Cognitive Development 27 (2012) 349–366

Contents lists available at SciVerse ScienceDirect

Cognitive Development

Why won’t you do what I want? The informative failures of children and models

Christopher H. Chatham[a],[*], Benjamin E. Yerys[b], Yuko Munakata[c]

[a] Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, RI, USA
[b] Department of Neurosciences, Children’s National Medical Center, George Washington University School of Medicine, Washington, DC, USA
[c] Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA

ARTICLE INFO

Keywords:
Connectionism
Basal ganglia
Gating
Error-driven learning
Striatum
Cognitive flexibility

ABSTRACT

Computational models are powerful tools – too powerful, according to some. We argue that the idea that models can “do anything” is wrong, and we describe how their failures have been informative. We present new work showing surprising diversity in the effects of feedback on children’s task-switching, such that some children perseverate despite this feedback, other children switch as instructed, and yet others play an “opposites” game without truly switching to the newly instructed task. We present simulations that demonstrate the failure of an otherwise-successful neural network model to capture this failure. Simulating this pattern motivates the inclusion of updating mechanisms that make contact with a growing literature on frontostriatal function, despite their absence in theories of the development of cognitive flexibility. We argue from this and other examples that computational models are more constrained than is typically acknowledged and that their resulting failures can be theoretically illuminating.

Models are powerful. Computational modeling enables researchers to formalize their theoretical assumptions, quantitatively test the implications of those assumptions, assess the causal impact of factors that may be difficult to manipulate experimentally, and derive novel, falsifiable and often counterintuitive predictions. These and other benefits of computational modeling have been widely acknowledged (Elman, 2005; Marcovitch & Zelazo, 2009; McClelland, 2009; Munakata, Snyder, &

[*] Corresponding author at: Dept. of Psychology and Neuroscience, Brown University, 89 Waterman Street, Providence, RI 02912, United States. Tel.: +1 410 591 0083.
E-mail address: chathach@gmail.com (C.H. Chatham).

0885-2014/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cogdev.2012.07.003

Chatham, 2011; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Thomas, McClelland, Richardson, Schapiro, & Baughman, 2009; Weng et al., 2001).

However, computational models can also be seen as powerful in a more troublesome way. It is sometimes said that models can be made to "do anything," so that the ability to simulate phenomena of interest tells us nothing. This concern is not without cause; multilayer neural network models can in theory approximate any continuous function to arbitrary accuracy (Cybenko, 1989), and are therefore sometimes castigated as too powerful to be psychologically meaningful (Massaro, 1988). Many models are criticized for having too many free parameters (Feldman, 2010). And even when models are appropriately constrained – for example, using constraints relating to psychology, biology and both task and input structure (Regier, 2003) – there is always the lingering possibility that a model captures behaviors of interest for epiphenomenal reasons. Indeed, models may not do things in the same way the brain does them (McClelland, 2009) – a problem that may have led some modelers to explicitly eschew their models' mechanistic implications and to focus instead on the way models formalize the problems that the brain must solve (Jones & Love, 2011).

We take a different tack at least with respect to neural network modeling, in which phenomena are modeled as arising from the interactive activation of a set of neuron-like units organized into layers. Some units correspond to the inputs of the model, others to its outputs, and yet others to its "hidden" processes (such as cognitive processing, when neural networks are used to model psychological phenomena). The activations of these units are determined by interconnections that vary in strength (the so-called "weights" of the model); these connection weights are adjusted through learning algorithms, such that activation patterns in the units of the input layer will yield the correct patterns of activation in the units of the output layer.

We argue that concerns about the flexibility of this class of models reflect at least two misunderstandings. The first is that the connection weights of a neural network model are equivalent to the beta weights of a statistical model – that is, that they may be used as free parameters to enable a direct fit to the data of interest. Indeed, in statistical machine learning applications, this is sometimes the case: The weights of a neural network model can be adjusted to fit the data of interest by minimizing the error between a model's predictions and empirical observation or by maximizing the likelihood of those observations. Used in this way, neural network models of sufficient size can indeed fit any well-defined continuous function to arbitrary accuracy (Cybenko, 1989). But in psychological applications, the weights of a neural network model typically are *not* adjusted to directly fit the data of interest. Instead of being trained directly to reproduce human-like data, the weights of neural network models are adjusted either according to the co-occurrence statistics of stimuli in the environment (in the case of Hebbian learning), or toward yielding optimal – but not necessarily human-like – performance (in the case of error-driven learning). In extreme cases, this can mean that the weights of neural network models are actually adjusted toward a *bad* fit to human data, for example toward perfect performance in a task on which humans perform poorly (Chatham et al., 2011). In any case, the weights of a neural network model applied to psychological phenomena are not free parameters in the sense that the beta weights of a statistical model are.

The second misunderstanding, which is the focus of the remainder of this article, is the perception that neural network models are more flexible and powerful that they actually are. To many a modeler's dismay, models cannot simply "do anything." In some cases computational models are *under*powered, e.g., they do not explicitly contain the mechanisms for hypothesis testing that are sometimes viewed as central to development (Gopnik, Wellman, Gelman, & Meltzoff, 2010). In other cases, the overestimation of neural network models' flexibility may reflect the published successes of connectionism and perhaps also a file-drawer problem (i.e., unpublished failures). But relative to the published successes of connectionism, its failures are often more illuminating: The computational power of neural network models can suggest that if such a model can't solve a problem in a particular way, the brain might not be able to solve the same problem in that particular way either. The failures of neural network models can thus be just as informative as their successes, particularly with respect to issues surrounding functional organization of the brain and processes of learning and development.

We describe three examples of such informative failures within computational models of cognitive development. In the first, the failure of a neural network model led to insights into why the brain has multiple memory systems (O'Reilly, 1996), with implications for understanding developmental phenomena like category learning. In a second, the failure of a model points to an important role for selective attention in the domain of physical reasoning, where selective attention is an often-unacknowledged factor in stage-like developmental transitions (Schapiro & McClelland, 2009). Our final example reports a novel task-switching phenomenon in children that our own model failed to capture. This failure motivates a revision to our account of the development of cognitive flexibility, which we report here, and may point the way toward an important point of convergence across competing accounts of cognitive flexibility. In each case, the failures of models have provided insights into important distinctions among the multiple neural mechanisms that cooperate to support cognitive development.

1. Computational tradeoffs in learning

Although abstract concepts are often thought to be the most difficult for children to acquire, the acquisition of abstract semantic knowledge can sometime precede that of more concrete knowledge. For example, children show a coarse-to-fine grained progression in their differentiation of items from various semantic categories (e.g., living vs. non-living things, or animals vs. non-animals). This differentiation can be assessed by asking children whether it is “silly” or “okay” to say, for example, that “This milk is alive” (Keil, 1979). While kindergartners will make only two or three distinctions across tested items dividing items into very coarse categories like “living” or “non-living,” children progressively construct more fine-grained distinctions as reflected by the increasing number of restrictions they place on appropriate attributions for items of different categories (e.g., that it is silly to say “This milk is tall”).

An elegant neural network model demonstrated how this kind of coarse-to-fine grained conceptual differentiation might occur (Rumelhart & Todd, 1993). This model was trained on an assortment of basic propositions about items like animals, plants, birds, flowers, robins, daisies, and other basic and subordinate level items from the category “living things.” For example, the model was trained to activate the output units corresponding to “GROW”, “MOVE”, and “FLY” in response to the inputs “ROBIN” and “CAN.” In this manner, initially random weights within the model were iteratively adjusted over training to increase the likelihood that the correct set of output units was activated by every combination of input units. Strikingly, the internal connection weights of the model underwent a coarse-to-fine grained differentiation of semantic categories that was not enforced by the training procedure, thus emergently capturing an essential feature of the developmental trajectory of semantic knowledge.

However, children learn many facts in a staged fashion, and not in the randomly permuted order that facts were learned by this model. For example, children are taught to add one to a given number – i.e., to count – before they learn to add by twos. Similarly, children generally learn to add before learning to multiply. Yet the models that so elegantly captured the developmental process of conceptual differentiation failed profoundly in learning arithmetic in this way (McCloskey & Cohen, 1989; Ratcliff, 1990) because they would entirely forget how to count prior to ever learning to add by twos. In other words, subsequent stages of learning would retroactively eliminate any earlier stage of learning; such “catastrophic interference” could be ameliorated only at the cost of dramatically reducing the models’ ability to extract the coarse generalities that enabled its fit to developmental data. The failure of these models to capture the benefits of staged learning were thought to undermine not only the explanatory value of such models to development, but also to learning more generally (McCloskey & Cohen, 1989).

