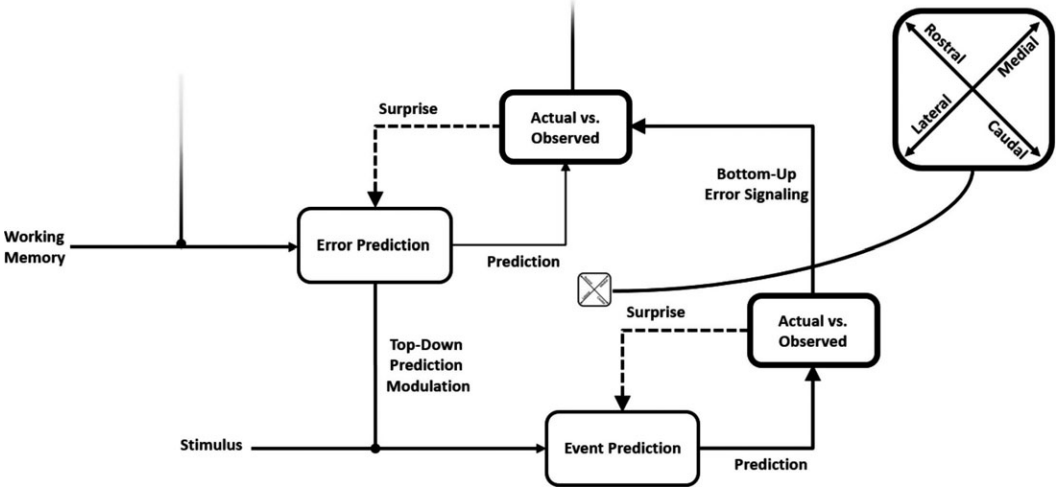

Fig. 3. The HER model. The HER model uses the prediction/error function of the PRO model (cf. Fig. 1) as a computational motif, that is, hierarchically iterated. At lower hierarchical layers, predictions and errors related to concrete event-event associations. Errors generated at the lower layers are used as outcome signals for higher layers, which learn predictions regarding the expected errors reported by lower layers. These error predictions are associated with task-relevant stimuli maintained in working memory. Subsequently, reactivation of these error predictions can be used to modulate predictive activity at lower layers in order to refine predictions regarding likely outcomes.

events, hierarchical levels in the HER model learn to predict errors resulting from the predictions of lower levels—essentially predicting the prediction errors. This motif can be repeated an arbitrary number of times, with increasingly abstract error signals passed upward through the hierarchy.

The error predictions elicited at each level during a task are passed to the next *lowest* hierarchical level in order to modulate the predictions of the lower level. Because the model learns the likely errors that are reported by a lower hierarchical level, top-down modulation of predictions by prediction error representations improves the model's estimate of likely events when a particular prediction error can be reliably predicted by a stimulus. This stands to reason insofar as a system that can predict its own failures can also modify its behavior to avoid those failures.

The HER model's treatment of prediction errors can thus be seen as building on a number of earlier models that incorporate surprise (Table 1). Among the earliest biological models was the RW model, but the surprise signals in that model were somewhat crude as they did not correspond to the prior probability of the actual outcome. The TD model calculated surprise on the difference between the magnitude of the actual versus predicted reward, and probability is included only implicitly in that the predictions reflect the expected value of reward, which includes probability. The PRO model (and, by extension, the HER model) incorporates surprise in several ways, both as a modified TD error and a prediction error similar to a delta rule (Widrow & Hoff, 1960). This variety of surprise signals in turn provides a rich basis for effects that can be driven by surprise, including updated predictions, modulation of learning rates, the initiation of foraging, and other reactive control events (Table 1).

