0.1 Introduction

Autism is a complex developmental disorder diagnosed by the presence of a triad of symptoms: social impairments, communication impairments, and repetitive stereotyped behaviors. The severity, and sometimes even the presence, of these symptoms varies greatly across those afflicted with the disorder. Due, in part, to this variability, autism is generally seen as a spectrum of disorders, collectively known as *autism spectrum disorders* (ASD). Steady progress has been made in early diagnosis, as well as in the design of interventions that mitigate problematic behaviors. However, no consensus has been reached concerning the neural basis of autism. This chapter presents a theoretical framework that explains many aspects of autistic behavior in terms of particular neurological differences. Specifically, *computational cognitive neuroscience* modeling methods are used to demonstrate how dysfunctional interactions between the midbrain dopamine (DA) system and the prefrontal cortex (PFC) could give rise to many of the behavioral patterns seen in ASD.

People with autism exhibit difficulties on a range of cognitive tasks. These tasks assess such capabilities as flexible adaptation, planning toward a goal, the generation of novel ideas, and determining the mental states of others Bennetto *et al.* (1996); Ozonoff and Jensen (1999); Turner (1999); Baron-Cohen *et al.* (1985). Abnormal gaits, problems initiating movements, abnormal sleep patterns, and an increased likelihood of developing a seizure disorder all accompany an autism diagnosis Rinehart *et al.* (2006, 2001); Stores and Wivggs (1998); Tuchman and Rapin (2002). Juxtaposed against the impairments of ASD exists a collection of spared, and sometimes enhanced, abilities. For example, superior perceptual discrimination is regularly seen on the *embedded figures task* (EFT) Shah and Frith (1983). Also, improved (and sometimes even savant) abilities have been observed in domains as diverse as mathematics, map memorization, music, artistic abilities, and date calculations Pring *et al.* (1995); Nelson and Pribor (1993); Happé (1999).

This diverse behavioral profile poses a daunting challenge to the development of a unified theory of ASD. Indeed, the most widely acknowledged theories are generally circumscribed to account for only specific behavioral phenomena. For instance, the "Theory of Mind" (TOM) hypothesis Baron-Cohen *et al.* (1985) asserts that people with autism lack the ability to understand mental states in others, offering an explanation of observed social difficulties. It is unclear, however, how the TOM hypothesis might explain non-social patterns of deficits and spared abilities. In comparison, the "Weak Central Coherence" (WCC) hypothesis Happé (1999), which posits a more "piecemeal" style of cognitive processing in ASD, rather than one that is more "holistic", explains the appearance of enhanced performance on tasks that require attention to detail, along with difficulties utilizing more global, contextual, and gestalt information, but it does not address the full range of observed phenomena. Similarly, problems with planning, the flexible adaptation of behavior, and the generation of novel ideas are the focus of "Executive Dysfunction" (ED) theory, which highlights changes in executive processing as a central feature of autism Hughes *et al.* (1994); Hill (2004).

Combining multiple theories of this kind might cover the behavioral landscape of ASD, but it is not clear that this approach will foster our understanding of the neural basis of the condition. There is good reason to be skeptical that the different broad domains addressed by these theories arise from distinct brain systems. Also, any neural account will need to explain exactly how biological differences, over the course of development, give rise to the diverse range of observed behavioral patterns.

Given these concerns, one might opt to pursue an explanation of ASD that begins with identified differences in brain structure and/or function, using the existent literature on functional localization in order to associate these differences with observed patterns of behavior. One of the primary obstacles to this approach is the fact that a diverse range of neurological differences have been discovered in ASD, including abnormalities in the cerebellum Rodier *et al.* (1996), increased brain volume Aylward *et al.* (2002), abnormalities in the PFC Casanova *et al.* (2003), differences in the hippocampus, amygdala, and hypothalamus (as well as other parts of the limbic system) Kemper and Bauman (1998), as well as findings suggesting differences in major nuerotransmitter systems, including serotonin, dopamine, glutamate, and cholinergic abnormalities Carlsson (1998); Rubenstein and Merzenich (2003); Perry *et al.* (2001); Martineau *et al.* (1992); Tsai (1999).

Given this extensive array of neurological differences in ASD, it is not surprising that strictly neuroscientific approaches, to date, have had little success in providing a unifying view of the biological mechanisms responsible for the patterns of behavior observed in autism. Even with extensive knowledge concerning differences in the developing brain of people with autism, it is difficult to understand exactly how these neuroscientific differences give rise to the complex behavioral profile of ASD.

Computational cognitive neuroscience modeling can be a useful tool for addressing this difficulty. What is needed is the fabrication and analysis of computational models of neural processes that can simulate the generation of behaviors comparable to those observed in the laboratory. The methods of computational cognitive neuroscience produce formal characterizations of the relationship between brain and behavior, entailing precise and testable hypotheses involving both neuroscientific and psychological measures. By offering explicit mechanistic accounts of the underlying neurobiology, while capturing actual behavioral patterns, computational cognitive neuroscience models provide a means of bridging the conceptual gap between cognitive psychology and cognitive neuroscience in ASD research.

One question previously explored using computational cognitive neuroscience techniques is how deliberate control over behavior (cognitive control) is instantiated within neural circuitry and how this control is adjusted as environmental contingencies change (cognitive flexibility). The prefrontal cortex (PFC) has been broadly implicated in both cognitive control and cognitive flexibility Stuss *et al.* (2000, 2001). Under some accounts, cognitive control involves the active maintenance of abstract rule-like representations in PFC Noelle (2012). These PFC representations provide a top-down task-appropriate processing bias to more posterior brain areas Cohen *et al.* (1990); Miller and Cohen (2001). Biologically, the active maintenance of frontal control representations is supported by dense patterns of recurrent excitation in the PFC, as well as intrinsic maintenance currents Goldman-Rakic (1987); Levitt *et al.*

(1993); Wang *et al.* (2004). Computational analyses have shown that cognitive control and cognitive flexibility are, in a sense, at odds. Cognitive control requires robust maintenance of a control representation, while cognitive flexibility requires the ability to quickly adapt these representations as task contingencies change. This processing conflict suggests the need for a mechanism to intelligently toggle the PFC between a maintenance mode and an updating mode. The fact that we learn to control our behavior in different ways depending on the current situation means that this toggling process must be learned. This need for learning has drawn the attention of researchers to the dopamine system.

Dopamine (DA), a neurotransmitter with diffuse projections throughout the brain, plays a central role in contemporary models of PFC function. The mesolimbic DA system is seen as implementing a reinforcement learning algorithm, driving the learning of action sequences that lead to reward Montague *et al.* (1996); Barto (1994). In PFC models, the DA system learns to adjust the state of PFC pyramidal cells, determining when cognitive control should be maintained and when it should be flexibly modified in order to succeed at the current task Braver and Cohen (2000). A useful analogy is that of a "gate" in a fenced enclosure. When cognitive control must be strong, the gate is closed, keeping out distracting inputs that might compromise the current PFC control signals. When the current control state is no longer appropriate, the gate opens, allowing the old control state to escape and permitting a new control representation to enter the PFC via its inputs. Computational models of PFC function suggest that intelligent "gating" of control representations in PFC can be learned, through experience, via the DA system O'Reilly *et al.* (2002).

We propose that these computational accounts of PFC/DA interactions are highly relevant for understanding the neural basis of ASD. There is growing evidence of abnormal DA functioning in people with autism. Aberrant levels of DA have been discovered in studies measuring DA via PET Fernell *et al.* (1997), as well as more indirect measures such as HVA metaboloites Martineau *et al.* (1992). Clinical trials of drugs that modulate levels of DA in the brain have shown some behavioral benefits, as well Posey and McDougle (2000). Studies have also found evidence of genetic differences in some kinds of dopamine receptors and morphological differences in the basal ganglia in people with ASD Staal *et al.* (2012); Qiu *et al.* (2010). DA system dysfunction is also associated with behaviors related to ASD symptoms, incuding increased prevalence of seizures Tuchman and Rapin (2002), repetitive behaviors Canales and Graybiel (2000), and problems with skilled motor function Rinehart *et al.* (2001). While the precise causal role that DA may play in autism is still unknown, the ties between DA and ASD are both numerous and compelling.

In this chapter, we report computational cognitive neuroscience simulation results supporting the conjecture that deficits in PFC/DA interactions are responsible for many of the interesting behavioral patterns observed in ASD. The basic idea is that reduced efficacy of the DA-based PFC gating mechanism results in overly perseverative top-down control. This has three results. First, the inability to properly adapt PFC control representations produces inflexible behavior. This is exemplified by the executive dysfunction profile observed in autism. The second consequence is more

subtle. We suggest that the flexible updating of PFC plays an important role in shaping associational areas of cortex, influencing synaptic plasticity in these areas so as to support the appropriate generalization of learned behaviors across contexts. The hypothesized lack of flexible updating of PFC in ASD results in the learning of cortical representations that are overly specific, hindering generalization. These learning deficits can be seen in measures of prototype extraction during category learning. Third, failure to appropriately update PFC can limit the use of temporally extended context information, explaining observed deficits in sequential implicit learning tasks and in the use of sentential context to disambiguate the meaning of words.

We begin by reviewing the processes of DA-mediated PFC updating. We then present a series of computational cognitive neuroscience models that demonstrate how dysfunctional PFC/DA interactions can account for experimental data concerning ASD behavioral patterns in the diverse domains of executive function, implicit learning, lexical dismbiguation, and prototype formation.

0.2 Background

0.2.1 Dopamine & Temporal Difference Learning

Our account of the role of the midbrain DA system in autistic behavior builds on findings and theories concerning the role played by DA in learning, as well as its relationship to PFC. Analyzing the response profile of DA neurons in the basal ganglia of monkeys, Schultz et al. (1997) demonstrated that DA cells appear to encode a prediction error in the amount of future reward to be delivered during the performance of a task. In other words, these cells seem to encode a *change in expected future reward*. This is interesting because a measure of change in expected future reward is a key variable in a powerful reinforcement learning algorithm known as Temporal Difference (TD) learning. In TD learning, the change in expected future reward is know as the TD Error.