Fairly exhaustive attempts to circumvent the catastrophic interference dilemma (French, 1991, 1992, 1999) ultimately suggested that an inherent computational tradeoff might underlie these failures. Indeed, these failures subsequently formed the basis of a successful theory for why the brain evolved separate mechanisms for semantic knowledge (neocortex) and episodic memory (hippocampus; McClelland, McNaughton, & O’Reilly, 1995). Put simply, the computational demand to simultaneously keep earlier learning experiences relatively distinct from subsequent ones – a characteristic of both staged learning and episodic memory – is irreconcilable with the demand to integrate across those experiences and thereby extract those generalities that seemed to characterize semantic memory development. More specifically, the successful separation of earlier and subsequent experiences required for both staged learning and episodic memory requires that activation be spread across many neurons (a so-called “sparse, distributed representation”), so that the resulting activity patterns are highly distinct from one another. In contrast, more overlapping representations are necessary to support generalization across many experiences, e.g., of the coarse-to-fine-grained kind that characterizes categorical knowledge.

The theory arising from these failures to capture developmental phenomena has, in turn, permitted a deeper understanding of what otherwise appear to be conceptually distinct phenomena. One example is the counterintuitive consequence of disrupting hippocampal neurogenesis, which can actually improve episodic memory (Saxe et al., 2007). The model makes sense of the finding that new hippocampal neurons are more generally excitable than their elder counterparts; they tend to “muddle” the otherwise clear separation of earlier and subsequent experiences that is necessary for encoding new experiences into episodic memory without suffering catastrophic interference.

In sum, a model that successfully captured one aspect of development (the coarse-to-fine-grained acquisition of semantic knowledge) failed to capture another (the benefits of staged learning). These failures led to better, broader understanding of the computational tradeoffs that may have led the brain to evolve separate memory systems for episodic and semantic memory, as well as a formalized theory of the differences between those memory systems. In this way, the failures of a model to capture developmental findings ultimately led to a theory that illuminates a relationship between disparate developmental phenomena.

1.1. Selective attention in physical reasoning: the balance scale task

Young children can show remarkable difficulty in tasks involving reasoning about the physical world, particularly those that involve mathematical concepts such as conservation of number or proportion. For example, when presented a balance with variable weights placed at variable distances on either side of a central fulcrum, young children pass through a number of discrete stages in reasoning about which side of the balance scale will fall. These stages are well characterized by response rules (Siegler, 1981), such that children will initially reason only on the basis of the number of weights on either side of the fulcrum (“Rule I”) and will subsequently consider distance as well, but only when the number of weights are equal (“Rule II”). (Subsequent stages are either less clear-cut cases of response rules or are not achieved even by most adolescents and adults; Siegler & Chen, 2002).

Surprisingly, discrete transitions like these can nonetheless be captured by neural network models that lack any explicit rules. A simulated version of the balance scale task was presented to a neural network model trained to activate the correct output units corresponding to “left side will fall” or “right side will fall,” in response to input units containing information about the distance and quantity of weights on each side of the fulcrum (McClelland, 1989). Networks become predisposed to attend to the quantity of weights on the balance scale due to increased training on problems in which quantity provides the critical information but subsequently transition to utilize distance when quantity is uninformative.

Subsequently it was suggested that transition from Rule I to Rule II was more discontinuous than the model could explain (Jansen and van der Maas, 2001). In particular, stage-like transitions among children using Rule I can be provoked by increasing the distance of items on the scale, with some children abruptly and stably adopting Rule II, and others reverting to Rule I only when distance was decreased *below* the point at which they had first begun to adopt Rule II. These patterns were heralded as evidence for the necessity of invoking rule use to explain behavior in the balance scale task.

Extensions to the model – allowing it to learn from its own behavior in the task – did yield these additional phenomena but also yielded spatial response biases typically not observed in this task. In particular, the model learned to respond that a particular side of the balance scale would drop, irrespective of the number and distance of weights from the fulcrum. Heroic attempts to eliminate this spatial bias failed to provide an adequate match to children’s behavior (Schapiro & McClelland, 2009), whether explicitly forcing spatial symmetry into the weights of the model (so that any increased tendency to respond on one side was balanced by an increased tendency to respond on the other), encouraging spatial symmetry using spatially balanced inputs (so that any learned tendencies to respond on one side of the fulcrum would be balanced by equivalent inputs from the opposite side), wholly restructuring the form of input received by the model so that the quantity of weights was encoded as a continuous function of unit activation rather than the number of units, or by other changes to the architecture. These failures, however, were informative; they suggested that the stage-like developmental transitions in the balance scale task may take a different form than the changes to the connection weights that typify the learning and development of neural network models.

Indeed, when a qualitatively different form of learned change was used, the model succeeded in capturing all of the essential phenomena characterizing transitions between Rule I and Rule II. Instead of adjusting connection weights, learning algorithms were used to adjust a “gain” parameter thought to correspond to the effect of selective attention (Kruschke & Movellan, 1991). Failure of the computational model led to the idea that attentional factors might be crucial to performance, rather than changes to long-term knowledge structures about proportion encoded by the model’s weights. This idea is consistent with the fact that children who perform well on the balance scale task may not perform well on formally similar physical reasoning problems, such as the balance beam task (Messer, Pine, & Butler, 2008).

The model allows us to make sense of these findings because it suggests that what drives change in the balance scale task is not the acquisition of discrete response rules about balance or proportion in general, nor only developments in graded long-term knowledge structures that approximate response rules about balance and proportion. Also implicated is how selective attention may operate on this knowledge. This insight arising from the model’s failure illuminates other developmental phenomena, such as the finding that children often regress from mature to immature strategies during repeated task performance (Siegler, 2007). Such regressions are difficult to reconcile with the idea that children acquire long-term knowledge about proportion during performance or discrete response rules, but could reflect the dynamic fluctuations that operate on longer-term knowledge structures.

1.2. Cognitive flexibility: the dimensional change card sort

Young children show a robust tendency to perseverate – to repeat old behaviors that are no longer appropriate. A particularly striking demonstration of this cognitive inflexibility comes from the Dimensional Change Card Sort (DCCS; Zelazo, Frye, & Rapus, 1996), in which children must sort bivalent cards first by one rule (e.g., color) and subsequently switch to another rule (e.g., shape). Three-year-olds often fail to switch to the second “postswitch” sorting rule, despite performing at ceiling on the “preswitch” rule. This perseveration occurs despite repeated instructions about the new rule and despite the fact that children can answer simple queries about this rule (Kirkham & Diamond, 2003; Munakata & Yerys, 2001; Perner & Lang, 2002; Zelazo & Frye, 1998). Even explicit negative feedback fails to eliminate perseveration (Bohlmann & Fenson, 2005). Of children who seem to flexibly switch in response to this negative feedback, some *still* fail to adopt the postswitch rule, opting instead to simply sort the cards in a way that opposes the preswitch rule. This “opposites” pattern has substantial implications for theories of cognitive flexibility.

One such theory posits a distinction between active and latent knowledge representations (Morton & Munakata, 2002). Only switchers are thought to have sufficiently strong actively maintained representations within working memory; these provide top-down support for the newly relevant dimension, helping to overcome a latent bias learned during the preswitch phase. Perseverators, by contrast, fail to sufficiently represent the current rule in this actively maintained form of working memory. Thus, improvements in working memory ability are supported in part by the increasing strength of such actively maintained representations – representations that also support cognitive flexibility.