While a detailed discussion of the mechanisms by which the HER model operates is beyond the scope of this paper, the principal mechanisms discussed above—bottom-up

Table 1
Model surprise signals

Surprise Signal	Effect of Surprise	References
Rescorla–Wagner $\Delta V_X^{n+1} = \alpha_X \beta_1 (\lambda_1 - V_{\text{total}}^n)$	Update CS-US association strength	(Miller et al., 1995; Rescorla & Wagner, 1972)
Temporal Difference $\delta(t) = r(t) + \lambda \hat{V}(t+1) - \hat{V}(t)$	Update prediction of scalar reward value	(Barto et al., 1983)
PRO model negative (i.e., omission of expected event) surprise $\omega_t^N = \sum_i \lfloor V_{i,t} - O_{i,t} \rfloor^+$	Modulate learning rate for proactive control action-outcome associations (Could also drive reactive control, foraging, or strategy switch events)	(Alexander & Brown, 2011, 2014)
PRO model vector Temporal Difference $\delta_{i,t} = r_{i,t} + \gamma V_{i,t+1} - V_{i,t}$	Update vector of prediction of likely events	(Alexander & Brown, 2011, 2014)
HER model prediction error $\mathbf{e} = \mathbf{a}(\mathbf{o} - \mathbf{p})$	Update stimulus-outcome predictions Update working memory gating probability Serve as event to be predicted by higher levels of the HER model	(Alexander & Brown, 2015) (Widrow & Hoff, 1960)

error signaling, top-down prediction modulation—allow the model not only to perform a wide range of tasks reported in the literature but to learn these tasks autonomously and in a manner consistent with human behavior during learning, even performing comparably well or better than current machine learning methods in some cases. Moreover, the HER model can provide a number of testable predictions regarding empirical effects in the frontal lobe (Alexander & Brown, 2015). The HER model explains activity in dlPFC as the maintenance, update, and modulation of prediction error representations, and it captures effects observed in dlPFC including differences in activity related to the information content of contextual cues, signals related to tonic and transient activity during WM maintenance and updating, effects of temporal and relational abstraction, mismatch enhancement and suppression in single neurons during a WM task, the influence of dlPFC lesions on the ERN, and many more (Gehring & Knight, 2000; Koechlin et al., 2003; Miller, Erickson, & Desimone, 1996; Nee et al., 2014; Reynolds, O'Reilly, Cohen, & Braver, 2012).

The HER model provides a novel perspective on existing subregional parcellations of the medial PFC. First, the finding of a rostro-caudal hierarchy of prediction errors (Fig. 2) (Zarr & Brown, 2016) fits with the HER model predictions as well as with a broader set of findings that the medial PFC has a number of functional subregions. In particular, the finding that more abstract task errors lead to more rostral activation than more concrete response errors (Desmet, Fias, Hartstra, & Brass, 2011) accords well with the HER model predictions and our fMRI findings (Zarr & Brown, 2016). Other studies have found that the same regions involved in prediction error are also involved in anti-tasks such as a countermanding task (Nachev, Rees, Parton, Kennard, & Husain, 2005), with a rostro-caudal distinction such that more posterior regions are more active when an action must be freely chosen as an act of volition rather than simply generated as instructed by the experimenter. Both effects of anti-task responding and free choice may be accounted for as prediction error effects, as follows. For anti-tasks, it has been found that in oculomotor tasks, the dorsal medial PFC in macaque monkeys does not respond in time to play a causal role in driving movement (Stuphorn, Brown, & Schall, 2010). This suggests that the dorsal medial PFC responds to the surprise of the anti-task cue but may not drive anti-task performance. Likewise, we have found that when a response must be freely chosen as an act of volition, the act of choosing may entail not only the choice-related activation but also internal predictions about which choice will be made, which leads to greater medial PFC activation, especially more posteriorly (Jahn, Nee, Alexander, & Brown, 2014). This is consistent with earlier findings (Nachev et al., 2005).

