Chapter 16
Reciprocal Interactions of Computational Modeling and Empirical Investigation

William H. Alexander and Joshua W. Brown

Abstract Models in general, and computational neural models in particular, are useful to the extent they fulfill three aims, which roughly constitute a life cycle of a model. First, at birth, models must account for existing phenomena, and with mechanisms that are no more complicated than necessary. Second, at maturity, models must make strong, falsifiable predictions that can guide future experiments. Third, all models are by definition incomplete, simplified representations of the mechanisms in question, so they should provide a basis of inspiration to guide the next generation of model development, as new data challenge and force the field to move beyond the existing models. Thus the final part of the model life cycle is a dialectic of model properties and empirical challenge. In this phase, new experimental data test and refine the model, leading either to a revised model or perhaps the birth of a new model. In what follows, we provide an outline of how this life cycle has played out in a particular series of models of the dorsal anterior cingulate cortex (ACC).

16.1 Introduction

A popular, though probably apocryphal, characterization of the geocentric model of the solar system is that, before it was replaced by the heliocentric model, it required "epicycles on epicycles on epicycles" in order to describe the movement of the planets and stars through the sky. Initial attempts explained the path of these heavenly bodies as revolving about the earth at a fixed distance along celestial spheres. While this simple model was sufficient to describe a majority of the data available, it was observed that certain planets exhibited retrograde motion, appearing to reverse the

J. W. Brown (✉)
Department of Psychological and Brain Sciences,
Indiana University, Bloomington, USA
e-mail: jwmbrown@indiana.edu

W. H. Alexander
Department of Experimental Psychology,
Ghent University, Henri Dunantlaan 2, B-9000 Gent, Belgium
e-mail: william.alexander@ugent.be

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_16

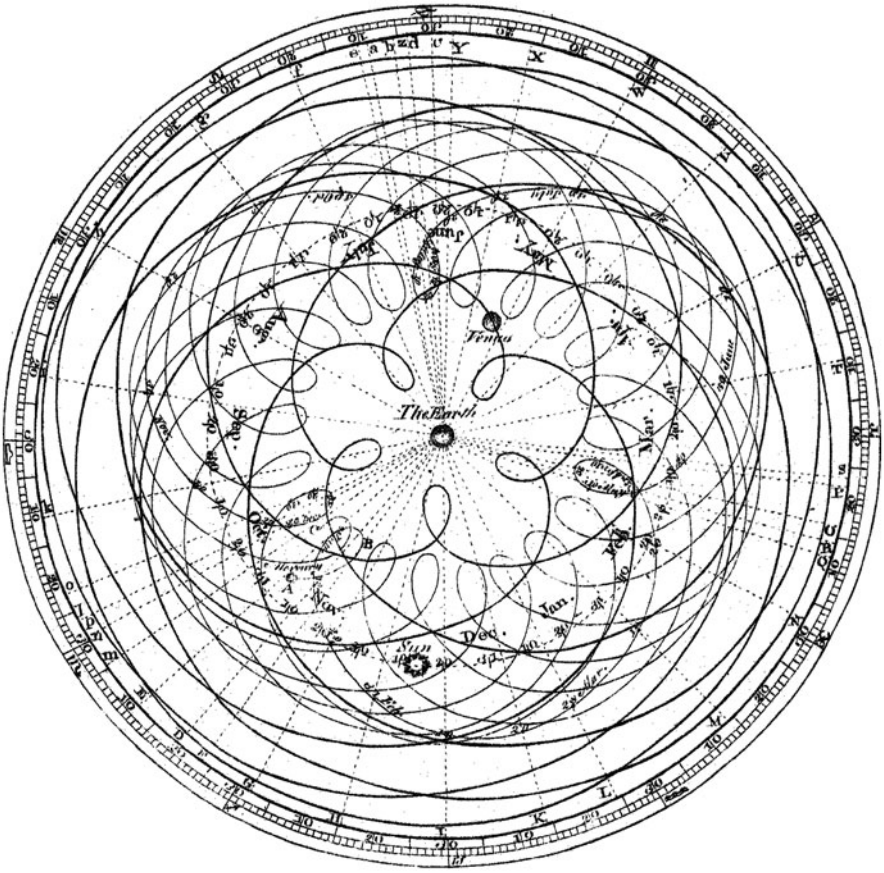

Fig. 16.1 The Ptolemaic model of the solar system. (Public domain)

direction of their path at certain points. In order to account for these changes, Ptolemy introduced epicycles into the geocentric model: in addition to following a circular path around the earth, planets also followed a second revolution around that path. Although epicycles could explain the apparent retrograde motion, the system was imperfect. In order to explain further anomalies between the Ptolemaic system and observations of the planets, the story goes, it was necessary to include ever more epicycles into the model, multiplying the complexity of the system for only modest gains in explanatory ability. Ultimately, as Kuhn argued, scientific progress came as the Copernican heliocentric model replaced the Ptolemaic model in a revolutionary (rather than evolutionary) way [1] (Fig. 16.1).

In a comparable manner to the Ptolemaic model of the solar system, the advent of sophisticated brain imaging techniques has advanced empirical knowledge at a pace that has outrun model building. Rather than epicycles, however, neuroimaging studies appear to assign ever more functions and modules to areas of the brain that show

increased BOLD activity ($p < 0.05$, corrected or not) for one condition over another. Early work in cognitive neuroscience sought to identify areas of the brain supporting cognitive processes whose existence had been inferred through psychological experimentation [2], and exuberant studies in scholarly journals proclaimed on a weekly basis that the area of the brain underlying a highly specific cognitive function had been identified. As research progressed, and the number of independent functional modules in the brain proliferated, a kind of weary cynicism set in, leading some to regard the new-fangled research methods as a modern form of phrenology. Although accusations regarding its status as a pseudoscience may be premature, a significant challenge to the field of cognitive neuroscience is presented by the seemingly endless supply of new data that, while saying much, explain little. The number of distinct effects observed under various experimental paradigms has made the brain appear to be a very crowded place indeed, seeming to include regions coding for everything up to and including the proverbial kitchen sink.

