Running head: DOPAMINE AND AUTISM

Autism Chapter

Trenton Kriete

Department of Psychology & Neuroscience

University of Colorado, Boulder

David C. Noelle

Cognitive & Information Sciences

University of California, Merced

Please address correspondence to:

David C. Noelle, Ph.D.
dnoelle@ucmerced.edu

Abstract

Do we need abstract for book chapter?

Autism Chapter

Introduction

Autism is an extremely complex developmental disorder diagnosed via the presence of a triad of symptoms including a qualitative impairment in social interactions, qualitative impairments in communication skills, and repetitive / stereotyped movements and behaviors (DSM-IV-TR, 2000). The severity, and sometimes even presence, of the various symptoms in autism is highly variable between those afflicted with the disorder. Due partly to this high variability, autism is generally seen as a spectrum of disorders known as autism spectrum disorders (ASD). Alongside the requirements for an autism diagnosis are diverse and complex physical and behavioral profiles which continue to challenge every theories seeking to explain autistic behavior to date. Steady progress is being made in early identification of behavioral characteristics of the disorder, as well as intervention techniques seeking to mitigate problematic behaviors. However, no consensus has been reached concerning the neural basis of autism. In the following we present a formal theoretical framework capable of explaining many aspects of autistic behavior in terms of specific neurological differences. Specifically, a computational account is used to demonstrate how dysfunctional interactions between the midbrain dopamine system (DA) 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 abilities as flexible adaptation, planning in order to reach a pre-specified goal, the generation of novel ideas, and determining the mental states of others (Bennetto, Pennington, & Rogers, 1996; Ozonoff & Jensen, 1999; Turner, 1999; Baron-Cohen, Leslie, & Frith, 1985). Physically, 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; Rinehart, Bradshaw, Brereton, & Tonge, 2001; Stores & Wivggs, 1998; Tuchman & Rapin, 2002).

Juxtaposed against the impairments of ASD exists a collection of spared, and sometimes enhanced, abilities in tasks such as the Embedded Figures Task (Witkin, Oltman, Raskin, & Karp, 1971) (EFT), where enhanced perceptual discrimination is regularly demonstrated people with autism (Jolliffe & Baron-Cohen, 1997; Shah & Frith, 1983). Along with the perceptual processing, superior and sometimes even savant abilities have been demonstrated in areas as diverse as mathematics, map memorization, music, artistic abilities, and date calculations (Pring, Hermelin, & Heavey, 1995; Mottron, Peretz, & Menard, 2000; Nelson & Pribor, 1993; Happe, 1999a).

The amazingly diverse profile demonstrated by people with ASD poses an incredibly daunting task for any single theory seeking to explain behavior in people with autism. Indeed, the most widely acknowledged theories of ASD are generally circumscribed to account for very specific aspects of behavior. For instance, an extremely pervasive theory in autism research has been the "Theory of Mind" (TOM) hypothesis, championed by Simon Baron-Cohen and Uta Frith (Baron-Cohen et al., 1985), among others. TOM posits that people with autism lack the ability to understand mental states, particularly intentional states ("I want", "I need", "I believe"), in others. TOM is presented as a primary reason people with autism struggle socially. It is unclear, however, how the TOM hypothesis could be used to explain the patterns of deficits and spared abilities outside the social domain in people with ASD. Instead, a separate theory is needed to account for this phenomenon. Often, the Weak Central Coherence (WCC) hypothesis is used to account for such abilities (Happe, 1999a). In WCC, people with autism are viewed as enacting a unique processing style, rather than possessing a deficit, per se. This style results in a "piecemeal" style of processing, rather than one that is more "holistic". It is argued that a bias in favor of processing the pieces, or parts, of objects and situations can enhance performance when the task requires attention to the smaller details. However, this bias comes at the price of a reduction in the ability to utilize more global, contextual, and gestalt information. WCC posits that in ASD, there is a fundamental problem integrating the parts and pieces of a situation in order to understand the more global "gist" or context. While WCC provides an account of some phenomena neglected by

TOM, there are still other behavioral patterns that fall outside the explanatory range of both of these theories. A third theory is needed to account for problems in planning, the flexible adaptation of behavior, and the generation of novel ideas. All of these processes have been viewed as depending upon an executive processing system. As such, dysfunction of these processes, known as Executive Dysfunction (ED), is believed by some to be a central feature of autism (Hughes, Russell, & Robbins, 1994; Hill, 2004; Ozonoff, Pennington, & Rogers, 1991).

The combination of multiple theories, in this manner, have allowed theorists to cover the behavioral landscape of ASD. It is not clear, however, that having a collection of relatively disjoint theories to explain different aspects of autism will foster the discovery of the underlying neural basis of this condition. Furthermore, even if reliable underlying neural differences are identified and correlated with specific behavioral patterns exhibited by people with autism, we will still not understand how the specific neurological differences give rise to the behavior. Knowing where the problem resides will not tell us why it results in autistic behavior.

In the following we address this gap by employing the methods of computational cognitive neuroscience. Computational models of cognition force the researcher to be explicit concerning the assumptions that are made, as well as the mechanisms employed, during scientific conjecture. The formal nature of these models allow us to form precise and testable hypotheses concerning the mechanisms responsible for the phenomena of interest. By incorporating explicit mechanistic characterizations of the underlying neurobiology, while capturing actual behavioral patterns, computational cognitive neuroscience models provide a means of bridging the conceptual valley between cognitive psychology and cognitive neuroscience in the domain of ASD research.

One general psychological question that has been explored using computational cognitive neuroscience techniques is how deliberate control over behavior (cognitive control) is instantiated within neural circuitry, as well as how this control is adjusted as environmental contingencies change (cognitive flexibility). The PFC has been broadly implicated in both cognitive control and cognitive flexibility (Stuss et al., 2000; Stuss, Floden, Alexander, Levine, & Katz, 2001). Under some accounts, cognitive control is enacted via the active maintenance of abstract rule-like representations in PFC. These sustained PFC representations provide a top-down task-appropriate processing bias to more posterior brain areas (J. D. Cohen, Dunbar, & McClelland, 1990). 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). Computational analyses have shown that these two cognitive processes, cognitive control and cognitive flexibility, are at odds with one another. Control requires robust and ongoing maintenance of a representation, where flexibility requires the ability to quickly and easily adapt these representations as task contingencies change. This processing conflict suggests the need for a mechanism that intelligently toggles the PFC between a maintenance mode and an updating mode. In order to avoid the introduction of an underspecified homunculus to control the PFC mode, computational accounts have sought a means for the PFC to learn when updating is appropriate. This focus on learning has drawn attention to the dopamine system.

Dopamine (DA), a neurotransmitter with diffuse projections throughout the brain, plays a vital role in contemporary models of prefrontal cortex function. The dopamine system is seen as implementing a reinforcement learning algorithm, driving the learning of action sequences that lead to reward (Montague, Dayan, & Sejnowski, 1996; Barto, 1994). The mesolimbic dopamine system also modulates frontal pyramidal cells. The DA projections to PFC are used to learn *through experience* when cognitive control should be maintained and when it should be flexibly modified in order to succeed at the current task (Braver & Cohen, 2000; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). A useful analogy for this process 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 needed 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. Recent computational models of PFC function suggest that intelligent "gating" of control representations in PFC can be learned, through experience, via the specific error signal provided by DA neurons, as uncovered by electrophysiological studies (Rougier et al., 2005; Rougier & O'Reilly, 2002).

These computational accounts of PFC function, and interactions between PFC and the DA system, are potentially highly relevant for understanding the neural basis of behavioral patterns in ASD. There is growing evidence for 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, Barthelemy, Jouve, Muh, & Lelord, 1992). Clinical trials testing different drugs that modulate levels of DA in the brain have shown behavioral benefits as well (Posey & McDougle, 2000; Tsai, 1999). Further strengthening the importance of DA dysfunction in people with ASD are the numerous links between DA and behaviors tied to ASD symptomology. These include increased prevalence of seizures, repetitive behaviors, and problems with skilled motor learning and control (Starr, 1996; Ralph-Williams, Paulus, Xiaoxi, Hen, & Geyer, 2003; Ralph, Paulus, Fumagalli, Caron, & Geyer, 2001; Graybiel, 2000; Matsumoto, Hanakawa, Maki, Graybiel, & Kimura, 1999). In the following we use computational modeling methods to investigate and formalize what effect a DA deficit would have on the behavior of a developing individual, relating simulated behavioral results to actual data from studies of people with autism.

In this chapter we investigate the degree to which a single deficit, a dysfunctional dopaminergic system, can account for many of the patterns of behavior demonstrated by people with autism. This investigation makes use of the methods of computational cognitive neuroscience, producing formal models that demonstrate how dopamine dysfunction can produce the behavioral patterns of interest. By focusing on a single neurological mechanism with diverse behavioral effects, this approach provides a level of inter-theoretic reduction not found in many current theories seeking to explain autism. The implications of the research expand beyond fundamental theorizing, potentially providing an improved understanding of the successes and failures of interventions utilized to reduce overselectivity and foster the generalization of learned behaviors in people with ASD.

In the following we propose that deficits in PFC / DA interactions are responsible for many of the interesting behavioral patterns observed in ASD, including executive dysfunction, stimulus overselectivity, and aspects of weak central coherence. The reduced efficacy of the DA-based gating mechanism on control representations stored within the prefrontal cortex results in overly perseverative attention. This has two results. First, the inability to properly adapt the control representations of the PFC produces inflexible behavior. This behavioral perseveration is manifested in inflexible actions and is exemplified by the executive dysfunction profile in autism. The second consequence of perturbed DA / PFC interactions is more subtle. We suggest that flexible adjustment of PFC plays an important role in shaping associational areas of cortex in a manner which affords generalization of behavior. The lack of flexible updating of the PFC results in cortical representations that associate an overly restricted subset of environmental features with appropriate behaviors (e.g. overselectivity). With only the restricted subset of cues dominating behavior, generalization will be hindered in situations not containing the associated and restricted subset of features. The resulting learning deficits can be seen in both tests of stimulus overselectivity and in measures of prototype extraction during category learning. Failure to appropriately update the contents of PFC can also result in an inability to appropriately use temporally extended context information, explaining observed deficits in sequential implicit learning tasks and in the use of sentential context to disambiguate words.

Next we provide important background information about autism spectrum disorders, a a more detailed description of the interactions between mid-brain dopamine system and the PFC, as well as a brief survey of previous computational modeling efforts seeking to explain behavior of people with autism. After the relevant background is covered, we summarize using computational modeling techniques how dysfunctional interactions between the DA system and the PFC can account for a wide range of behaviors in people with ASD including Executive Dysfuncion, overselectivity, implicit learning deficits, difficulty with lexical disambiguation, as well as prototype formation.

Background

Theories of Autism

Many current theories attempting to account for behavior demonstrated by people with autism can coarsely be divided into two categories, psychological and neuroscientific. Psychological theories attempt to explain behavior in terms of a deficit in a cognitive mechanism, while neuroscientific approaches look towards differences in the anatomy and biology first. Both approaches have recently begun incorporating information from the other in order to provide a more complete account. Psychological theories are looking to abnormalities in regions of the autistic brain which are known to correlate with the deficits that they posit. On the other hand, differences in the neuroanatomy are being matched to behavioral phenomena by identifying similar cognitive deficits in the adult neuropsychological literature, describing how damage to specific brain areas can affect behavior in a developed system. While the attempt at synergy and general acknowledgment of the need to explain autism in more reductionistic terms is extremely encouraging, there is still no consensus, nor any momentum, towards a particular approach or theory. The following background is intended to provide context on the current state of theorizing in autism, as well as providing further support and needed details to better support my specific theory of dysfunctional DA / PFC interactions in people with autism.

Psychological Approaches

Psychological approaches can generally be categorized as either hypothesizing core problems in the social domain or in a more domain general processing mechanism (Happe, Ronald, & Plomin, 2006). Deficits in the ability to attribute mental states to others, or theory of mind (TOM), has generally been used to explain social difficulties in people with autism (Baron-Cohen et al., 1985). More domain general processing mechanisms, such as deficits in executive processing or a drive to process information in a “piecemeal” manner, are used to explain other aspects of ASD, including perseverative behavior and spared abilities (Hill, 2004). Generally, science favors reducing explainable phenomena to its most basic components — to a single explanation if possible. The current trend in autism research sees this goal as, at best, distant. The theoretical competition is currently playing out within restricted problem domains in ASD, instead of across domains. For instance, it is argued by some that we should give up on finding any single mechanism explaining the triad of diagnostic impairments (social, communication, and rigid / perseverative behaviors) in autism (Happe et al., 2006). Instead, hypotheses are geared towards a subset of behaviors and not viewed as competing with theories attempting to explain behavior outside of their intended scope (Hill & Frith, 2003). This is, of course, not always the case, but, does appear to be a general guideline utilized by many theorists. Thus, as the following sections are read, providing a review of the current major psychological theories of autism, it is important to keep in mind that many researchers do *not* see these as competing positions.

Theory of Mind. The theory of mind (TOM) (Baron-Cohen et al., 1985) hypothesis suggests that social problems in people with autism are driven largely by an inability to understand the mental states of themselves and of others. “Mental states” are used here to refer to things such as our “beliefs”, “desires”, and “intentions”. The ability to utilize “mental state” information is casually referred to as “mindreading”. “Mindreading” is argued by proponents of TOM to be vital to the success of complex and dynamic social situations. In these situations, it is often the case that I may need to infer and interpret another person’s beliefs, desires, and/or intentions in order to respond in a socially appropriate manner. Without the ability to “mindread”, a person will likely fare poorly in social settings. The absence of this ability in people with ASD is hypothesized, according to TOM, to be at the core of their social difficulties.

Weak Central Coherence. TOM deficits are used to provide a possible explanation for a large range of the social deficits found in people with autism, but have little to say about other aspects of the cognitive profile in ASD, such as attentional abnormalities and problems in generalization. For instance, children diagnosed with autism tend to focus on parts of play objects, often at the cost of more functional or conventional ways of utilizing the toys (Bruckner & Yoder, 2007; Joseph, 1999). It is unclear how a deficit in attributing mental states would lead to such a pattern of behavior. Other theories, such as Weak Central Coherence (WCC), are focused on explaining such attentional abnormalities demonstrated by people with autism.

Strong coherence can be thought of as a tendency to integrate pieces of information into a coherent whole, or global interpretation, of the information. Weak Central Coherence (WCC) (Happe, 1999b; Frith, 2003) can be described as the opposite of this tendency, where the local parts are not gathered into a coherent "gestalt", but, instead, are left as atomic elements for processing. In Frith's account of WCC, it is suggested that people with autism exhibit a weak central coherence, processing the world in a local and "piecemeal" manner, rather than integrating the local pieces into more coherent and global wholes. It is important to note that this can be seen as a difference in processing styles, rather than a cognitive deficit, per se. This distinction is important because it affords WCC the ability to account for the spared, or even enhanced, abilities found in ASD, while still providing an explanation for the differences between normally functioning individuals and those with autism. This is a major strength of WCC. An example of this unique processing style can be found in the embedded figures test (EFT) (Witkin et al., 1971).

Executive Dysfunction. The Executive Dysfunction hypothesis views autism as emerging from a deficit in executive control over behavior (Hughes et al., 1994; Ozonoff et al., 1991). This hypothesis is used to account for the rigid, inflexible, and perserverative "stuck-in-set" behavior found in autism (Hill, 2004). Executive functioning is used as an umbrella term for a variety of deliberate and modulatory processes, such as planning, cognitive control, and cognitive flexibility. These processes are traditionally associated with frontal neural circuits, evidenced by deficits in tasks thought to measure executive processing in frontally damaged patients (Stuss et al., 2000, 2001). This theory is bolstered by impaired performance on many executive function tasks such as those believed to measure planning (e.g., Tower of Hanoi (Hughes et al., 1994; Ozonoff & Jensen, 1999)) and cognitive flexibility (e.g., Wisconsin Card Sort Test (Bennetto et al., 1996)). However, there are unaffected areas of executive functioning found in people with autism, as well. For instance, cognitive control seems to be relatively unaffected, as measured by the classic Stroop task. This raises into question the general Executive Dysfunction hypothesis as it has traditionally been cast. It is possible, however, that the executive problems found in ASD are not necessarily due to damage to the PFC, proper, but arise from problems with other brain structures that have connections with, and affect the functioning of, the frontal lobes (Robbins, 1997). It is just these kinds of questions —whether executive problems can be explained in terms of the dysfunction of specific neural circuits interacting with PFC— which computational models are well suited to help us explore.

Stimulus Overselectivity. Since Kanner's original description of "early infantile autism" in 1943, it has been noted that people with autism seem to be preoccupied with specific and sometimes peculiar parts of objects and situations (Kanner, 1943). In 1971 this phenomena was operationalized and termed "stimulus overselectivity". In the seminal study in this paradigm, a compound stimulus comprised of auditory, visual, and tactile components was presented to both low functioning children with autism and age matched control subjects. (See Figure **??**.) Initially, the subjects were trained to respond to the compound stimulus via an operant conditioning paradigm. Participants were rewarded when they made a specific action (e.g. lever press) when the compound stimulus was presented. After the acquisition of this initial stimulus / response pairing, each individual component was presented separately to assess the degree to which the individual components, themselves, had acquired control of the behavior. In the normally functioning control group, the participants responded equally to each of the individual components of the stimulus, demonstrating a lack of overselectivity. The group consisting of people with autism, however, responded to only one component of the three tested, thus demonstrating overselectivity. No systematic preference between the components was noted across subjects. This important result demonstrated how behavior in people with autism may be dominated by a small, restricted feature set, as compared to what is actually available in the environment. This result has been replicated across various modalities (Reynolds, Newsom, & Lovaas, 1974; Ploog & Kim, 2007) and even when varying the number of features (Lovaas & Schreibman, 1971).