A neural network model has been used to instantiate the active-latent account (Morton & Munakata, 2002; Munakata, 1998; Stedron, Sahni, & Munakata, 2005; for related models, see Cohen, Dunbar, & McClelland, 1990; Marcovitch & Zelazo, 2009). The model activates one of the two output units in response to inputs corresponding to both visual features of the cards and verbal instructions (Fig. 1). An internal representations layer corresponds to posterior cortical regions and represents specific colors and shapes. Activation in the units within this layer is competitive to the extent that any given

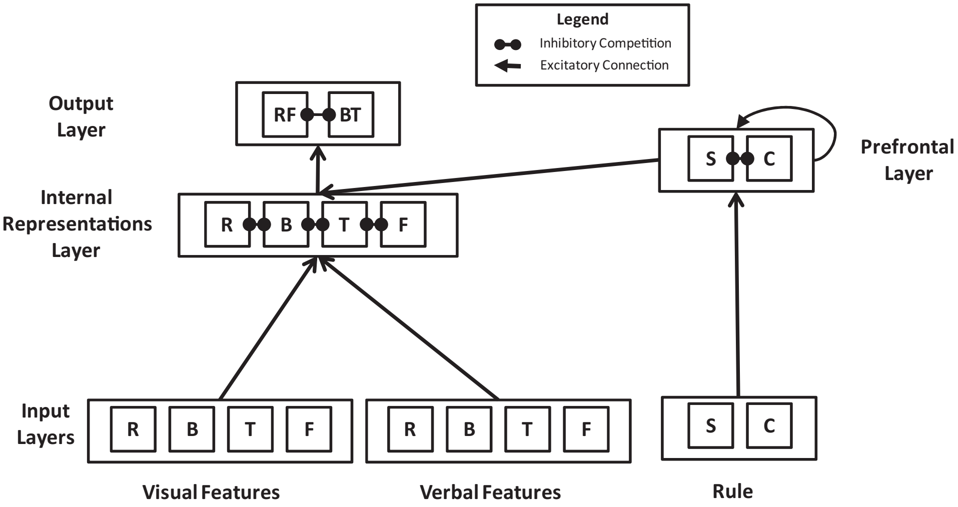

Fig. 1. Architecture of a neural network model of children's flexibility in sorting cards. The model contains three input layers, one corresponding to the visual features of the to-be-sorted cards (T = truck, F = flower, R = red, and B = blue), one to the verbal descriptions of the cards, and one to instructions provided by the experimenter (S = Shape, C = Color). Stimulus-specific information in these input layers directly activates corresponding stimulus-specific representations in the internal representations layer (connections not shown), where activation is subject to competition across units. This layer sends activation to the output layer, which has units corresponding to the target cards. A prefrontal layer receives input from the experimenter's instructions regarding the task-relevant dimension. This prefrontal layer contains self-excitatory projections that enable active processing of abstract information (shape and color) that is conveyed to corresponding units in the internal representations layer (i.e., the "S" unit activates both the "T" and "F" units, whereas the "C" unit activates both the "R" and "B" units). If represented with sufficient strength, this information drives the network to respond on the basis of the currently relevant dimension.
Adapted from Morton and Munakata (2002).

unit being more active makes it more difficult for other units within the layer to also become active. This posterior cortical layer mediates the flow of activation from corresponding input and output units. Finally, a layer corresponding to the prefrontal cortex persistently maintains information over time, thus constituting a form of working memory maintenance (Goldman-Rakic, 1995) that is here used to represent the current sorting rule. Activation in this prefrontal layer biases not only the competition unfolding within the internal representation layer (so as to support efficient processing of the currently relevant dimensions) but also itself – thereby providing an actively maintained representation that supports abstract, dimensional processing.

During preswitch, the posterior pathway supporting the processing of the preswitch dimension is strengthened through associative (Hebbian) learning. Thus, only those networks with strong prefrontal representations can subsequently bias the posterior cortical or "internal representations" layer to process the postswitch dimension in the face of the strengthened and competing associations learned during preswitch. For this reason the model correctly predicts the relatively bimodal distribution of performance on the DCCS: Children initially switch and continue to do so or wholly perseverate. The model also predicts that children should succeed on simple queries about the postswitch rules, but that they should fail when those queries contain the same conflict as the to-be-sorted cards, as shown by Munakata and Yerys (2001), because only the latter queries will activate the strengthened and competing but now inappropriate preswitch dimension. The model further predicts that switchers should show a reaction time advantage in responding to queries about one-dimensional stimuli due to their superior prefrontal representations of abstract rules, confirmed by Blackwell, Cepeda, and Munakata (2009), as well as the concomitant benefits that these abstract representations should confer on the ability to generalize a sorting rule to new stimuli, confirmed by Kharitonova, Chien, Colunga, and Munakata (2009). Finally, the model can be encouraged to switch through scaffolding of the input in postswitch, which strengthens the posterior cortical pathway for the postswitch dimension, such scaffolding also improves children's performance (Brace, Morton, & Munakata, 2006).

Despite simulating many phenomena from the DCCS, and making numerous predictions that have been verified behaviorally in children, the model has limitations. In particular, it has no mechanism for simulating the effect of feedback. This decision was made in part because feedback does not seem to affect perseveration in other developmental domains. For example, in the A-not-B task, infants reliably perseverate in reaching for toys in previous hiding locations, regardless of whether they are rewarded by being allowed to play with the toy anyway or "punished" by being prevented from playing with the toy (Smith, Thelen, Titzer, & McLin, 1999). Models have thus successfully simulated performance on this task despite lacking mechanisms for feedback processing (Marcovitch & Zelazo, 2009; Munakata, 1998). For this reason, the role of feedback in the DCCS is worthy of careful consideration.

Although feedback substantially improves performance in the DCCS (Bohlmann & Fenson, 2005) and related tasks (Chevalier, Dauvier, & Blaye, 2009), it might do so in a counterintuitive way: feedback could merely encourage children to perform the mapping opposite to the preswitch mapping(Morton, Trehub, & Zelazo, 2003), rather than encourage adoption of the new postswitch rule. For example, if the postswitch rule is to sort by shape, children can correctly sort a blue flower card in the "red flower" pile either by perseverating on the color dimension while *reversing* their behavior (i.e., putting blue things with red things; an intradimensional reversal), or by properly switching to the newly relevant dimension of shape (i.e., putting flowers with flowers; an extradimensional shift). These possibilities can be distinguished by determining whether children place a blue flower in the blue flower pile; children performing an intradimensional reversal should fail this simple test. The opposites strategy thus reflects another form of perseveration; children perseverate on the preswitch dimension and reverse their responses based on negative feedback, rather than truly switching to the postswitch dimension as instructed.

We assessed this possibility empirically and found that after receiving feedback during postswitch, some children indeed failed to sort such "identity" cards correctly (i.e., they sorted red flowers with blue trucks, and blue trucks with red flowers), reflecting an "opposites" strategy in response to feedback. Subsequent simulations indicate that this phenomenon could not be captured by the original model. The elaborations to the model that are required to capture this finding involve the use of working memory updating mechanisms, potentially supported by the basal ganglia, which have not been previously considered a major factor in developmental progression on this task.

2. Method

2.1. Participants

Twenty-four 3-year-olds ($M = 36$ months, range ± 6 days; 12 female; 87% Caucasian) were recruited through a university pool. Parents received a small gift for their child's participation and $5 to cover travel costs. Two children were excluded from the analyses because of experimental error. One additional child failed to sort by the first rule, and was removed from all relevant analyses of switching ability.

2.2. Design and procedure

Children were asked to sort cards into trays according to one rule (based on shape or color) in a preswitch phase, and according to a different rule in a postswitch phase. In contrast to previous studies that provided feedback only in the preswitch phase (Munakata & Yerys, 2001; Zelazo et al., 1996), children received negative feedback in both phases. Ordering of the rules, the target cards, and their locations were counterbalanced across participants.

The procedure was adapted from the one used by Munakata and Yerys (2001). Each child sat at a table across from the experimenter (E). There were two trays on the table, each with a target card fastened above it (e.g., a red truck above one tray, and a blue flower above the other). The cards to be sorted (blue truck, red flower) matched each target card on one dimension. In the preswitch phase, E provided a rule to sort cards by one dimension (e.g., “In the color game, all the red ones go here, but all the blue ones go there”) and proceeded to sort one card (a blue truck or red flower) face down into each tray. On the following six trials, E reminded the child of the rules, labeled a card by the relevant dimension (e.g., “Here’s a red one”) and asked, “Where does this go in the color game?” E provided feedback on every trial. If the child sorted correctly, the experimenter said, “Good job” in a happy tone. If the child sorted incorrectly, E corrected the sort with the card facing down and said, “This is not right. This one goes here” in a soft tone.

In the postswitch phase, E said, “Now we’re going to switch and play a new game, the shape game. We’re not going to play the color game anymore. No way. We’re going to play the shape game, and the shape game is different.” Then, E provided the new rules (e.g., “In the shape game, all the flowers go here, but all the trucks go there”). On the following six trials, E reminded the child of the postswitch rules, labeled a card by the relevant dimension (e.g., “Here’s a truck”) and asked, “Where does this one go in the shape game?” Feedback was provided on every trial as in the preswitch phase. Finally, the child was asked to sort two “identity” cards, each of which matched one target card on both dimensions – in other words, red trucks merely needed to be sorted with red trucks and blue flowers with blue flowers.