One region of the brain in particular, the anterior cingulate cortex (ACC), has become associated with this kitchen-sink effect [3]. ACC, by virtue of its high interconnectivity, is promiscuously active in almost any task that involves some level of engagement and action. It has been noted that the rate at which cingulate activity is reported in fMRI studies has increased exponentially since the 1990s, and it is projected that by the end of the Twenty-first century neuroscience will achieve the "cingularity", the point at which there are more scholarly works investigating cingulate activity than there are cells in the cingulate itself [4]. Although the Gage, Parikh and Marzullo article is emphatically tongue in cheek, the authors are not far off the mark when they note that, given the diversity of functions that have been attributed to the cingulate, it appears that this region of the brain does everything. Functions attributed to ACC include detection and processing of error [5, 6], resolving behavioral conflict [7, 8], detecting and predicting reward [9, 10], anticipating and indicating painful stimuli [11, 12], signaling negative affect [13], deploying attention [14], learning the value of actions [15, 16], and a host of others.

Investigation of the function of ACC has been of particular interest in the area of cognitive control. In typical cognitive control tasks, subjects are required to inhibit a prepotent, stimulus-driven response in favor of a less-automatic response. A classic example is the Stroop task [17], in which subjects are presented with color words that are displayed in various font colors. The subject is instructed to indicate the color in which the word is written, and to ignore the denotative meaning of the word itself. For trials in which the meaning of the word and the font color both indicate the same response ('congruent'), the task is trivially easy. However, on trials in which the meaning of the word and the font color differ ('incongruent'), successful performance of the task requires the subject to make only one of two cued responses. The processes by which an individual interprets stimuli in order to select and execute a response are collectively referred to as cognitive control.

16.2 The Conflict Model

One highly influential interpretation of ACC activity is the conflict monitoring model [7, 8]. In this interpretation, task stimuli which cue multiple, mutually incompatible responses induce a state of conflict that needs to be resolved in order to successfully generate appropriate responses. ACC activity indexes conflict as the summed multiplicative interaction of cued responses

$$\text{Conflict} = \Sigma W_{ij} a_i a_j \tag{16.1}$$

where a_i and a_j are neural activities representing mutually incompatible response cues, and W_{ij} represents the degree of mutual incompatibility between them. In the case of the Stroop task, when only one response is cued on congruent trials, conflict, and by extension ACC activity, remain low. For incongruent trials, two competing responses are cued, resulting in increased ACC activity.

Although the conflict model is notable for its ability to account for a range of effects observed in ACC from EEG and fMRI studies, the model does not address how the arbitrary stimuli used in many cognitive control tasks might come to be associated with conflicting behavioral responses. Conflict in the Stroop task, for example, only exists due to the demands placed upon the subject to respond only with the color of the font a word is printed in rather than with the color denoted by the word itself. If the subject were instead instructed that it would be acceptable to make a response indicating either the word identity OR the font color, it is difficult to imagine how this would lead to a state of behavioral conflict. In short, conflict is not a property that inheres in external stimuli, but one that is constructed through the interaction of stimuli and the context in which they are experienced.

Situations in which a particular stimulus can cue entirely different behavioral responses are prevalent in day-to-day life. Take the example of encountering a stop sign as one is traveling along a road. Depending on the particular mode of transport, the responses one needs to generate in order to comply with the denotative meaning of the sign may vary a great deal. If one were traveling by bicycle, for instance, the appropriate response would be to apply brakes by operating levers located on both hand grips. If one were traveling by motorcycle, however, operating the controls on the handlebars would result either in increased speed or disengaging the clutch, neither of which would be optimal for coming to a stop.

It is easy to imagine that a proficient bicyclist who is learning how to ride a motorcycle may experience a high degree of conflict as she adjusts to the differences between the two modes of transport. Initial attempts to stop while on a motorcycle might require increased vigilance and attention (i.e., cognitive control) to ensure that the appropriate response is made. As the individual gains experience, however, there is less need to exercise control when switching from a bicycle to a motorcycle, implying that information regarding conflicting responses can be learned. In the case of our novice motorcyclist, the prepotent response when confronted with a stop sign, trained through many years of bicycling, is to apply hand brakes. When riding a motorcycle, as outlined above, this response leads to a sub-optimal result