Surprise signals in the medial PFC may occur particularly more dorsally. Beyond the rostro-caudal distinctions, here also appears to be a dorsal-ventral distinction in the medial PFC in which dorsal regions (straddling the pre-SMA and dorsal ACC, Brodmann's area 32) represent prediction error, while the more ventral regions (BA 24/32) represent pain and control signals (Jahn, Nee, Alexander, & Brown, 2016). The dorsal regions show apparent task difficulty effects (Shenhav, Straccia, Cohen, & Botvinick, 2014) and have been implicated in licensing effortful behavior (Holroyd & Yeung, 2012; Shenhav, Botvinick, & Cohen, 2013; Verguts, Vassena, & Silvetti, 2015), but these may actually reflect

prediction error signals rather than difficulty or effort per se, consistent with the PRO and HER models (Brown & Alexander, 2017; Vassena, Deraeve, & Alexander, 2017). These prediction error (i.e., surprise) signals can effectively slow responding (Forstmann, van den Wildenberg, & Ridderinkhof, 2008; Wessel & Aron, 2017). The more ventral regions may be involved in more proactive control (Braver, Gray, & Burgess, 2007) and in driving foraging behavior as well (Kolling, Behrens, Mars, & Rushworth, 2012; Kolling, Behrens, Wittmann, & Rushworth, 2016). This dorsal/ventral distinction may account for regional differences in conflict versus error signals, with error, that is, prediction error found more dorsally and conflict signals found more ventrally (Desmet et al., 2011).

2. Conclusion

More generally, the HER model suggests that prediction error, long a critical component of theories regarding behavior and neural function, is not merely a mechanism that is useful for driving learning about the manner in which the world works. While countless theories and computational models take error calculation and minimization as key components of a learning system, prediction error itself tends to be only a means to an end, a cost function to be minimized in order to learn "rules" or "task sets." Instead of taking the representation of such quantities as given, that there exist specific units whose activity not only represents but also instantiates a particular rule governing behavior to avoid error, the HER model proposes that rules themselves are error representations. From this perspective, then, error and surprise become the fundamental neural currency that is elaborated through multiple stages from early sensory processing to high-level cognitive behaviors. The HER model further supplies an existence proof for theories that propose a unifying framework for understanding the organization of neocortex such as predictive coding. Sophisticated behavior in the model derives from a hierarchically iterated motif of prediction and prediction error computations that engages in passing error information up the hierarchy, and prediction information (in the form of expected prediction error) downward. The HER model thus fills a critical void in our understanding the steps by which the brain processes information from the earliest sensory areas in order to drive goal-directed behavior.

Acknowledgments

WHA was supported in part by FWO-Flanders Odysseus II Award #G.OC44.13N.

References

Alexander, W. H., & Brown, J. W. (2010). Computational models of performance monitoring and cognitive control. *Topics in Cognitive Science*, *2*, 658–677.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14* (10), 1338–1344. https://doi.org/10.1038/nn.2921.

Alexander, W. H., & Brown, J. W. (2014). A general role for medial prefrontal cortex in event prediction. *Frontiers in Computational Neuroscience*, *8*, 1–11. https://doi.org/10.3389/fncom.2014.00069.

Alexander, W. H., & Brown, J. W. (2015). Hierarchical error representation: A computational model of anterior cingulate and dorsolateral prefrontal cortex. *Neural Computation*, *27*, 2354–410. https://doi.org/10.1162/NECO_a_00779.

Amador, N., Schlag-Rey, M., & Schlag, J. (2000). Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field. *Journal of Neurophysiology*, *84* (4), 2166–2170. Available at http://jn.physiology.org/content/84/4/2166.short. Accessed Oct. 27, 2017.

Amiez, C., & Petrides, M. (2012). Neuroimaging evidence of the anatomo-functional organization of the human cingulate motor areas. *Cerebral Cortex*, *56*, 3–78. https://doi.org/10.1093/cercor/bhs329.

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *19*(12), 2082–99. https://doi.org/10.1162/jocn.2007.19.12.2082.