A major implication of overselectivity is a reduced ability to generalize learned behaviors and general associations to novel settings. A study investigating why people with autism fail to generalize a newly learned behavior to a novel setting initially had each individual with autism taught a new behavior (e.g. to raise their right arm when the phrase, “raise your right arm” was spoken). After the initial behavior acquisition phase, the individuals were moved to a new location, which included a new experimenter, and tested the ability to generalize and perform the recently learned behavior in the novel setting. Next, for those participants who failed the generalization phase, items from the original setting were systematically introduced in order to determine exactly what had been learned by the participants with autism. Each of the participants who had failed to generalize appeared to be reliant on very specific, often idiosyncratic, pieces of the original training situation where the behavior was initially learned. For instance, one individual required the exact same hand movements, made by the original experimenter, to be made by the new experimenter in the new situation for any transfer to occur. Another person with autism needed the table and chairs from the original room to be present before he would transfer the learned behavior to the novel environment. Generalization is problematic in people with autism, based on these and similar reports, due to learning associations between the desired behavior and a restricted, possibly irrelevant, set of information. Failure to generalize learned behaviors is a major concern for many, if not all, intervention techniques used in autism treatment. Indeed, reducing overselectivity is a key ingredient in many of the most used therapies for children with autism (Lovaas, Schreibman, & Koegel, 1974; Koegel et al., 1989).

Neurobiology

Autistic deficits are widely believed to have a neurological origin. However, neuroscientific frameworks to date have had little success in providing a unifying view of the neural mechanisms responsible for behavior in autism. Indeed, the vast amount of variance in brain regions implicated as possible underlying neural substrates in ASD makes the task of identifying a unified neuroscientific account somewhat daunting. Caveats aside, there are many neurobiological differences thought to exist in ASD that are worth further exploration. In the end, all consistent underlying differences in the neurobiology need to be accounted for by any complete theory, or combination of theories, seeking to explain autism.

*** ADD PARAGRAPH HERE SUMMARIZING CURRENT FINDINGS ***

Cerebellum.

Mirror Neurons.

Brain Volume and Structure.

Limbic System.

Prefrontal Cortex (PFC).

Neurotransmitters.

Previous Computational Models of Autism

The formal and explicit nature of computational cognitive modeling suggests 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. It is not at all clear that the current computational models attempting to provide explanations for the behavior of people with autism have accomplished these goals. A brief review of existing computational modeling efforts focusing on the anomalous behaviors found in autism is presented in this section.

****** I THINK WE SHOULD GREATLY SHORTEN THIS, MAINLY PROVIDING A SUMMARY AND HEAD NOD TO CURRENT AND PAST MODELING WORK ******

Issues with Previous Models of Autism

Most of the existing models of autism reviewed above are fairly abstract in nature, making little contact with specific neurobiological considerations (I. L. Cohen, 1994; McClelland, 2000; O'Loughlin & Thagard, 2000). Even those models of autism which have incorporated biology into their framework have thus far only matched qualitative patterns of behavior in people with ASD, not attempting to account for any quantitative behavioral data (Gustafsson, 1997; Grossberg & Seidman, 2006). Models more tightly coupled with observed functional properties of neurobiological systems and constrained by actual behavioral data will be able to more precisely inform theories of ASD.

Dopamine & Autism Spectrum Disorders

Why Dopamine?

Given the ample choices of brain areas to investigate, many with intriguing correlations to behavior in autism, why champion a closer look at the role of dopamine in opposition to any of the other implicated areas? In truth there is, at this point, no "smoking gun" marking dopamine as a central cause of ASD. However, dopamine does provide us with a unique opportunity to build a bridge between seemingly disparate and complex observed patterns of behavior in ASD. Neurally, dopamine affects nearly all of the brain areas associated with autistic behavior, including the cerebellum, amygdala, PFC, parietal lobes, and the hippocampus (Barik & de Beaurepaire, 1996; Jackson & Moghaddam, 2001; Tessitore et al., 2002; Miller & Cohen, 2001; Mehta et al., 2000; Li, Cullen, Anwyl, & Rowan, 2003). Behaviorally, numerous strong links exist between autism and dopamine. This is impressive, given the diversity and range of physical (motor difficulties, stereotypies, etc.) and cognitive (social problems, executive dysfunction, etc.) differences demonstrated by people with ASD.

Evidence for Dopamine Dysfunction in ASD. The evidence for dopamine abnormalities in ASD is strong. Both PET imaging studies and urinalysis studies reveal differences in levels of dopamine in people with autism (Fernell et al., 1997; Martineau et al., 1992). However, clinical trials investigating the effects of both dopamine agonists and antagonists have had mixed results (Posey & McDougle, 2000; Tsai, 1999). Many studies of drugs affecting brain levels of DA have shown improvements in stereotypic behaviors and other problem behaviors, but, some have shown an increase. Studies involving clinical trials can be very difficult to evaluate. For instance, at low doses, DA antagonists have been shown to actually have an agonistic effect on phasic dopamine release (Frank & O'Reilly, 2006). While it maybe difficult to interpret the clinical results, the data does strongly suggest a role for dopamine in the behavior of people with autism.

There are a number of different empirical results that support the idea that DA function plays an important role in autism. These will be discussed next. As isolated arguments, these results provide only weak support tying DA to ASD. Taken together, however, they provide a strong case for seeing the DA system as key to understanding many of the behavioral patterns in autism.

DA Ties to ASD Symptomology. Dopamine has a role in many of the problematic behaviors demonstrated by people with autism. These behaviors range from lower-level, non-intentional behaviors (such as seizures), to those at a much higher-level, such as learning to follow eye gaze and the ability to control and flexibly adapt our behavior. The breadth of the links described below are, perhaps, the strongest argument for a closer examination of the causal role of dopamine in autism.

Approximately 1 in 4 people with a diagnosis of autism will develop seizures during adolescence, significantly higher than the prevalence observed in the general population (Tuchman

& Rapin, 2002). In normally developing individuals, seizures occur approximately 1 in every 200 people. This indicates a more than ten-fold increase in ASD as compared to the general population. For many decades, researchers have believed that dopamine plays a major role in epilepsy (Starr, 1996). Anti-convulsant medications are known to have direct affects on the dopamine system, helping mitigate the seizures caused in epilepsy. The link between DA and seizures further strengthens an argument for the possibility of a role of DA in ASD.

People with autism also demonstrate a wide spectrum of motor abnormalities and problems. These problems range from problems initiating behaviors, repetitive movements, as well as abnormal gaits (Rinehart et al., 2001, 2006; APA, 2000). There are possibly large benefits to investigating the underlying cause of the motor and movement abnormalities found in ASD, as well as understanding the possible social manifestations of these movement and motor difficulties. Not the least of these possible benefits could be a diagnostic criteria for autism containing specific observable behaviors to compliment the existing socially based standards (Leary & Hill, 1996). The basal ganglia and mesolimbic dopamine system are widely accepted as a vital component in learning and initiating motor movements. Problems within these brain areas are believed to be at the root of disorders, such as Parkinson's and Huntington's disease, which manifest problems in motor control, as well as other more cognitive symptoms. Also, stimulation of the dopamine receptors located within the striatum has been shown to induce motor stereotypies, and ameliorated / abolished by blocking the dopamine transmission within the striatum (Canales & Graybiel, 2000; Ralph-Williams et al., 2003; Ralph et al., 2001). The link of motor control, movements, and motor problems to dopamine and dopamine producing areas is strong. Indeed, as mentioned, even stereotypic behavior, one of the triad of impairments currently needed for an autism diagnosis, has direct ties to dopamine.

One of the earliest and most reliable predictors in autism-related language impairments and social performance is abnormalities in shared joint attention (SJA) (Dawson et al., 2004). SJA can be defined as the ability to coordinate and follow attention between one's self, an object, and another person. The sharing of attention is often achieved by following eye gaze and pointing gestures in order to find a rewarding object in the environment. Problems with SJA are believed to be a key reason for the communicative and social deficits demonstrated by people with ASD. Recently, a formal account of how gaze following, and subsequently shared attention, is learned via interactions between an infant and caregiver has been proposed (Triesch, Teuscher, Deak, & Carlson, 2006). According to this account, SJA emerges naturally given only a very basic set of assumptions and mechanisms. At the core of these assumptions lies a learning account, inspired by reinforcement learning paradigms and precise neurobiological data. The reward based learning method employed by Triesch and colleagues, known as Temporal Difference or (TD) learning, has been strongly linked to the firing patterns of midbrain dopamine cells by other researchers (Barto, 1994; Montague et al., 1996). Leveraging this dopamine inspired learning method, Triesch et al. demonstrate how gaze following emerges naturally, with experience, given a simple structured environment and the aforementioned learning paradigm. Importantly, it is also shown how manipulating the reward structure reproduces problems in eye gaze following and shared attention as seen in people with ASD. The authors are careful to not make any unnecessary claims or assumptions as to the exact underlying biological mechanism responsible for the theorized change. However, as mentioned, the reward structure in models such as these is most often associated with the firing patterns of cells located within the midbrain dopamine system. It is also worth noting that a recent extension of the model has been used to show how these same basic learning mechanisms can lead to cells, within a modeled supplementary motor area, with receptive field properties similar to “mirror neurons” (Triesch, Jasso, & Deak, 2007).

The ties of DA to autism are both numerous and compelling. Further investigations into precisely how these differences affect behavior has great potential in providing a common language, that of neurobiology, for linking many disparate behaviors in people with ASD.

Dopamine & Temporal Difference Learning

Our account of the role of dopamine in autistic behavior builds on past findings concerning the midbrain dopamine system and its relationship to the prefrontal cortex. Analyzing the response profile of DA neurons in the basal ganglia of monkeys, Schultz et al. (1997) demonstrated DA cells can encode a prediction error in the amount of future reward expected to be given to the monkey. In other words, these cells seem to encode a *change in expected future reward*. Figure 1 shows results from a population of midbrain DA cells during one of Schultz's experiments. The top panel represents the situation in which the monkey is not expecting reward, but then receives reward (e.g., a sip of juice). Notice that the DA cells fire upon receiving the reward (signified by "R" on the graph), encoding a positive change in what the monkey was expecting. In the bottom left panel, the monkey has now been conditioned to associate a flash of light with the delivery of the juice, after a short delay. In other words, the monkey now knows that the flash of light predicts future reward. When the flash of light is seen (represented as "CS", for "conditioned stimulus", in the graph), the DA cells fire. This can be explained as the monkey expecting future reward once the light comes on, signaling that juice is expected to be coming soon: a positive change in expected future reward. However, when the reward is delivered ("R") the cells to do not fire, since the monkey was already expecting reward. When the juice is delivered there is no change in expected future reward, in this case, and, therefore, no increase in the rate of DA firing. In the panel located at the bottom right, the DA cells again fire for the flash of light ("CS", conditioned stimulus) , but this time the experimenters *withhold the juice* at the time when the monkey is expecting the juice to be delivered. The monkey is *expecting* reward, but no reward is delivered. Thus, at the time that juice is expected, there is a negative going change in expected future reward. Notice that the firing rates of the DA cells around the expected delivery time of reward ("R") actually dip below their baseline firing rate and, indeed, appear to encode this negative change in expected future reward.

This is very interesting because change in expected future reward is also the key variable in a very powerful reinforcement learning algorithm known as Temporal Difference (TD) learning. In

TD learning, the change in expected future reward, the same value the DA cells appear to be encoding, is know as the TD Error. Across two consecutive time steps the TD Error is given by:

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

Where $r(t)$ is a continuous reward value that is delivered at each time step based on system performance (e.g., $r(t) = 1$ for correct performance and $r(t) = 0$ on time steps when no reward is presented), $V(t)$ and $V(t+1)$ are the expected future rewards at times t and $t+1$ respectively, $\delta(t)$ is the change in expected future reward, or TD Error, and γ is a constant discounting factor, where $0 < \gamma \leq 1$. Adjusting γ changes the amount by which temporally distant rewards are discounted as compared to rewards that can be attained in the temporally near future.

This connection has led researchers to formalize the role of midbrain DA neurons in the brain's learning mechanisms (Barto, 1994; 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 (Montague et al., 1996), driven by the reward-prediction DA signal.

Computational Models of PFC

Our work builds on previous computational modeling work of formal accounts of DA's affect on PFC functioning. The effect of DA is formalized by equating the firing rate of midbrain DA neurons to the key variable, the TD Error, of the powerful TD learning algorithm. Using this connection between biology and machine learning, researchers have been able to provide models of how motor systems can learn sequences of overt actions leading to reward. One of the primary insights of some recent models of PFC functioning 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 & Cohen, 2000). As described earlier, it is helpful to think of the maintenance versus updating of PFC in terms of a gating mechanism. When the gate is closed, the PFC representations are robustly maintained and protected from interference. When the task contingencies change, the gate can be opened to allow for a more useful PFC goal or rule to be maintained to influence further processing. The key insight is that, if TD can be used to learn sequences of *overt* actions, it might be possible to use this same error signal 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, researchers have shown that this account is plausible (Braver & Cohen, 2000; O'Reilly, Noelle, Braver, & Cohen, 2002). A layer of processing units representing the PFC is included in these models, and this layer is used to actively maintain abstract task dimensions across the firing patterns of the units. For instance, the PFC layer can encode, and actively maintain, a representation such as "pay attention to the color of the stimuli". This maintained pattern of activity can then provide a "top-down" bias or up-modulation of pathways in posterior brain areas associated with the processing of stimulus color (J. D. Cohen et al., 1990). The extra biasing provided by the PFC can be used to drive weaker, less automatic, behaviors (e.g. naming the color as opposed to reading the word in the Stroop task) when appropriate. This activation based modulation is thought to be key to our ability to provide cognitive control over behavior (J. D. Cohen & Servan-Schrieber, 1992). The DA based adaptive gating mechanism can be used, within this context, as a way to 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 the system is doing better than expected, close the gate on PFC representations so we are more likely to keep doing the same thing. Conversely, when the network starts performing worse than expected, possibly due to task contingencies changing, this will result in a negative TD Error, signaling that the system is not performing as well as expected, and indicating that the system should adapt its behavior to perform more optimally. The negative TD error can be used as a gating signal on the PFC representations, signaling the gate to open, allowing a new representation to replace the old, thereby allowing the network to flexibly adjust its control over behavior.

Along with providing a neural mechanism that can learn to appropriately and adaptively gate PFC representations, these models have also been successful in tying frontal disturbances, such as those found in schizophrenia, to deficits in cognitive control (J. D. Cohen & Servan-Schrieber, 1992) and cognitive flexibility (Braver & Cohen, 1999; O'Reilly et al., 2002). A recent elaboration of this model, XT (Rougier et al., 2005), is the first neuroscientific model able to provide quantitative fits to a hallmark task of cognitive control, the Stroop task, and a widely used measure of cognitive flexibility, WCST, in both neurologically intact and frontally damaged people.

General Modeling Approach

One reasonable modeling strategy would be to create one expansive model that encompasses all phenomena of interest, demonstrating both how normal and autistic-like behavior may arise in the same framework given specific dopamine-related parameter adjustments. However, the range of relevant behaviors that are affected by autism is incredibly varied making this approach nearly untennable. Rather than building one massive model from scratch, separate previously published models, each considered to be one of the strongest representatives in their respective domains, were manipulated in a manner which is consistent with our general hypothesis. Supplementing the previously published models, one completely novel and unpublished model, encompassing an important aspect of behavioral data in autism, was developed as well. This approach allows me to demonstrate —using the best models currently available— how inflexible attentional switching, caused by deficient DA / PFC interactions, can facilitate and possibly cause the behavior seen in people with autism. 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 demonstrate how cognitive inflexibility, arising from improper DA modulation of PFC, can account for a wide variety of behavioral phenomena observed in autism, including: poor generalization in concept learning tasks (e.g. prototype formation), a lack of context sensitivity in language processing, stimulus overselectivity, executive dysfunction, and deficits in implicit learning.

Executive Dysfunction

Patterns of Function and Dysfunction

People with autism are impaired across a broad range of cognitive tasks, including planning (Bennetto et al., 1996), flexibly adapting behavior (Bennetto et al., 1996; Ozonoff & Jensen, 1999), and tasks requiring spontaneous generation of novel behaviors and ideas (Turner, 1999). These particular tasks have been associated with executive control processes, so the observed impairments have led some researchers to view executive dysfunction as a central feature of autism. Indeed, the Executive Dysfunction (ED) theory of autism seeks to explain many of the behavioral patterns exhibited by these individuals in terms of a failure of executive control over behavior (Hughes et al., 1994).

There is extensive evidence that the prefrontal cortex plays an important role in executive control. Along with the central claim of ED, this suggests that the root cause of many autistic behavioral patterns may lie in abnormalities in this region of the brain. This is an interesting hypothesis because, while substantial progress has been made in many areas of autism research, no consensus has been reached concerning the neural basis of the disorder. This view of ED suggests that the irregular development of prefrontal cortex may underly the patterns of cognitive performance seen in autism.

A more detailed examination of autistic behavior reveals that not all forms of executive processing are impaired, however. A somewhat perplexing aspect of the executive profile demonstrated by people with autism is that cognitive flexibility has been shown to be impaired while fundamental cognitive control remains robust and relatively unaffected. Cognitive control describes my ability to enact a behavior in the presence of a distracting or more automatic competing response. In contrast, cognitive flexibility can be described as our 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 & Jensen, 1999). This dichotomy challenges the notion that autistic behavior is the result of a global impairment of executive processes, perhaps mediated by frontal abnormalities.

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 matched for age and verbal ability (Griffith, Pennington, Wehner, & Rogers, 1999), calling into question the role of ED in the etiology of autism. Any theory intending to explain executive dysfunction in autism must account for the relative “strengths” and “weaknesses” that have been observed, as well as for this lack of observable deficits early in development.