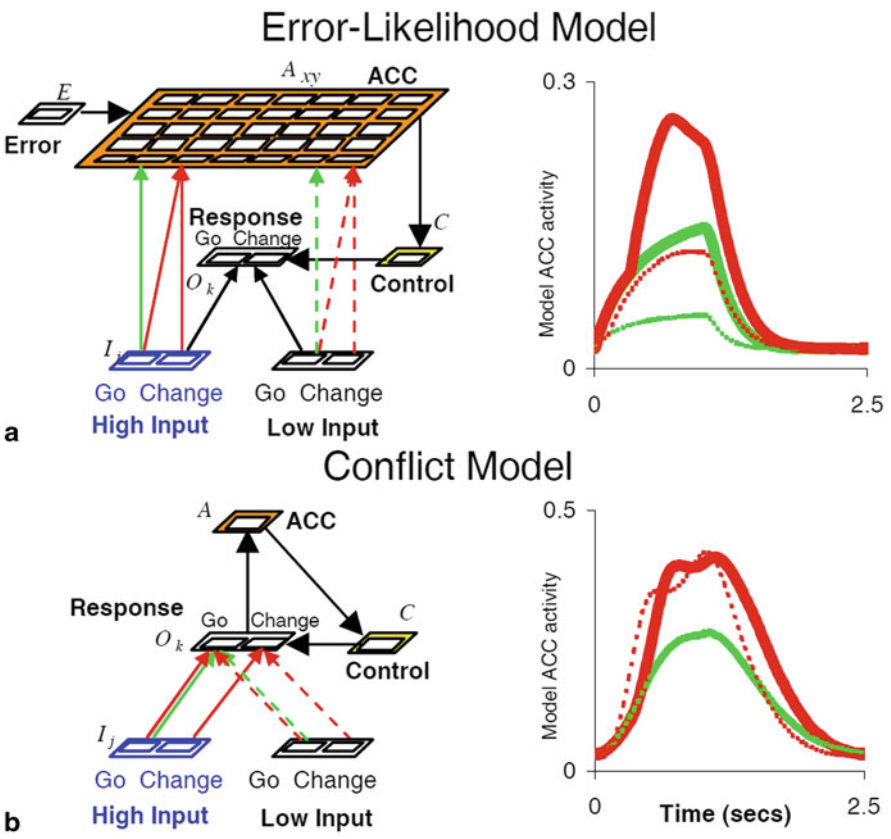

Fig. 16.2 The error likelihood model, as contasted with the conflict model in a change signal task. *Red curve*: conflict. *Green curve*: no-conflict. *Solid line*: high error likelihood. *Dashed line*: low error likelihood. (Adapted by permission of the AAAS from [18])

("not stopping"), and thus constitutes a behavioral error. The intuition, then, is that high-conflict situations, such as choosing between using hand or foot controls on a motorcycle, are those that are associated with an increased likelihood of error, rather than conflict *per se*, because a particular combination of movements may be mutually incompatible in one context but not in another.

16.3 Stage 1: Birth—The Error Likelihood Model

The error likelihood model of ACC [18] (Fig. 16.2) proposed a mechanism by which information regarding behavioral error may contribute to ACC activity. The principal component of the model consists of a self-organizing map (SOM), representing ACC, that receives excitatory projections from representations of task-related stimuli.

Initially, adjustable weights representing the excitatory influence of stimuli on ACC are low. Over the course of experience with a task, the model learns to associate single units in the SOM with the presentation of a particular stimulus. This association is learned by the hypothetical mechanism of dopaminergic (DA) disinhibition of ACC. Tonic activity of midbrain DA neurons may actively inhibit neurons in ACC; transient depressions in baseline DA activity, associated with negative reward prediction errors [19], disinhibit ACC and contribute to Hebbian-type learning between active stimulus representations and stochastically active single units in the SOM. Over the course of training, units within ACC are activated by stimulus representations in proportion to the frequency with which each stimulus is associated with error.

Although the intuition underlying the error likelihood model, that stimulus-dependent behavioral conflict may be learned, is largely consistent with the conflict monitoring theory of ACC, simulations of the two models revealed divergent predictions regarding ACC activity for contexts in which behavioral conflict is absent (Fig. 16.2) yet the likelihood of behavioral error differs between conditions. This discrepancy between the two models informed the design of a new behavioral task designed to investigate whether ACC activity reflected differences in anticipation of error. In the change signal task, subjects are asked to respond according to the direction indicated by an arrow presented to them. For most trials ("Go" trials), the cue remains valid and the subject makes the initially cued response. On a subset of trials ("Change" trials), however, the presentation of the arrow is followed by the presentation of a second arrow, indicating that the subject should cancel the initial response and instead make the alternate response. The timing between the presentation of the first and second arrows can be manipulated to enforce specific error rates, and the color of the arrows serves as an implicit cue indicating whether there is a high or low likelihood of committing an error.

Both the error likelihood and conflict models predict increased ACC activity for Change trials in which behavioral conflict is present. However, for Go trials, in which conflict is absent, the error likelihood model predicts increased ACC activity for conditions in which the likelihood of an error is higher, as indicated by the color of the arrow. These predictions were tested using fMRI, and revealed that, consistent with the error likelihood model, ACC activity was significantly greater for trials with a high likelihood of error, despite the absence of conflicting cues [18]. This suggested that the conflict model was incomplete, as it could not account for the error likelihood effect, and that the error likelihood model may provide a more complete account of the data. We note however that the conflict model has had (and continues to have) a very useful life—it provides a simple, elegant account of existing phenomena, it provided testable predictions, and it provided inspiration for subsequent model development, in this case the error likelihood model. Similarly, the error likelihood model was inspired by an earlier model in which dopaminergic reward omission signals drive activity in the ACC [19].

16.4 Stage 2: Maturity—Empirical Tests of the Error Likelihood Model

While the error likelihood model is able to account for observed error likelihood effects within ACC, the hallmark of a mature model lies in its ability to guide further empirical investigation. Additional simulations of the error likelihood model revealed additional predictions regarding ACC activity that directly informed a new set of experiments. We reasoned that if ACC activity reflects the prediction of an error rather than conflict, then perhaps two different response cues would lead to greater ACC activity regardless of whether or not they led to a state of conflict. We tested this prediction and found evidence in favor of it [20]. Still, while the error likelihood model accounts for a number effects observed within ACC, additional empirical evidence suggests that the error likelihood model itself is incomplete. The first challenge was an apparent failure to replicate the error likelihood effect empirically. A subsequent paper tested for error likelihood effects with fMRI and ERP and reported a null result [21]. This challenged us to ask whether individuals differ in their sensitivity to likely errors, and whether those who are more sensitive to error likelihood at the neural level may likewise tend to be more careful to avoid errors and risky behavior in general. We tested this empirically and found it to be the case [22], and this also led to a refined error likelihood model in which the learning rate from errors could vary across the population [23].