This formal connection has led researchers to propose a specific role for DA neurons in the brain's learning mechanisms Montague *et al.* (1996), equating the firing rate of the DA cells with the amount of change in expected future reward, or TD Error. Neurally plausible implementations of TD learning have been implemented and have been used to model the learning of motor sequences in the striatum Barto (1994), driven by the reward-prediction DA signal. Reinforcement learning models of this kind have been highly successful at accounting for a broad range of both behavioral and neuroscientific data Dayan and Niv (2008).

0.2.2 Computational Models of Prefrontal Cortex

The modeling work presented in this chapter is built upon an established computational framework for understanding interactions between PFC and the DA system during tasks requiring cognitive control and cognitive flexibility. One of the primary insights of this framework is that the DA based TD learning mechanism might be used to learn, from experience, when to robustly maintain current representations in the PFC versus allowing updating to occur Braver and Cohen (2000). As described earlier, it is helpful to think of the maintenance versus updating of PFC in terms of a gating mechanism. The key insight is that, in addition to driving the learning of *overt* motor sequences, the TD Error, encoded in the firing rate of DA cells, can be used to learn *covert* actions, such as when to open and when to shut the gate on PFC representations. By building computational models of PFC function using this framework, researchers have shown that this account is plausible Braver and Cohen (2000). In these models, PFC control representations are actively maintained in the sustained firing patterns of the modeled PFC pyramidal cells. For example, the PFC can encode, and actively maintain, a representation of "pay attention to the color of the stimuli". This maintained pattern of activity can then provide a "top-down" bias, up-modulating pathways in posterior brain areas associated with the processing of stimulus color Cohen *et al.* (1990). The biasing provided by PFC can be used to drive weaker, less automatic, behaviors (e.g., naming the ink color as opposed to reading the word in the Stroop task Stroop (1935)), as appropriate. This activation based modulation is thought to support our ability to provide cognitive control over behavior Cohen and Servan-Schrieber (1992). The DA based adaptive gating mechanism can signal to PFC when it is appropriate to strengthen the maintenance of the representation currently encoded (i.e., close the gate). This occurs when a positive TD Error arises, signifying a positive change in expected future reward. In other words, when we are doing better than expected, close the gate on PFC representations, so we are more likely to keep doing the same thing. Conversely, if we start to perform worse than expected, possibly due to task contingencies changing, this is signaled by a negative TD Error (i.e., DA cells firing below their base rate). The negative TD error can be used as a gating signal on PFC representations, causing the gate to open, allowing a new control representation to replace the old, thereby allowing for the flexible adjustment of controlled behavior.

Along with providing a neural mechanism that can learn to appropriately and adaptively gate PFC representations, these models have also been successful at relating frontal disturbances, such as those found in schizophrenia, to deficits in cognitive control Cohen and Servan-Schrieber (1992) and cognitive flexibility Braver and Cohen (1999). One computational model that included this form of PFC/DA interaction, called *XT* Rougier *et al.* (2005), was the first neuroscientific model able to provide quantitative fits to a hallmark task of cognitive control, the Stroop task Stroop (1935), and a widely used measure of cognitive flexibility, the Wisconsin Card Sort Task (WCST) Berg (1948), in both neurologically intact and frontally damaged people. We will explain how the XT model may be used to

capture patterns of executive dysfunction in autism, but a note about our general computational modeling strategy is in order, first.

0.3 General Modeling Approach

This work focuses on using the methods of computational cognitive neuroscience to demonstrate how changes in PFC/DA interactions, leading to difficulties in updating PFC representations, can explain a broad array of behavioral patterns observed in autism. One reasonable strategy would involve the fabrication of a single complex biological model that is capable of performing all of the laboratory behaviors of interest, showing that this model matches the behavior of typically developing people but produces the behavioral patterns observed in ASD when it is modified to include PFC/DA dysfunction. Unfortunately, the range of relevant behaviors is broad, making the production of a single model capable of all of the behaviors of interest untennable.

As an alternative, we have pursued a strategy that builds directly upon the rich existing modeling literature. Separate previously published models, each successful in capturing behavioral patterns in a particular domain, were modified in a manner that reflects our general hypothesis that PFC/DA interactions are disrupted in ASD so as to reduce flexibility in PFC updating. In each case, we have found that hindering PFC updating causes the modified model to produce patterns of performance seen in ASD. It is important to note that no additional mechanisms were introduced into any of the previously published models. Only their existing, previously justified, mechanisms for cognitive flexibility were manipulated in order to capture the behavior of people with autism. Using this approach, we can show how cognitive inflexibility, arising from improper DA modulation of PFC, can account for a wide variety of behavioral phenomena observed in autism, including executive dysfunction, deficits on implicit learning tasks, problems with word sense disambiguation, and reduced prototype formation.

0.4 Executive Dysfunction

0.4.1 Executive Task Performance

People with autism are impaired at a variety of tasks involving planning Bennetto *et al.* (1996), flexible adaptation of behavior Bennetto *et al.* (1996); Ozonoff and Jensen (1999), and spontaneous generation of novel behaviors Turner (1999). Tasks of this kind have been associated with executive control processes. This has led some researchers to view executive dysfunction as a central feature of autism Hughes *et al.* (1994).

A more detailed examination of autistic behavior reveals that not all forms of executive processing are impaired, however. A perplexing aspect of the ASD executive profile is that cognitive flexibility is impaired while fundamental cognitive control remains relatively unaffected. Cognitive control describes the ability to enact a behavior in the presence of a distracting or more automatic competing response. In contrast, cognitive flexibility is the ability to fluently adjust cognitive control as contingencies change. A classic measure of cognitive control is the Stroop task Stroop (1935), and a common measure of cognitive flexibility is performance on the Wisconsin Card Sort Test (WCST) Berg (1948). Persons with autism have been shown to exhibit poor WCST performance, but they exhibit no more interference on the Stroop task than healthy controls Ozonoff and Jensen (1999). This dichotomy challenges the notion that autistic behavior is the result of a global impairment of executive processes.

A second challenge appears in the developmental trajectory of executive deficits in autism. In young children with autism, executive abilities are intact when compared with controls Griffith *et al.* (1999). Differences in cognitive flexibility arise over the course of development.

These behavioral findings can be explained by positing separate mechanisms for cognitive control and for the flexible adaptation of control. In autism, the mechanism for control may be intact, but the flexibility mechanism may be compromised. Interestingly, this segregation of function is captured by the *Cross-Task Generalization Model (XT)* Rougier *et al.* (2005). Driven by broad neurocomputational considerations, XT casts PFC as central to cognitive control, while PFC/DA interactions mediate cognitive flexibility.

0.4.2 The XT Model

The architecture of the XT model is shown in Figure 1. This model makes use of the Leabra framework O'Reilly and Munakata (2000). The input of XT consists of two layers of neural units used for the presentation of up to two stimulus objects. It is natural to think of the rows of each input layer as representing different dimensions (e.g., color, shape, texture) and the columns indexing features across each dimension (e.g., red, orange, green, blue). The Response layer has essentially the same structure as an input layer, but includes one additional unit, which codes for "no response".

A collection of Hidden layers map from stimuli to responses. Activity in these layers can be modulated through top-down biasing signals from an actively maintained representation in the PFC layer. Unlike previous models, the actively maintained representations in PFC are learned through a developmental process that involves training the model to perform a variety of simple tasks. These tasks share a need to selectively attend to individual dimensions of the stimuli. The standard synaptic plasticity mechanisms of the Leabra framework, applied over the course of developmental training, produces PFC representations that support focusing attention on a single stimulus dimension at any one time Rougier *et al.* (2005).

The Task input indicates which task is to be performed, with one input unit coding

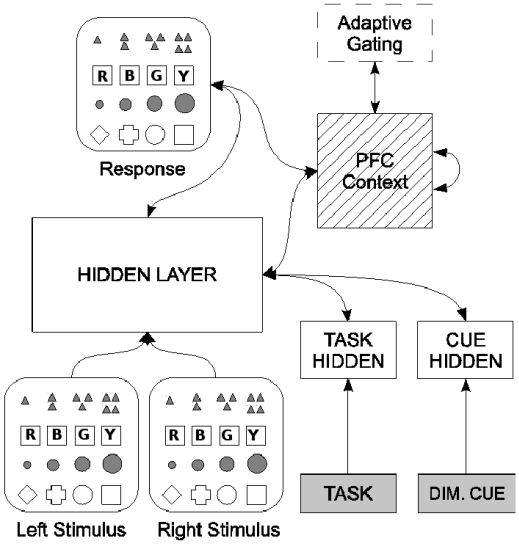


Figure 1 XT Model Architecture. The AG unit implements "adaptive gating" by modeling the effects of DA on PFC. Adapted from Rougier et al. (2005).

for each task. The Dimension Cue layer is used to indicate the currently relevant stimulus dimension, with each unit in the layer corresponding to a specific dimension. All of the Dimension Cue units are turned off for tasks in which the model must discover the relevant dimension on its own.

The flexible adjustment of cognitive control is implemented using a DA-based adaptive gating (AG) mechanism. The model receives a positive scalar reward signal whenever it produces a correct response, and the AG calculates the change in expected future reward: the TD Error. Importantly, the AG, modeling phasic DA responses, not only modulates the learning of reward expectation but also manipulates the "gate" on PFC. When the model performs better than expected, the current PFC representation is strengthened. When the model performs worse than expected the current PFC representation is destabilized, allowing a new, possibly more appropriate PFC representation to be entertained Rougier *et al.* (2005).

0.4.3 Modeling Autism Using XT

Our theory suggests that a deficit in DA functioning can account for the impaired cognitive flexibility seen in people with autism, while leaving cognitive control robust and relatively unaffected. We have tested this theory by reducing the effect of the DA signal in the XT model by scaling the TD Error (AG) by a constant factor, κ. The model of typically developing individuals uses $\kappa = 1$, and autism is modeled using $\kappa < 1$. This scaling of the TD Error by κ is the only modification from the original XT model that we made. A κ value of 0.54 was found to produce the best fit to human performance. This reduction of the DA signal decreases the efficacy of the PFC gating system, resulting in less efficient destabilization of PFC when errors are made.