One clear approach to explaining the executive profile seen in ASD involves positing separate mechanisms for cognitive control and for the flexible adaptation of control. In autism, the mechanism for control might be intact, but the mechanism responsible for the flexible adjustment of control might be compromised. Interestingly, this segregation of function is captured by an existing computational model of the prefrontal cortex and its role in executive processing: the *Cross-Task Generalization model (XT)*. Driven by broad neurocomputational considerations, XT casts the prefrontal cortex (PFC) as central to cognitive control, while interactions between the PFC and the mesolimbic dopamine (DA) system mediate the flexible adaptation of control. XT has been used to capture the performance of both frontally damaged individuals and healthy controls on both the Stroop task and WCST (Rougier et al., 2005).

In this chapter, we demonstrate that XT also offers a possible explanation for the executive processing profile exhibited by persons with autism. Specifically, we have found that simply weakening the influence of DA on PFC in the model is sufficient to both qualitatively and quantitatively capture autistic performance on both Stroop and WCST. This computational modeling result suggests that executive deficits in autism may be mediated by PFC/DA interactions. Importantly, XT is a learning model, in which the development of neural representations and associated behavioral performance can be tracked as the model matures. Leveraging this property of XT, we show that the late appearance of executive deficits might be explained by the late maturation of PFC representations and PFC/DA interactions. According to the model, early performance is driven largely by non-frontal, more posterior, brain systems which are largely unaffected by the posited DA-related abnormalities in autism. As the PFC becomes more effective, differences in PFC/DA interactions are unmasked.

Modeling Prefrontal Cortex

Gating in the Prefrontal Cortex. As described in Section 2.3, the PFC has been broadly implicated in cognitive control and cognitive flexibility (Stuss et al., 2000, 2001). Under some accounts, cognitive control is enacted via the active maintenance of abstract rule-like representations in PFC. These sustained PFC representations provide a top-down task-appropriate processing bias to more posterior brain areas (J. D. Cohen et al., 1990). 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). Computational analysis of these neural circuits have shown that active maintenance and the flexible adaptation of control are at odds, with the mechanisms that maintain PFC representations, and protect them from distracting inputs, acting as an obstacle to the rapid updating of PFC contents in response to shifting contingencies. Thus, in order to achieve flexible behavior, a separate mechanism is needed to intelligently and rapidly update the actively maintained PFC control representations in a task appropriate manner. This can be seen as a “gating” mechanism, toggling between a state of maintenance and a state of updating, as appropriate for the task. XT suggests that the gating decision is learned from experience, and this learning process critically involves the midbrain dopamine system, reified as the TD learning algorithm (Braver & Cohen, 2000; Rougier et al., 2005; Barto, 1994).

The XT Model. The architecture of the XT model is shown in Figure 2. This computational cognitive model makes use of the biologically grounded Leabra framework (O'Reilly & Munakata, 2000). The input of XT consists of two layers of neural units that can be used to specify 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, with strong lateral inhibition among response units allowing the network to output a single stimulus feature (e.g., red) at a time. The Response layer includes one additional unit, which codes for "no response".

A collection of Hidden layers model posterior cortical circuits that map from stimuli to responses. Activity in these layers can be adjusted, however, through top-down biasing signals from an actively maintained representation in the PFC layer. Unlike previous models, the representations that can be maintained in PFC are learned through a "childhood development" process that involves extensive training on the performance of a variety of related tasks. These developmental tasks share a need to selectively attend to individual dimensions of the stimuli. For instance, the network might be presented with a large red ball and a small red duck, and might be asked to identify the feature that is the "same" across these two stimuli. In this case, the network must focus upon the "color" dimension and respond "red". On each trial, corrective feedback is provided to drive Leabra's synaptic learning processes, and correct responses from the network produce an external reward signal ($r(t)$) of one (as opposed to the usual zero reward). Over time, the developmental training of the network allows it to identify the separate dimensions of stimuli, and it causes the network to hone its skill at focusing attention on a single dimension at a time (Rougier et al., 2005).

The Task input to the network indicates which task is to be performed (e.g. the “Naming Feature” task, the “Matching Feature” task, the “Smaller Feature” task, or the “Larger Feature” task), with one input unit coding for each task. The Dimension Cue layer is used to inform the network of the currently relevant stimulus dimension. For example, the Dimension Cue input is used to model the Stroop task by informing the network when it should attend to the color of the stimulus rather than its word form, or vice versa. Each unit in the Dimension Cue layer corresponds to a stimulus dimension, and all of these inputs are turned off for tasks in which the network must discover the relevant stimulus dimension on its own, as in WCST.

The flexible adjustment of cognitive control is implemented using a DA-based adaptive gating (AG) mechanism. The synaptic weights feeding the special AG unit calculate an estimate of $V(t)$, and these weights are adjusted based on $\delta(t)$ using a neural implementation of the TD learning algorithm. Importantly, the TD error, which encodes a DA signal, is also used to manipulate the “gate” on the PFC layer. When the model performs better than expected ($\delta(t) > 0$) the current PFC representation is strengthened using an intrinsic maintenance current to stabilize the existing pattern of activity. When the model performs worse than expected ($\delta(t) < 0$), the current PFC representation is destabilized, allowing a new, possibly more appropriate PFC representation, based on the inputs into the PFC layer, to be entertained. In the model, the $\delta(t)$ value directly modulates excitatory ionic maintenance currents (g_m below). Large maintenance currents drive the membrane potential of simulated neurons in the PFC up, pushing them towards their maximal firing rate. These currents are not allowed to become negative, being clipped at zero instead. The maintenance currents, g_m, of simulated neurons in PFC are computed by:

$$g_m(t-1) = 0\, if\ |\delta(t)| > \theta \tag{2}$$

$$g_m(t)_j = g_m(t-1) + \delta(t)a_j \tag{3}$$

$$where\ a_j\ is\ the\ current\ activation\ value\ of\ PFC\ unit\ j$$

Therefore, a positive $\delta(t)$ will result in an increase in active maintenance of PFC representations, while a negative $\delta(t)$ will destabilize PFC. The value θ represents a threshold value for the ionic currents. If the TD error, $\delta(t)$, exceeds this amount ($\theta = .5$ in all simulations), then the maintenance currents, g_m, are effectively reset. Over time, the network learns to maintain PFC representations that are likely to result in reward.

XT is the first computational cognitive neuroscience model to explore the development of PFC representations, and it is the first to provide good quantitative fits to both Stroop and WCST data, for both neurologically intact and frontally damaged people, based on a biologically informed architecture.

Modeling Autism Using XT

General Approach. Based on the XT framework, my 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 by a constant factor, κ, where $\kappa = 1$ for healthy individuals and $\kappa < 1$ for people with autism. The scaling of $\delta(t)$ by κ is the only modification from the original XT model that we have made in these simulations. A κ value of 0.54 was found to produce the best fit to human performance, so this value was used in all of my simulations. This reduction of the DA signal can be seen as decreasing the efficacy of the PFC gating system, resulting in the less efficient destabilization of PFC when errors are unexpectedly made. The scaling was only used to modulate the simulated maintenance currents within the PFC layer, and not used to modify the learning of the weights into the Adaptive Gating unit. An important point of future work is to also test the effects of this manipulation on these weights, weakening the overall influence that dopamine has on learning within circuits involved in the computation of future expected reward, as well.

Modeling WCST. The WCST consists of a deck of cards, which contain stimuli varying along three dimensions (e.g., color, shape, quantity) and across four different features per dimension (e.g., for color dimension: red, orange, green, & blue). Participants are told to sort the cards into piles, but they are not given any instructions concerning how to do this correctly. Instead, only sparse feedback —"Correct" or "Incorrect"— is given upon the placement of each card, until the proper sorting strategy is discovered. After the sorting rule (e.g., sort by color) is learned by the participant, and 10 consecutive correct sorts are accomplished, the rule is changed without informing the subject. This procedure repeats until either 6 correct categories (sets of 10 correct consecutive sorts) are achieved, or all 127 cards in the deck are exhausted. Errors are recorded as incorrect sorts, with perseverative errors scored as an incorrect sort that used the last correct sorting rule. Success at WCST requires the ability to flexibly change the dimension being maintained by PFC as the sorting rules change. Modeling WCST in the XT framework required use of only three of the five possible input dimensions. The same methods for administrating WCST to human participants were used in these simulations.

The network was presented with stimuli at the input or "stimulus array" layer by activating individual units, one for a single feature across each of three dimensions. (See Figure 3.) This input represented the current card to be sorted. The task of the network was to name the currently relevant feature (e.g., the feature "red" if color is the currently relevant sorting dimension). No information is provided via the Dimension Cue layer concerning which dimension should be used as the currently correct sorting rule, leaving the network to use a more-or-less random search strategy until the correct rule is discovered. The network only receives sparse feedback — "reward" or "no reward" — receiving reward on trials when correct performance is achieved, and no reward when an incorrect guess is made. Left to the random search strategy, the network's performance would tend towards chance, leading to grossly deficient performance on WCST. XT is able to leverage the DA based AG mechanism coupled with the active maintenance and top-down influencing properties of the PFC layer in order to successfully perform the task. The

AG mechanism strengthens the PFC's intrinsic ionic maintenance currents when the network is performing well, allowing PFC to actively maintain currently relevant information (e.g., pay attention to the color dimension), biasing the processing pathways which are part of the currently maintained stimulus dimension so as to give them a competitive advantage over rival pathways. The actively maintained PFC representations form a "memory" of the rule. When the rule switches (e.g., after 10 consecutive correct sorts), the actively maintained representation becomes invalid. If the network allows the invalid representation to influence subsequent processing, a large amount of perseverative errors will result. The AG prevents this by providing a gating signal to PFC when reward is expected but not delivered, allowing a new representation to be acquired by PFC.

Modeling Stroop. Stroop tests cognitive control by measuring the ability to inhibit a prepotent response. In Stroop, the stimuli are different words presented in various colored fonts. Participants are asked to either read the word or to name the color of the font in which the text is presented. People are faster overall at reading the word as opposed to naming the color of the word. Furthermore, when comparing the neutral (e.g., the word "house" in red font) versus the incongruent (e.g., the word "green" written in red font) conditions, people are slower in the incongruent case for color naming, but not for word reading. This is known as Stroop interference.

Cohen and Servan-Schreiber (1990) provided a computational account of the Stroop task. This model incorporated multiple associative pathways from stimulus features to possible responses. They used a stronger, more automatic word reading pathway through posterior cortex and a relatively weak color naming pathway. In their model, a task representation, maintained in PFC, provided top-down biasing, injecting extra activity into the color naming pathway when doing so was necessary to overcome the prepotent word reading pathway (i.e., when asked to name the font color). The resulting competition between the two strong pathways resulted in an increase in response time in the color naming incongruent condition. Response time was not increased in the word reading incongruent condition, however, as the weaker color naming pathway, when unsupported by PFC, offered little competition.

The XT model addressed the Stroop task in much the same way, leveraging competition between pathways with different levels of strength or prepotency. In order to simulate the relative prepotency of a stimulus dimension in my model, we manipulated the frequency in which one dimension (font color) was experienced during training, making the dimension relevant only 25% as often as the other dimensions (word reading). The settling time, in “cycles”, of the network was linearly scaled to “milliseconds” using a single free parameter, allowing a direct comparison between model results and human data.

Model Simulation Results

Simulations. A total of 100 networks were prepared using the XT developmental training procedure, stopping when the network achieved a stringent performance criterion, or after a maximum of 100 epochs. Each network was then tested under both conditions of DA modulation ($\kappa = 1$ for healthy networks and $\kappa = .54$ for modeled autistic performance), on each of the WCST and the Stroop tasks. The 100 networks were each treated as individual subjects for the purposes of data analysis.

WCST Results. My WCST simulations of both healthy individuals ($\kappa = 1$) and persons with autism ($\kappa = 0.54$) provided results matching those reported in the literature. The differences between the simulated performance of normally functioning individuals and the simulated autistic performance were statistically reliable ($p < 0.001$), and consistent with previous studies (Prior & Hoffman, 1990; Ozonoff & Jensen, 1999; Minshew, Meyer, & Goldstein, 2002). In particular, the perseverative error measure was significantly higher in the DA modulated version of my model as compared to the model of normal function. (See Figure 4.)

Stroop Results. Model performance on the Stroop task provided a good quantitative fit to human performance. (See Figure 4.) The model with intact DA function displayed the classic Stroop reaction time results. The prepotent word reading dimension showed uniform reaction times across both congruent and conflict conditions, while the weaker color naming dimension exhibited a slowing in reaction times when the stimuli were incongruent. The performance of the autistic model was virtually identical, with no significant increase in the overall Stroop effect ($F(1, 198) = 0.62$; $p > 0.43$), which is consistent with past findings (Ozonoff & Jensen, 1999).

Developmental Results. My initial 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. A second set of simulations was conducted to address this issue and to examine the developmental time course of cognitive control and cognitive flexibility in these models. PFC development and model performance were analyzed over the entire developmental period of the model (100 epochs). Two groups of 10 networks (individuals) 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, the developmental time course data provided some insight into why executive deficits might appear late in autism, as described in the literature (Griffith et al., 1999). This insight can be had by carefully examining how PFC representations change in the model over time, and how these representations affect Stroop and WCST performance.

PFC Representations. Figures 5 and 6 plot the synaptic strengths from the PFC layer to the Response layer. Each large box corresponds to a PFC unit, and each encapsulated small box corresponds to a Response unit, with the strength of the connection from the given PFC unit to the given Response unit being reflected in the brightness of the box (lighter means stronger). Note that each row designates connections to Response layer units representing features in the same stimulus dimension. (See the upper left corner of Figure 2.) Thus, these plots can be used to examine the degree to which the PFC layer developed a representational scheme that allowed for the selective modulation of individual stimulus dimensions. For a given PFC unit (large box), a bright row of connections (small boxes) indicates that that unit selectively supports attention to the stimulus dimension corresponding to that row. In Figure 5, which was generated early in developmental training, both the autistic network and the control network lack strong dimensional weights. However, in Figure 6, both networks have developed strong dimensional representations in PFC, as evidenced by the extremely salient horizontal bands of "strong" weights. Thus, appropriate PFC representations were acquired only slowly over the course of developmental training, and the DA manipulation did not hinder the formation of these representations.

Executive Dysfunction Development. Introducing a deficit in the DA-based gating mechanism from the beginning of development still allows the network to capture autistic performance on tasks requiring cognitive flexibility and control. In Figures 7 & 8, the Stroop interference effect and the number of perseverative errors in WCST are plotted as a function of developmental training time. Figure 7 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 ($p < .003$), demonstrating robust cognitive control throughout the entirety of development. Figure 8, however, demonstrates a significant increase in the number of perseverative errors made by the DA-modulated network as early as epoch 12. During the first 53 epochs there was a significant difference ($p < 0.05$) during only 26.4% of the epochs. However, later in the development of the model (epochs $54 - 100$) a significant difference was reached 93.6% of the time. Interestingly, neither "healthy" models nor DA-deficient models showed a distinct advantage or disadvantage during the earliest stages of development. We believe that the lack of executive deficits early in the development of these models was largely due to the fact that both models lacked strong PFC representations early in training. Without strong, dimensional PFC representations, the models were forced to rely more heavily on the Hidden layers (posterior cortex) to perform the tasks. Since weakening the DA-based gating mechanism had no effect on these "posterior" areas of the model, neither model showed any advantage over the other at this early stage.

Summary

Given the XT account of the role of PFC in executive control, we have shown that a single manipulation — reducing the efficacy of the DA signal — is sufficient to capture the pattern of performance exhibited by people with autism on basic tests of cognitive flexibility (WCST) and cognitive control (Stroop). Furthermore, we demonstrated that weakening the DA-based gating mechanism over the entire course of PFC development continues to reflect both the "strengths" and "weaknesses" of autistic performance, while also providing some insight into the late appearance of cognitive flexibility deficits in autism. Specifically, the absence of strong PFC representations early in development masks the problematic PFC/DA interactions. Early in development, both normally developing and autistic individuals face these tasks without much useful support from PFC, so the PFC gating mechanism matters little. As the model continues to develop, however, and the PFC representations strengthen, the weakened DA based gating mechanism manifests itself behaviorally as impaired cognitive flexibility.

Stimulus Overselectivity

Lacking Appropriate Context

Various contemporary theories hypothesize that people with autism possess deficits in integrating contextual information in an appropriate manner. Problems integrating information, it is argued, result in a processing style which highlights the specific pieces of the environment at the cost of more general high-level information (Happe, 1999b). This "piecemeal" style is capable of explaining an impressive variety of both the advantageous and detrimental behavior demonstrated by people with autism. One possible mechanism that could give rise to information integration difficulties would be a tendency to restrict attention to an excessively small number of features, with difficulty in shifting attention to other features. In this chapter, we explore this possibility using a simple computational model of the role of prefrontal cortex in attention.

Stimulus overselectivity, where a restricted set of components within a compound stimulus tend to dominate behavior, was first documented in the early 1970s in people with autism (Lovaas, Schreibman, Koegel, & Rehm, 1971). This effect has since been demonstrated both within and across different modalities as well as by varying the number of features composing the compound stimulus (Reed & Gibson, 2005). The paradigmatic task involves conditioning responses to a stimulus made of multiple, simultaneously presented, components. After the initial association of the compound stimulus with reward/response, each individual component is tested separately, assessing the degree each has acquired control over the subject's behavior. (See Figure **??**.) Normally developing individuals respond equally to all components. People with autism, to the contrary, are more likely to respond to a single component, thus demonstrating overselectivity. Overselective behavior in people with autism is a plausible explanation for the observed problems in generalizing learned behaviors to novel situations. In such situations, restricted, often irrelevant, portions of the environment become tightly coupled with the performance of the desired behavior. If this restricted portion is not consistently available to the individual, generalization to new settings will suffer. This is a major focus of many behavioral and intervention techniques.