A second challenge to the error likelihood model consists of surprise effects, e.g., the finding that unexpected errors lead to greater ACC activity than less surprising errors [24]. The error likelihood model simply could not account for this effect, because the error likelihood signals constituted a prediction that did not depend on the actual outcome, but the surprise effect depends heavily on the nature of the outcome, e.g., an error. Furthermore, we and others found that the well-known error effects in ACC can actually invert themselves, so that when errors are more likely than correct outcomes, then the correct outcomes yield greater ACC activity than errors [25–27]. Still more challenges confronted the error likelihood model. More recent papers argued that error likelihood effects did not exist or were subsumed by simple correlations between ACC activity and response time [28, 29]. We also found that unexpected delays in feedback could lead to ACC activity, which suggested that the timing of feedback was as important as its valence [30].

Meanwhile, monkey neurophysiology studies of ACC provided their own challenge to both the conflict and error likelihood models. In an earlier study, we failed to find evidence of conflict monitoring cells in ACC, despite having recorded over 450 cells [31]. Instead, ACC cells generally seemed to reflect the value of actions, integrate recent reward history [32], indicate the value of explorative vs. exploitative behavior [33], and signal the prediction [9, 34] and detection [31, 35] of reward. These findings taken together present a significant challenge to the effort to develop a unifying theory regarding ACC activity across species, and, indeed, have been taken as evidence that no such unification is possible [36]. Nevertheless, recent studies

[37] have sought to test this idea by recording BOLD activity in monkeys performing tasks that have been observed to produce conflict-type effects in humans. These studies show increases in BOLD response in similar regions of monkey and human brains commonly associated with cognitive control, including ACC, providing evidence that ACC in monkey and human are functionally similar. This suggests that a reconciliation of human and monkey ACC results is possible.

16.5 Stage 3: Dialectic Methods of Model Development

Confronted with mounting evidence that the error likelihood model is, at best, an incomplete account of ACC function, our goal was to develop a new computational account that was *less incomplete*. In doing so, we sought to resolve two dialectic tensions, that of the theoretical-empirical dialectic, and the empirical-empirical dialectic. In the first, we questioned whether the theory underlying the error likelihood model was disproved by contradictory evidence, or whether minimal extensions could expand the theory without becoming a process of adding ever more epicycles. In the second, we question how empirical evidence in apparent contradiction with itself might reflect a common underlying process.

16.5.1 The Theoretical-Empirical Dialectic

In the face of empirical findings that challenge one's model, what is a modeler to do? One of the biggest challenges we face as modelers, aside from the challenges of developing models, is the parental affection that we feel towards the models we develop. Our modeling efforts have been guided in this regard by a pair of classic works that should be required reading for all scientists, and especially modelers. The first of these papers argues that we must beware of parental affection for a theory—rather than trying to extend the reign of an aged monarch of a theory, or even favoring a working hypothesis, we should entertain multiple competing hypotheses and sincerely ask which is the best account of the phenomenon, even if this means abandoning our own theoretical progeny [38]. This may seem an obvious point, but parental affection, and an associated desire to guard one's reputation as a modeler, is likewise a strong and pervasive instinct. The second paper argues that the best experiment is not one that is designed to prove a theory or model, but rather one that is designed to *dis*-prove a theory. As Platt puts it, our experiments should answer "the Question", which is this: What model or models do your experiments rule out [39]? In this way, as the space of plausible models is reduced, a field may converge on a more faithful model of the phenomena in question. All of this requires that parental affection for one's own theory must be set aside. Thereafter, the modeler is free to embrace the dialectic of Hegel, to wrestle with both the theory and the empirical

results, and find a synthesis that moves the field forward toward a new generation of model.

16.5.2 The Empirical-Empirical Dialectic

A key process in model building is finding ways to reconcile and integrate apparently contradictory findings. It is especially conducive to progress if relevant data can be considered from a range of species (human, monkey) and modalities (behavior, fMRI, ERP, single unit neurophysiology, etc.). In the case of the ACC, if the function is indeed similar across species, how does one reconcile, for example, the apparent involvement of single neurons in ACC in reward prediction and processing with the apparent involvement of the region as a whole with principally error prediction and processing? We approach this overarching question by addressing two related questions. First, why are single neurons whose activity reflects reward-related processes observed many times more frequently than neurons whose activity reflects error and error prediction? Recall that in the error likelihood model, simulated ACC activity prior to the presentation of task-related feedback is proportional to the frequency with which errors are observed for a given stimulus context. Individual units in the SOM representing ACC in the error likelihood model, analogous to single neurons within ACC, are selectively associated with task stimuli that predict error; at the limit, as the number of trials goes to infinity, the entire SOM is partitioned such that the number of units responding selectively to a particular stimulus will be proportional to the frequency with which errors are observed for that stimulus relative to other task stimuli. If ACC neurons were in fact learning associations between stimuli and subsequent error alone, one would expect that single-unit neurophysiology studies would find ubiquitous error-related neurons.

One possible explanation for the predominance of reward-related neurons observed in single-unit studies, in the framework of the error likelihood model, is that *single units within an SOM learn to associate task related stimuli not only with error, but also with correct outcomes.* In the context of the literature on cognitive control, both with monkeys and humans, experimental conditions are generally arranged such that trials on which errors are observed generally represent a minority of the total number of trials in an experiment. While this explanation may account for the prevalence of reward-related neurons within ACC, it poses an additional challenge to the error likelihood model. The error likelihood model learns to associate task cues with subsequent error through the putative mechanism of dopaminergic disinhibition. In this scheme, ascending projections from midbrain DA neurons tonically inhibit activity within ACC. The occurrence of an error produces a transient decrease in the baseline firing rate of DA neurons, resulting in disinhibition of ACC, with an attendant increase in activity in neurons that receive excitatory input from representations of task-related stimuli, resulting in Hebbian-type learning of conjointly active units in ACC and units representing task-cues. Since the activity of DA is generally

depressed only when an outcome is worse than expected (e.g., withholding of a predicted reward), this mechanism cannot explain how neurons in ACC may come to represent reward.