3. Results

3.1. Behavioral data

Due to non-normal distributions of both the present sample (91.7% of children sorted all cards correctly in the preswitch phase and 87.5% of children sorted both identity cards correctly or incorrectly) and a “no-feedback” sample drawn from an earlier study (Munakata & Yerys, 2001), we used a chi-square test (with Yates correction for small cell entries) to compare them. Children were classified as passing each sorting phase if they sorted at least four out of the six cards correctly. Children were classified as passing the identity sort if they sorted both cards correctly.

Feedback improved children’s postswitch sorting substantially. Nineteen of 24 children in the feedback condition passed the postswitch phase, compared to 6 of 16 children in the no-feedback condition, $\chi^2(1, N=40)=5.44$, $p=.02$. Children who sorted perfectly did not receive any negative feedback, so they were treated identically in the feedback and no-feedback conditions. A more direct test of the effects of feedback thus excluded these children. Again, more children in the feedback condition (15 of 20) sorted correctly in postswitch compared to the no-feedback condition (1 of 11), $\chi^2(1, N=31)=9.187$, $p=.002$.

Although feedback helped children sort correctly in the postswitch phase, we do not know from this analysis whether they truly switched from the preswitch rule to the postswitch rule, or if they used an opposites strategy with the preswitch rule. Children in the feedback condition were classified as truly switching if they passed both the postswitch phase and identity sorting. Children in the no-feedback condition were classified as truly switching if they passed postswitch sorting. Identity sorting was not tested in the no-feedback condition because children would have no reason to play an opposites game in the postswitch phase when they were receiving no feedback. Fifteen of 24 children in the feedback condition truly switched, compared to 6 of 16 in the no-feedback condition. This comparison did not reach significance $\chi^2(1, N=40)=1.508$, $p=.219$. However, with a more direct comparison (removing children who sorted perfectly in postswitch), more children truly switched in the feedback condition (11 of 20) than in the no-feedback condition (1 of 11), $\chi^2(1, N=31)=4.518$, $p=.034$. Thus, among the 15 of 20 children who appeared to switch to the second rule in response to negative feedback, more than 25% were actually using an opposites strategy.

Moreover, feedback did not completely eliminate perseveration. Even with feedback, 9 of 24 children still failed to truly pass the postswitch phase. This performance was significantly worse than children’s true preswitch performance, whereas one would expect similar performance on preswitch and postswitch if feedback eliminated all perseveration. Instead, 9 children truly passed preswitch while failing postswitch, whereas only one child showed the reverse pattern, indicating that children perseverate even with feedback, $\chi^2(1, N=25)=6.4$, $p=.021$. The more direct comparison (removing children who sorted perfectly on postswitch) yielded the same result, since all 4 perfect postswitch sorters fall in the irrelevant preswitch-pass/postswitch-pass cell.[1]

3.2. Simulations

Empirically we observed that feedback improves DCCS performance while falling short of completely eliminating perseveration and surprisingly also encourages some children to adopt an opposites strategy. In an attempt to capture these phenomena, we extended the previous neural network models of these tasks (Morton & Munakata, 2002; Stedron et al., 2005), which were based on an exclusively feedforward architecture, to include mechanisms for supporting feedback processing. We implemented feedback as an error-driven learning signal, such that the model's weights were adjusted following each card sort to make the correct answer more likely. This entailed the use of bidirectional connectivity between the output layer (where the correct answer is provided) and the internal representations layer, so that error signals could propagate backwards into the rest of the network.[2]

This model, and substantial variations thereof, categorically failed to produce the observed patterns of data. As described below, this failure persisted throughout a systematic search across parameter settings in the model, including variations in the strength of feedback connections, the learning rate, the relative mixture of associative and error-driven learning, and changes to network connectivity. These failures were nonetheless highly informative for reasons we will describe.

With low learning rates, no model could learn to associate red with blue (and blue with red) – that is, models failed to learn the requisite stimulus-response mappings for executing the “opposites” strategy. With higher learning rates, models began to learn the *correct* mapping more quickly and failed to learn the opposites mapping to a degree that would dominate processing in the identity card sort. High learning rates also led all models to switch after just an instance or two of negative feedback – contrary to the finding that feedback does not eliminate all perseveration. Decreasing the sensitivity of the models to negative feedback, either by reducing the strength of the feedback connections, or by reducing the amount of error-driven learning in particular, also failed; models then showed too little sensitivity to negative feedback. To establish that these failures were due to computational tradeoffs inherent to the model, we performed a brute force search through 562,500 networks; none of these demonstrated an acceptable match to empirical observation.

These failures prompted us to reconsider the computational demands. For an opposites strategy to be learned, representation of the old rule must be sufficiently robust for error-driven learning to remap the preswitch dimensions (i.e., red goes with blue), as opposed to merely strengthening the processing of postswitch dimensions. Yet the level of prefrontal recurrent connectivity that is required for this kind of robustness is at least as large as that supporting switching in the first place, even in the absence of feedback.

We therefore reasoned that while perseverators may often have insufficient prefrontal representations to support processing of the new rule, those children that use the opposites strategy may be better characterized in a different way. One possibility is that the opposites strategy is driven by a failure to update the prefrontal cortex with the currently relevant rule. This updating function is thought to be subserved by striatal mechanisms; indeed, striatal mechanisms are becoming an important component of network models of higher-level cognition for precisely this reason (Chatham et al., 2011; Hazy, Frank, & O'Reilly, 2007; Kruger & Dayan, 2009; O'Reilly & Frank, 2006; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). In addition to these biological and the aforementioned computational reasons for including an updating mechanism in our model, previous work indicates that updating mechanisms may be particularly important for capturing switch cost phenomena in adults (Reynolds,

[1] These are the only analyses to include the single child who failed preswitch, as is required to accurately compare performance across preswitch and postswitch; to be conservative, this child was classified as passing postswitch (despite not actually needing to switch in the postswitch phase).

[2] Not only is such bidirectional connectivity a ubiquitous feature of cortex (Felleman & Van Essen, 1991); it also enables the use of a biologically-plausible form of error-driven learning in the current model (known as “Generalized Recirculation;” O'Reilly, 1996).

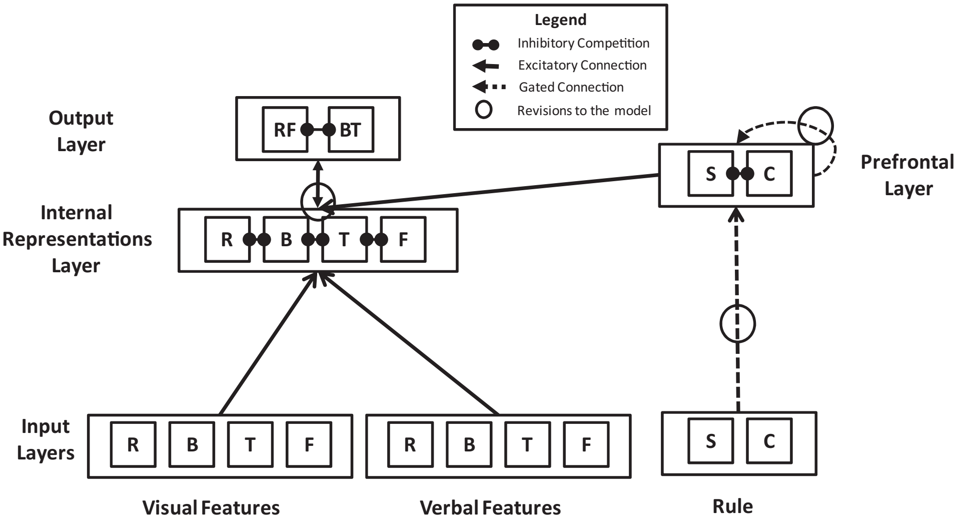

Fig. 2. Architecture of the revised model. Revisions to the original model (Fig. 1) are circled here. One is the inclusion of bidirectional connectivity from the output layer to the internal representations layer; these connections allow feedback signals to affect the activation patterns elsewhere in the network. The new model also includes gated projections to the prefrontal layers (dotted lines). Thus, a prefrontal layer can receive input from the task instructions, if the network is allowed to update this information. This prefrontal layer now also contains intrinsic maintenance currents that are momentarily cleared as the network is updating information, if allowed to do so.

Braver, Brown, & van der Stigchel, 2006). This finding naturally raises the possibility that updating mechanisms are also important for capturing task-switching phenomena in children, such as the one we have observed.