0.4.4 Modeling WCST

As in the original XT work, each trial of the WCST was modeled by presenting a single stimulus object "card" at the model's inputs, with one unit activated for each feature of the stimulus across three stimulus dimensions. (See Figure 2.) The model "sorted" the "card" by outputing the feature of the current stimulus relevant for sorting. For example, when cards were to be sorted by color, the model was expected to output the color of the stimulus (e.g., "red"). Importantly, the model was not instructed concerning the sorting rule (i.e., the Dimension Cue inputs were all off). Thus, the model needed to search for the relevant stimulus dimension. The model received a positive scalar "reward" signal when its output was correct. The use of TD learning to produce AG activity, along with the top-down modulation from PFC, allowed XT to successfully learn to focus on the relevant stimulus dimension. The AG mechanism strengthened the maintenance currents of modeled PFC pyramidal cells during good performance, causing the active maintenance of the sorting rule in PFC. When the sorting rule was switched, the actively maintained PFC representa-

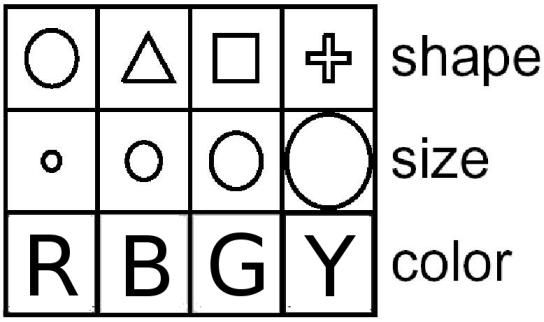


Figure 2 XT WCST Stimulus Input

tion became invalid. Continued active maintenance of PFC firing rates would result in perseverative errors. In the standard XT model, the AG alleviated this problem by providing a gating signal to PFC when reward was expected but not delivered, allowing PFC contents to change Rougier *et al.* (2005).

0.4.5 Modeling Stroop

The XT model performed the Stroop task in much the same way as Cohen's and Servan-Schreiber's seminal model Cohen *et al.* (1990), leveraging competition between pathways of varying strength: the strong word reading pathway and the weak color naming pathway. In order to simulate this competetion, the frequency of trials in which one dimension (font color) was experienced during development was manipulated. The font color dimension was relevant only 25% as often as the other dimension, word reading. The time needed for output activity to stabilize was taken as a response time measure Rougier *et al.* (2005).

0.4.6 Simulation Results

0.4.6.1 Simulations

Due to stochasticity in model initial conditions and in the sequence of developmental trials, 100 distinct network models were prepared using the XT training procedure, stopping when a performance criterion was reached, up to a maximum of 100 training epochs. Each network was tested under both conditions of DA modulation ($\kappa = 1$ for healthy networks and $\kappa = .54$ for networks modeling ASD), on both WCST and Stroop. Each of the 100 networks was treated as an individual subject for the purposes of data analysis.

0.4.6.2 WCST Results

WCST results matched data reported in the literature. The differences between simulated performance of normally functioning individuals and simulated autistic performance were statistically reliable ($p < 0.001$) and consistent with previous studies Prior and Hoffman (1990); Ozonoff and Jensen (1999); Minshew *et al.* (2002). Specifically, perseverative errors, marking a failure to flexibly discard the initial sorting rule when it was no longer rewarded, were significantly more numerous in the reduced DA modulation version of XT. (See Figure 3.)

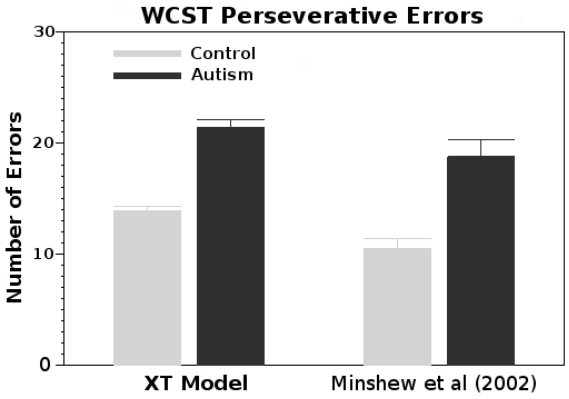
0.4.6.3 Stroop Results

Model performance on the Stroop task provided a good quantitative fit to human performance. (See Figure 3.) The model with intact DA function displayed the classic Stroop reaction time results — slowing for conflict stimuli when color naming. The performance of the ASD model showed no significant increase in Stroop interference in comparison to the healthy model ($F(1, 198) = 0.62$; $p > 0.43$), which is consistent with past findings Ozonoff and Jensen (1999).

0.4.6.4 Developmental Results

These simulations involved the introduction of a DA deficit only after the model was fully developed. Thus, these simulations ignored the possibility that an early manifestation of a DA deficit might hinder the proper learning of PFC representations, introducing an impairment in cognitive control in the model that is not observed in autistic subjects. To address this issue, the DA deficit in the ASD model was introduced prior to developmental training. PFC development and model performance were analyzed over the entire developmental period of the model (100 epochs). Two groups of 10 networks were used: an autistic group with $\kappa = 0.54$ and a control group with $\kappa = 1.00$. At the end of developmental training, the networks exhibited the same pattern of results as seen in the initial simulations. Furthermore, we found that the DA deficit did not hinder the learning of useful PFC representations, allowing for focused attention on a single stimulus dimension.

Interestingly, these simulations offer a potential explanation for the observation that cognitive flexibility deficits, in comparison to controls, appear late in development Griffith *et al.* (1999). In Figures 4 & 5, the Stroop interference effect and the number of perseverative errors in WCST are plotted over developmental training time. Figure 4 shows no effect of the DA manipulation across development. Stroop interference was significantly greater for the autistic networks during only 1 training epoch out of the 100 ($p < 0.003$), demonstrating robust cognitive control throughout development. Figure 5, however, shows a significant difference in perseverative errors only after an initial period of developmental training. During the first 53 epochs there was a significant difference ($p < 0.05$) during only 26.4% of the epochs, but later in development (epochs $54 - 100$) a significant difference was reached 93.6% of the time. Importantly, neither healthy nor autistic models showed a distinct advantage or disadvantage during the earliest stages of development. We conjecture that early poor performance by both model types was largely due to the fact that strong, dimensionally selective, PFC representations had yet to learned. Without such PFC



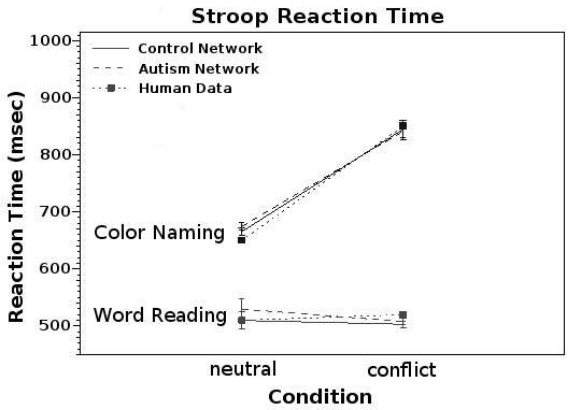


Figure 3 Top: Comparing performance on WCST perseveratitive errors for typically developing and autistic individuals (human data from Minshew et al., 2002). Error bars are standard errors of the mean. Bottom: Stroop reaction time, comparing simulated performance to human data (healthy human data from Dunbar et al. (1984), and autistic subjects perform no differently than controls (Ozonoff et al., 1999)).

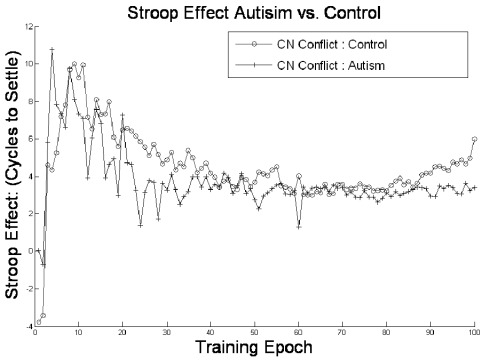


Figure 4 Model Stroop Interference Over the Course of Development.

representations, the networks were forced to rely more heavily on synaptic plasticity in the Hidden layers (posterior cortex) to perform the tasks. Later in development, both healthy and autistic networks acquired good PFC representations, but the models with reduced DA influence on PFC displayed difficulties in updating those representations when expected reward stopped being delivered (i.e., when the sorting rule was changed).

0.4.7 Summary

These simulations show that, given the XT account of the role of PFC in executive control, reducing the influence of DA on PFC adaptive gating is sufficient to capture the pattern of performance exhibited by people with autism on tests of cognitive flexibility (WCST) and cognitive control (Stroop). Furthermore, we demonstrated that weakening the DA signal over the course of PFC development continues to reflect autistic performance, while also providing some insight into the late appearance of cognitive flexibility deficits in autism. More information about these simulations may be found in Kriete & Noelle (2015).

0.5 Implicit Learning

0.5.1 Implicit Learning Deficits in Autism

Some researchers have suggested that people with autism display a deficit in the implicit learning of relationships that exist between objects and situations in the world Mostofsky *et al.* (2000). Klinger, Klinger, & Pohlig (2006) argue that im-

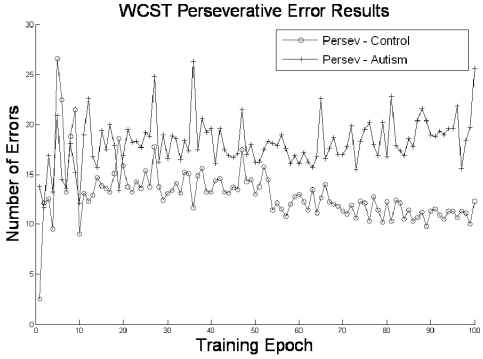


Figure 5 Model WCST Perseverative Error Count Over the Course of Development.

paired implicit learning results in difficulties in recognizing relationships across experiences, leading to problems with acquiring general knowledge. People with autism commonly have difficulties generalizing learned knowledge to new situations, hindering the learning of behaviors needed for independent living.