Regardless of the specific mechanisms contributing to the distribution of reward and error-related single neurons within ACC, a second question concerns the time course of neural activity observed within the region. Specifically, is reward and error related activity in ACC predictive in nature, or is the region involved principally in evaluation of actual outcomes? In this regard, evidence from single-unit neurophysiology and fMRI/EEG studies alike is mixed. By far the most robust effect observed in ACC from fMRI and EEG data is that of error. ACC was initially implicated in the processing of behavioral error based upon observations of a negative-going ERP component specifically on trials in which an error occurred, the well-known error-related negativity (ERN) [5, 6] an effect which has since been replicated numerous times across different recording techniques. Additionally, single neurons in monkey have been observed whose activity is specific to the delivery of reward or the occurrence of an error. Together, these findings suggest that ACC activity is primarily evaluative in nature, and depends on feedback from the environment or a subject's own behavior to elicit a response. On the other hand, substantial evidence suggests that ACC activity frequently anticipates or precedes feedback of this nature. Neurons in monkey ACC have been observed whose activity increases as an expected reward draws temporally closer, indicating a role for the area in prediction [34, 40]. Similarly, findings such as the error likelihood effect, wherein differences in activity are observed following a cue that is associated with future outcomes, provide evidence for ACC's involvement in prediction.

Previous theories of ACC have variously assigned primarily evaluative or predictive functions to the region. Evaluative theories of ACC have suggested it signals the discrepancy between actual and intended responses [41], discrepancies in value [19], corrective actions following behavioral error [42], or the multiplicative interaction of multiple cued responses in the conflict model [8, 43]. Although these theories can be described as evaluative in the sense that the computations necessary to perform them operate on information that is present within the system, rather than information that is anticipated but not present, it is not clear that they distinguish between anticipatory and reactive activity in ACC. In the case of Scheffers and Coles [41] and Steinhauser et al. [42], ACC activity is reactive in nature, depending on the comparison between responses or detection of sequences that have already occurred. In contrast, the conflict model suggests ACC activity may be observed prior to the actual generation of a response as two incompatible responses compete. Similarly, the RL-ERN theory [44] suggests that ACC activity may precede the actual generation of a response due to value discrepancies induced by changes in response unit activity in the model. For all of these models, however, the primary computation performed by ACC explicitly evaluates present information; anticipatory or predictive activity is merely implicit, reflecting continuous, ongoing evaluation. In contrast, the error likelihood model is based on explicit prediction; activity prior to feedback reflects the likelihood of an event (behavioral error) that has yet to be experienced, while increased activity in the model following incorrect performance implicitly indicates error as the disinhibition

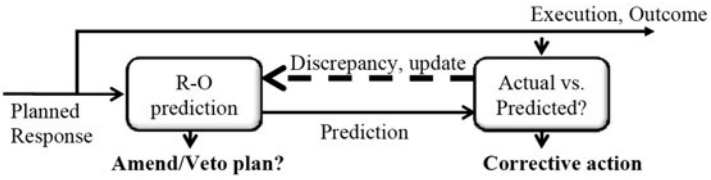

Fig. 16.3 The PRO model, showing distinct response-outcome (R-O) prediction unit (*left*), and evaluation unit (*right*). (Reprinted by permission of the Cognitive Science Society, from [45])

of predictive units. More generally, we find that theories regarding ACC function can be categorized as being either explicitly evaluative/implicitly predictive, or explicitly predictive/implicitly evaluative. A notable lack, therefore, in theorizing about ACC, is the possibility that *ACC activity reflects both explicit predictive and evaluative functions*.

16.6 The PRO Model—A Synthesis

16.6.1 PRO Model Stage 1: Birth

In the preceding section, we identified two questions that informed our thinking regarding the function of ACC in the context of cognitive control and decision making. First, how might individual neurons in ACC become associated principally with the anticipation and detection of reward? And second, is activity in ACC consistent with either explicit or implicit prediction and evaluation? In both cases, we identified potential solutions to these questions, leading to a concise hypothesis regarding possible ACC function that forms the basis of the *predicted response-outcome* (PRO) model (Fig. 16.3), that ACC learns explicit predictions regarding future outcomes, regardless of their affective valence, and signals unexpected deviations from predicted events [45, 46].

Similar to previous models of ACC, the PRO model builds on standard models of reinforcement learning, and extends these models in two ways. First, standard RL is concerned with learning an optimal behavioral policy given an underlying value function, with the goal of selecting at each opportunity the behavior that leads to the highest long-term value. In typical formulations, value is represented as a scalar quantity, reflecting the average reward (positive or negative) that can be obtained from a given state. In contrast, the PRO model learns separate predictions of likely conjunctions of responses and outcomes, regardless of whether the outcome is affectively positive or negative. The rationale behind this extension is that in the context of cognitive control, behavior is not selected, as such, but is governed by the interaction of prepotent stimulus-driven behaviors with top-down goals that may conflict with automatic responses. It is not, for example, sufficient to learn that naming the color of the font is more valuable than reading the word in the Stroop task, since value is

only one factor in determining behavior. In order to successfully deploy cognitive control, it is necessary to learn the likelihood that a certain, automatic response will be produced, independent of how valuable that response may be. Secondly, and related to the first point, is that while prediction errors in standard RL reflect better-than or worse-than expected outcomes, prediction errors in the PRO model indicate the extent to which an observed outcome was expected to occur (again, independent of the affective valence of the event itself). In standard RL both the occurrence of a reward that was not predicted as well as the non-occurrence of a predicted aversive stimulus both result in positive prediction errors—better-than expected outcomes. In the PRO model, these outcomes result in positive and negative prediction errors, respectively, indicating an outcome occurred that wasn't expected (positive surprise) and an outcome that was predicted failed to occur (negative surprise).