Cognitive Flexibility & Overselectivity

My modeling work on executive dysfunction clearly demonstrates how impaired interactions between the mesolimbic dopamine (DA) system and the prefrontal cortex (PFC) can result in perseverative attention to an overly restricted set of stimulus dimensions and features, as exhibited by perseveration in the WCST (Kriete & Noelle, 2006). (See also Chapter 3.) Utilizing an abstraction of this same mechanism, my next computational model indicates flexible updating of PFC is needed to capture healthy performance on the stimulus overselectivity task. Importantly, introducing perseverative attention within the PFC control layer of my model (possibly enacted via perturbed DA/PFC interactions) results in a significant increase in stimulus overselectivity, and this provides a mechanism for the appearance of stimulus overselectivity in people with autism.

An intriguing recent study suggests that stimulus overselectivity can be induced in healthy individuals by requiring the concurrent performance of a working memory task (Reed & Gibson, 2005) with overselectivity training consisting of learning a multi-component stimulus to response mapping. Working memory tasks are widely believed to enlist the resources of PFC, providing additional support for the conjecture that healthy individuals utilize this area when performing this task.

An Overselectivity Task

An operant conditioning paradigm traditionally has been used when assessing the degree of overselective behavior in people with ASD. In this psychological task, a compound stimulus, consisting of multiple separate components, is associated with an action that leads directly to reward. The components can differ in modality (e.g. auditory, visual, and tactile) or be different stimuli within the same modality. After the initial stimulus / action / reward association has been learned, each individual component of the compound stimulus is presented separately in order to assess the degree each able to elicit a response. The severity of overselective behavior is measured by noting the number of the compound components capable of eliciting a response in isolation from the others. Overselectivity occurs when the number of components capable of driving a response in isolation is lower than the total number which comprise the compound stimuli. If an individual responds to all components equally, this indicates that attention has been distributed across all components during learning and no overselectivity is demonstrated.

Modeling Overselectivity

My investigations into overselectivity utilized a very simple artificial neural network model constructed using the biologically inspired Leabra framework (O'Reilly, 1996). In this simple model (see Figure 9), an input layer represents the components of the compound stimulus. It is useful to consider each unit of this layer as an individual component of a compound stimulus. For example, the first three units can be thought of as representing an auditory, visual, and tactile component respectively. To represent the compound, all three individual units are clamped to a high value simultaneously. The Hidden layer learns stimulus to response mappings, and provides a modeled abstraction of posterior brain systems. A response layer encodes the “decision” of the network based on information received from the computations performed within the network. The two possible outputs represented within the response layer of the network are “Respond” and “Do-Not-Respond”. Additionally, a PFC layer provides a top-down influence on processing within the “Hidden” (posterior) layer. The PFC layer has one extra unit which is utilized in the working memory load condition described below. (Note, however, that this working memory load unit is not shown in Figure 9.). The input layer also contains one extra unit, this unit can be thought of as representing a “No Stimulus” condition, analogous to when the participant is not being asked to respond to a stimulus object. Each unit in the PFC layer is associated with exactly one unit in the input layer. These input / PFC “pairs” project to a unique pool of hidden layer units, producing an isolated processing pathway for each stimulus component and its corresponding PFC unit. This enables each individual PFC unit to have selective influence upon a unique individual component of the compound stimulus. In other words, each hidden unit is directly modulated by only one of the three possible pathways. Note, however, that there is full recurrence between all of the hidden layer units. This does allow for one processing pathway to possibly influence the computations performed by another pathway. Additionally, the unit representing the working memory load selectively projects to another isolated pool of hidden units. These hidden units are also fully recurrent, and thus connected to all other hidden layer units as described previously. Only one difference exists when comparing the working memory load unit to the other modeled PFC units and the unique processing pathways of which each is a part. Namely, the hidden units projected to by the working memory load unit do not receive any projections from the input layer. This roughly corresponds to, and provides a way to model, the necessary lack of external stimuli during the performance of working memory tasks.

When modeling both normal and autistic performance, the network learned to “Respond” to the compound stimulus by simultaneously activating the three appropriate input units, representing an auditory, tactile, and visual component respectively. This is an extremely easy task for the network to learn, as it only needs to associate a response with the stimulus that is presented. This duplicates the simplicity of the original behavioral study by Lovaas et al. During the “healthy” normal condition, the PFC is allowed to switch between all three possible states, simulating a fully functional and flexible PFC. When modeling autism, however, only a single unit of the PFC layer was activated, and remained so throughout the entirety of training. Perseveration on the single PFC dimension simulates a deficit in flexibly updating the representations maintained within the PFC. This difference, the ability to flexibly adapt representations actively maintained by the PFC, is the only parameter manipulated from the model of normal performance in order to simulate autistic-like performance, all other parameters remained constant between models.

After the initial stimulus / response training with the compound stimulus, each component was presented individually to the network. The measure of interest is the number of individual components capable of correctly producing a “Respond” output from the network. The results for simulated autistic and control performance are average results for 100 separate runs of the model in each condition.

One additional condition was tested within the model framework just described. A recent study provides support for the role of PFC in overselective performance (Reed & Gibson, 2005). In this study, Reed et al. demonstrated that the addition of a working memory (WM) load, where additional items must be maintained and remembered at the end of a trial, actually leads to overselective responding in normally developing individuals. This is interesting because WM tasks are traditionally associated with processes localized within the frontal lobes, which we argue is a vital cortical area involved in the development of overselectivity in people with autism. In order to investigate whether my simple model can capture these results, an irrelevant additional item was maintained in the PFC layer during the initial learning phase, simulating a WM load. This was achieved by keeping one “extra” PFC unit constantly actively throughout the entirety of training, simulating maintenance of extra information with the PFC. All other parameters were identical to the "Normal" control condition. This includes the flexible updating of the PFC, which was allowed to flexibly switch between all three other items during training.

Overselectvity Simulation Results

My modeling efforts qualitatively match human performance and provide evidence that rapid and flexible updating of the PFC is necessary to prevent a restricted cue set from gaining control over behavior. (See Figure 10).

The model of autistic performance responded to significantly fewer components ($p < .05$), compared to the model allowed to flexibly update its PFC representations, demonstrating overselective behavior. Providing additional support for the hypothesis that the PFC influences learning in other cortical areas in interesting and important ways, a modeled WM load during training of a "healthy" network results in significantly more overselective responding ($p < .05$), also capturing recent behavioral results. Overselectivity arises in this model due an effect of PFC-directed attention on the learning of associations between stimulus features and the response output. When PFC activity is directed to a particular stimulus pathway, the activation levels of the hidden units of that pathway are increased. Lateral inhibition within the Hidden Layer, driven by this increased activity, reduces activity in the pathways corresponding to the other stimulus components. Thus, learning primarily takes place within the selected pathway, as synaptic plasticity is strongest in Leabra in the presence of presynaptic activity. In the autism network, which remains focused on a single pathway throughout training, the synaptic weights grow strong only within the selected pathway, with the connections in the other pathways remaining relatively weak. Thus, the autism network fails to learn to generate responses to the unattended stimulus features, even if attention is later directed to those features. In contrast, the healthy model flexibly adjusts PFC activity during training, allowing different pathways to be strengthened on different trials, eventually producing strong associations between all of the stimulus features and the need to respond. Leabra's Hebbian learning mechanism further ensures that these associations are robust.

(See Figure 11.)

Summary

The modeling results presented here suggest that, in people with autism, overselectivity may be driven largely by abnormalities in DA/PFC interactions, causing inflexibility in the shifting of top-down attention. When the PFC is unable to flexibly and appropriately update its contents, representations in cortical areas downstream from the PFC develop which are dominated by an overly restricted, or possibly even irrelevant, subset of features from the environment. Poor generalization occurs, under this account, due to the same abnormal cortical representations. The inability to flexibly update the PFC increases the likelihood that the only environmental associations that will be learned in a given situation will involve spurious correlations (e.g., idiosyncratic features of the training process), with other, more relevant, factors escaping attention. Subsequent dependence on such spurious correlations can cripple generalization performance. We suggest that learning over extended developmental timescales with this impairment may lead to behavior which looks like an integration problem on the surface, but, is actually just integrating a limited amount of information. The results presented here, coupled with the previous research on how DA/PFC impairments can explain executive dysfunction in autism, provide further support for a common neurological cause underlying a variety of behaviors observed in autism.

Implicit Learning

Implicit Learning Deficits in Autism

In addition to executive dysfunction and stimulus overselectivity, is it plausible that abnormal PFC/DA interactions can also account for the deficits in implicit learning observed in people with autism? Implicit learning is learning that occurs without any awareness of the specific knowledge acquired during the process. Researchers have suggested that people with autism have a core deficit in their ability to implicitly learn about the inherent relationships that exist between objects and situations in the world (Mostofsky, Goldberg, & Landa, 2000; Klinger, Klinger, & Pohlig, 2006). Klinger et al. argue that impaired implicit learning results in difficulties in recognizing the relationships that exist across experiences, likely leading to problems forming general knowledge about categories of items and types of situations. Difficulties in generalizing learned knowledge to new situations are commonly observed in people with autism, and these difficulties frequently act as a central obstacle to the development of behaviors needed for autonomy and independent living. Thus, a precise characterization of the mechanisms responsible for these generalization deficits would be very valuable to any effort to design ways to mitigate these serious issues in people with autism.

The Serial Response Time Task (SRTT)

Poor performance on tasks such as the learning of artificial grammars (Reber, 1967) and an apparent lack of learning during serial response time tasks (SRTT) have been used as evidence supporting implicit learning deficits in people with ASD (Mostofsky et al., 2000; Klinger et al., 2006). In the following we specifically investigate patterns of behavior for the SRTT.

In a common version of SRTT, participants are presented with four buttons, with exactly one button illuminated at any one time. Participants are asked to simply press the currently illuminated button as quickly and accurately as possible. Once a button is depressed, a new button is illuminated, prompting the participant to press the new button, and this sequence of cued button presses continues for a block of 80 responses, with an experimental session consisting of five of these blocks. The illumination order of the buttons is the key manipulation of the SRTT. During the first and the final (fifth) block the order in which the buttons are illuminated is random. However, during blocks 2, 3, and 4 there is a hidden pattern in the responses that are required. This hidden structure is apparently detected by many healthy participants, as there is a significant reduction in the reaction time required to press the correct button across blocks 2, 3, and 4. Importantly, this reduction in reaction time does not occur during the randomized first and fifth blocks. The common interpretation of these results is that learned knowledge of the hidden sequential pattern allows participants to better “anticipate” which button will be illuminated next, allowing them to prepare this upcoming action and, thereby, speed their response. Knowledge of the hidden structure is seen as “implicit”, however, as most participants claim no explicit knowledge of the sequential pattern (Cleeremans & McClelland, 1991).

People with autism, however, do not show marked improvement during the intermediate blocks of the SRTT, providing support to the claim that autism impairs implicit learning abilities (Mostofsky et al., 2000). While this behavioral result is interesting in its own right, we still do not have any sort of understanding of the underlying biological mechanism(s) behind this deficit.

Some insight might be gained from the neuropsychological literature involving the SRTT. Specifically, deficits in tasks assessing implicit learning have been linked to damage to the cerebellum. This is intriguing, as there is ample evidence of cerebellar abnormalities in people with autism (Courchesne et al., 1994). However, other tasks traditionally associated with the cerebellum, such as judgement of timing, show no differences between people with autism and normally developing controls (Mostofsky et al., 2000). Recently, evidence has emerged suggesting that PFC and the basal ganglia may be important players in implicit learning as well (Matsumoto et al., 1999; Pascual-Leone, Wassermann, Grafman, & Hallett, 2004). It is this latter connection that we will pursue, here, using an established computational model of the SRTT to investigate the possibility that PFC/DA abnormalities may give rise to the implicit learning problems observed in people with autism.

Modeling Implicit Learning

Modeling SRTT Performance. Seminal work on modeling healthy SRTT performance has been conducted by Cleeremans et al. (1991). In these neurocomputational models, a simulated neural circuit is presented with an input that encodes the currently illuminated button, and the output of this circuit is read as the system's expectation for the next button to be illuminated. This input to output mapping is mediated by a collection of hidden units, and synaptic learning methods are used to improve this mapping with experience. Importantly, these networks also include a "context layer" of neural units which can learn to actively maintain information about the history of previously presented inputs, allowing the model to base its predictions on more than the currently illuminated button. The activation levels of the neural units in the context layer are set to be "copies" of the hidden unit activation levels whenever a new input is presented, making the Cleeremans model of SRTT performance essentially a simple recurrent network (SRN) (Elman, 1990) trained to predict the next button press. (See Figure 12.) The schematic network architecture is shown in Figure 13. This model has provided good fits to healthy human performance on the SRTT (Cleeremans & McClelland, 1991).

Since the hidden sequential structure in the intermediate blocks of the SRTT is often complex, the information provided by the context layer is vital for the success of the model. Importantly, the context layer in this model plays an identical functional role to the PFC in other models, actively maintaining information that can be used to modulate an input-output mapping. In our previous executive dysfunction model, the PFC actively maintained information about the currently relevant stimulus dimension (e.g., "focus on the ink color" in Stroop or "sort cards based on shape" in WCST), so as to modulate performance. In our stimulus overselectivity model, the PFC actively maintained information about the current stimulus feature to receive executive attention, modulating how the stimulus was processed. In this model, the context layer actively maintains information about the preceding button presses, allowing that information to modulate the prediction of the next button. The main difference between our previous PFC models and this SRTT model involves the timing with which the contents are updated. In our previous models, the PFC was updated in a dynamic fashion, based on learned task contingencies. In this model, the context layer is updated with each new input presentation. Thus, the SRN context layer is analogous to the PFC in our previous models, with the updating "gate" forced to open with each new input, as previously suggested by other researchers (O'Reilly & Frank, 2006).

It is important to note that, in order to capture the relevant sequential information, the SRN 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 expect to capture the performance of people with autism.

Model Modifications. We made small modifications to a previous implementation of the Cleermans SRTT model which uses the biologically grounded Leabra framework, reducing the original implementation's 10-unit inputs and outputs to only 4, to capture the structure of the 4-button SRTT (O'Reilly & Munakata, 2000). The resulting network is shown in Figure 13. In this model, an Input Layer represents the four distinct buttons. The Hidden Layer learns the prediction mapping and provides a modeled abstraction of posterior brain systems. A Response Layer encodes the prediction output of the network. Additionally, a Context Layer provides a top-down influence on processing within the Hidden Layer, reflecting the role of PFC.

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. Normally, the Context Layer is updated with each input, but the autism model only updated its Context Layer with some fixed probability which was less than one. This 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 normally contained within this layer much less reliable, making the learning of complex sequential structures much more difficult for the network.

Computing Reaction Time in a SRN. The measure of interest in the SRTT is the response time of the participants throughout the training blocks. In order to compare model performance to human reaction times, Cleeremans et al. (1991) translated network outputs into a probability distribution over the four buttons using a Luce choice ratio (Luce, 1963) and then linearly scaled the error between this prediction distribution and the actual outcome (i.e., the next button illuminated) to produce a modeled response time. This procedure assumes that there is a linear increase in response time with prediction error. We used this method, as well, introducing three free parameters for fitting the model to data: a linear scaling parameter from error to milliseconds, a base response time (when error is zero) for the healthy model, and a base response time for the autism model. Note that different base response times were used for the normally developing and autism cases in order to capture the difficulty people with autism regularly exhibit when initiating movements (Rinehart et al., 2001).

Implicit Learning Simulation Results

Network simulations were repeated 100 times in each of the experimental conditions, with initial synaptic weights randomized for each repetition. Average performance results for each block were compared to previously reported response time data for both people with autism and normally developing controls (Mostofsky et al., 2000). A grid search was performed over the possible probabilities of updating the Context Layer, testing probabilities from 0.0 to 1.0 in steps of 0.1, with the three linear scaling parameters optimized to reduce sum-squared deviation from the human response time data. The updating probability, and associated scaling parameters, that produced the lowest sum-squared deviation from the human data was identified as the best fit model.

The simulation results match human performance both qualitatively and quantitatively, providing evidence that impairments in PFC updating can result in implicit learning deficits like those seen in people with autism. When the healthy network is restricted to perfectly update its Context Layer (i.e., with probability one), the corresponding best-fit probability for the autism network is 0.5, with an SSE of 652.[1] The corresponding scaling parameter from error to reaction time is 261.4, the healthy base time is 458.6 msec, and the autism base time is 534.5 msec. The resulting modeled reaction times, along with human data from the literature, is shown in Figure 14.

A repeated measures ANOVA was conducted on blocks 1, 2, 3, and 4 of the model results, and a significant Group by Block interaction ($F(3, 97) = 62.007$; $p < 0.000001$) was detected. From these results we can conclude that the networks simulating autistic performance demonstrated significantly less learning over the crucial training blocks of 2, 3, and 4, as compared to the networks allowed to properly update their PFC-like Context Layers. Thus, clear implicit learning deficits were present in the autism model.

Summary

The modeling results presented in this chapter suggest that, in people with autism, implicit learning deficits may be driven largely by abnormalities in DA/PFC interactions, causing inflexibility in the updating of contextual information. Without the proper updating of actively maintained contextual information, it is essentially impossible to properly integrate temporally separated pieces of information, such as the order of items in a sequence. Thus, our computational account highlights how PFC/DA dysfunction can lead to problems with information integration. This is particularly interesting, since one prominent behavioral theory of autism, *Weak Central Coherence*, posits that deficits in integrating contextual information lay at the core of this disorder (Happe, 1999b).