Our revised model with these updating mechanisms is portrayed in Fig. 2. The model is conceptually similar to the previous one except that the prefrontal layers receive input from the instructions input layer *only* if the network is allowed to update. Also, the units of the prefrontal layer now also contain intrinsic maintenance currents that provide support for self-sustaining activity but are momentarily disabled (such that there are momentarily no intrinsic maintenance currents) during the time when the network is allowed to update.[3] This latter component to the updating mechanism allows networks to more stably represent the current rule but also prevents representations from being so stable as to also be inflexible. These prefrontal mechanisms are also conceptually similar to those used in the prefrontal layers of larger-scale models with basal ganglia-mediated updating mechanisms trained via reinforcement learning. In the current model we omit these details of the basal ganglia circuit, opting instead for a more abstract implementation, where updating is manipulated as an individual differences variable (Reynolds et al., 2006) akin to the way that the strength of prefrontal recurrent connections is here modeled as an individual differences variable (Morton & Munakata, 2002). In addition to capturing the phenomena simulated by the previous model, this revised model readily succeeded in capturing the effects of feedback.

As in prior work, all models performed at a 100% correct level during preswitch. However, only those networks with a sufficiently strong level of prefrontal recurrent weights were capable of switching without feedback when allowed to update; all other networks failed to perform well on the postswitch phase of the task (Fig. 3A). This phenomenon occurs for the same reason as in the original model: At low recurrent weights, the prefrontal representation of the new rule is not sufficiently strong to overcome the associative learning that occurred during preswitch, such that the output layer as well as the

[3] Additional changes were necessary to the model, including changes to the learning rate and a compensatory change to the contrast function on weight change. The authors can be contacted for details.

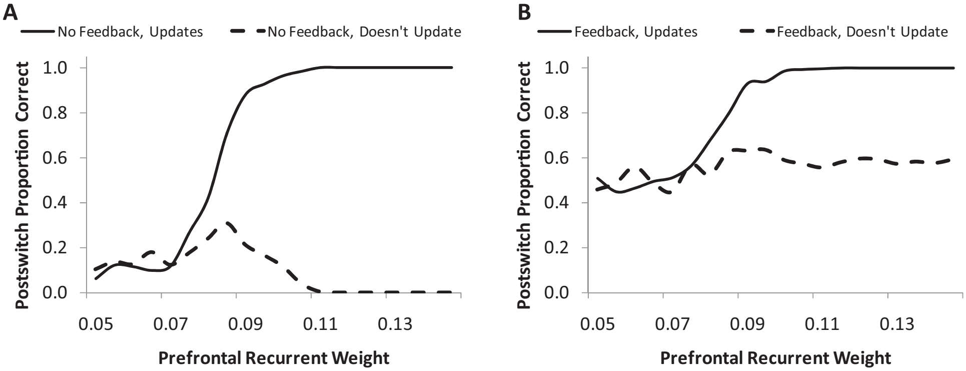

Fig. 3. Postswitch performance of the revised model, both (A) without feedback and (B) with feedback, for networks that successfully updated the new rule (solid lines) and those that did not (dotted lines). A. Postswitch performance was a function of prefrontal recurrent weight, such that networks that successfully switched were those networks with strong prefrontal recurrent weights and which could successfully update the new rule. In contrast, prefrontal recurrent weight was largely without effect among networks that failed to update the new rule. B. Postswitch performance was much improved when feedback was provided during postswitch, in keeping with empirical data (Bohlmann & Fenson, 2005; original data).

internal representations layer have come to more strongly represent stimuli in terms of their features along the preswitch dimension than in terms of their features along the postswitch dimension. Because these dimensions are in conflict, and because they compete in the model, strong recurrent weights are necessary to produce the active representations that can overcome these latent habits.

When feedback was provided to these networks, postswitch performance was much improved (Fig. 3B). The improvement due to feedback comes largely from the fact that feedback tends to increase the model's weights along the postswitch dimension. The difference between these activations is then used to adjust the model's weights. As such, feedback has its effect on reducing the habits built up through preswitch and strengthening the processing of postswitch features, as opposed to strengthening more active and abstract prefrontal representations. In this way, the model explains why feedback on one sorting rule in the DCCS does not improve performance on sorting according to subsequent rules (Bohlmann & Fenson, 2005).

However, and as observed in children, feedback did not eliminate perseveration. Indeed, a number of networks failed to achieve at least 4 of 6 cards correct in postswitch despite receiving feedback on every trial (Fig. 4A; right side, black bar), consistent with our observations in children. Feedback sometimes fails to yield criterial postswitch performance because networks often require multiple experiences of feedback before the resulting weight changes are sufficiently strong to overcome the associative learning which occurred in preswitch.

Thus, although not all models pass the typical criterion for postswitch performance, these learning signals do have a measurable effect. For example, feedback increases the likelihood that models will transition from an incorrect sort to a correct sort relative to the reverse transition (Fig. 4B). A more subtle version of the same pattern can also be observed in models performing the DCCS without feedback, consistent with empirical observations (van Bers, Visser, van Schijndel, Mandell, & Raijmakers, 2011), because each trial is associated with verbal instructions and a restatement of the correct sorting rule, both of which give networks an opportunity to strengthen their representations of features along the postswitch dimension.

In some cases, this feedback did lead networks to fail the identity card sort. This behavior was highly conditional on recurrent weights of the prefrontal units (Fig. 5A) and present only among those models that failed to update, such that models with high recurrent weights tended to fail the subsequent identity card sort if they failed to update the new rule. This phenomenon occurs because the continued strong representation of the preswitch dimension causes the models to learn an opposites mapping, rather than to strengthen its processing of the postswitch sorting rule. Models with lower recurrent weights tended not to become so strongly "stuck" on the preswitch dimension

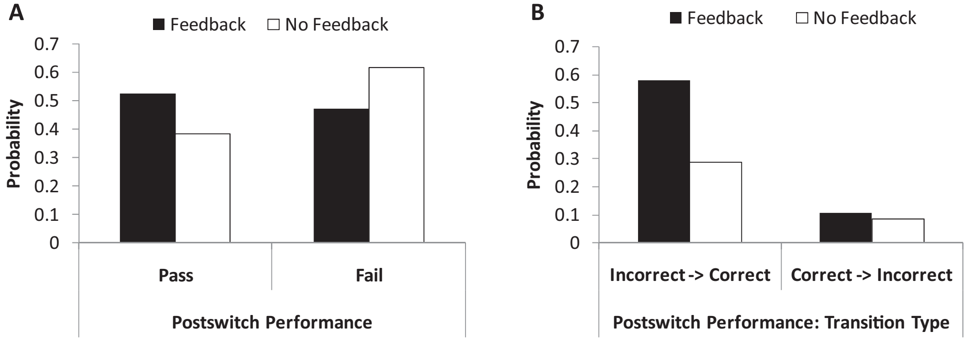

Fig. 4. The effects of feedback on criterial postswitch performance (A) and transition probabilities (B). A. Feedback substantially improved model performance: Networks tended to get fewer trials correct without feedback (white bars) than with feedback (black bars). However, some networks failed to reach the criterion for passing postswitch despite feedback. B. Across all networks, feedback increased the likelihood of transitioning from incorrect sorting to correct sorting. Nonetheless, even without feedback, models were more likely to make that transition than to show the opposite transition (from correct to incorrect) across subsequent trials in postswitch, consistent with empirical data (van Bers et al., 2011).

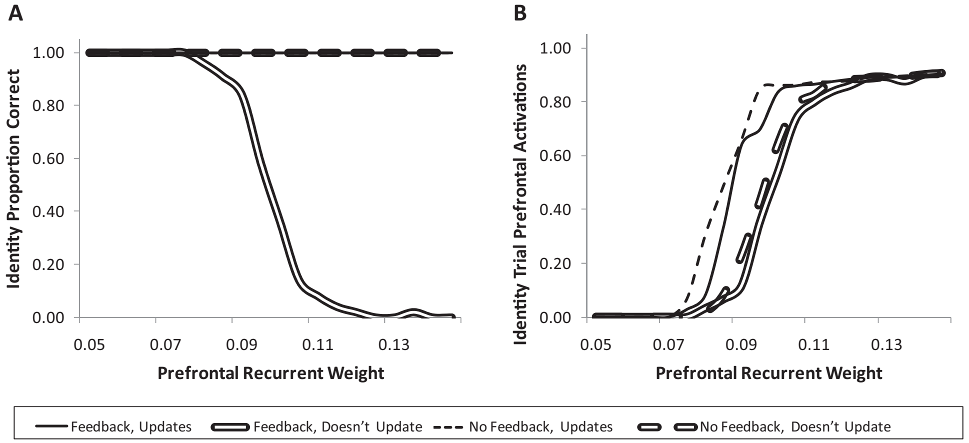

Fig. 5. Identity card sort performance (A) and prefrontal activations (B). (A) Identity card sort performance was largely at ceiling, with the exception of networks that failed to update the new rule but had strong prefrontal recurrent weights. (B) The activation within the prefrontal layers was highest overall for networks that updated, regardless of whether they received feedback (dotted) or not (solid). In contrast, networks that failed to update showed lower prefrontal activations across all recurrent weight values, regardless of whether they received feedback (dashed hollow line) or did not (continuous hollow line).

and therefore learn to correctly use this feedback to strengthen their representations of the features relevant to the postswitch rule. The model thus predicts that the “opposites” approach observed in children results from relatively poor updating of working memory, in conjunction with relatively robust working memory maintenance; as such, the children who display the opposites approach are only a subset of those who would normally perseverate on the no-feedback DCCS.