Although the PRO model departs from the error likelihood model in adopting a RL-based formulation, a key intuition underlying the error likelihood model, that ACC activity reflects the frequency with which errors are observed in a given stimulus context, is preserved. The PRO framework extends this intuition by allowing for the representation not only of error frequency, but also the frequency with which correct trials are observed. The activity of individual units within the PRO model, each representing a specific response-outcome conjunction, reflects the frequency with which that conjunction is observed, analogous to the SOM that formed the basis of the error likelihood model. Indeed, for typical cognitive control tasks in which correct trials are more frequently observed than error trials, activity in prediction-related units is greater than activity in units related to predicting error outcomes. This aspect of the PRO model corresponds well with the intuition developed in the previous section regarding how ACC neurons may predominantly represent reward and reward-related processes: in the course of experience with a task, neurons learn to represent events based on the frequency with which those events are observed.

16.6.2 Reconciling Existing Data

How then can the PRO model simultaneously account for the predominance of reward-related neurons in ACC as well as effects of error, error likelihood, and conflict observed at the level of populations of neurons? Above, we introduced the notions of positive and negative surprise as they relate to standard RL and the extensions introduced in the PRO model. In particular, negative surprise, the unexpected non-occurrence of a predicted outcome, is found to provide a plausible mechanism by which the ensemble activity of reward-related neurons produces apparent error effects. As previously discussed, correct trials are far more frequently observed in typical cognitive control tasks, leading—in the PRO framework—to proportionally greater activity in units predicting reward. Negative surprise, calculated as

$$\omega_t^N = \sum_i MAX\left(Expected_i - Actual_i, 0\right) \tag{16.2}$$

i.e., the predicted outcome minus the observed outcome, suggests that when a highly reliable event such as a correct, rewarded trial fails to occur, observed neural activity should be greater than when a marginally probable event, such as an error, fails to occur. The PRO model thus reinterprets error effects within ACC as the surprising non-occurrence of a predicted (rewarding) outcome. Furthermore, the same logic can additionally reconcile previous findings that did not previously fit within any explanatory framework, including findings of greater activity in ACC for infrequent vs. frequent errors [18, 24], and error effect-like activity for unexpected successes on low probability gambles [26].

Is activity in ACC primarily predictive or evaluative? The negative surprise signal which we deploy to explain the diverse array of effects observed in ACC is, by definition, evaluative in that it compares predicted outcomes to observed outcomes. In order to perform such an evaluation, however, it is necessary that predictions be maintained somewhere within the brain; one possibility is that ACC is involved in both evaluation and prediction, and that these two functions may be either spatially segregated or that individual units performing each function may be intermingled. Another possibility is that extracingulate areas maintain predictions regarding likely outcomes which are used by ACC to calculate the deviation between predicted and observed outcomes. Distinguishing between these two hypotheses may be problematic, however, in that the timed prediction signal assumed by the PRO model as the basis for computing negative surprise is correlated with the negative surprise signal. Following the presentation of a stimulus that reliably predicts future outcomes, both the prediction and the negative surprise signal are identical. This follows from equation (2), in which negative surprise is calculated as the current predicted outcome ($Expected_i$) minus the current observed outcome ($Actual_i$); prior to the occurrence of an outcome, $Actual_i = 0$, and negative surprise is equal to the current prediction. Following the occurrence of an outcome, however, the signals are predicted to diverge, particularly in the case of the occurrence of a likely outcome, where *Expected* and *Actual* values are similar, resulting in low negative surprise but high predictive activity. This may suggest one manner in which areas of the brain involved distinctly in prediction, and not evaluation, may be distinguished from areas within ACC whose activity is consistent with the negative surprise signal suggested by the PRO model.

16.6.3 PRO Model Stage 2: Maturity—Motivating New Experiments

A key goal for computational models is not only to account for previously observed data, but also to suggest additional research questions which may be empirically tested. In both respects, the PRO model has performed well. The PRO model has motivated tests of surprise effects in ACC by others [25], as well as a number of experiments in our own research group. In work currently in preparation, we identify overlapping areas in ACC that reflect the surprising absence of a predicted painful stimulus as well as the unexpected non-occurrence of a difficult cognitive task. In additional work, we find that ACC activity in substance-dependent individuals, rather

than reflecting an overall lack of engagement in cognitive tasks [47], is best explained as increased attention to rewarding outcomes relative to aversive outcomes, biasing the overall predictions learned by the PRO model to reflect attenuated predictions regarding possible negative consequences of one's actions. A recent study also suggests that predictive signals within the ACC can be distinguished from evaluative processes [48]. All of these studies were motivated in part by the predictions of the PRO model.

16.6.4 A Unifying Framework

In the PRO formulation, activity in ACC is associated with the relatively straightforward mechanism of "negative surprise." While we show that this mechanism does well at accounting for observed effects in ACC in the context of cognitive control, ACC is a promiscuous region of the brain, being implicated in social interaction, affective and emotional response, and processing aversive stimuli. An open question, then, is whether the simple mechanism of negative surprise can be applied more generally to questions beyond the specific area of cognitive control. Put another way, is negative surprise one of several specialized functions, each devoted to a particular cognitive function, or does ACC implement some form of negative surprise across the many cognitive functions it is involved in?