Poor performance when learning artificial grammars Reber (1967) and learning deficits on serial response time tasks (SRTT) have been used as evidence for general implicit learning impairments in ASD Mostofsky *et al.* (2000); Klinger *et al.* (2006). In the following, we focus on SRTT phenomena.

0.5.2
The Serial Response Time Task (SRTT)

In a common version of SRTT, participants press buttons, one at a time, as they are illumnated or highlighted. The order of illumination is the key manipulation. During the first and the final (fifth) block of trials, the order in which the buttons are illuminated is random. However, during blocks 2–4 there is a hidden sequential pattern in the buttons that are illuminated. An observed reduction in the reaction time of correct button presses during blocks 2–4 indicates that healthy participants become sensitive to the hidden pattern. Importantly, this reduction is not observed during blocks 1 and 5. Knowledge of the hidden structure is seen as "implicit", however, as most participants claim no awareness of the sequential pattern Cleeremans and McClelland (1991). In contrast, people with autism do not show marked improvement during blocks 2–4, suggesting that autism impairs implicit learning abilities Mostofsky *et al.* (2000).

While this behavioral result is interesting in its own right, it also raises questions concerning the biological mechanism(s) behind this deficit. There is some evidence that PFC and the basal ganglia are important for implicit learning Matsumoto *et al.* (1999); Pascual-Leone *et al.* (2004). We have explored this connection using an established computational model of the SRTT, investigating the possibility that

PFC/DA abnormalities may give rise to the implicit learning problems observed in ASD.

0.5.3 Modeling SRTT Performance

Seminal work on modeling healthy SRTT performance has been conducted by Cleeremans & McClelland (1991). In this work, simulated neural circuits were repeatedly presented with an input encoding the currently illuminated button, and the output of the circuit was read as a prediction of the next button to be illuminated. Importantly, these neural networks included a "context layer" of neural units which learned to actively maintain information about the sequence of previous inputs, allowing the model to base its predictions on more than the currently illuminated button. These models were essentially simple recurrent networks (SRNs) Elman (1990) trained to predict the next button to be be pressed. We recreated the Cleeremans & McClelland SRTT model with one small modification. In order to match available SRTT data for people with autism, we reduced the original implementation's 10 buttons (inputs and outputs) to 4 buttons. The schematic network architecture is shown in Figure 6.

The Cleermans & McClelland model assumed that button press reaction times are linearly reduced with button prediction accuracy. Network outputs were converted into a probability distribution over the buttons using a Luce choice ratio Luce (1963), and the difference between this distribution and the actual next illuminated button was linearly scaled to produce a simulated response time. We used exactly the same method to simulate reaction times, introducing three free parameters: a scaling constant from prediction error to milliseconds, a base response time (when error is zero) for our healthy model, and a base response time for our autism model.

The context layer in the Cleeremans & McClelland model played an identical functional role to the PFC in other models, actively maintaining information to modulate an input-output mapping. In this case, the context layer actively maintained information about the preceding button presses, influencing the prediction of the next button. In our executive dysfunction model, described previously, the PFC was updated in a dynamic fashion, based on learned task contingencies. In this SRTT model, the context layer is updated with each new input. Thus, the SRN context layer is analogous to the PFC in XT, with the updating "gate" forced to open with each new input O'Reilly and Munakata (2000).

In order to capture the relevant sequential information, the SRN model must update the context layer in a fast and appropriate manner. This flexible updating of contextual information is precisely the cognitive mechanism we hypothesize to be suspect in people with autism. By restricting the ability of the SRN to update the context layer, mirroring the PFC updating failures that arise with weakened PFC/DA interactions in our other models, we aimed to capture the performance of people with autism. We implemented this updating restriction with a single new parameter: a probability that context layer (PFC) updating will be successful with each new input. Healthy behavior was modeled by setting this probability to one, and the probability was reduced

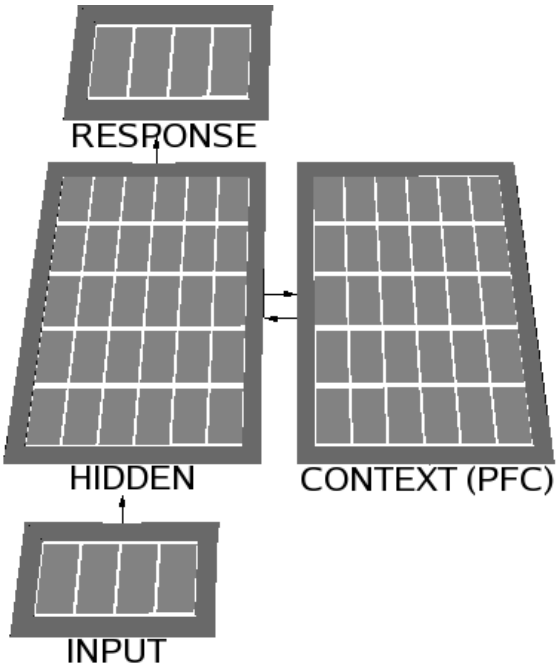


Figure 6 Network Diagram of SRTT Model

to model ASD performance. This manipulation is analogous to reducing the efficacy of the DA-based gating signal to the PFC. Restricting the updating of the PFC, in this manner, makes the temporally extended information stored there much less reliable, hindering the learning of complex sequential structures in the ASD model.

0.5.4 Implicit Learning Simulation Results

Model simulations were repeated 100 times in each of the experimental conditions, with initial synaptic connection strengths randomized for each repetition. Average performance results for each block were compared to previously reported response time data for both people with autism and typically developing controls Mostofsky *et al.* (2000). The context layer updating probability and the response time scaling parameters that produced the lowest sum-squared deviation from the human data were identified as the best fit model.

The resulting modeled reaction times, along with previously reported human data, appear in Figure 7. The best-fit updating probability for the autism model was found to be 0.5. A repeated measures ANOVA on blocks 1, 2, 3, and 4 of the model results showed a significant Group by Block interaction ($p < 0.000001$). This indicates that

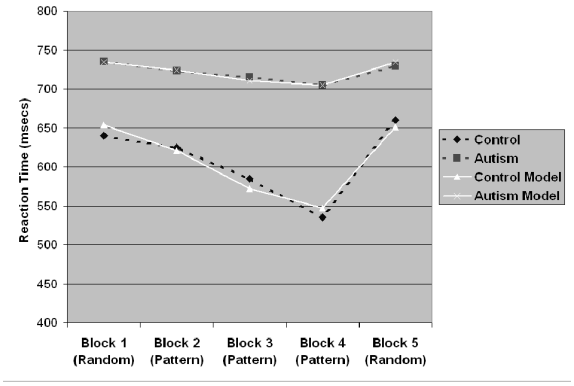


Figure 7 Model Results & Human Behavioral Data from Mostofsky et al. (2000)

the ASD networks demonstrated significantly less learning over the crucial training blocks (2–4) than the networks allowed to reliably update their PFC-like context layers. Thus, clear implicit learning deficits were found in the autism model.

0.5.5 Summary

Our simulation results matched human performance both qualitatively and quantitatively, providing evidence that impairments in PFC updating can result in SRTT deficits like those seen in ASD. It is interesting to note that this account posits deficits in learning temporal patterns rather than in implicit learning, per se. More information about these simulations may be found in Kriete & Noelle (2009).

0.6 Lexical Disambiguation

0.6.1 Sentential Context

In the previous section, we showed how weakened PFC/DA interactions can hinder the integration of information over time. Such temporal integration is also important for determining the meaning of ambiguous words in a sentence. Homographs are words that share spelling but have different meanings and pronunciations (e.g., "bow", "tear"). To identify the word referenced by a homograph, we typically must rely on sentential context. People with autism have difficulties utilizing context when interpeting homographs. Instead, they tend to produce the most frequent pronunciation Happé (1997). Some neurocomputational models of sentence processing use an

SRN architecture, like that used to model the SRTT. The context layer integrates a sequence of words into an evolving representation of the sentence. As before, the SRN must update the context layer in a fast and appropriate manner in order to integrate information from the full sequence of words.

0.6.2 Lexical Ambiguity in Schizophrenia

Interestingly, people with schizophrenia also have problems utilizing context in language disambiguation tasks Cohen and Servan-Schrieber (1992). The specific pattern of deficits is different than those observed in ASD, however. Coarsely, schizophrenics lose context with delay, while people with autism often fail to use sentential context even when that information was recently provided.

Also, there is evidence that many schizophrenia symptoms arise from abnormal DA functioning Cohen and Servan-Schrieber (1992). How might the DA dysfunctions in schizophrenia and autism differ so as to produce observed differences in lexical disambiguation? One possibility involves the differential effects of "tonic" DA, which changes slowly in cortex, and "phasic" DA, which rapidly influences processing. Cohen and Servan-Schreiber (1992) argue that schizophrenia symptoms arise from abnormal tonic DA levels. In contrast, it is phasic DA signaling that is thought to capture TD Error, as in adaptive gating models of PFC. This leads to the hypothesis that degraded PFC updating, as in the SRTT model, should produce ASD-like patterns of disambiguation deficits, while degraded active maintenance of PFC representations should result in the error pattern seen in schizophrenia.

0.6.3 Testing Context Sensitivity

There is a psychological test has been used to evaluate both people with schizophrenia and ASD on their ability to utilize context when disambiguating words Cohen and Servan-Schrieber (1992); Happé (1997). The test uses ambiguous words with both a high frequency interpretation and a low frequency one. On each trial, a sentence fragment is presented which contains one ambiguous word (e.g., "with a BOW") along with a disambiguating fragment (e.g., "the juggler ended his act"). The sentence fragments are presented in various orders. Specifically, there are four conditions: (1) the low frequency meaning is correct and the context comes last, (2) the low frequency meaning is correct and the context comes first, (3) the high frequency meaning is correct and the context comes first, and (4) the high frequency meaning is correct and the context comes last.