If some of those children who perseverate on the no-feedback version of the DCCS in fact have good working memory maintenance abilities, why should perseverators show *poorer* generalization of their sorting rules to novel stimuli, when good working memory maintenance abilities are thought to support good generalization (Kharitonova et al., 2009)? One possibility is that children who show the opposites strategy in response to feedback constitute a relatively small proportion of those who perseverate on the no-feedback DCCS. It is also possible that this proportion do not as robustly represent their sorting rule as switchers, despite the good working memory maintenance abilities of both groups. We assessed this latter possibility in the model by examining the absolute level of prefrontal activation in all networks, which corresponds more directly to maintenance of the sorting rule than prefrontal recurrent weights.

Indeed, although networks that utilized an "opposites" strategy had relatively high levels of recurrent weights among the prefrontal units, the activation within these units after postswitch was significantly lower than those occurring in other networks with equivalent recurrent weights (Fig. 5B). Specifically, activation in these units was lower in networks that failed to update and received feedback than within those networks that also failed to update but did not receive feedback, $t(20) = 2.59$, $p = .018$, and lower than within those networks that did update, regardless of whether they received feedback (both $t > 3.09$, $p < .007$ for both). This pattern arises from the benefit accrued by the provision of the rule on every trial among networks that can update; networks that fail to update also fail to benefit from this information. Moreover, this pattern indicates that those children who would display an opposites strategy in the feedback version of the DCCS – but who perseverate in the no-feedback DCCS – would in general be at a disadvantage in tests of generalization relative to switchers in the no-feedback DCCS, due to less robust maintenance of their sorting rule.

4. Discussion

While feedback does not seem to strongly influence some perseverative behaviors in childhood (Smith et al., 1999), feedback has been shown to substantially reduce perseveration in children's card-sorting (Bohlmann & Fenson, 2005). To clarify this apparent discrepancy, we administered a feedback version of a card-sorting task that enabled us to assess whether feedback improved children's performance by supporting "true switching" to the new rule (i.e., the instructed extradimensional shift) or by supporting the adoption of an "opposites" strategy (i.e., an intradimensional reversal). We found that while feedback does encourage some children to truly switch to the new rule, others persistently fail to correctly sort the majority of cards despite this feedback. Yet another group of children adopt an "opposites" strategy and thereby fail to correctly match cards identical to one another (i.e., they sort red flowers in the blue truck pile, and blue trucks in the red flower pile). This diversity importantly constrains the interpretation of prior reports of the effects of feedback in the DCCS (Bohlmann & Fenson, 2005) and provides a strong challenge for computational models of the development of cognitive flexibility.

Our previous model of this task was extended to simulate the effects of feedback but failed to capture the diverse patterns observed empirically. This failure reflected several computational tradeoffs. First, the high learning rates necessary to yield sensitivity to feedback tended to support true switching instead of the opposites approach (i.e., reversed processing of features from the preswitch dimension). Second, the degree of working memory maintenance required to support sustained processing of the preswitch features – with sufficient strength so that those stimulus-response mappings could be *reversed* by feedback – was at least as great as that required to support switching in the first place.

These failures prompted us to consider the possible importance of working memory updating in the DCCS. Indeed, working memory ability is now widely acknowledged to be an interaction of both working memory maintenance mechanisms (similar to those recurrent prefrontal connections implemented in prior instantiations of the active-latent framework) and working memory updating. To this end we included an updating mechanism, putatively subserved by striatal regions in a revised version of the model. Although updating is not explicit in theories of the development of cognitive flexibility, this revision nonetheless makes contact with a growing literature on the functional recruitment of striatal regions during cognitive flexibility tasks (Casey et al., 2004; Rubia et al., 2006; Wager, Jonides, Smith, & Nichols, 2005), as well as with the abnormal recruitment of striatum in psychopathologies characterized by poor flexibility (Baym, Corbett, Wright, & Bunge, 2008; Britton et al., 2010). This revised model succeeded in simulating the effects of feedback in the DCCS, yields testable predictions, and provides novel insight into the neurocognitive processes that may underlie the development of cognitive flexibility.

In this revised model, the opposites approach was found to emerge among networks that failed to update in response to the postswitch rule but also had sufficient working memory maintenance to strongly bias their processing in favor of the preswitch rule. In contrast, models that also failed to update in response to postswitch but had *reduced* working memory maintenance failed to achieve criterial postswitch accuracy despite feedback. All other models behaved in accordance with the use of the postswitch rule, but the effects of feedback were uniformly stimulus-specific in nature, thereby explaining why feedback does not tend to improve sorting performance according to subsequent rules (Bohlmann & Fenson, 2005).

In theory, a more abstract, rule-based representation could support the opposites strategy (O'Reilly, Noelle, Braver, & Cohen, 2002). We consider it unlikely that children of this age utilize abstract, rule-based representations for "opposites," given their poor performance on tasks like the Day-Night Stroop (Gerstadt, Hong, & Diamond, 1994; Diamond, Kirkham, & Amso, 2002) and Luria's tapping task (Diamond & Taylor, 1996). Nonetheless, our hypothesis is testable by virtue of a distinguishing and novel prediction from our model: Children who use the opposites strategy should nonetheless succeed in sorting novel identity cards correctly. However, if the strategy entails utilizing an abstract rule-based representation of an "opposites rule," they should not.

Our model makes two additional novel predictions. First, the model predicts that children who use the opposites strategy should show better working memory maintenance abilities than many of the children who truly switch in response to feedback in the DCCS. Thus, these children may be at an advantage on speeded tasks that do not involve working memory updating, for example those tasks that require rapidly sorting one-dimensional cards or answering non-conflict questions (Blackwell et al., 2009). Second, the model predicts that these same children should perform particularly poorly in tasks that strongly tax working memory updating, even those where an "opposites" strategy might be expected to yield good performance. One example of this situation is the aforementioned tapping task (Diamond & Taylor, 1996), in which children must tap a surface once if the experimenter taps it twice (and vice versa). Although use of an opposites strategy might be thought to yield good performance on this task, our model predicts that children who use it in the DCCS will nonetheless perform poorly here; they should be particularly prone to difficulty due to failures in updating working memory with the number of taps produced by the experimenter. If such predictions were falsified, this would pose a significant challenge to the current account.

It is always possible that a model gives rise to phenomena of interest for the wrong reasons; computational, biological, and psychological constraints are necessary ingredients for a more compelling case (Massaro, 1988; McClelland, 2009; O'Reilly and Munakata, 2000; Regier, 2003). Our model is strongly constrained at each of these levels. First, as described earlier, several computational tradeoffs motivate revision of the model to include updating mechanisms.[4] Second, our model is also constrained biologically: We utilize a biologically plausible form of error-driven learning, mechanisms for actively maintained working memory that are based on prefrontal anatomy, and an updating mechanism that is conceptually similar to that used in computational models of detailed neurophysiological phenomena (O'Reilly & Frank, 2006). Finally, our model is psychologically constrained by its adherence to the procedures of the DCCS and by its grounding in recent theorizing that difficulty in the DCCS must arise at least in part from failures to appreciate the importance of instructions to switch tasks, and not merely from the insufficient maintenance of abstract, rule-based representations (Kloo, Perner, Kerschhuber, Dabernig, & Aichhorn, 2008).