Preliminary evidence suggests that the PRO model may provide a unifying framework for understanding ACC function not only across different recording methodologies, but also across different subdisciplines of neuroscience. Since its publication, the PRO model paper [46] has been cited in studies taking affective, social, and clinical neuroscience perspectives. Previously, ACC has been variously attributed roles in the processing of painful stimuli, indicating negative affect related to social exclusion, and disengagement of cognitive control in substance-dependent behaviors. Under the interpretation of the PRO model, however, these effects may reconciled as reflecting facets of a single underlying mechanism, that of negative surprise. Single unit neurophysiology and fMRI studies have identified neurons in ACC, as well as ensemble activity, showing increased activity in the region when monitoring the decisions made by others [49, 50], especially when the outcome of those decisions violates expectations. If we were to extend our notion of what constitutes an outcome to include any predictable event, whether it be terminal feedback at the end of a trial, or intermediate states such as the appearance of additional stimuli, we might expect ACC activity to more generally to reflect predictions about states and to signal surprising state transitions [51]. Simulations of the PRO model incorporating this more general definition of "outcome" suggest that the principle mechanism of negative surprise can account for activity observed in ACC related to the prediction of task-related stimuli, as well as activity related to the onset of such stimuli. Perhaps the best example of this is the mismatch negativity (MMN), a negative ERP component elicited when a periodic stimulus, reliably and repeatedly

presented to a subject, is unexpectedly withheld. EEG studies have identified generators of the MMN in visual and auditory cortex, depending on the modality of the stimulus. More recent studies have identified a generator within medial PFC/ACC, consistent with the anatomical localization of the ERN, and independent of modality.

These findings suggest a general function of ACC in signaling unexpected state transitions, consistent with a potential role in providing learning signals for a model-based RL algorithm implemented across distributed across multiple brain regions. If such is the case, we might expect the activity of areas of the brain with a high degree of connectivity with ACC to also reflect additional aspects of such an algorithm. In this regard, dorsolateral PFC (DLPFC) is a likely target for additional modeling and empirical efforts. Previous work has identified DLPFC as being critically involved in maintaining task-dependent rules in working memory, deploying top-down control, and, like ACC, activity in DLPFC correlates with state prediction errors (SPEs) [52]. The PRO model provides critical constraints on the types of information and computations that may plausibly be observed within DLPFC. Although the PRO model undoubtedly is incomplete at some level of detail, the central intuition that ACC maintains multiple, simultaneous predictions regarding future events, is more consistent with existing evidence than competing accounts. In much the same way the heliocentric model of the solar system supplanted the over-complex geocentric model through a single, unifying premise, the goal of future work investigating the function of brain is not only to identify new roles to assign to individual regions, but also to reconcile how those roles may fit within a general framework.

16.7 Conclusion

In the above discussion, we have laid out some principles of model development that have served us well, and we illustrate these principles by tracing out three generations of models of the ACC. The themes we have highlighted include the life cycle of a model, the dialectic between competing models, the dialectic between models and data, and the dialectic between apparently contradictory data. All of these can be viewed as processes that lead to progress and the development of better models. In tracing out the example ACC models here, we do not mean to imply that they are the only fruitful approaches to modeling the ACC, as there are other models that we do not have space to treat fully. Neither do we intend to tell a triumphalist story that the PRO model is the best model. While we find that it provides the best account of the data to date, it is still a model, and is therefore by definition incomplete. The PRO model will have served its purpose well if it inspires a new set of experiments, inspires a deeper understanding of the ACC, and ultimately leads to an even better model.

Exercises

1. What is the advantage of having two different computational neural models of a phenomenon rather than one?
2. Read Platt [39] and Chamberlin [38]. How could you apply the methods they advocate to your own research?
3. What is the life cycle of a model as described here?
4. Name and describe two key dialectics in the process of model-building.
5. Which is a better experimental effort: to try to prove that a computational neural model is true, or to try to prove that it is false or incomplete?
6. Read the paper on the PRO model [46], or another current computational neural model. Try to think of an experiment (or existing data) that could falsify or otherwise challenge the model.

Further Reading

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. *Nat Neurosci*, 14(10), 1338–1344. doi:10.1038/nn.2921

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. C. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624–652.

Brown, J. W., & Braver, T. S. (2005). Learned Predictions of Error Likelihood in the Anterior Cingulate Cortex. *Science*, 307(5712), 1118–1121.

Acknowledgments Supported by the Intelligence Advanced Research Projects Activity (IARPA) through Department of the Interior (DOI) contract D10PC20023. 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.

References

1. Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago
2. Poldrack RA (2008) The role of fMRI in cognitive neuroscience: where do we stand? Curr Opin Neurobiol 18(2):223–7
3. Poldrack RA (2012) The future of fMRI in cognitive neuroscience. Neuroimage 62(2):1216–20
4. Gage GJ, Parikh H, Marzullo TC (2008) The cingulate cortex does everything. Ann Improbable Res 14(3):12–15
5. Falkenstein M et al (1991) Effects of crossmodal divided attention on late ERP components: II. Error processing in choice reaction tasks. Electroencephalogr Clin Neurophysiol 78:447–455
6. Gehring WJ et al (1990) The error-related negativity: an event-related potential accompanying errors. Psychophysiology 27:34