Lexical Disambiguation

Using Sentential Context

We have presented examples of how inflexible attentional modulation of posterior brain areas by the PFC, caused by deficient DA / PFC interactions, can account for an increasingly diverse set of behaviors including executive dysfunction, overselective behavior, and implicit learning differences in people with ASD. In the previous chapter on implicit learning in autism, we showed how this deficit can lead to problems integrating information over temporally extended time frames. One particularly relevant situation that requires the ability to integrate information across experiences is our ability to understand the proper meaning of ambiguous words in a sentence. Homographs are words with one spelling, but possessing different pronunciations and meanings, such as “bow” and “tear”. In order to determine the proper pronunciation (and meaning) of a homograph, we must rely on the preceding sentential context. For instance, the word “tear” is ambiguous until it is used in the sentence, “The incredibly moving and sad story brought a tear to my eye”. Only after observing the word “tear” in the context provided by the sentence are we able to determine the proper meaning and pronunciation. Pronouncing homographs is a task used by proponents of WCC as evidence supporting the conjecture that people with autism process the pieces of situations at the cost of more global “gestalt”-like information. People with autism appear unable to properly utilize this context in order to correctly pronounce homographs. Instead, they will rely on the statistically most frequent pronunciation (Jolliffe & Baron-Cohen, 1997; Happe, 1997).

Contemporary connectionist models of sentence processing utilize an information processing mechanism termed a “context layer”. As described in Chapter 5, the context layer provides a means to integrate past information, such as the previous words read, into an evolving representation of a sentence. The information provided by the context layer is vital for the success of the model on tasks such as homograph pronunciation. A simple recurrent network (SRN) is most often used as the basic framework for such modeling endeavors (Elman, 1990). It is important to note that, in order to provide the appropriate contextual information, the SRN must update the context layer in a fast and appropriate manner. The flexible updating of contextual information is precisely the cognitive mechanism we hypothesize to be suspect in people with autism.

Lexical Ambiguity in Schizophrenia

A possible clinical comparison group for lexical disambiguation performance in people with autism can be found in people with schizophrenia. It is known that schizophrenics demonstrate problems utilizing context across a number of tasks, including language disambiguation tasks (J. D. Cohen & Servan-Schrieber, 1992). This is a very interesting comparison for a two reasons:

1. While people with schizophrenia demonstrably possess problems utilizing context when disambiguating words in natural language processing, the specific pattern of deficits is qualitatively different than the same behavior observed in people with autism, as discussed in more detail below.

2. Many of the symptoms of schizophrenia are believed to arise due to abnormal dopamine functioning (J. D. Cohen & Servan-Schrieber, 1992).

If, as would be consistent with my theory, schizophrenia is believed to have similar underlying deficits to autism, namely a dopamine abnormality, how can the two disorders produce qualitatively different deficits on this psychological task? The answer lies in the details of the *kind* of dopamine deficit that is posited by theorists in both disorders. Dopamine is believed to have at least two different kinds of effects on cortical processing. The first, known as “tonic”, is believed to enact its effects over a relatively long time scale. The second, known as “phasic”, is argued to have rapid influence on cortical processing. According to Cohen and Servan-Schreiber (1992), abnormal dopamine functioning in schizophrenia is associated with the slower effects of tonic DA. However, the precise firing and timing of the mesolimbic dopamine cells that inspire the TD-Learning based account upon which my theory is based, are of the second type, phasic dopamine. In the following we explore, computationally, how the inability to properly update the PFC can explain the psychological profile of people with autism on lexical disambiguation tasks, as well as how different kinds of DA processing may result in the qualitatively different, but still deficient, behavioral profiles of schizophrenia and autism.

Differences Disambiguating Context in Schizophrenia and ASD

The same general psychological test has been used to test people with schizophrenia and ASD on their ability to utilize context when disambiguating the meaning of words (J. D. Cohen &

Servan-Schrieber, 1992; Jolliffe & Baron-Cohen, 1997; Happe, 1997), the disambiguation of homonyms. One difference to note is that homographs were used in the ASD population, whereas homonyms were used in the assessment of schizophrenics. Homonyms are homographs which are pronounced identically, but still possess different meanings based on the context in which they are utilized. Since context is vital for disambiguation in both homonyms and homographs, they were modeled identically, essentially treating the phonological representation as unimportant for the question at hand. The ambiguous homographs have two different interpretations, one is the most common or high frequency interpretation, and the other is the least common, or low frequency version. In this task, a sentence fragment is presented which contains one ambiguous word (e.g. "without a PEN") along with a sentence fragment which can disambiguate the meaning of the ambiguous word (e.g. "You can't keep your chickens"). The sentence fragments are created to enable the contextually relevant information to be presented before or after the ambiguous word.

The sentences are distributed across 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 (4) The high-frequency meaning is correct and the context comes last.

People with autism demonstrate problems utilizing context in both conditions (1) and (2), providing a significantly higher percentage of incorrect high-frequency responses compared with normally developing control groups (Jolliffe & Baron-Cohen, 1997; Happe, 1997). (See Figure 15.) The autism group was not significantly different than controls on conditions (3) and (4), showing no obvious bias in determining the meaning of high-frequency words. The profile for schizophrenics is slightly different. People with schizophrenia do show an impairment in utilizing context, but only during condition (2), when the context is presented first. (See Figure 16.) Theorists have interpreted this result as evidence for problems using temporally extended context in schizophrenia. To clarify, this suggests that in condition (1), the contextual information was temporally close enough to the required response to enable the subjects diagnosed with schizophrenia to correctly utilize this information and make a proper response. Condition (4) was not tested in the schizophrenic population, and therefore no human data is available.

Modeling Lexical Disambiguation Effects in Autism

A Model of Word Sense Disambiguation. The connectionist model utilized in Cohen and Servan-Schreiber (1992) was modified to investigate lexical disambiguity performance in people with autism and schizophrenia. (See Figure 17.) A localist code is used to represent words as inputs to the network. For instance, one unit each represents the ambiguous word “PEN”, one unit for “BANK”, etc. The input layer also contains words that are used as the contextual cues to assist the network in the disambiguation task. The output of the network (the “Meaning Output Module” in Figure 17), also uses a localist code to represent the various possible meanings of the network. Specifically, one unit is used to represent the interpretation of the word “BANK” as a “financial Institution”, and one unit for the alternative meaning “area next to a river”, and so forth. The context layer (labeled “Discourse Module” in Figure 17), was modified from its original version. The original “Discourse Module” uses hand coded representations to encode sentential context information, meaning the very important process of learning representations which facilitate the integration of previous experience is hard wired. we replaced the original “Discourse Module” version with a SRN layer, providing a means for the network to learn how to integrate temporally extended information through repeated experience. (See Figure 18.) In addition to providing contextual information to the “Semantic Module”, the SRN layer is also locally recurrent (Jordan, 1986). This helps to encourage a sentential, as opposed to single word, representation in the context layer, mimicking the original form of the “Discourse Module”.

Training. The original training procedures utilized in Cohen and Servan-Schreiber (1992) was modified in this computational exploration of lexical ambiguity. This is a necessary change due to the architectural switch from hand-coded context representations to those learned by an SRN as described previously.

The corpus used for training involved 50 input units representing 50 ambiguous homographs. Also used as input to the network was 50 disambiguating context units. Two context units were assigned to each of the 50 homograph units, resulting in each homograph unit being paired with two different context units. Also, each context unit is paired with two different homograph units, ensuring that every context unit is involved in two completely different disambiguation trials. Setting up the input corpus in this manner prevents the network from learning a one-to-one association of a context unit to a specific homograph word meaning, removing any possibility that the network could solve the task while ignoring all of the homograph units, instead relying only on the information provided by the contextual units to disambiguate the meaning of a homograph. One context unit was used for the more frequent use (strong) and one unit for the less frequent (weak) use of each homograph. This will results in 150 possible outputs for the network (100 possible homograph interpretations and 50 context words used for disambiguation of the homographs). Each possible output can be thought of as representing the proper "meaning" input words presented to the network. (See Figure 18.) The network also used 100 units each within the Hidden layer and the Context Layer, it should be noted, however, that varying this value had little effect on the networks overall performance.

One word (input unit) is presented to the network at a time, requiring the network to make a response to every word independent of whether the input is supplied via a contextual or homograph unit. On each trial, the network is presented with a mini-clause consisting of three words, including a context/homograph pair followed by a homograph probe. The pieces of each mini-clause should be thought of as representing the sentences presented to the subjects in the original study (the context and homograph units), followed by questioning the subjects on the meaning of the ambiguous word (the probe homograph unit). The order of presentation for the context and homograph units were counterbalanced across all trials, ensuring equal experience with utilizing context at both the beginning and the end of mini-clauses. The final probe homograph unit was always presented last, and was needed to assess if the network could properly disambiguate the meaning of the homograph after reading the mini-clause. The network was trained to respond with the correct "meaning" for both input types, homograph and context units, by activating a single unit within the output layer which represents the proper semantic interpretation of the input word presented to the network. (See Figure 18.) Supervised error-correction learning, as implemented in the Leabra modeling framework, was used in order to teach the model to respond appropriately. The network was presented with the strong interpretation of the homographs on 70% of the trials and the weak interpretation on 30%, a gross approximation of the differential exposure humans receive with common and uncommon meanings of homographs.

Testing. The same testing procedures as Cohen and Servan-Schreiber (1992) were utilized. The model was presented with a context and ambiguous word unit pair, one unit at a time. The order of presentation was again counterbalanced across all testing trials. After the presentation of the pair, the model was presented with the probe ambiguous homograph word unit. Note that these are the same "mini-clauses" that were utilized throughout the training procedure. The main measure of interest is percentage of incorrect uses of the strong interpretation by the model during the presentation of the probe homograph unit, when the weak interpretation is correct based on the context provided in the mini-clause. The results were separated into two groups, "Context Presented First" and "Context Presented Last" to enable comparison to the human subjects data.

Modeling Schizophrenia. In the original model of schizophrenic performance of Cohen & Servan-Schreiber, a hypothesized tonic DA deficit was instantiated by reducing the gain of the activation function on all modeled PFC units. Figure 19 shows the overall effect of this manipulation. Coupled with the dynamics of the network, the result was a less stable PFC representation of the contextual information. Thus, if the context were to come at a point temporally distant from homograph "probe", the contextual information maintained within the modeled PFC was more likely to degrade and be of little use to the network. However, if the context was presented temporally close to the"probe", the contextual information maintained within the PFC could still be used to disambiguate the meaning. This manipulation successfully matched the pattern of behavior seen in schizophrenia. Functionally analogous to the gain manipulation of Cohen & Servan-Schreiber, we systematically reduced the ability of the Context Layer (PFC) to hold on to information over an extended period of time in order to model the performance of people with schizophrenia. Under normal circumstances, the Context Layer has the ability to effectively hold onto previous information over multiple time steps. In a standard SRN, a complete copy of the previous time steps Hidden Layer activity pattern is copied directly into the Context Layer upon every time step. However, within the Leabra framework there are two mixing parameters responsible for allocating what percentage of the activation is a pure copy of the previous Hidden Layer activity and what percentage is maintained from the previous state of the Context Layer itself. These parameters can be seen as manipulating how well the network updates the contents of the PFC and how well the network can actively maintain this information within the PFC respectively. By systematically reducing the parameter that controls the amount of previous Context Layer activity retained from time step to time step, we can functionally reduce the stability of the contextual representations maintained within the modeled PFC. This manipulation provides an extremely close approximation to the gain reduction performed by Cohen & Servan-Schreiber. Also, schizophrenia most often emerges after a significant amount of development has occurred, therefore this deficit was only instantiated after the network had been trained to criterion. In the healthy model, the context maintenance parameter is set to ensure that 30% percent of the previous Context Layer's activation was retained from the previous time step. In order to investigate schizophrenic behavior we reduced this parameter to 20%, 10%, and 0%, systematically destabilizing the modeled PFC Layer's maintained pattern of activity.

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. Normally, the Context Layer is updated with each input, but the model of autistic performance only updated the layer with some fixed probability which was less than one. Specifically, we investigated successful PFC update probabilities of 100% (control), 90% and 80%. Restricting the ability of the PFC to update its contents is analogous to reducing the efficacy of the DA-based gating signal to the PFC, and is the exact same manipulation previously utilized and argued for in Chapter 5. Preventing the updating of the PFC in this manner causes the temporally extended information normally contained within this layer to be much less reliable, making the learning of the sentential context difficult for the model to accomplish. Without a reliable contextual representation of the previously presented information, the model will be forced to rely heavily the statistical frequency of the input corpus in order to learn the task. In other words, in lieu of reliable contextual information, the most common interpretation of the homographs will be the most reliable indicator that the model can utilize in order to maximize it's performance. Importantly, autism is believed to be a developmental disorder, with key neural differences likely present from very early on or even since birth. To capture this fact, we restricted the ability of the modeled PFC layer to update throughout the entirety of training, simulating the deficit existing throughout key developmental periods.

Lexical Dysambiguation Simulation Results

Overview. Network simulations were repeated 10 times in each of the experimental conditions, with initial synaptic weights randomized for each repetition. Each network experienced the entire training corpus 20 times. The training corpus contains every potential context word / ambiguous homograph word pair, for a total of 200 different "mini-clauses", 100 with the contextual word presented first and 100 with the homograph occurring first in the clause. Each of the "mini-clauses" are presented in a random order during each of the 20 blocks of training. Lateral inhibition within the Output layer of network strongly encouraged the network to respond with a single "word". The unit with the highest overall activation level was used as the model's response. The key error measure across all conditions is the number of errors committed when assigning meaning to the ambiguous probe word and is visually presented in Figures 20 and 21.

The higher this measure, the worse the models performed on the disambiguation task.

Schizophrenia Model Results. The model of homonym reading in schizophrenia was able to qualitatively match the behavioral data and reproduce the seminal modeling results of Cohen & Servan-Schreiber (1992). (See Figure 20.) The "Control" network performed well regardless of the frequency of the meaning word ("Rare" vs."Common"). Also, the ordering of contextual information had virtually no effect on model performance. The model performed well regardless of if the context was presented early in the mini-clause (Context-First) or if it came later (Ambiguous-First).[2] However, as the PFC representations were systematically destabilized, as is hypothesized to occur in schizophrenia due to tonic-DA dysfunction, the error rate rose significantly. Importantly, this decrease in task performance only occurred when the context was presented first, but not when the context was presented last. (See Figure 20.) Two-way ANOVAs were performed separately on each destabilization level – 0% Destabilized (Control), 33% Destabilized, 66% Destabilized, and 100% Destabilized –, with frequency ("Rare" vs. "Common") and ordering ("Context First" vs. "Meaning First") as factors. An effect of frequency was observed as the PFC was destabilized 33%, 66%, and 100% (respectively, F(1,36) = 18.29, $p < .05$; F(1,36) = 57.48, $p < .05$; F(1,36) = 78.98, $p < .05$). Also, an effect of ordering was observed for each condition (respectively, F(1,36) = 14.97, $p < .05$; F(1,36) = 77.72, $p < .05$; F(1,36) = 109.42, $p < .05$). Importantly, a there was a significant frequency by ordering interaction effect in all cases where the PFC was destabilized (respectively, F(1,36) = 11.99, $p < .05$; F(1,36) = 54.84, $p < .05$; F(1,36) = 73.49 $p < .05$). This supports my observation that only the context first error rate rose significantly as a result of PFC destabilization. No effects were found in the control network results for either frequency F(1,36) = .64, $p > .05$; or ordering F(1,36) = 3.47, $p > .05$, and the frequency by ordering interaction was also non-significant F(1,36) = 1.77, $p > .05$. This matches patterns of behavior reported in previous schizophrenia research (J. D. Cohen & Servan-Schrieber, 1992), which demonstrated that people with schizophrenia have difficulty utilizing context when it is presented temporally distal from the homograph it is intended to disambiguate.

Autism Model Results. By restricting the ability of the Context Layer to update in a flexible manner, we was able to capture patterns of behavior consistent with previous findings of lexical disambiguation research in people with ASD (Jolliffe & Baron-Cohen, 1997; Happe, 1997). (See Figure 21.) Specifically, as the probability of updating the content of the modeled PFC layer was systematically reduced, the network became increasingly more reliant on the statistical frequency of the words. Behaviorally this was manifested in a significantly higher error rate when the model needed to utilize contextual information, *regardless of the temporal distance from the required response*, in order to correctly interpret the "Rare" meaning of the homograph in the sentence. Two-way ANOVAs were performed separately on each level of PFC updating – 100% probability to update PFC (Control), 90% probability to update PFC, and 80% probability to update PFC – with frequency ("Rare" vs. "Common") and ordering ("Context First" vs. "Meaning First") as factors. An effect of frequency was observed in the "autistic-like" condition, where the PFC is restricted to update only 90% and 80% of the time (respectively, F(1,36) = 64.52, p < .05; F(1,36)=116.58, p < .05). However, an effect of ordering was not observed (respectively, F(1,36) = 0.35, p > .05; F(1,36) = 0.31, p > .05). Also, the frequency by ordering interaction was not significant in either case (respectively, F(1,36) = 0.87, p ..05; F(1,36) = 0.31, p > .05). These results suggest that the model of autistic performance has difficulty utilizing contextual information to overcome a more frequent interpretation of a homograph, regardless of when it is presented in the sentence. No effects were found in the control network results for either frequency F(1,36) = .64, p > .05; or ordering F(1,36) = 3.47, p > .05, and the frequency by ordering interaction was also non-significant F(1,36) = 1.77, p > .05. It is important to note that as the probability of updating the Context Layer was reduced, the model of autistic behavior performed worse overall when compared to the control network, including common interpretations of the homographs. This pattern was not reported in the behavioral study of people with ASD. However, the extremely low number of sentences in each condition (5) may be resulting in ceiling effects, limiting the ability to reliably detect this difference in their sample. (See Figure 15.)