There are certainly important caveats to the current work. First, we simulate feedback using an associative learning signal that is confined largely to the processing of stimulus-response mappings and stimulus features (i.e., to the internal representations layer of the model). Variations of the current model in which feedback directly affects prefrontal representations fail to reliably give rise to the opposites strategy – learning within the prefrontal layers simply occurs too slowly (Rougier et al., 2005) for the opposites strategy to emerge. Nonetheless, feedback does have effects on rule-like prefrontal representations, effects that may be mediated by the kind of striatally based reinforcement learning mechanisms over which we have abstracted here (Reynolds et al., 2006). Thus, implementation of these more complex mechanisms is an important direction for future work. Second, our data are limited not only in that we observed a small number of opposites strategy users but also in that it is possible a similar proportion of children would use an opposites strategy even without feedback; such a result

[4] To our knowledge, the particular computational tradeoffs we report are novel, but they are not the only computational tradeoffs that suggest the necessity of updating mechanisms (the stability-flexibility dilemma; Goschke, 2003).

would pose a challenge to many theories of cognitive flexibility, including our own. However, such concerns are significantly tempered by the fact that the opposites strategy can be robustly observed in other tasks, but only under conditions of feedback (Morton et al., 2003). Third, there are inherent limitations in using a comparison group from a previous study, although we have endeavored to minimize these limitations by using the same recruiting methods, materials, and even experimenter across the two studies compared here.

Despite these constraints and limitations, our model not only succeeds in capturing numerous phenomena from the DCCS, but also highlights a possible point of convergence across multiple competing accounts of the development of cognitive flexibility. In particular, our inclusion of a mechanism that may fail to flexibly update abstract, actively maintained and graded representations in prefrontal cortex provides one formal mechanistic basis for the attentional inertia account of perseveration (Kirkham, Cruess, & Diamond, 2003), by which children may literally become “stuck” on the preswitch features of the task. According to our model, this attentional inertia could in part reflect a failure of striatal mechanisms in updating an interconnected prefrontal area with the current rule.[5]

Functional neuroimaging of children performing the DCCS indicates that subcortical and various prefrontal areas are recruited during postswitch (Morton, Bosma, & Ansari, 2009). In adults, strikingly similar foci have been observed in tasks of similar complexity to the DCCS (Badre, Kayser, & D’Esposito, 2010), including dorsal premotor cortex (pMD), an area just anterior to pMD (so-called “pre-pMD”), and a striatal region in the posterior caudate. Frontostriatal loops of this kind are thought to be arrayed along a rostro-caudal hierarchy, such that more rostral regions of the prefrontal cortex influence the updating signal between more caudal frontostriatal loops (Badre & Frank, 2012; Frank & Badre, 2012). An extension of our model to include a more rostral prefrontal area could be used to investigate the developing control over the gating process that our model suggests is critical and to test the role of insufficiently strong abstract representations within these more rostral prefrontal regions in updating failures (and in turn, perseveration).

Importantly, this extension highlights a possible link between our account of cognitive flexibility and the Cognitive Complexity and Control theory of perseveration (CCC theory; Bunge & Zelazo, 2006; Zelazo & Frye, 1998; Zelazo, Muller, Frye, & Marcovitch, 2003). According to CCC theory, perseverators fail to sufficiently represent a “superordinate” rule that supports switching between the rules of preswitch and postswitch. If rostral areas of prefrontal cortex represent such superordinate rules, then insufficient maintenance of these superordinate rules could contribute to updating failures. In this way, our model points toward a biological mechanism that might correspond to the CCC theory of perseveration (Bunge & Zelazo, 2006), and a possible mechanistic integration of this account and our own.

Updating failures will of course almost certainly take multiple forms. Updating failures could arise from the insufficient maintenance of superordinate rules but may also arise from poor monitoring of the environment for the demand to switch tasks (Chevalier, Wiebe, Huber, & Espy, 2011). In addition, striatally based gating mechanisms of the kind used here for updating may also be important for processes like rule selection (Chevalier & Blaye, 2009). More broadly, our discussion has so far focused only on ways in which updating might not occur when it should, but updating may also fail by occurring when it should not. For example, the model posits that updating will briefly disable the intrinsic maintenance currents that support rule representation; thus, erroneous updating could momentarily interrupt the intrinsic maintenance of currently relevant goals and might thus be viewed as a source of the “goal neglect” that is known to occur within the DCCS (Marcovitch et al., 2007; Marcovitch et al., 2010; Towse et al., 2007).

This multiplicity of plausible updating dynamics poses a clear challenge for specifying a precise account of how updating may change with age. For example, updating failures could initially take the form explored in the current model, by failing to occur when they would be desirable (i.e., a

[5] We note that the original model provided an alternative mechanistic basis for the attentional inertia account, such that insufficiently strong, active representations of the postswitch rule could fail to overpower the latent habit-based memories that were associatively learned during preswitch. In contrast, the current model demonstrates a more active and perhaps “executive” form of the attentional inertia account. Future work may seek to discriminate between these two qualitatively different forms of attentional inertia.

"miss"). Subsequent developmental changes might lead to greater updating (i.e., "false positives," where updating occurs when it should not) and thus possibly to a greater incidence of goal neglect at the same time as prevalence of the opposites strategy is declining. Alternatively, maturation may be associated with increasingly discriminative updating, rather than a simple tradeoff between updating misses and false positives. In this case, both goal neglect and phenomena like the opposites strategy would be expected to decrease with age at roughly similar rates. Finally, both patterns might be observed, such that children increasingly suffer from updating false positives as opposed to misses until some criterial stage of development, at which point superordinate rules can be used to drive more discriminative updating. Future work should seek to distinguish between these possibilities to identify if and how updating differentially contributes to cognitive flexibility across development.

5. Conclusions

Models are powerful tools in part *because* they cannot simply "do anything" – they often fail, for theoretically and biologically informative reasons. Computational tradeoffs in learning have led to a richer understanding of the division of labor between neocortex and hippocampus, and tradeoffs in a proportional reasoning model highlight the often-unacknowledged role of selective attention in that domain. Our model, too, suffered from computational tradeoffs in simulating a novel task-switching effect, in which feedback sometimes encourages children to adopt an "opposites" strategy instead of truly switching to the newly instructed rule. A model revised to include updating mechanisms – a factor not typically considered in developmental investigations of task-switching – not only captures our findings but suggests a possible point of convergence across multiple theories of the development of cognitive flexibility.

Acknowledgements

We thank members of the Cognitive Development Center, Trent Kriete, and the National Institutes of Health (P50-MH079485 and RO1-HD37163).

References

Badre, D., & Frank, M. J. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 2: Evidence from fMRI. *Cerebral Cortex*, *22*, 527–536.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*, 315–316.

Baym, C. L., Corbett, B. A., Wright, S. B., & Bunge, S. A. (2008). Neural correlates of tic severity and cognitive control in children with Tourette syndrome. *Brain*, *131*, 165–179.

Blackwell, K. A., Cepeda, N. J., & Munakata, Y. (2009). When simple things are meaningful: Working memory strength predicts children's cognitive flexibility. *Journal of Experimental Child Psychology*, *103*, 241–249.

Bohlmann, N. L., & Fenson, L. (2005). The effects of feedback on preservative errors in preschool aged children. *Journal of Cognition and Development*, *6*, 119–131.

Brace, J. J., Morton, J. B., & Munakata, Y. (2006). When actions speak louder than words: Improving children's flexibility in a card-sorting task. *Psychological Science*, *17*, 665–669.

Britton, J. C., Rauch, S. L., Rosso, I. M., Killgore, W. D. S., Price, L. M., Ragan, J., et al. (2010). Cognitive inflexibility and frontal-cortical activation in pediatric obsessive–compulsive disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*(9), 944–953.

Bunge, S., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, *15*, 118–121.

Casey, B. J., Davidson, M. C., Hara, Y., Thomas, K. M., Martinez, A., Galvan, A., et al. (2004). Early development of subcortical regions involved in non-cued attention switching. *Developmental Science*, *7*(5), 534–542.

Chatham, C. H., Herd, S. A., Brant, A. M., Hazy, T. E., Miyake, A., O'Reilly, R., et al. (2011). From an executive network to executive control: A computational model of the n-back task. *Journal of Cognitive Neuroscience*, *23*, 3598–3619.

Chevalier, N., & Blaye, A. (2009). Setting goals to switch between tasks: Effect of cue transparency on children's cognitive flexibility. *Developmental Psychology*, *45*, 782–797.

Chevalier, N., Dauvier, B., & Blaye, A. (2009). Preschoolers' use of feedback for flexible behavior: Insights from a computational model. *Journal of Experimental Child Psychology*, *103*, 251–267.