People with autism demonstrate problems identifying the contextually appropriate meaning of the ambiguous word in both conditions (1) and (2), providing a significantly higher percentage of incorrect high frequency responses compared with typically developing controls Happé (1997). (See Figure 8.) ASD participants show no significant differences from controls in conditions (3) and (4). The profile for schizophrenics is slightly different. People with schizophrenia show an impairment

Participant Group		Common Pronunciations		Rare Pronunciations	
		Before	After	Before	After
Normal	Mean	4.88	5.00	4.88	5.00
	Std. Dev	0.33	0.00	0.33	0.00
	Range	(4-5)	(5-5)	(4-5)	(5-5)
Autism	Mean	4.77	4.88	3.82	4.06
	Std. Dev	0.44	0.33	1.19	0.97
	Range	(2-5)	(4-5)	(2-5)	(2-5)

Figure 8 Lexical Disambiguation in ASD. "Common" refers to the high frequency interpretation of the ambiguous word being appropriate, while "Rare" refers to cases in which the low frequency interpretation is correct. "Before/After" refers to the order of sentence fragments, with the ambiguous word occurring before/after the contextual information, respectively. Table reproduced from Jolliffe and Cohen (1999).

in utilizing context, but only during condition (2), when the context is presented first. (See Figure 9.) This has been interpreted as evidence for problems maintaining contextual information over time in schizophrenia. (Condition (4) was not tested in the schizophrenic population, so no human data are available in that case.)

0.6.4 Modeling Lexical Disambiguation

The connectionist model utilized by Cohen and Servan-Schreiber (1992) was modified to investigate lexical disambiguation in people with autism and schizophrenia. (See Figure 10.) A localist code was used to represent words as inputs to the network. These inputs included ambiguous words and words that could act as contextual cues. The output of the network (the "Meaning Output Module" in Figure 10) also used a localist code to represent the various possible meanings of the current word being presented to the network. For example, separate output units would be used to represent the meaning of "BANK" as a "financial institution" and the meaning of "BANK" as "land alongside a river".

While, in the original model, the context layer ("Discourse Module" in Figure 10) used a researcher specified code for sentential context information, we allowed this layer to learn its own distributed representations by training the network as an SRN on the task of producing the correct output meaning for each input word form, in the context of previously presented inputs. Specifically, there were 100 input units encoding 50 ambiguous words and 50 disambiguating context words. The output contained two units for each of two meanings of the 50 ambiguous words, plus 50 units for the meanings of the context words, for a total of 150 output units. Of the two meanings for each ambiguous word, one was considered "strong" (high frequency), and the second was considered "weak" (low frequency). The hidden and context layers contained 100 units, though varying this value had little effect.

Each trial consisted of presenting the model with a sequence of three words, with

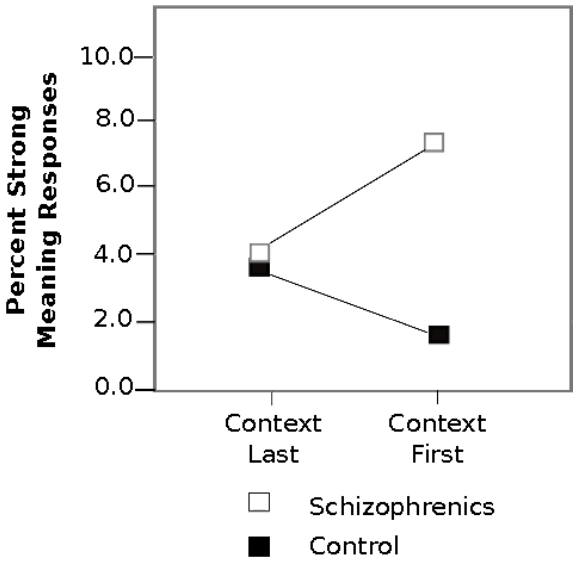


Figure 9 Lexical Disambiguation in Schizophrenia. Vertical axis shows percentage of errors using the most common interpretation of the ambiguous word, when the rare interpretation was correct. Figure adapted from Cohen and Servan-Schreiber (1992).

the model expected to output the appropriate meaning of each word as it was presented. Each sequence began with a context/ambiguous word pair, with the order of these two words counterbalanced to produce "Context Presented First" trials and "Context Presented Last" trials. The third word in each sequence was a repetition of the ambiguous word, with the model expected to produce the context appropriate meaning for this probe word regardless of the order of the initial word pair. During SRN training, the strong word meaning was used on 70% of trials, and the weak meaning was used on 30% of trials. On every trial, the context word perfectly disambiguated the otherwise ambiguous word. (See Figure 11.)

After training, the network was tested on the same three word sequences on which it was trained, following the procedure used by Cohen and Servan-Schreiber (1992) without modification. The primary measure of interest was the fraction of trials in which the model output the strong meaning during the final probe word presentation in the sequence, but the weak meaning was contextually appropriate.

0.6.5 Modeling Schizophrenia

In the original model of schizophrenic performance, Cohen & Servan-Schreiber implemented their hypothesized reduction in tonic DA levels by reducing the gain of the activation function on all modeled PFC units. Coupled with the dynamics of the

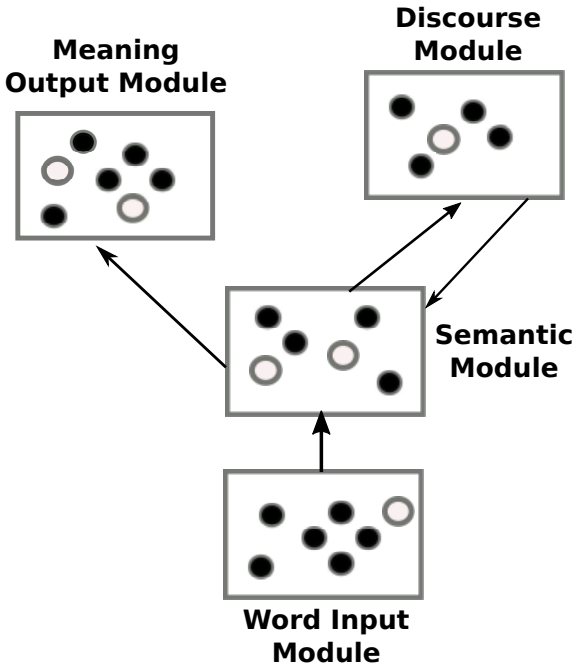


Figure 10 The Original Lexical Disambiguation Model. Image adapted from Cohen and Servan-Schreiber (1992).

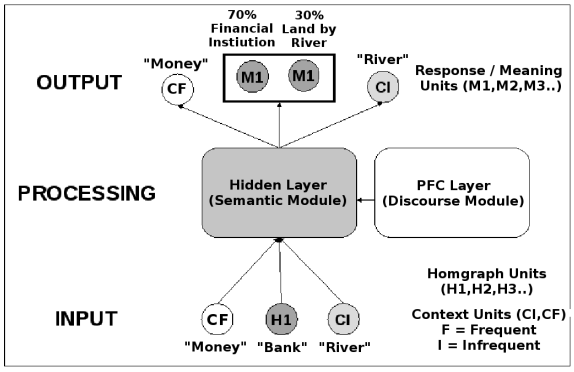


Figure 11 Lexical Disambiguation Model and Task. The task was to correctly respond to any input word by activating the output unit encoding the appropriate meaning for that word. The context units were used to retain sentential context, needed for disambiguation.

network, the result was a less stable PFC representation of context. Thus, presenting disambiguating context early could result in the gradual loss of information needed to determine the correct meaning of the probe. This manipulation successfully matched the pattern of behavior seen in schizophrenia. We approximated the gain reduction used in the original model by decaying context layer activation by a fixed multiplicative constant. In our model of healthy performance, 30% of context layer activity was retained on each time step. To model schizophrenia at various levels of severity, this context maintenance parameter was reduced to 20%, 10%, and 0%. Since schizophrenic symptoms typically appear after a substantial amount of development, the context activation decay was only instantiated after the model had completed its SRN training.

0.6.6 Modeling Autism

In order to model the performance of people with autism, we restricted the probability of successfully updating the context layer (PFC) upon each input presentation. Specifically, we investigated successful PFC updating probabilities of 100% (healthy control), 90%, and 80%. The reduction of this probability made the temporally extended information in PFC less reliable, hindering the learning of appropriate use of sentential context. To capture the developmental nature of autism, we restricted updating of the modeled PFC layer from the beginning of SRN training.

0.6.7 Simulation Results

Model simulations were repeated 10 times in each of the experimental conditions, with initial synaptic weights randomized for each repetition, and results were aggregated over these 10 repetitions. The possible word triplets were presented in a random order during training, and synaptic strengths were not allowed to change during testing. We measured the number of errors committed when assigning a meaning to an ambiguous probe.

The schizophrenia model qualitatively matched the behavioral data and reproduced the seminal modeling results of Cohen & Servan-Schreiber (1992). (See Figure 12.) Control networks performed well regardless of the frequency of the correct meaning ("Rare" vs. "Common"). Also, the ordering of contextual information had virtually no effect on healthy model performance. However, as the PFC representations were systematically destabilized in order to model schizophrenia, the error rate rose significantly. Importantly, this increase in error only occurred when the context was presented first, but not when the context was presented last.

By restricting the ability of the context layer (PFC) to appropriately update, the ASD model captured previously observed patterns of behavior Happé (1997). (See Figure 13.) Specifically, as the probability of successful updating was reduced, the ASD networks became increasingly reliant on the frequency of word meanings. This manifested in higher error rates when the model needed to utilize contextual infor-

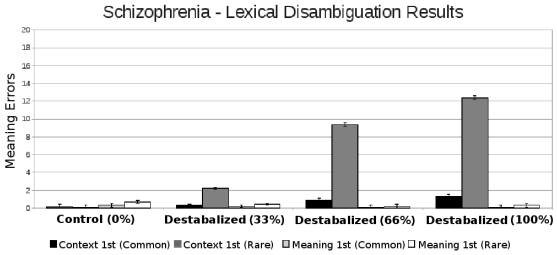


Figure 12 Schizophrenia Model Results. As the modeled PFC is destabilized, the model performs selectively worse on condition (2). In this condition, the contextual information is presented first and the model must use this information to override the more common interpretation of the ambiguous word. (The control model maintained 30% of each context/PFC unit's activity between time steps, and destabilization was produced by reducing this value to 20%, 10%, and, finally, 0% for complete destabilization.)