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, *10*(9), 659–69. https://doi.org/10.1038/nrn2667.

Barbas, H., & Pandya, D. N. (1989). Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. *The Journal of Comparative Neurology*, *286*, 353–375. Available at http://onlinelibrary.wiley.com/doi/10.1002/cne.902860306/abstract. Accessed Oct. 27, 2017

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, & Cybernetics*, *13* (5), 834–846.

Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. C. (1998). Evaluating the demand for control: Anterior cingulate cortex and conflict monitoring (Report). (Technical Report 98.1) Pittsburgh, PA: Center for the Neural Basis of Cognition.

Botvinick, M. M., Nystrom, L., Fissel, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402* (6758), 179–181.

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In C. J. A. Conway, M. Kane, A. Miyake, & J. Towse (Ed.), *Variation of working memory*. Oxford, UK: Oxford University Press.

Brown, J. W., & Alexander, W. H. (2017). Foraging value, risk avoidance, and multiple control signals: How the anterior cingulate cortex controls valuebased decision-making. *Journal of Cognitive Neuroscience*, *29*, 1656–1673.

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307* (5712), 1118–1121. https://doi.org/307/5712/1118[pii]10.1126/science.1105783

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477.

Cole, M. W., Yeung, N., Freiwald, W.a, & Botvinick, M. (2009). Cingulate cortex: Diverging data from humans and monkeys. *Trends in Neurosciences*, *32*(11), 566–74. https://doi.org/10.1016/j.tins.2009.07.001.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15* (4-6), 603–616. https://doi.org/s0893-6080(02)00052-7[pii]

de Spinoza, B. (1910). *Short treatise on god, man, and his well-being*. A. Wolf (Ed.). London: A and C Black.

Desmet, C., Fias, W., Hartstra, E., & Brass, M. (2011). Errors and conflict at the task level and the response level. *Journal of Neuroscience*, *31*(4), 1366–1374. https://doi.org/10.1523/JNEUROSCI.5371-10.2011.

Dosenbach, N. U., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A., & ... Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *104* (26), 11073–11078. https://doi.org/0704320104[pii]10.1073/pnas.0704320104

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(49), 16601–8. https://doi.org/10.1523/JNEUROSCI.2770-10.2010.

Forstmann, B. U., van den Wildenberg, W. P. M., & Ridderinkhof, K. R. (2008). Neural mechanisms, temporal dynamics, and individual differences in interference control. *Journal of Cognitive Neuroscience*, *20*(10), 1854–1865. https://doi.org/10.1162/jocn.2008.20122.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*, 385–390.

Gehring, W. J., & Knight, R. T. (2000). Prefrontal-cingulate interactions in action monitoring. *Nature Neuroscience*, *3* (5), 516–520. Available at: https://doi.org/10.1038/74899. Accessed Oct. 27, 2017

Gemba, H., Sasaki, K., & Brooks, V. B. (1986). "Error" potentials in limbic cortex (anterior cingulate area 24) of monkeys during motor learning. *Neuroscience Letters*, *70* (2), 223–227. Available at: https://doi.org/10.1016/0304-3940(86)90467-2. Accessed Oct. 27, 2017

Hayden, B. Y., & Platt, M. L. (2010). Neurons in anterior cingulate cortex multiplex information about reward and action. *Journal of Neuroscience*, *30* (9), 3339–3346. https://doi.org/30/9/3339[pii]10.1523/jneurosci.4874-09.2010.

Hohnsbein, J., Falkenstein, M., & Hoorman, J. (1989). Error processing in visual and auditory choice reaction tasks. *Journal of Psychophysiology*, *3*, 32.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, *16*(2), 122–8. https://doi.org/10.1016/j.tics.2011.12.008.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2* (5), 580–593. https://doi.org/10.1002/wcs.142.

Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance monitoring by anterior cingulate cortex during saccade countermanding. *Science*, *302*, 120–122.