Chevalier, N., Wiebe, S. A., Huber, K. L., & Espy, K. A. (2011). Switch detection in preschoolers' cognitive flexibility. *Journal of Experimental Child Psychology*, *109*, 353–370.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*(3), 332–361.

Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*, 303–312.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to 'Do as I say, not as I do'. *Developmental Psychobiology*, *29*, 315–334.

Diamond, A., Kirkham, N. Z., & Amso, D. (2002). Conditions under which young children CAN hold two rules in mind and inhibit a prepotent response. *Developmental Psychology*, *38*, 352–362.

Elman, J. L. (2005). Connectionist models of cognitive development: Where next? *Trends in Cognitive Science*, *9*, 111–117.

Feldman, J. A. (2010). Cognitive science should be unified: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences*, *14*, 341.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*, *22*, 509–526.

French, R. M. (1991). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the 13th Annual Cognitive Science Conference* (pp. 173–178). Erlbaum: Hillsdale, NJ.

French, R. M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, *4*(3–4), 365–377.

French, R. M. (1999). Catastrophic forgetting in neural networks. *Trends in Cognitive Sciences*, *3*(4), 128–135.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 31/2–7 years old on a Stroop-like day–night test. *Cognition*, *53*, 129–153.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485.

Gopnik, A., Wellman, H. M., Gelman, S. A., & Meltzoff, A. N. (2010). A computational foundation for cognitive development: comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences*, *14*, 342–343.

Goschke, T. (2003). Voluntary action and cognitive control from a cognitive neuroscience perspective. In S. Maasen, W. Prinz, & G. Roth (Eds.), *Voluntary action: Brains, minds, and sociality*. Oxford University Press.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Toward an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B*, *362*, 1601–1613.

Jansen, B. R. J., & van der Maas, H. L. J. (2001). Evidence for the phase transition from Rule I to Rule II on the balance scale task. *Developmental Review*, *21*, 450–494.

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–188.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.

Kharitonova, M., Chien, S., Colunga, E., & Munakata, Y. (2009). More than a matter of getting "unstuck": Flexible thinkers use more abstract representations than perseverators. *Developmental Science*, *12*, 662–669.

Kirkham, N. Z., Cruess, L., & Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimensions witching task. *Developmental Science*, *6*, 449–476.

Kirkham, N. Z., & Diamond, A. (2003). Sorting between theories of perseveration: Performance in conflict tasks requires memory, attention and inhibition. *Developmental Science*, *6*(5), 474–476.

Kloo, D., Perner, J., Kerschhuber, A., Dabernig, S., & Aichhorn, M. (2008). Sorting between dimensions: Conditions of cognitive flexibility in preschoolers. *Journal of Experimental Child Psychology*, *100*, 115–134.

Kruger, K. A., & Dayan, P. (2009). Flexible shaping: How learning in small steps helps. *Cognition*, *110*(3), 380–394.

Kruschke, J. K., & Movellan, J. R. (1991). Benefits of gain: Speeded learning and minimal hidden layers in back-propagation networks. *IEEE Transactions on Systems, Man and Cybernetics*, *21*, 273–280.

Marcovitch, S., Boseovski, J. J., & Knapp, R. J. (2007). Use it or lose it: examining preschoolers' difficulty in maintaining and executing a goal. *Developmental Science*, *10*, 559–564.

Marcovitch, S., Boseovski, J. J., Knapp, R. J., & Kane, M. J. (2010). Goal neglect and working memory capacity in 4- to 6-year-old children. *Child Development*, *81*(6), 1687–1695.

Marcovitch, S., & Zelazo, P. D. (2009). A hierarchical competing systems model of the emergence and early development of executive function. *Developmental Science*, *12*, 1–18.

Massaro, D. (1988). Some criticisms of connectionist model of human performance. *Journal of Memory and Language*, *27*, 213–234.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 8–45). New York: Oxford University Press.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 109–165). New York: Academic Press.

Messer, D. J., Pine, K. J., & Butler, C. (2008). Children's behaviour and cognitions across different balance tasks. *Learning and Instruction*, *18*(1), 42–53.

Morton, J. B., Bosma, R., & Ansari, D. (2009). Age-related changes in brain activation associated with dimensional shifts of attention: An fMRI study. *NeuroImage*, *46*, 336–358.

Morton, J. B., Trehub, S. E., & Zelazo, P. D. (2003). Sources of inflexibility in 6-year-olds' understanding of emotion in speech. *Child Development*, *74*, 1857–1868.

Morton, J. B., & Munakata, Y. (2002). Are you listening? Exploring a knowledge action dissociation in a speech interpretation task. *Developmental Science*, *5*, 435–440.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the A-not-B task. *Developmental Science*, *1*(2), 161–184.

Munakata, Y., & Yerys, B. E. (2001). All together now: When dissociations between knowledge and action disappear. *Psychological Science*, *12*(4), 335–337.

Munakata, Y., Snyder, H. S., & Chatham, C. H. (2011). Mechanistic accounts of frontal lobe development. In D. T. Stuss, & R. T. Knight (Eds.), *Principles of frontal lobe function*. Oxford University Press.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*, 895–938.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the frontal cortex and basal ganglia. *Neural Computation*, *18*, 283–328.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Noelle, D., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex*, *12*, 246–257.

Perner, J., & Lang, B. (2002). What causes 3-year-olds, difficulty on the dimensional change card sorting task? *Infant and Child Development*, *11*, 93–105.

Ratcliff, R. (1990). Connectionist models of recognition and memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 205–308.

Regier, T. (2003). Constraining computational models of cognition. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (2nd ed., Vol. 24, pp. 611–615). London: Macmillan.

Reynolds, J. R., Braver, T. S., Brown, J. W., & van der Stigchel, S. (2006). Computational and neural mechanisms of task switching. *Neurocomputing*, *69*, 1332–1336.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of United States of America*, *102*, 7338–7343.

Rubia, K., Smith, A. B., Woolley, J., Nosarti, C., Heyman, I., Taylor, E., et al. (2006). Progressive increase of frontostriatal brain activation from childhood to adulthood during event-related tasks of cognitive control. *Human Brain Mapping*, *27*(12), 973–993.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Saxe, M. D., Malleret, G., Vronskya, S., Mendez, I., Garcia, A. D., Sofroniew, M. V., et al. (2007). Paradoxical influence of hippocampal neurogenesis on working memory. *Proceedings of the National Academy of Sciences of United States of America*, *104*, 4642–4646.

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, *110*(1), 395–411.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*(2), 1–74.

Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, *10*, 104–109.

Siegler, R. S., & Chen, Z. (2002). Development of rules and strategies: Balancing the old and the new. *Journal of Experimental Child Psychology*, *81*, 446–457.

Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: The task dynamics of the a-not-b error. *Psychological Review*, *106*(2), 235–260.

Stedron, J. M., Sahni, S. D., & Munakata, Y. (2005). Common mechanisms for working memory and attention: The case of perseveration with visible solutions. *Journal of Cognitive Neuroscience*, *17*, 623–631.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Thomas, M. S. C., McClelland, J. L., Richardson, F. M., Schapiro, A. C., & Baughman, F. (2009). Dynamical and connectionist approaches to development: Toward a future of mutually beneficial co-evolution. In J. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), *Toward a new unified theory of development: Connectionism and dynamical systems theory re-considered*. Oxford: Oxford University Press.

Towse, J. N., Lewis, C., & Knowles, M. (2007). When knowledge is not enough: The phenomenon of goal neglect in preschool children. *Journal of Experimental Child Psychology*, *96*, 320–332.

van Bers, B. M., Visser, I., van Schijndel, T. J., Mandell, D. J., & Raijmakers, M. E. (2011). The dynamics of development on the dimensional change card sorting task. *Developmental Science*, *14*, 960–971.

Wager, T. D., Jonides, J., Smith, E. E., & Nichols, T. E. (2005). Toward a taxonomy of attention shifting: Individual differences in fMRI during multiple shift types. *Cognitive Affective and Behavioral Neuroscience*, *5*(2), 127–143.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., et al. (2001). Autonomous mental development by robots and animals. *Science*, *291*, 599–600.

Zelazo, P. D., & Frye, D. (1998). II. Cognitive complexity and control: The development of executive function. *Current Directions in Psychological Science*, *7*, 121–126.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development*, *11*, 37–63.

Zelazo, P. D., Muller, U., Frye, D., & Marcovitch, S. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development*, *68*(3, Serial No. 274).