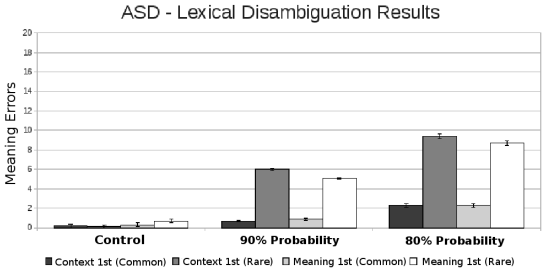


Figure 13 ASD Model Results. Systematically reducing the probability of PFC updating success resulted in worse performance on conditions (2) and (4). These are the conditions that require the use of contextual information in order to override the more common interpretation of the ambiguous word.

mation, regardless of the temporal distance from the context to the probe, as seen in people with autism.

0.6.8 Summary

Using a computational model very similar in structure to that used to capture SRTT performance in ASD, we have shown that dysfunction in the updating of PFC can explain deficits in word sense disambiguation observed in people with autism. The pattern of these deficits is distinct from that seen in schizophrenia, but the same model captures schizophrenic data by incorporating a reduction in PFC active maintenance stability, due to lower tonic DA levels.

0.7 Prototype Formation

0.7.1 Category Learning in Autism

Prototype formation is invaluable for learning concepts and categories. A category prototype is a kind of representational average of the features of category members that have been seen. Some studies have suggested that, when learning a new concept or category, children with autism are less likely to form a prototype representation than typically developing children Klinger and Dawson (2001); Gastgeb *et al.* (2009). Difficulty identifying an abstract prototype is argued to underly the generalization problems found in ASD.

Typically developing children exhibit a "prototype effect" when they are shown objects in a category and are then asked to categorize novel objects. This effect involves more reliably and/or more quickly identifying a prototype object as a member of a target category, in comparison to other test objects. Klinger & Dawson (2001) found that people with ASD do not exhibit the "prototype effect".

In the Klinger and Dawson (2001) experiment, paricipants viewed cartoon pictures of imaginary animals, with each animal belonging to one of a small set of fictional animal categories. The members of each category of animal varied in four features. For example, animals in the "Mip" category varied in the sizes of horns, wings, mouths, and feet. Each feature could take on one of five discrete values, ordered 1 to 5 (e.g., horns always were of one of five specific sizes). Thus, individual cartoons could be specified by four feature values (e.g., "5115" for largest horns, smallest wings, smallest mouth, and largest feet). Different animal categories had distinct sets of four features, as illustrated in Figure 14.

The children initially learned to classify a target category of animals (e.g., "Mip") in contrast to a single non-target category (e.g., "Pev"). A member of the target category was first dislayed, and the target category name was provided (e.g., "This is a Mip."). Participants were then presented with a series of target and non-target stimuli, one at a time, and they were asked to identify the members of the target category. As is clear from Figure 14, this category learning task was easy, as different animal categories possessed different sets of features. Corrective feedback was provided after each response. During this learning process, the presented target category stimuli only used feature values 1, 2, 4, and 5, but never 3. (See Figure 15.) Only a subset of the possible combinations of feature values appeared, but the average value used for each feature in the target category was 3, making the unobserved "3333" stimulus object the prototype for the category.

Immediately following this learning process, participants were tested for the "prototype effect". On each test trial, two animals from the target category were juxtaposed, and participants were asked to select one as the "best" category member (e.g., "Which is the best Mip?"). The trials of interest involved pairing the prototype ("3333") with either a previously viewed animal or one that was novel but was composed only of previously viewed feature values (e.g., "1425"). Typically devel-

oping children chose the prototype at a rate significantly higher than chance (79%). However, children with autism selected the prototype at a rate indistinguishable from chance (54%). Klinger and Dawson argued that these results demonstrated a lack of prototype formation in ASD.

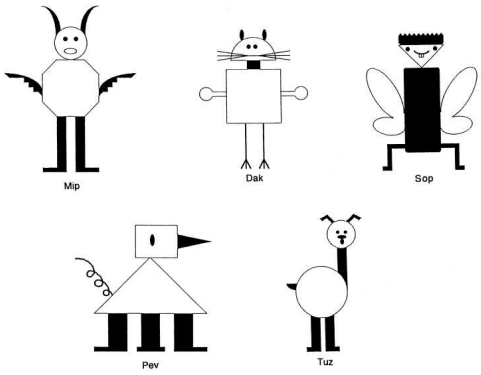


Figure 14 Prototypes for Different Stimulus Categories. Note that each participant observed members from only two categories. Image adapted from Klinger & Dawson (2001).

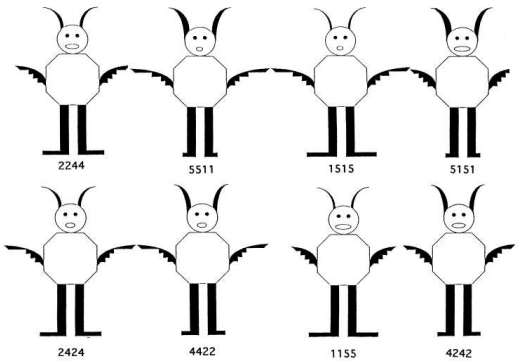


Figure 15 Variation Within a Category. Image adapted from Klinger & Dawson (2001).

0.7.2 Modeling Category Learning with ALCOVE

ALCOVE is a highly successful computational model of human performance on category learning tasks Kurschke (1992). ALCOVE learns to categorize stimuli from

corrective feedback on practice trials. There are three main processing layers in ALCOVE. An input layer represents the features of a current stimulus object. A hidden layer contains exponentially decaying radial basis function units, each of which becoming active to the degree that the current input is similar to that unit's preferred stimulus. The output layer contains one unit per category, with activation calculated as a weighted sum of hidden unit activity. The output activation levels are transformed into a probability distribution over category choices using a Luce choice ratio Luce (1963). Learning arises from the adjustment of connection weights so as to reduce error Kurschke (1992).

Importantly, ALCOVE also learns a set of attentional "weights" that reflect the relative importance of the various stimulus features for producing accurate categorization decisions. There is one weight value per stimulus feature (or "dimension"), constrained to be between 0 and 1, with larger values indicating more sensitivity to variation in the given feature. ALCOVE's attentional weights determine the relative focus of the model on specific aspects of the stimulus, and are, therefore, analogous to the hypothesized role of PFC in directing top-down attentional control. (Consider the role of PFC in the previously described XT account of Stroop and WCST performance, focusing processing on specific aspects of the stimulus.) In the ALCOVE framework, PFC perseveration would involve restricting the attentional weights, limiting focus to one or a small number of the stimulus features.

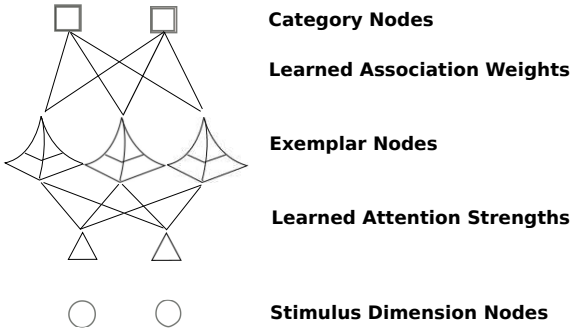


Figure 16 Structure of ALCOVE (Attention Learning CoVEring map). Image adapted from Kruschke (1992).

0.7.3 Prototype Learning in ALCOVE

We used ALCOVE to model the prototype formation study of Klinger & Dawson (2001). Two output category units were used (e.g., "Mip" & "Non-Mip"). Four inputs, each with its own attentional weight, captured the four features that varied across members of the target category, with the 1–5 feature levels linearly scaled to 0.2–1.0 activations. (An activation of 0.0 was used to indicate the absence of a fea-

ture.) The model learned the target category from the same stimuli that were presented to children in Klinger & Dawson (2001), with the standard ALCOVE error correction learning mechanism used to determine attentional and connection weights. Learning was disabled during subsequent testing for prototype formation.

Testing involved presenting the trained model with certain key stimulus objects and recording the activation of the target category output unit for each of them. The stimuli of interest included the prototype ("3333"), a stimulus seen during training (e.g., "1515"), and a novel stimulus (e.g., "1551"). For each pair of the prototype and another stimulus object, the recorded output activations were transformed into selection probabilities using a Luce choice ratio Luce (1963). The mean probability of selecting the prototype over non-prototype stimuli was the measure compared to available experimental results.

0.7.4 Modeling Autism in ALCOVE

The ALCOVE model was modified in two ways to simulate neural differences in autism. First, we biased one randomly selected attentional weight to be high by initializing its value to 0.9. The other attentional weights were initialized to 0.1. This manipulation captures, at a gross level, our conjecture that deficient DA based gating of PFC results in overly perseverative control over top-down attention, effectively restricting the features used during processing. In the control version of the model, the attentional weights were all initialized to 0.1, as is standard in ALCOVE. Attentional weights were adjusted during the category learning process, using the standard ALCOVE method for doing so. The second modification reflected the claim that ASD involves hyper-specific perception and behavior Happé (1999). The ALCOVE "specificity" parameter, which controls the rate of exponential decay in the activation function of the hidden units, was increased from its standard value of 1.0 to 2.5 in the autism model. This caused each hidden unit in the autism model to become strongly active for a more restricted range of stimuli.