Jahn, A., Nee, D. E., Alexander, W. H., & Brown, J. W. (2014). Distinct regions of anterior cingulate cortex signal prediction and outcome evaluation. *NeuroImage*, *95*, 80–9. https://doi.org/10.1016/j.neuroimage.2014.03.050.

Jahn, A., Nee, D. E., Alexander, W. H., & Brown, J. W. (2016). Distinct regions within medial prefrontal cortex process pain and cognition. *The Journal of Neuroscience*, *36*(49), 12385–12392. https://doi.org/10.1523/JNEUROSCI.2180-16.2016.

Jehee, J. F. M., & Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Computational Biology*, *5*(5), e1000373. https://doi.org/10.1371/journal.pcbi.1000373.

Jehee, J. F. M., Rothkopf, C., Beck, J. M., & Ballard, D. H. (2006). Learning receptive fields using predictive feedback. *Journal of Physiology-Paris*, *100*(1–3), 125–132. https://doi.org/10.1016/j.jphysparis.2006.09.011.

Jessup, R. K., Busemeyer, J. R., & Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *Journal of Neuroscience*, *30* (9), 3467–3472.

Kiehl, K. A., Liddle, P. F., & Hopfinger, J. B. (2000). Error processing and the rostral anterior cingulate: An event-related fMRI study. *Psychophysiology*, *37*, 216–223.

Kim, C., Johnson, N. F., Cilles, S. E., & Gold, B. T. (2011). Common and distinct mechanisms of cognitive flexibility in prefrontal cortex. *Journal of Neuroscience*, *31*(13), 4771–4779. https://doi.org/10.1523/JNEUROSCI.5923-10.2011.

Klopf, A. (1972). Brain function and adaptive systems – A heterostatic theory. Air Force Cambridge Res. Lab. Res. Rep., AFCRL-72-0164.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, *302*(5648), 1181–5. https://doi.org/10.1126/science.1088545.

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, *336*(6077), 95–98. https://doi.org/10.1126/science.1216930.

Kolling, N., Behrens, T., Wittmann, M., & Rushworth, M. (2016). Multiple signals in anterior cingulate cortex. *Current Opinion in Neurobiology*, *37*, 36–43. https://doi.org/10.1016/j.conb.2015.12.007.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434. https://doi.org/10.1364/JOSAA.20.001434.

Lieberman, M. D., & Eisenberger, N. I. (2015). The dorsal anterior cingulate cortex is selective for pain: Results from large-scale reverse inference. *Proceedings of the National Academy of Sciences*, *112*(49), 15250–15255. https://doi.org/10.1073/pnas.1515083112.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal cortex and anterior cingulate cortex in cognitive control. *Science*, *288*, 1835–1838.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117* (3), 363–86. Available at http://www.ncbi.nlm.nih.gov/pubmed/7777644. Accessed Oct. 27, 2017

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16* (16), 5154–5167.

Nachev, P., Rees, G., Parton, A., Kennard, C., & Husain, M. (2005). Volition and conflict in human medial frontal cortex. *Current Biology*, *15*(2), 122–128. https://doi.org/10.1016/j.cub.2005.01.006.

Nee, D. E., & Brown, J. W. (2012). Rostral-caudal gradients of abstraction revealed by multi-variate pattern analysis of working memory. *NeuroImage*, *63*(3), 1285–94. https://doi.org/10.1016/j.neuroimage.2012.08.034.

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A meta-analysis of executive components of working memory. *Cerebral Cortex*, *23*(2), 264–282. https://doi.org/10.1093/cercor/bhs007.

Nee, D. E., Jahn, A., & Brown, J. W. (2014). Prefrontal cortex organization: Dissociating effects of temporal abstraction, relational abstraction, and integration with fMRI. *Cerebral Cortex*, https://doi.org/10.1093/cercor/bht091.