7. Botvinick M et al (1999) Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature 402(6758):179–181
8. Botvinick MM et al (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624–652
9. Shidara M, Richmond BJ (2002) Anterior cingulate: single neuronal signals related to degree of reward expectancy. Science 296(5573):1709–11
10. Shima K, Tanji J (1998) Role of cingulate motor area cells in voluntary movement selection based on reward. Science 282:1335–1338
11. Chandrasekhar PVS et al (2008) Neurobiological regret and rejoice functions for aversive outcomes. Neuroimage 39(3):1472–84
12. Rainville P (1997) Pain affect encoded in human anterior cingulate but not somatosensory cortex. Science 277(5328):968–971
13. Eisenberger NI, Lieberman MD, Williams KD (2003) Does rejection hurt? An FMRI study of social exclusion. Science 302(5643):290–2
14. Posner MI, Petersen SE, Fox PT, Raichle ME (1988) Localization of cognitive operations in the human brain. Science 240(4859):1627–1631
15. Rudebeck PH et al (2008) Frontal cortex subregions play distinct roles in choices between actions and stimuli. J Neurosci 28(51):13775–85
16. Walton ME, Devlin JT, Rushworth MFS (2004) Interactions between decision making and performance monitoring within prefrontal cortex. Nat Neurosci 7(11):1259–1266
17. Stroop JR (1935) Studies of interference in serial verbal reactions. J Exp Psychol 18(6):643–662
18. Brown JW, Braver TS (2005) Learned predictions of error likelihood in the anterior cingulate cortex. Science 307(5712):1118–1121
19. Holroyd CB, Coles MG (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psych Rev 109(4):679–709
20. Brown JW (2009) Multiple cognitive control effects of error likelihood and conflict. Psychol Res 73(6):744–50
21. Nieuwenhuis S et al (2007) Error-likelihood prediction in the medial frontal cortex: a critical evaluation. Cereb Cortex 17:1570–1581
22. Brown JW, Braver TS (2007) Risk prediction and aversion by anterior cingulate cortex. Cogn Affect Behav Neurosci 7(4):266–77
23. Brown JW, Braver TS (2008) A computational model of risk, conflict, and individual difference effects in the anterior cingulate cortex. Brain Res 1202:99–108
24. Holroyd CB, Krigolson OE (2007) Reward prediction error signals associated with a modified time estimation task. Psychophysiology 44(6):913–7
25. Ferdinand NK et al (2012) The processing of unexpected positive response outcomes in the mediofrontal cortex. J Neurosci 32(35):12087–92
26. Jessup RK, Busemeyer JR, Brown JW (2010) Error effects in anterior cingulate cortex reverse when error likelihood is high. J Neurosci 30(9):3467–3472
27. Oliveira FT, McDonald JJ, Goodman D (2007) Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. J Cogn Neurosci 19(12):1994–2004
28. Grinband J et al (2011) The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. Neuroimage 57(2):303–311
29. Yeung N, Nieuwenhuis S (2009) Dissociating response conflict and error likelihood in anterior cingulate cortex. J Neurosci 29(46):14506–14510
30. Forster SE, Brown JW (2011) Medial prefrontal cortex predicts and evaluates the timing of action outcomes. Neuroimage 55(1):253–65
31. Ito S et al (2003) Performance monitoring by anterior cingulate cortex during saccade countermanding. Science 302:120–122
32. Kennerley SW et al (2006) Optimal decision making and the anterior cingulate cortex. Nat Neurosci 9(7):940–947
33. Hayden BY, Pearson JM, Platt ML (2011) Neuronal basis of sequential foraging decisions in a patchy environment. Nat Neurosci 14(7):933–9

34. Amador N, Schlag-Rey M, Schlag J (2000) Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field. J Neurophysiol 84(4):2166–70
35. Matsumoto M et al (2007) Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10(5):647–656
36. Cole MW et al (2009) Cingulate cortex: diverging data from humans and monkeys. Trends Neurosci 32(11):566–74
37. Ford KA et al (2009) BOLD fMRI activation for anti-saccades in nonhuman primates. Neuroimage 45(2):470–6
38. Chamberlin TC (1965) The method of multiple working hypotheses. Science 148(3671):754–759
39. Platt JR (1964) Strong inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. Science 146(3642):347–353
40. Amiez C, Joseph J-P, Procyk E (2005) Anterior cingulate error-related activity is modulated by predicted reward. European J Neurosci 21(12):3447–52
41. Scheffers MK, Coles MG (2000) Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J Exp Psychol Hum Percept Perform 26(1):141–51
42. Steinhauser M, Maier M, Hübner R (2008) Modeling behavioral measures of error detection in choice tasks: response monitoring versus conflict monitoring. J Exp Psychol Hum Percept Perform 34(1):158–76
43. Yeung N, Cohen JD, Botvinick MM (2004) The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol Rev 111(4):931–59
44. Holroyd CB et al (2005) A mechanism for error detection in speeded response time tasks. J Exp Psychol Gen 134(2):163–191
45. Alexander WH, Brown JW (2010) Computational models of performance monitoring and cognitive control. Top Cogn Sci 2:658–677
46. Alexander WH, Brown JW (2011) Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci 14(10):1338–1344
47. Goldstein RZ et al (2009) Anterior cingulate cortex hypoactivations to an emotionally salient task in cocaine addiction. Proc Natl Acad Sci U S A 106(23):9453–8
48. Jahn A, Nee DE, Brown JW (2011) The neural basis of predicting the outcomes of imagined actions. Front Neurosci 5:128–128
49. Hillman KL, Bilkey DK (2012) Neural encoding of competitive effort in the anterior cingulate cortex. Nat Neurosci 15(9):1290–7
50. Suzuki S et al (2012) Learning to simulate others' decisions. Neuron 74(6):1125–37
51. Wessel JR et al (2012) Surprise and error: common neuronal architecture for the processing of errors and novelty. J Neurosci 32(22):7528–37
52. Glascher J et al (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66(4):585–595