0.7.5 Simulation Results

Model simulations were repeated 80 times in each experimental condition, with each repetition treated as an individual subject for data analysis. The control network exhibited the "prototype effect" observed in typically developing children, choosing the prototype over the non-prototype 70.52% of the time. For consistency, we used the same analysis as Klinger & Dawson (2001), a one-sample T-test, in order to determine if the model's preference for the prototype was reliably different than chance (50%). Analysis of the control model's performance indicated that the "prototype effect" was reliably larger than chance ($p < 0.0001$). The ASD model, however, preferred the prototype only 52.91% of the time, which was not reliably different than chance ($p > 0.30$). This matches the lack of a "prototype effect" in people with autism reported in previous studies Klinger and Dawson (2001); Gastgeb *et al.*

(2009).

0.7.6 Summary

These results show that failures at prototype formation in autism can be explained in terms of our hypothesized DA/PFC deficit. Restricting ALCOVE's attentional mechanism to hinder the flexible consideration of all stimulus features results in a lack of prototype preference. It is worth noting that matching the specificity parameter of the control model to match that of the autism model (2.5) does not change the presented pattern of results. This is evidence that it was the adjustment of initial attentional weights in the autism model, reflecting perseveration in PFC guided attention, that drove the observed difference.

0.8 The Utility of Computational Models for Understanding Autism

0.8.1 Computational Cognitive Neuroscience

An important contribution of this work involves the fact that it uses a relatively novel tool in ASD research: the methods of computational cognitive neuroscience. These methods provide a way to formalize how differences in the underlying neural circuitry give rise to the patterns of behavior found in people with autism. Specifically, we modified previously published and validated computational models of human behavior in accordance with the hypothesis that autism involves dysfunctional PFC/DA interactions, hindering updating of PFC contents. These modified models were shown to capture the behavioral performance of people with autism. This approach allowed us to offer a unified biological explanation for autistic behavior in the diverse areas of executive dysfunction, an implicit learning task, lexical disambiguation, and category learning. Indeed, the strength of the work presented here is not in any single model, but in providing a unified, plausible, precise neural mechanism that is capable of providing a kind of intertheorhetic reduction previously absent in ASD research. This work provides an example of the important role that computational cognitive neuroscience can play in improving our understanding of autism and other developmental disorders.

0.8.2 Previous Computational Models of Autism

The formal and explicit nature of computational cognitive modeling offers a novel approach to autism research. In order for computational models to be useful in this endeavor, they must be constrained by both bottom-up (neurobiological mechanisms) and by top-down (observed behavior) considerations, providing a formal character-

ization of the relationship between these levels of analysis. While there have been some previous computational models of ASD, it is not at all clear that they have offered such an explicit and detailed connection between biology and behavior.

Some previous computational models have attempted to address specific behavioral aspects of autism, including poor generalization Cohen (1994); Gustafsson (1997), Weak Central Coherence O'Loughlin and Thagard (2000), and overselectivity McClelland (2000). One ambitious computational framework, developed by Grossberg & Seidman (2006) , offers an explanation of multiple aspects of autism, including poor generalization, as well as cognitive and emotional issues.

A shortcoming of many of the existing models of autism is their fairly abstract nature, making little contact with specific neurobiological properties or measures Cohen (1994); McClelland (2000); O'Loughlin and Thagard (2000). Models that have incorporated biology have, thus far, only matched qualitative patterns of behavior rather than attempting to account for any quantitative behavioral data Gustafsson (1997); Grossberg and Seidman (2006). Models that are more tightly constrained by both neurobiological mechanisms and quantitative behavioral data, such as the models presented in this chapter, may have a more profound impact on our understanding of autism.

0.9 Conclusion

Autism diagnoses are on the rise. At the same time, ASD continues to pose a serious challenge to researchers. To date, there is no consensus concerning the neural underpinnings of ASD. Further complicating our understanding of autism is the staggeringly diverse behavioral profile of the disorder, as well as multiple physical abnormalities that often accompany a diagnosis. In this chapter, we have presented an approach intended to address some of these issues. Dopamine has diffuse and widespread effects throughout the brain. It has strong ties to multiple clinical populations. Increased seizure rates, motor abnormalities, stereotyped and repetitive behaviors, executive dysfunction, abnormal gaits, problems learning to follow eye gaze, and attentional abnormalities are all key components of behavior in autism, and all are linked tightly to the midbrain dopamine system. These things make dopamine dysfunction an intriguing candidate mechanism to consider. This chapter provides additional reasons to suspect a role for dopamine in autism. We have argued that perturbed DA/PFC interactions may lead to overly perseverative top-down control in autism, providing a neurally precise and plausible mechanism that might link previously disjoint theories in autism research.

By separating the mechanisms responsible for cognitive control and the flexible adjustment of control, perplexing aspects of the executive dysfunction profile in ASD are nicely captured. Cognitive control is instantiated via actively maintained control representations in PFC. Cognitive flexibility is implemented via interactions between PFC and the DA system. These interactions are suspect in autism, resulting in problems specifically in the flexible and appropriate updating of control and capturing the

problematic profile of executive dysfunction seen in ASD Hill (2004). Developmentally, executive dysfunction appears late in childhood. Our modeling efforts indicate that this may occur due to the protracted development of the PFC. In our model, early performance is driven largely by non-frontal, more posterior, brain systems which are largely unaffected by the posited DA abnormalities in ASD. As the PFC becomes more effective, differences in PFC/DA interactions are unmasked.

This work also speaks to Weak Central Coherence theory, which posits that people with autism have difficulties integrating pieces of information into a coherent "gestalt" Frith (1989); Happé (1999). The simulations reported in this chapter demonstrate how top-down PFC modulation of neural processing can influence the representations learned in other cortical areas. As such, WCC may be recast from being a general problem of information integration to being a problem of integrating the *wrong* information, due to the inflexible updating of PFC representations. Problems with implicit learning in the SRTT, as well as using sentential context to disambiguate word meanings, can both be explained by this account. These tasks depend on previously experienced information to be readily available at a later time in order for normal learning to occur, including learning that occurs at developmental time scales. Without reliable contextual information, neural systems struggle to integrate past information in an appropriate manner, driving learning to depend on other, less reliable, cues (e.g., frequency of word meaning).

We have also discussed differences in how people with autism learn category structures Klinger and Dawson (2001); Gastgeb *et al.* (2009). It is reasonable to assume that, in order to correctly form and use a prototype, we must have the ability to spread our attention across the relevant features of category examples. Learning to "overvalue" a restricted subset of features results in a failure to discern a valid prototype, and inflexible updating of top-down attentional control from PFC can produce such a restricted focus. In this way, our account addresses at least some phenomena of *stimulus overselectivity* observed in ASD Lovaas *et al.* (1971); Reed and Gibson (2005).

The convergence of evidence supporting dopamine's role in autism, combined with the possibility of providing a conceptual bridge spanning multiple theories in autism, is extremely encouraging. The work presented in this chapter helps to demonstrate the potential of computational cognitive neuroscience methods when investigating the links between biology and behavior in people with ASD.

References

Aylward, E.H., Minshew, N.J., Field, K., Sparks, B.F., and Singh, N. (2002) Effects of age on brain volume and head circumference in autism. *Neurology*, **59** (2), 175–183.

Baron-Cohen, S., Leslie, A.M., and Frith, U. (1985) Does the autistic child have a "theory of mind"? *Cognition*, **21** (1), 37–46.

Barto, A.G. (1994) Adaptive critics and the basal ganglia, in *Models of Information Processing in the Basal Ganglia* (eds J.C. Houk, J.L. Davis, and D.G. Beiser), MIT Press, Cambridge, Massachusetts, pp. 215–232.

Bennetto, L., Pennington, B.F., and Rogers, S.J. (1996) Intact and impaired memory functions in autism. *Child Development*, **67** (4), 1816–1835.

Berg, E.A. (1948) A simple objective test for measuring flexibility in thinking. *Journal of General Psychology*, **39**, 15–22.

Braver, T.S. and Cohen, J.D. (1999) Dopamine, cognitive control, and schizophrenia: The gating model. *Progress in Brain Research*, **121**, 327–349.

Braver, T.S. and Cohen, J.D. (2000) On the control of control: The role of dopamine in regulating prefrontal function and working memory, in *Control of Cognitive Processes: Attention and Performance XVIII* (eds S. Monsell and J. Driver), MIT Press, Cambridge, Massachusetts, chap. 31, pp. 713–737.

Canales, J.J. and Graybiel, A.M. (2000) A measure of striatal function predicts motor stereotypy. *Nature Neuroscience*, **3** (4), 377–383.

Carlsson, M.L. (1998) Hypothesis: Is infantile autism a hypoglutamatergic disorder? Relevance of glutamate - serotonin interactions for pharmacotherapy. *Journal of Neural Transmission*, **105** (4–5), 525–535.

Casanova, M.F., Buuxhoeveden, D., and Gomez, J. (2003) Disruption in the inhibitory architecture of the cell minicolumn: Implications for autism. *The Neuroscientist*, **9** (6), 496–507.

Cleeremans, A. and McClelland, J.L. (1991) Learning the structure of event sequences. *Journal of Experimental Psychology*, **120** (3), 235–253.

Cohen, I.L. (1994) An artificial neural network analogue of learning in autism. *Biological Psychiatry*, **36** (1), 5–20.

Cohen, J.D., Dunbar, K., and 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.

Cohen, J.D. and Servan-Schrieber, D. (1992) Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, **99** (1), 45–77.

Dayan, P. and Niv, Y. (2008) Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, **18**, 185–196.

Elman, J. (1990) Finding structure in time. *Cognitive Science*, **14**, 179–211.

Fernell, E., Watanabe, Y., Adolfsson, I., Tani, Y., Bergstrom, M., Hartvig, P., Lilja, A., von Knorring, A.L., Gillberg, C., and Langstrom, B. (1997) Possible effects of tetrahydrobiopterin treatment in six children with autism – clinical and positron emission tomography data: A pilot study. *Developmental Medicine and Child Neurology*, **39** (5), 313–318.

Frith, U. (1989) *Autism: Explaining the Enigma*, Blackwell, Oxford.

Gastgeb, H.Z., Rump, K.M., Best, C.A., Minshew, N.J., and Strauss, M.S. (2009) Prototype formation in autism: Can individuals with autism abstract facial prototypes? *Autism Research*, **2** (5), 279–284.