Summary

In this chapter modeling results were presented demonstrating how a specific neural difference, dysfunctional interactions between the mid-bran DA system and PFC, can capture patterns of dysfunction in people with autism when utilizing sentential context to disambiguate the meaning of homographs. My work suggests that improper DA / PFC interactions lead to problems flexibly updating the contents of the PFC, and difficulties arise due to the lack of reliable contextual representations. This causes the model to rely more heavily on statistical regularities in the environment. In other words, the model adopts a policy of responding with the most common interpretation of a homograph in order to minimize errors, in lieu of appropriate contextual guidance from the Context Layer. A previously published and theoretically justified computational model of lexical disambiguation was modified in a manner consistent with my hypothesis in order to capture the behavior of people with autism (J. D. Cohen & Servan-Schrieber, 1992). Interestingly, the Cohen & Servan-Schreiber model was originally developed as an investigation into lexical disambiguation difficulties in people with Schizophrenia. This provided my investigation with an interesting clinical comparison group, especially considering the qualitatively different behavioral profiles of the two disorders coupled with the suggestion of DA dysfunction underlying each respective pattern. By positing the disorders actually have different *kinds of DA dysfunction*, namely tonic DA dysfunction in schizophrenia and phasic DA abnormalities in ASD, the model is capable of explaining both the original findings of Cohen & Servan-Schreiber and the patterns of behavior found in people with ASD (Jolliffe & Baron-Cohen, 1997; Happe, 1997).

Prototype Formation

Category Learning in Autism

We have shown how the overly perseverative top-down attentional influence of the PFC could underlie an extensive array of behaviors in people with ASD. In this last chapter before concluding and summarizing my work, we investigate if differences in concept formation in people with ASD can be included in the list of behaviors encompassed by my theory.

Prototype formation is an important skill and is believed to be invaluable in learning to categorize information in our environment. The prototype for a specific category is a representational average of all the features of examples previously seen for that category. The ability to learn an abstract representation of a category can greatly reduce the memory demands on an individual. Recent studies suggest when learning a new concept or category, children with autism are less likely to form and use a prototype representation when compared to normally developing individuals (Klinger & Dawson, 2001; Gastgeb, Rump, Best, Minshew, & Strauss, 2009). Difficulty in forming and properly using an abstracted prototype is argued by some researchers to underly many of the generalization problems found in people with ASD. Both of the aforementioned studies investigated the ability to form a prototype via a phenomena known as the “prototype effect”. The “prototype effect” is a phenomena in which the prototype is preferably remembered as a member of the category when compared to other possible examples, even though it was not experienced previously. The results of these studies suggest that people with ASD do not demonstrate the “prototype effect” during a standard category learning task. My modeling efforts will concentrate on the study by Klinger and Dawson (2001), which we describe next.

Investigating Prototype Formation

In order to investigate prototype formation in people with ASD, Klinger and Dawson (2001) had subjects perform a basic categorization task. There were two phases to the experiment. The first phase was a familiarization task in which subjects were exposed to a variety of examples of imaginary animals. This phase was used to teach the subjects a single animal category. (See Figure 22.) The second phase was a testing phase requiring the subjects to identify the animal which was the best example of the imaginary category that they had learned. Each imaginary animal belonged to a specific category, and members of each category possessed four category-specific features of interest, each of which could have a range of different discrete values.

For instance, an item belonging to the “Mip” category would display horns, wings, mouths, and feet, with different Mips having different horn widths, wing widths, mouth widths, and foot lengths. Each feature ranged in five discrete steps from smallest to largest values. For example, the horn widths smallest value could be 2.5 cm and each subsequent length increasing by a constant .50 cm resulting in five discrete lengths ranging from 2.5 cm to 4.5 cm. The smallest length for each feature was labeled with a value of 1, the second longest was labeled with a value of 2, up to the largest length which was labeled with a value of 5 for coding purposes. This coding scheme allows for the four digit encoding of individual stimuli, with the different numbers representing the respective values of the features they represent (e.g. “5115” where each number corresponds to a value of a distinct feature of the imaginary animal).

During the familiarization phase subjects were shown eight examples of a category compared to eight non-member examples (e.g. Mip vs. Pev; Mip vs. Dak; etc.). On the first familiarization trial, subjects were given the name of the target category (e.g. “This is a Mip”), and asked to identify the target in all subsequent trials. Target categories were able to be identified by simply noting the presence of features specific to each respective category. Specifically, it was not necessary to attend to the different feature values for correct categorization. For instance, an instance of the “Mip” category has horns and an octangular body, no other imaginary animal category possessed these features. Feedback was provided on each trial, providing praise if the correct category was selected, and correcting the response otherwise. Familiarization stimuli were constructed such that the children would see target category examples that contained feature values 1, 2, 4, and 5, but never 3. (See Figure 23.) After all familiarization trials, the average value used for each feature in the target category was 3, resulting in an average prototype for the category possessing the feature values of 3333 (all average lengths and widths of the possible four features in the target category).

Testing immediately followed the familiarization phase. During testing, a prototypical example for the category (3333) was paired with a previously seen familiarization example (e.g.

5511) or a novel stimulus comprised of previously experienced feature values (e.g. 1551). Subjects were asked to select which example was the "best" category member. Normally developing individuals choose the prototype at a rate significantly higher than chance. However, people with autism selected the prototype exemplar at chance levels when comparing to non-prototype stimuli (e.g. the novel or familiar stimuli). (See Figure 24.) The authors argued that this effect demonstrates a lack of prototype abstraction in people with autism.

Modeling Category Learning with ALCOVE

ALCOVE is a formal computational model of human performance in category learning (Kruschke, 1992) and has been highly successful at capturing human category learning performance when classification feedback on example category members is provided. ALCOVE will learn, after repeated exposure to examples and explicit feedback, what stimulus dimensions are important for correct categorization of the examples. There are three main processing layers used in ALCOVE, an input layer with units representing stimulus features in psychological space. The hidden layer contains units that roughly represent the "location" of stimuli in psychological space. The activation profile of the hidden layer is similar to an exponentially decaying radial basis function, the closer in psychological space an input vector presented to ALCOVE is to the "location" of a hidden layer unit, the higher the activation level of the unit. The level of hidden unit activation decays in an exponential fashion as a function of the distance between the location of the hidden unit and the input to the network. The output units are standard linear units summing the activation across weighted connections from the hidden layer units. It is useful to view these units as representing different category labels (e.g. "Mip"). In order to match human performance, the activation levels of the output units are scaled to response probabilities using a Luce Choice Rule:

$$P(B) = exp(\phi a_B)/\sum(exp(\phi a_b)) \tag{4}$$

$P(B)$ computes the probability of selecting the category B for a stimulus being presented to the network, a_B is the activation level of the output unit representing category B, $\sum(exp(\phi a_b))$ is the sum over the exponential value of all of the output activations, and ϕ is a gain parameter used to scale the results.

Importantly, for our purposes, ALCOVE utilizes a set of dimensional attention "weights", or parameters, when learning to categorize stimulus objects. These parameters are constrained to the range between 0 and 1, and, conceptually, as the parameter value increases the more sensitive the model is to changes along this feature dimension. ALCOVE's dimensional attention weights capture the degree to which attention should be spread over stimulus features and, thus, are analogous to the hypothesized role of PFC in directing top-down attentional control. In the ALCOVE framework, perseverative attention would limit dimensional attention weights, resulting in an overweighting of a small number of the available features. As described below, limiting ALCOVE's dimensional attention weights in this way allows us to capture the performance of individuals with autism as documented by Klinger and Dawson (2001).

Prototype Learning in ALCOVE

The training and testing of ALCOVE was modeled after the Klinger et al. study described above. The training corpus was based on examples identical, in principal, to the target categories used during the familiarization trials. (See Figure 23.) As in the original study, eight familiarization trials were presented to the network during training. On each presentation, the model was required to label each target stimulus (e.g. this is a "Mip"). Only two response units ("Mip" & "Non-Mip"), were required in order for the network to perform the task. Four input units, each with dimensional attention weights, were activated for each example presented to the network. The four input units represent the four unique varying features used within each category in the behavioral study. The level of activation for each input varied in a systematic fashion in accordance with the feature value of the stimulus. For instance, if the stimuli presented has feature values (1551), the input vector presented to the network used the values (.20 1.0 1.0 .20). The same eight familiarization stimuli as were used in the Klinger et al. study were presented to the ALCOVE network, simulating the same learning environment as the subjects experienced. The standard ALCOVE error correction weight adjustments were used during the entirety of training. After training ALCOVE on the familiarization stimuli, testing proceeded with all learning disabled in the model.

During testing, the prototype (3333) and the non-prototype test items, a novel stimuli (e.g. 1551) and a previously seen familiarization stimuli (e.g. 1515), were presented to ALCOVE separately. The overall activation level of the output unit for the target category was recorded and used to determine which stimuli type was the "preferred" category example for ALCOVE. In other words, the higher the activation of the output unit corresponding to the correct category, the stronger the "vote" for the example presented to the network. The model always correctly categorized each stimuli regardless if it was a prototype, novel, or previously seen example. In order to compare model preference between prototype and non-prototype stimuli the activation levels were scaled to response probabilities using a Luce Choice Rule. (See Equation 4.). The average response probability for prototype versus non-prototype stimuli is the measure used in the comparison to the actual experiment results.

Modeling Autism in ALCOVE

In order to simulate autistic performance we artificially biased one randomly selected attentional dimension by strengthening the weighting of that dimension's attentional weight in ALCOVE. This was done by simply hand coding one of the attentional weights at an extremely high value (.9). The other dimensional attention weights were all initialized to low values (.1). This manipulation captures, at a gross level, my conjecture that deficient dopamine based gating of PFC representations will result in overly perseverative attention to a restricted set of features or dimensions. In the control version of the model, the dimensional attention weights were spread out evenly and all initially set to a low level (.1), as is standard in ALCOVE, allowing the network to allocate its attention in a more appropriate manner. These parameters were adjusted by the standard ALCOVE learning rule for dimensional attention weights throughout the training process in both the control and autistic-like conditions. One additional ALCOVE parameter was adjusted in order to better simulate the performance of people with autism on the categorization task. The standard ALCOVE equation for the activation levels of the exponentially decaying hidden units contains a multiplicative scaling parameter referred to by Kruschke as "specificity". As this parameter increases, the activation profile of the hidden unit becomes more narrow, resulting in a smaller range of feature values that are capable of highly activating the unit. In other words, increasing the specificity parameter makes the hidden units in ALCOVE hyper-specific, sensitive to a more-restrictive range of features in the environment. This is interesting because many researchers have argued that people with autism exhibit hyper-specific behavior in their everyday lives (McClelland, 2000; Happe, 1999b). In the results described next, the specificity parameter in ALCOVE was adjusted from 1.0 in the control network to 2.5 for the network modeling behavior of people with ASD. The manipulation of the specificity parameter in ALCOVE has been previously used and justified to capture performance of a clinical population during the investigation of categorization behavior in amnesic patients (Nosofsky & Zaki, 1998).

Prototype Formation Simulation Results

Network simulations were repeated 80 times in each of the experimental conditions, with each repetition treated as an individual subject for data analysis. The control network was able to reproduce the prototype effect as observed in normally developing individuals, choosing the prototype over the non-prototype 70.52% of the time. (See Figure 26.) For consistency we used to the same analysis as Klinger et al. (2001), a one-sample T-test, in order to determine if the model's preference for the prototype over non-prototype stimuli was statistically reliably different from chance (50%). The analysis of the control model's performance indicated that the prototype effect was indeed significantly different, and larger, than chance (T-Value: 4.05; $p < .0001$). Looking at the performance of the model of autistic behavior, however, we see a very different profile. (See

Figure 26.) The ASD network preferred the prototype over the non-prototype stimuli only 52.91% of the time according to our measure, and statistical analysis confirms that this is not significantly larger than chance (T-Value: 0.52; $p > .30$). This matches the lack of a prototype effect in people with autism found in previous behavioral studies (Klinger & Dawson, 2001; Gastgeb et al., 2009).

Summary

The results presented here support the feasibility of explaining the categorization performance of people with autism in terms of my hypothesized deficit, dysfunctional interactions between the mid-bran DA system and PFC. In this chapter, the widely used ALCOVE model was modified in order to simulate how inflexible updating of the contents of PFC would manifest within ALCOVE's framework. Restricting the ability of ALCOVE's global attentional mechanism to spread its influence evenly over all features of a stimulus results in a lack of prototype preference in the network. This manipulation is functionally the same as the deficit that has been posited in previous chapters to explain distinct behavioral patterns in people with ASD including executive dysfunction, overselective behavior, implicit learning deficits, and problems disambiguating the contextual information during sentence processing.

Future Directions

FILL ME IN, OR REMOVE THIS SECTION

The Utility of Computational Models for Understanding Autism

An important goal and contribution of my work is the use of 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 hardware 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 my hypothesis of dysfunctional PFC / DA interactions in an attempt to capture the performance of people with autism. Utilizing this approach we captured autistic behavior in the diverse areas of executive dysfunction, overselectivity[3], tasks of implicit learning, homograph disambiguation, and category learning. Indeed, the strength of the work presented here is not in any single explanation of autistic behavior, but in providing a single, plausible, precise neural mechanism that is capable of providing a level of inter-theorhetic reduction previously not seen in ASD research. However, at this point the concept of dysfunctional interactions between the mid-brain DA system and the PFC in people with autism is still just a theory, there is much work that is left to be done to either justify or modify my theory going forward.

Conclusion

As awareness and resources continue to grow, so does the overall number of diagnoses of people with ASD. At the same time, ASD continues to pose a massive challenge to researchers. While great progress is being made in areas such as early identification of ASD, as well as intervention techniques, to date there is no sign of a converging consensus as to the true neural underpinnings of ASD. Further complicating our understanding of the mechanisms that may underly autism is a staggeringly diverse behavioral profile as well as multiple physical abnormalities that often accompany a diagnosis. In this document, we have presented a program of research in an attempt to address some of these issues. Dopamine has diffuse and widespread effects of throughout the human cortex. This coupled with strong ties to multiple clinical populations as well as numerous ties to human behavior make it an intriguing initial candidate for a disorder possessing a profile like that of autism. However, it is not until we recognize the myriad of ties between DA and core autistic behavioral differences that we begin to see the true potential of the DA dysfunction hypothesis of ASD. 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 mid-brain dopamine system (Rinehart et al., 2006, 2001; Tuchman & Rapin, 2002;

Hill, 2004; Ozonoff et al., 1991; Starr, 1996; Ralph-Williams et al., 2003; Ralph et al., 2001; Graybiel, 2000; Matsumoto et al., 1999), supporting the argument for a role for the dopaminergic system in the etiology of autism. Importantly, we have argued that perturbed DA / PFC interactions may lead to overly perseverative attention 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 specific executive dysfunction profile demonstrated by people with autism are nicely captured. Cognitive control is instantiated via actively maintained control representations within the prefrontal cortex. Cognitive flexibility is implemented via interactions between PFC and the mid-brain dopamine system. These interactions are suspect in autism, resulting in problems in flexibly updating the control instantiated via the PFC and capturing the problematic profile demonstrated by people with ASD (Hill, 2004; Ozonoff et al., 1991). Developmentally, executive dysfunction does not appear until later in childhood. My modeling efforts indicate that this may occur due to the protracted development of the PFC. In my model, early performance is driven largely by non-frontal, more posterior, brain systems which are largely unaffected by the posited DA-related abnormalities in autism. As the PFC becomes more effective, differences in PFC/DA interactions are unmasked.

Stimulus overselectivity, where a restricted subset of possible items or features in the environment dominate behavior in people with ASD, can also be subsumed under the same theoretical framework. We hypothesize that frequent and flexible updating of the attentional and control representations stored in the PFC is necessary in order to prevent an overly restricted subset of items in the environment from gaining control over our behavior. Under this account, inflexible and infrequent updating results in a restricted subset of features from the environment dominating the contents maintained within the PFC. Subsequently, through an associative learning process, the restricted subset comes to possess stronger “association weights” and thus dominate responding compared to the other features in the environment.

In another area of interest, weak central coherence theorists posit that people with autism have difficulties integrating pieces of information into a coherent whole or "gestalt"(Frith, 1989; Happe, 1999b). A major contribution of this work is demonstrating how top-down PFC-like mechanisms may influence the representations learned in other cortical areas. As such, WCC may be recast not as a problem in the integration of information per se, but rather as integrating the *wrong* information, due to the inflexible updating of attentional / control representations stored within the PFC. Problems with implicit learning as well as using contextual information to disambiguate sentential context can both be explained utilizing this account. Tasks investigating implicit learning, such as the Serial Response Time Task (SRTT), depend on previous information about the sequence to be readily available on subsequent time steps in order for normal learning to occur. Without reliable contextual information, neural systems will struggle to integrate the past information in an appropriate manner. The same is true when determining the meaning of ambiguous words in a sentence, without an appropriate representation of the context, the best we can do is to rely on the statistical frequency of the words we experience.

Finally, people with autism also demonstrate an atypical prototype effect when learning category structures (Klinger & 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 out somewhat evenly across the relevant features of a stimulus or category example. If, as would be caused by inflexible attentional of the PFC, we highlight and learn to "over-value" a restricted subset of the features, the representation that is learned would likely not represent the standard mathematical average of psychological feature values as is argued to occur during prototype formation. In this case, an individual may become "overselective" and weight the restricted subset of feature values more highly during category determination.