Niki, H., & Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Research*, *171* (2), 213–224. Available at: https://doi.org/10.1016/0006-8993(79)90328-7 Accessed Oct. 27, 2017

Osterhout, L., & Holcomb, P. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*, 785–806.

Procyk, E., Wilson, C. R. E., Stoll, F. M., Faraut, M. C. M., Petrides, M., & Amiez, C. (2014). Midcingulate motor map and feedback detection: Converging data from humans and monkeys. *Cerebral Cortex*, *26*, 467–476. https://doi.org/10.1093/cercor/bhu213.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. & Prokasy, W. F. (Eds.) *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton- Century-Crofts.

Reynolds, J. R., O'Reilly, R. C., Cohen, J. D., & Braver, T. S. (2012). The function and organization of lateral prefrontal cortex: A test of competing hypotheses. *PLoS ONE*, *7*(2), e30284. https://doi.org/10.1371/journal.pone.0030284.

Riggall, A. C., & Postle, B. R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(38), 12990–8. https://doi.org/10.1523/JNEUROSCI.1892-12.2012.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80* (1), 1–27. Available at: http://jn.physiology.org/cgi/content/full/80/1/1. Accessed Oct. 27, 2017

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–40. https://doi.org/10.1016/j.neuron.2013.07.007.

Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Neuroscience*, https://doi.org/10.1038/nn.3771.

Shidara, M., & Richmond, B. J. (2002). Anterior cingulate: Single neuronal signals related to degree of reward expectancy. *Science*, *296* (5573), 1709–1711. Available at: https://doi.org/10.1126/science.1069504. Accessed Oct. 27, 2017.

Shipp, S., Adams, R. A., & Friston, K. J. (2013). Reflections on agranular architecture: Predictive coding in the motor cortex. *Trends in Neurosciences*, *36*(12), 706–716. https://doi.org/10.1016/j.tins.2013.09.004.

Stuphorn, V., Brown, J. W., & Schall, J. D. (2010). Role of supplementary eye field in saccade initiation: Executive, not direct, control. *Journal of Neurophysiology*, *103* (2), 801–816. https://doi.org/00221.2009 [pii]10.1152/jn.00221.2009.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science*, *314*(5803), 1311–1314. https://doi.org/10.1126/science.1132028.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press.

Vassena, E., Deraeve, J., & Alexander, W. H. (2017). Predicting motivation: Computational models of PFC can explain neural coding of motivation and effort-based decision-making in health and disease. *Journal of Cognitive Neuroscience*, *29*, 1633–1645.

Verguts, T., Vassena, E., & Silvetti, M. (2015). Adaptive effort investment in cognitive and physical tasks: A neurocomputational model. *Frontiers in Behavioral Neuroscience*, *9*, https://doi.org/10.3389/fnbeh.2015.00057.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412* (6842), 43–48. Available at: https://doi.org/10.1038/35083500. Accessed Oct. 27, 2017

Wager, T. D., Atlas, L. Y., Botvinick, M. M., Chang, L. J., Coghill, R. C., Davis, K. D., & Yarkoni, T. (2016). Pain in the ACC? *Proceedings of the National Academy of Sciences*, *113*(18), E2474–E2475. https://doi.org/10.1073/pnas.1600282113.

Wessel, J. R., & Aron, A. R. (2017). On the globality of motor suppression: Unexpected events and their influence on behavior and cognition. *Neuron*, *93*(2), 259–280. https://doi.org/10.1016/j.neuron.2016.12.013.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. *1960 IRE WESCON Convention Record* (pp. 96–104). New York: IRE.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46* (4), 681–692. https://doi.org/s0896-6273(05)00362-4[pii]10.1016/j.neuron.2005.04.026.

Zarr, N., & Brown, J. W. (2016). Hierarchical error representation in medial prefrontal cortex. *NeuroImage*, *124*, 238–247. https://doi.org/10.1016/j.neuroimage.2015.08.063.