Goldman-Rakic, P.S. (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory, in *Handbook of Physiology – The Nervous System* (ed. F. Plum), American Physiological Society, Bethesda, MD, pp. 373–417.

Griffith, E.M., Pennington, B.F., Wehner, E.A., and Rogers, S.J. (1999) Executive functions in young children with autism. *Child Development*, **70** (4), 817–832.

Grossberg, S. and Seidman, D. (2006) Neural dynamics of autistic behaviors: Cognitive, emotional, and timing substrates. *Psychological Review*, **113** (3), 483–525.

Gustafsson, L. (1997) Inadequate cortical feature maps: A neural circuit theory of autism. *Biological Psychiatry*, **42** (12), 1138–1147.

Happé, F. (1999) Autism: Cognitive deficit or cognitive style? *Trends in Cognitive Sciences*, **3** (6), 216–222.

Happé, F.G.E. (1997) Central coherence and theory of mind in autism: Reading homographs in context. *British Journal of Developmental Psychology*, **15** (1), 1–12.

Hill, E. (2004) Executive dysfunction in autism. *Trends in Cognitive Sciences*, **8** (1), 26–32.

Hughes, C., Russell, J., and Robbins, T.W. (1994) Evidence for executive dysfunction in autism. *Neuropsychologia*, **32** (4), 477–492.

Kemper, T.L. and Bauman, M. (1998) Neuropathology of infantile autism. *Journal of Neuropathology and Experimental Neurology*, **57** (7), 645–652.

Klinger, L.G. and Dawson, G. (2001) Prototype formation in autism. *Development and Psychopathology*, **13** (1), 111–124.

Klinger, L.G., Klinger, M.R., and Pohlig, R.L. (2006) Implicit learning impairments in autism spectrum disorders: Implications for treatment, in *New Developments in Autism: The Future is Today* (eds J.M. Perez, P.M. Gonzalez, M.L. Comi, and C. Nieto), Jessica Kingsley Publishers, London.

Kriete, T. and Noelle, D.C. (2009) Implicit learning deficits in autism: A neurocomputational account, in *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (eds N. Taatgen and H. van Rijn), Cognitive Science Society, Amsterdam, The Netherlands, pp. 309–314.

Kriete, T. and Noelle, D.C. (2015) Dopamine and the development of executive dysfunction in autism spectrum disorders. *PLoS ONE*, **10** (3). E0121605, doi:10.1371/journal.pone.0121605.

Kurschke, J.K. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99** (1), 22–44.

Levitt, J.B., Lewis, D.A., Yoshioka, T., and Lund, J.S. (1993) Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 46). *Journal of Comparative Neurology*, **338**, 360–376.

Lovaas, O., Schreibman, L., Koegel, R., and Rehm, R. (1971) Selective responding by autistic children to multiple sensory input. *Journal of Abnormal Psychology*, **77** (3), 211–222.

Luce, R.D. (1963) Detection and recognition, in *Handbook of Mathematical Psychology* (eds R.D. Luce, R.R. Bush, and E. Galanter), Wiley, New York, pp. 103–189.

Martineau, J., Barthelemy, C., Jouve, J., Muh, J.P., and Lelord, G. (1992) Monoamines (serotonin and catecholamines) and their derivatives in infantile autism: Age-related changes and drug effects. *Developmental Medicine and Child Neurology*, **34** (7), 593–603.

Matsumoto, N., Hanakawa, T., Maki, S., Graybiel, A.M., and Kimura, M. (1999) Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *Journal of Neurophysiology*, **82** (2), 978–998.

McClelland, J.L. (2000) The basis of hyperspecificity in autism: A preliminary suggestion based on properties of neural nets. *Journal of Autism and Developmental Disorders*, **30** (5), 497–502.

Miller, E.K. and Cohen, J.D. (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24** (1), 167–202.

Minshew, N.J., Meyer, J., and Goldstein, G. (2002) Abstract reasoning in autism: A dissociation between concept formation and concept identification. *Neuropsychology*, **16** (3), 327–334.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, **16** (5), 1936–1947.

Mostofsky, S.H., Goldberg, M.C., Landa, R.J., and Denckla, M.B. (2000) Evidence for a deficit in procedural learning in children and adolescents with autism: Implications for cerebellar contribution. *Journal of the International Neuropsychological Society*, **6** (7), 752–759.

Nelson, E.C. and Pribor, E.F. (1993) A calendar savant with autism and Tourette syndrome. Response to treatment and thoughts on the interrelationships of these conditions. *Annals of Clinical Psychiatry*, **5** (2), 135–140.

Noelle, D.C. (2012) On the neural basis of rule-guided behavior. *Journal of Integrative Neuroscience*, **11** (4), 453–475.

O'Loughlin, C. and Thagard, P. (2000) Autism and coherence: A computational model. *Mind and Language*, **15** (4), 375–392.

O'Reilly, R.C. and Munakata, Y. (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, MIT Press, Cambridge, Massachusetts.

O'Reilly, R.C., Noelle, D.C., Braver, T.S., and Cohen, J.D. (2002) Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex*, **12** (3), 246–257.

Ozonoff, S. and Jensen, J. (1999) Specific executive function profiles in three neurodevelopmental disorders. *Journal of Autism and Developmental Disorders*, **29** (2), 171–177.

Pascual-Leone, A., Wassermann, E.M., Grafman, J., and Hallett, M. (2004) The role of the dorsolateral prefrontal cortex in implicit procedural learning. *Experimental Brain Research*, **107** (3), 479–485.

Perry, E.K., Lee, M.L.W., Martin-Ruiz, C.M., Court, J.A., Volsen, S.G., Merrit, J., Folly, E., Iversen, P.E., Bauman, M.L., Perry, R.H., and Wenk, G.L. (2001) Cholinergic activity in autism: Abnormalities in the cerebral cortex and basal forebrain. *American Journal of Psychiatry*, **158** (7), 1058–1066.

Posey, D.J. and McDougle, C.J. (2000) The pharmacotherapy of target symptoms associated with autistic disorder and other pervasive developmental disorders. *Harvard Review of Psychiatry*, **8** (2), 45–63.

Pring, L., Hermelin, B., and Heavey, L. (1995) Savants, segments, art and autism. *Journal of Child Psychology and Psychiatry*, **36** (6), 1065–1076.

Prior, M.R. and Hoffman, W. (1990) Neuropsychological testing of autistic children through an exploration with frontal lobe tests. *Journal of Autism and Developmental Disorders*, **20** (4), 581–590.

Qiu, A., Adler, M., Crocetti, D., Miller, M.I., and Mostofsky, S.H. (2010) Basal ganglia shapes predict social, communication, and motor dysfunctions in boys with autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, **49** (6), 539 – 551.

Reber, A.S. (1967) Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, **6**, 855–863.

Reed, P. and Gibson, E. (2005) The effect of concurrent task load on stimulus over-selectivity. *Journal of Autism and Developmental Disorders*, **35** (5), 601–614.

Rinehart, N.J., Bradshaw, J.L., Brereton, A.V., and Tonge, B.J. (2001) Movement preparation in high-functioning autism and asperger disorder: A serial choice reaction time task involving motor reprogramming. *Journal of Autism and Developmental Disorders*, **31** (1), 79–88.

Rinehart, N.J., Tonge, B.J., Iansek, R., McGinley, J., Brereton, A.V., Enticott, P.G., and Bradshaw, J.L. (2006) Gait function in newly diagnosed children with autism: Cerebellar and basal ganglia related motor disorder. *Developmental Medicine & Child Neurology*, **48** (10), 819–824.

Rodier, P.M., Ingram, J.L., Tisdale, B., Nelson, S., and Romana, J. (1996) Embryological origin for autism: Developmental anomalies of the cranial nerve motor nuclei. *Journal of Comparative Neurology*, **370** (2), 247–261.

Rougier, N.P., Noelle, D.C., Braver, T.S., Cohen, J.D., and O'Reilly, R.C. (2005)

Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, **102** (20), 7338–7343.

Rubenstein, J.L.R. and Merzenich, M.M. (2003) Model of autism: Increased ratio of excitation/inhibition in key neural systems. *Genes, Brain, and Behavior*, **2** (5), 255–267.

Schultz, W., Dayan, P., and Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, **275** (5306), 1593–1599.

Shah, A. and Frith, U. (1983) An islet of ability in autistic children: A research note. *Journal of Child Psychology and Psychiatry*, **24** (4), 613–620.

Staal, W., de Krom, M., and de Jonge, M. (2012) The dopamine-3-receptor gene (DRD3) is associated with specific repetitive behavior in autism spectrum disorder (ASD). *Journal of Autism and Developmental Disorders*, **42** (5), 885–888.

Stores, G. and Wivggs, L. (1998) Abnormal sleep patterns associated with autism: A brief review of research findings, assessment methods and treatment strategies. *Autism*, **2** (2), 157–169.

Stroop, J.R. (1935) Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, **28**, 643–662.

Stuss, D.T., Floden, D., Alexander, M.P., Levine, B., and Katz, D. (2001) Stroop performance in focal lesion patients: Dissociation of processes and frontal lobe lesion location. *Neuropsychologia*, **39** (8), 771–786.

Stuss, D.T., Levine, B., Alexander, M.P., Hong, J., Palumbo, C., Hamer, L., Murphy, K.J., and Izukawa, D. (2000) Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia*, **38** (4), 388–402.

Tsai, L.Y. (1999) Psychopharmacology in autism. *Psychosomatic Medicine*, **61**, 651–665.

Tuchman, R. and Rapin, I. (2002) Epilepsy in autism. *Lancet Neurology*, **1**, 352–358.

Turner, M. (1999) Generating novel ideas: Fluency performance in high-functioning and learning disabled individuals with autism. *Journal of Child Psychology and Psychiatry*, **40** (2), 189–201.

Wang, M., Vijayraghavan, S., and Goldman-Rakic, P.S. (2004) Selective D2 receptor actions on the functional circuitry of working memory. *Science*, **303**, 853–856.