The convergence of data supporting dopamine's role in autism, combined with the possibility of providing a conceptual bridge spanning multiple theories in autism, is extremely encouraging. Computational modeling results presented in this document help to demonstrate the potential for formal computational models investigating the links between possible neural underpinnings and behavior in people with ASD. Using simulations, constrained and informed by both biology and observed behavior, precise and testable predictions of underlying mechanisms can be made, providing a theoretical bridge between psychological and anatomic theories of ASD. Namely, my modeling efforts suggest that dysfunctional interactions between the mid-brain DA system and the PFC may lead to overly perseverative top-down attentional effects in people with ASD. By casting the PFC as key player in both attention and in the shaping of posterior cortical areas, my theory provides a way to unify multiple previously disparate behavioral phenomena observed in people with ASD.

References

APA. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR (fourth edition, text revision)*. Washington DC: American Psychiatric Association.

Barik, S., & de Beaurepaire, R. (1996). Evidence for a functional role of the dopamine d3 receptors in the cerebellum. *Brain research*, *737*, 347-350(4).

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

Barto, A. G. (1994). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). MIT: MIT Press.

Bennetto, L., Pennington, B. F., & Rogers, S. J. (1996). Intact and impaired memory functions in autism. *Child Development*, *67*, 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., & Cohen, J. D. (1999). Dopamine, cognitive control, and schizophrenia: The gating model. *Progress in Brain Research*, *121*, 327–349.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 713–737). Cambridge, MA: MIT Press.

Bruckner, C. T., & Yoder, P. (2007, March 1). Restricted object use in young children with autism: Definition and construct validity. *Autism*, *11*(2), 161-171.

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

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

Courchesne, E., Townsend, J., Akshoomoff, N. A., Saitoh, O., Yeung-Courchesne, R., Lincoln, A., . . . Lau, L. (1994). Impairment in shifting attention in autistic and cerebellar patients. *Behavioral Neuroscience*, *108*(1), 848–865.

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., & Liaw, J. (2004). Early social attention impairments in autism: social orienting, joint attention, and attention to distress. *Developmental psychology*, *40*(2), 271.

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

Fernell, E., Watanabe, Y., Adolfsson, I., Tani, Y., Bergstrom, M., Hartvig, P., . . . 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.

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral neuroscience*, *120*(3), 497.

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

Frith, U. (2003). *Explaining the enigma*. Oxford: Blackwell.

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

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum (Ed.), *Handbook of pysiology - the nervous system* (pp. 373–417). Bethesda, MD: American Physiological Society.

Graybiel, A. M. (2000). Levodopa-induced dyskinesias and dopamine-dependent stereotypies: a new hypothesis. *Trends in Neurosciences*, *23*(10 Suppl), S71-S77. (This article provides two interesting contributions. 1. References and evidence for dopamine dependent stereotypies in animal models (as well as some anecdotal evidence in humans). 2. A hypothesis of a neural mechanism which could be responsible for the stereotypies. Namely, imbalance in the Matrix-Strisomes where the strisomes activity starts to overpower the matrix cells in the striatum.)

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

Grossberg, S., & 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.

Happe, F. (1997). Central coherence and theory of mind in autism: reading homographs in context. *Journal of Developmental Psychology*, *15*, 1–12.

Happe, F. (1999a). Autism: Cognitive deficit or cognitive style? *Trends in cognitive sciences*, *3*(6), 216-222.

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

Happe, F., Ronald, A., & Plomin, R. (2006, 2006/10//print). Time to give up on a single explanation for autism. *Nature neuroscience*, *9*(10), 1218-1220. (M3: 10.1038/nn1770; 10.1038/nn1770)

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

Hill, E., & Frith, U. (2003). Understanding autism: insights from mind and brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1430), 281-289.

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

Jackson, M. E., & Moghaddam, B. (2001, January 15). Amygdala regulation of nucleus accumbens dopamine output is governed by the prefrontal cortex. *Journal of Neuroscience*, *21*(2), 676-681.

Jolliffe, T., & Baron-Cohen, S. (1997). Are people with autism and asperger syndrome faster than normal on the embedded figures test? *Journal of Child Psychology and Psychiatry*, *38*(5), 527-534.

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach* (Tech. Rep. No. 8604). University of California, San Diego.

Joseph, R. M. (1999). Neuropsychological frameworks for understanding autism. *International Review of Psychiatry*, *11*, 309-324(16).

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*, 217–250.

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

Klinger, L. G., Klinger, M. R., & Pohlig, R. A. (2006). Implicit learning impairments in autism spectrum disorders: Implications for treatment. In J. M. Perez, P. M. Gonzalez, M. L. Comi, & C. Nieto (Eds.), *New developments in autism.* London: Jessica Kinglsey.

Koegel, R., Schreibman, L., Good, A., Cerniglia, L., Murphy, C., & Koegel, L. K. (1989). *How to teach pivotal behaviors to children with autism: A training manual* (Tech. Rep.).

Kriete, T., & Noelle, D. C. (2006). Dopamine and the development of executive dysfunction in autism. In *Proceedings of the 5th international conference on development and learning.*

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Leary, M. R., & Hill, D. A. (1996). Moving on: autism and movement disturbance. *Mental Retardation*, *34*(1), 39-53.

Li, S., Cullen, W. K., Anwyl, R., & Rowan, M. J. (2003). Dopamine-dependent facilitation of ltp induction in hippocampal ca1 by exposure to spatial novelty. *Nature neuroscience*, *6*(5), 526-531. (M3: 10.1038/nn1049; 10.1038/nn1049)

Lovaas, O., & Schreibman, L. (1971). Stimulus overselectivity of autistic children in a two stimulus situation. *Behaviour Research and Therapy*, *9*, 305-310.

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

Lovaas, O., Schreibman, L., & Koegel, R. L. (1974, 06/01). A behavior modification approach to the treatment of autistic children. *Journal of Autism and Developmental Disorders*, *4*(2), 111-129. (M3: 10.1007/BF02105365)

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), (pp. 103–189). New York: Wiley.

Martineau, J., Barthelemy, C., Jouve, J., Muh, J. P., & 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., & Kimura, M. (1999, August 1). 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.

Mehta, M. A., Owen, A., Sahakian, B., Mavaddat, N., Pickard, J. D., & Robbins, T. W. (2000, March 15). Methylphenidate enhances working memory by modulating discrete frontal and parietal lobe regions in the human brain. *Journal of Neuroscience*, *20*(6), 65RC.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167-202. (M3: doi:10.1146/annurev.neuro.24.1.167)

Minshew, N. J., Meyer, J., & 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., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

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

Mottron, L., Peretz, I., & Menard, E. (2000). Local and global processing of music in high-functioning persons with autism: Beyond central coherence? *Journal of Child Psychology and Psychiatry*, *41*(8), 1057-1065. (M3: doi:10.1111/1469-7610.00693)

Nelson, E. C., & 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.

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*.

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

O'Reilly, R. C. (1996). *The Leabra model of neural interactions and learning in the neocortex* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh.

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. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex*, *12*(3), 246–257.

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

Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991). Executive function deficits in high-functioning autistic individuals: Relationship to theory of mind. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *32*, 1081–1105.

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

Ploog, B., & Kim, N. (2007, 09/14). Assessment of stimulus overselectivity with tactile compound stimuli in children with autism. *Journal of Autism and Developmental Disorders*, *37*(8), 1514-1524. (M3: 10.1007/s10803-006-0244-5)

Posey, D. J., & 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., & Heavey, L. (1995). Savants, segments, art and autism. *Journal of Child Psychology and Psychiatry*, *36*(6), 1065-1076. (M3: doi:10.1111/j.1469-7610.1995.tb01351.x)

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

Ralph, R. J., Paulus, M. P., Fumagalli, F., Caron, M. G., & Geyer, M. A. (2001). Prepulse inhibition deficits and perseverative motor patterns in dopamine transporter knock-out mice: Differential effects of d1 and d2 receptor antagonists. *Journal of Neuroscience*, *21*(1), 305-313.

Ralph-Williams, R. J., Paulus, M. P., Xiaoxi, Z., Hen, R., & Geyer, M. A. (2003). Valproate attenuates hyperactive and perseverative behaviors in mutant mice with a dysregulated dopamine system. *Biological psychiatry*, *53*(4), 352-359.

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

Reed, P., & Gibson, E. (2005). The effect of concurrent task load on stimulus over-selectivity. *Journal of Autism and Developmental Disorders*, *35*, 601-614(14).

Reynolds, B. S., Newsom, C. D., & Lovaas, O. (1974, 12/01). Auditory overselectivity in autistic children. *Journal of abnormal child psychology*, *2*(4), 253-263. (M3: 10.1007/BF00919253)

Rinehart, N. J., Bradshaw, J. L., Brereton, A. V., & 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*, 79-88(10).

Rinehart, N. J., Tonge, B. J., Iansek, R., McGinley, J., Brereton, A. V., Enticott, P. G., & 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. (M3: doi:10.1111/j.1469-8749.2006.tb01229.x)

Robbins, T. W. (1997). Integrating the neurobiological and neuropsychological dimensions of autism. In J. Russell (Ed.), *Autism as an executive disorder* (pp. 21–53). Oxford: Oxford University Press.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & 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.

Rougier, N. P., & O'Reilly, R. C. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, *26*(4), 503–520.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward.

Science, *275*, 1593–1599.

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

Starr, M. S. (1996). The role of dopamine in epilepsy. *Synapse*, *22*(2), 159-194.

Stores, G., & Wivggs, L. (1998, June 1). 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., & 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., . . . 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.

Tessitore, A., Hariri, A., Fera, F., Smith, W., Chase, T., Hyde, T., . . . Mattay, V. (2002, October 15). Dopamine modulates the response of the human amygdala: A study in parkinson's disease. *Journal of Neuroscience*, *22*(20), 9099-9103.

Triesch, J., Jasso, H., & Deak, G. O. (2007, June 1). Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior*, *15*(2), 149-165.

Triesch, J., Teuscher, C., Deak, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, *9*(2), 125-157.

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

Tuchman, R., & Rapin, I. (2002). Epilepsy in autism. *Lancet Neurology*, *1*, 352-358(7).

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

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. (1971). *A manual for the embedded figures test*. Palo Alto, California: Consulting Psychologists Press.

Author Note

Place acknowledgements here.

Footnotes

[1] If the healthy network is allowed to update imperfectly, as well as the autism network, the best fit arises when the healthy network updates with 0.6 probability and the autism network updates with 0.2 probability, producing an SSE of 549. Unfortunately, since the variance of the human data was not reported in Mostofsky et al. (2000), we cannot assess if these parameters are reliably better at fitting the human data than the $1.0/0.5$ case.

[2] Note that the same Control network is used in both the schizophrenia and autism modeling results.

[3] The overselectivity model was the only model that was not previously published.

Figure Captions

Figure 1. Firing rates of midbrain dopamine neurons of the basal ganglia during classical conditioning (Adapted from Schultz et al., 1997)

Figure 2. XT Model Architecture. The upper left corner shows a caricature of the input to the XT network, with rows portraying stimulus dimension (color, shape, size, etc.) and columns indexing feature values across dimensions (small, medium, large, etc.). Each small box in this network diagram represents a Leabra processing unit, and arrows between layers represent complete unit-to-unit connectivity between layers. The AG unit implements “adaptive gating” by modeling the effects of the dopamine system on PFC.

Figure 3. XT WCST example stimulus input

Figure 4. Top: Comparing model performance on WCST perseveratitive errors for normally functioning and individuals with autism to a previous study (Minshew et al., 2002). Model results capture performance of people with autism, committing significantly more perseverative errors. Error bars are standard errors of the mean. Bottom: Stroop reaction time plot comparing the simulated autistic and normally functioning network’s performance to human data. Healthy human data from Dunbar et al. (1984). Autistic subjects perform no differently than controls (Ozonoff et al., 1999). Error bars are standard errors of the mean.

Figure 5. PFC Representations early in development (epoch 5): left - control model, right - autism model. Each large box corresponds to the connection strength from that PFC unit, and each small box corresponds to the connection strength from that PFC unit to a Response layer unit, with strength growing with brightness. Note the lack of any strong dimensional representations in both versions.

Figure 6. PFC Representations late in development (epoch 100): left - control model, right -

autism model. Each large box corresponds to the connection strength from that PFC unit, and each small box corresponds to the connection strength from that PFC unit to a Response layer unit, with strength growing with brightness. Strong dimensional representations have formed in both models, as exhibited by strong weights from individual PFC units to all of the Response layer features for a given dimension.

Figure 7. Stroop interference score for the model over the course of development.

Figure 8. WCST perseverative error count for the model over the course of developmental training.

Figure 9. Network diagram of stimulus overselectivity model.

Figure 10. Simulation results modeling overselectivity task (see Figure **??**) by manipulating the flexible updating of a PFC-like mechanism. Only the simulation where the PFC is allowed to flexibly adjust the activation of its representations ("normal condition") resulted in all three components of the compound stimulus gaining equal control over behavior. Both the inflexible updating (autism condition) and the addition of irrelevant information in PFC manipulations (working memory load condition) result in a restricted subset of components gaining control over behavior, demonstrating overselective behavior.

Figure 11. Top panel represents a cartoon model of the stimulus overselectivity task where the PFC is able to flexibly adjust its representations (normal condition). The resulting "weights" driving the response are all similar, resulting in each dimension or feature having equal influence over the response. The bottom panel represents inflexible updating of the PFC (autism condition). The red arrow indicates that the PFC is "stuck" on the first dimension or feature. The weights are biased in favor of the feature perseverated on by the PFC, resulting in behavior being dominated by a restricted subset of those available, demonstrating overselectivity when simulating autistic performance.

Figure 12. General structure of a Simple Recurrent Network (SRN) model. Image adapted from Cleeremans and McClelland (1991).

Figure 13. Network Diagram of SRTT Model

Figure 14. Scaled Model Results & Human Behavioral Data from Mostofsky et al. (2000)

Figure 15. Results from study on lexical disambiguation in people with ASD. Common Before/After refers to a common interpretation of the homograph, and it occurring Before/After the contextual information respectively. Similarly, Rare Before/After refers to a rare interpretation of the homograph, and it occurring Before/After the contextual information respectively. Participants were presented with five sentences in each condition. Table from Jolliffe and Cohen (1999).

Figure 16. Results from study on lexical disambiguation in people with Schizophrenia. Y-axis shows percentage of errors using the most common interpretation of the homograph, when the rare interpretation was correct. Figure adapted from Cohen and Servan-Schreiber (1992).

Figure 17. The original lexical ambiguity model. Image adapted from Cohen and Servan-Schreiber (1992).

Figure 18. Cartoon of the lexical ambiguity model and task (the number of inputs and outputs to the network is significantly larger than what is displayed). The task of the model is to correctly respond to any input word presented within the "INPUT" layer of the network. This is accomplished by activating the unit representing the appropriate meaning at the "OUTPUT" layer of the network. There are two types of input word units, ambiguous homograph units and context units. The context units are used to disambiguate the meaning of a specific homograph unit. Words are presented to the network one at a time and it is the networks job to learn, through experience, to respond with the correct meaning via the "OUTPUT" layer of the network. For instance, in this figure the ambiguous homograph input unit represents the word "Bank" and the frequent and

infrequent context input units represent “Money” and “River” respectively. When the infrequent context unit “River” is presented first to the network followed by the homograph unit for “Bank”, the network was expected to learn to use the context unit’s meaning to correctly respond “Land by River”. This response occurs during the final trial when the network is probed for the correct meaning of the ambiguous homograph unit after the initial mini-clause has been presented.

Figure 19. Modeled effects of tonic DA on a standard activation function. An increase in the amount of DA heightens the effects of the net input, effectively increasing the signal-to-noise ratio in the network, while lower DA decreases the effects, reducing the signal-to-noise ratio. Image adapted from Cohen and Servan-Schreiber (1992).

Figure 20. 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 homograph.

Figure 21. ASD Model Results. Systematically reducing the ability of the modeled PFC to update its contents 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 homograph.

Figure 22. Examples of different members of a category used in the prototype learning experiment. Image adapted from (Klinger and Dawson, 2001).

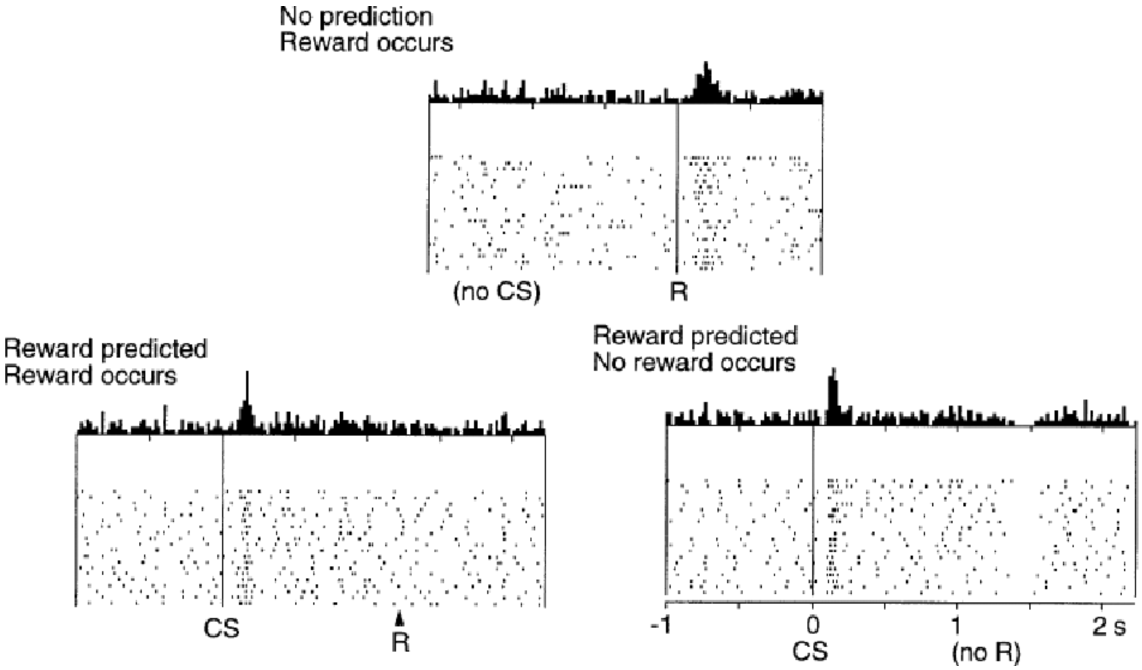
Figure 23. Examples of different categories used in the prototype learning experiment. Image adapted from (Klinger and Dawson, 2001).

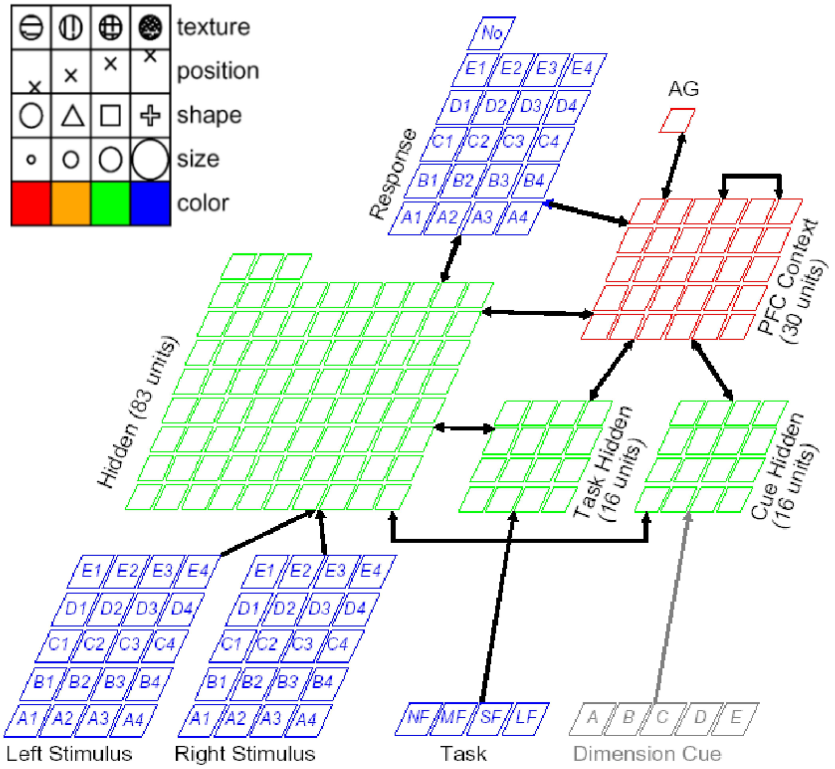
Figure 24. Results adapted from (Klinger and Dawson, 2001). The measure on the y-axis is the percentage of times that the prototypical animal was chosen over either a novel or familiar

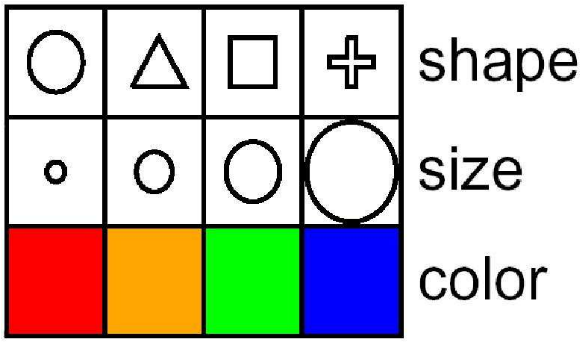
example. Their results indicate that the normally developing group chose the prototype at greater than chance levels, clearly demonstrating a prototype effect. However, the ASD group did not show a prototype effect, this is argued to support the authors conjecture that people with ASD do not form or utilize prototype representations in categorization tasks. (Please note: Only the “Prototype” condition is relevant to the discussion. The results for “Familiar” refer to a comparison in the original study investigating if there was a preference for previously experienced or novel animals.)

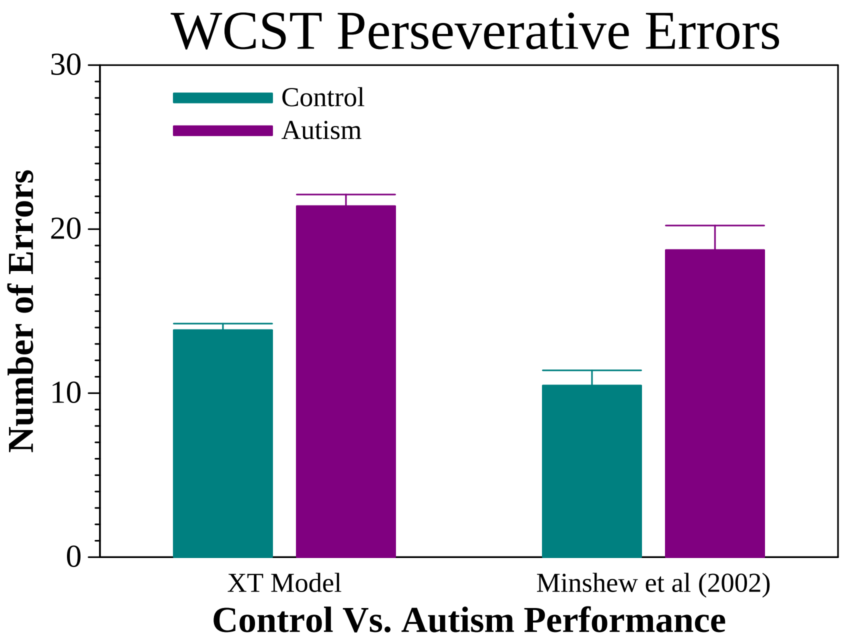
Figure 25. Structure of ALCOVE (Attention Learning CoVEring map). Image adapted from (Kruschke, 1992).

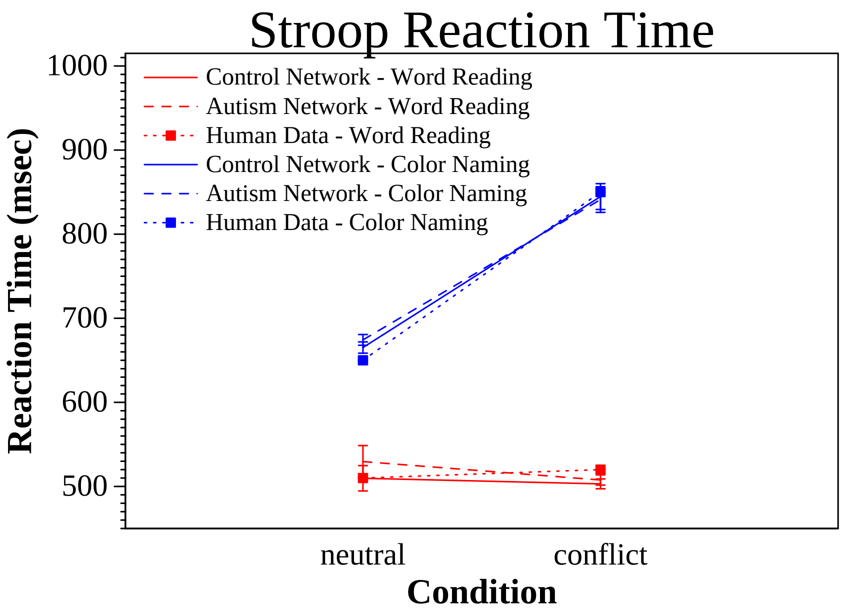
Figure 26. ALCOVE Model Results. The measure on the y-axis is the percentage of times that the prototypical animal was chosen by the model over the non-prototype choices. In the control network, the prototype was chosen 70.52% of the time on average, demonstrating a strong prototype effect and matching the human subject experiment results. However, the model of autistic performance chose the prototype on average only 52.91% of the time over the non-prototype stimuli. Statistical analysis reveal that the control network chose the prototype at a rate significantly greater than chance ($p < .0001$) while the ASD network was not statistically distinguishable from chance ($p > .30$).

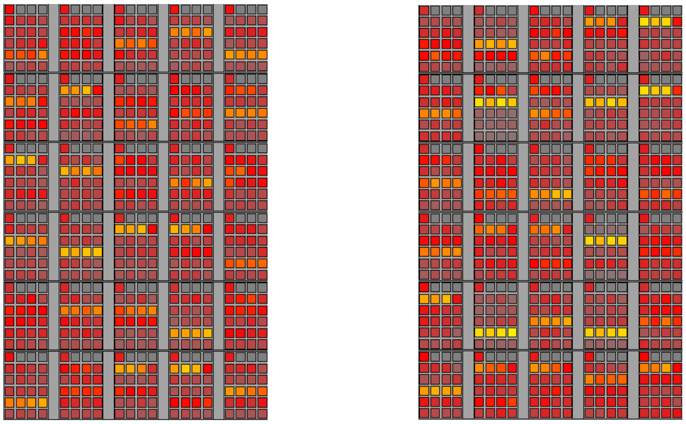

No prediction
Reward occurs
(no CS)
R
Reward predicted
Reward occurs
CS
R
Reward predicted
No reward occurs
-1
0
1
2 s
CS
(no R)

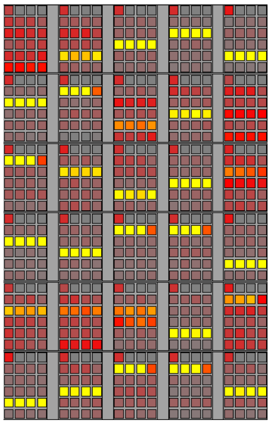

texture
position
shape
size
color
No
E1 E2 E3 E4
D1 D2 D3 D4
C1 C2 C3 C4
B1 B2 B3 B4
A1 A2 A3 A4
Response
AG
PFC Context (30 units)
Hidden (83 units)
Task Hidden (16 units)
Cue Hidden (16 units)
E1 E2 E3 E4
D1 D2 D3 D4
C1 C2 C3 C4
B1 B2 B3 B4
A1 A2 A3 A4
Left Stimulus
E1 E2 E3 E4
D1 D2 D3 D4
C1 C2 C3 C4
B1 B2 B3 B4
A1 A2 A3 A4
Right Stimulus
NF MF SF LF
Task
A B C D E
Dimension Cue

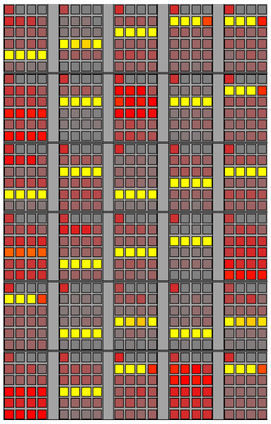

shape
size
color

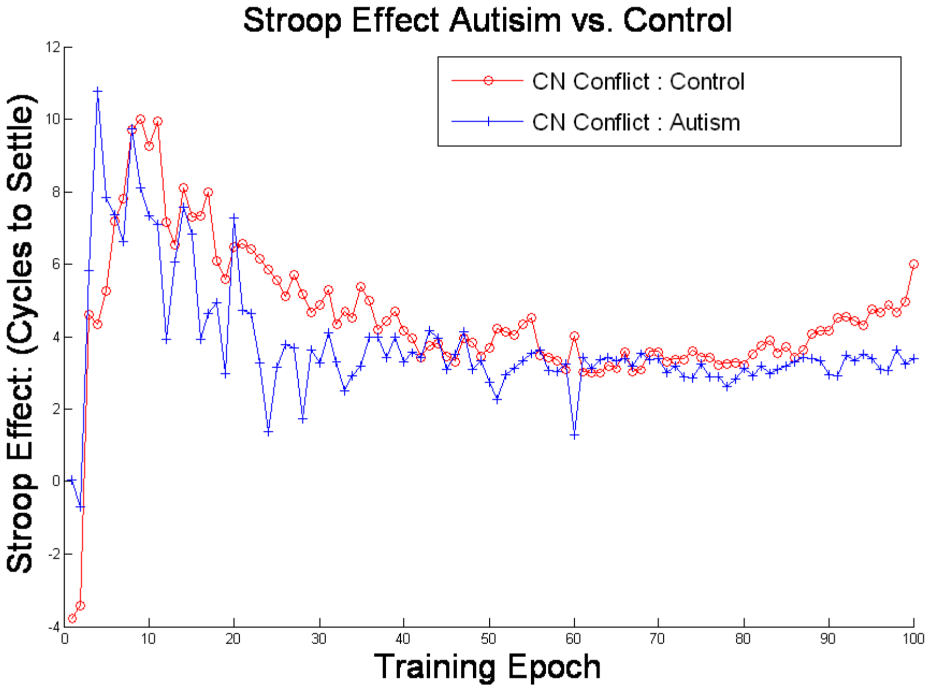

WCST Perseverative Errors
Control
Autism
Number of Errors
30
20
10
0
XT Model
Minshew et al (2002)
Control Vs. Autism Performance

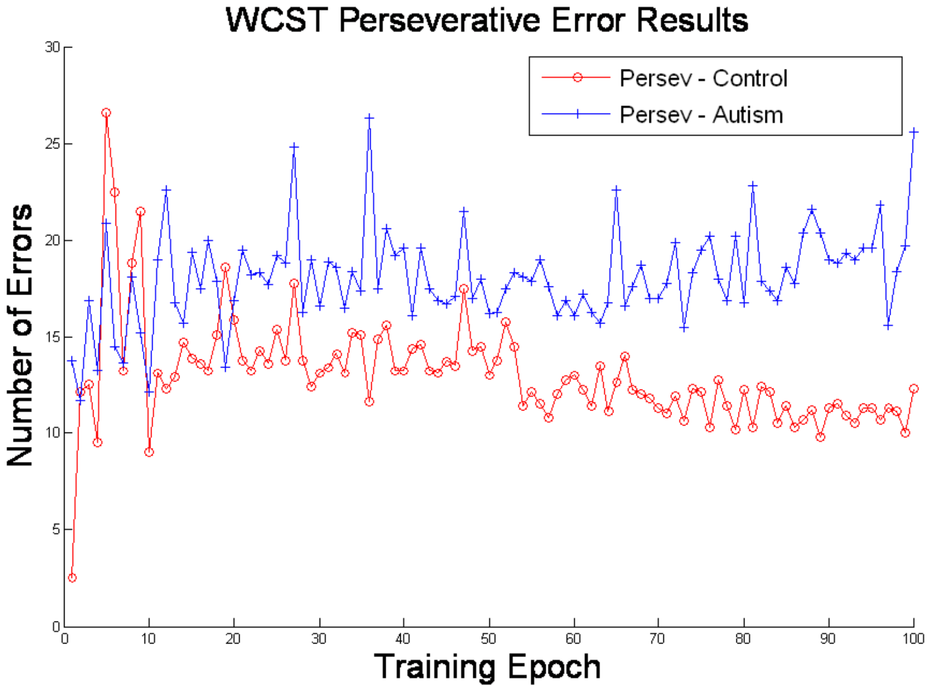

Stroop Reaction Time
Control Network - Word Reading
Autism Network - Word Reading
Human Data - Word Reading
Control Network - Color Naming
Autism Network - Color Naming
Human Data - Color Naming
Reaction Time (msec)
1000
900
800
700
600
500
neutral
conflict
Condition

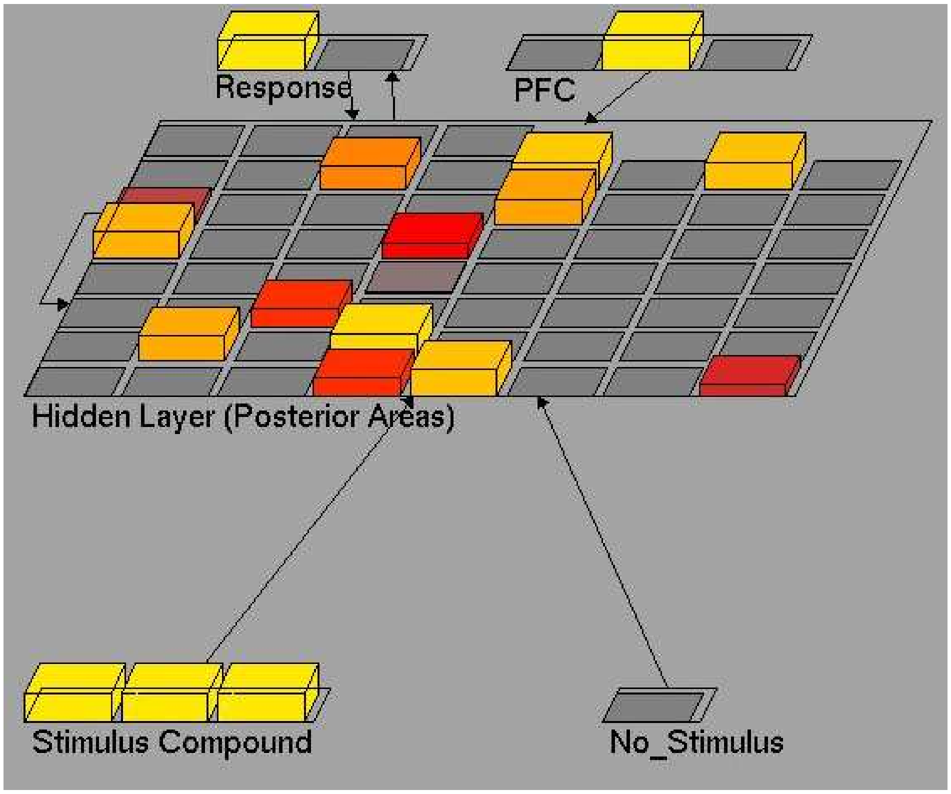


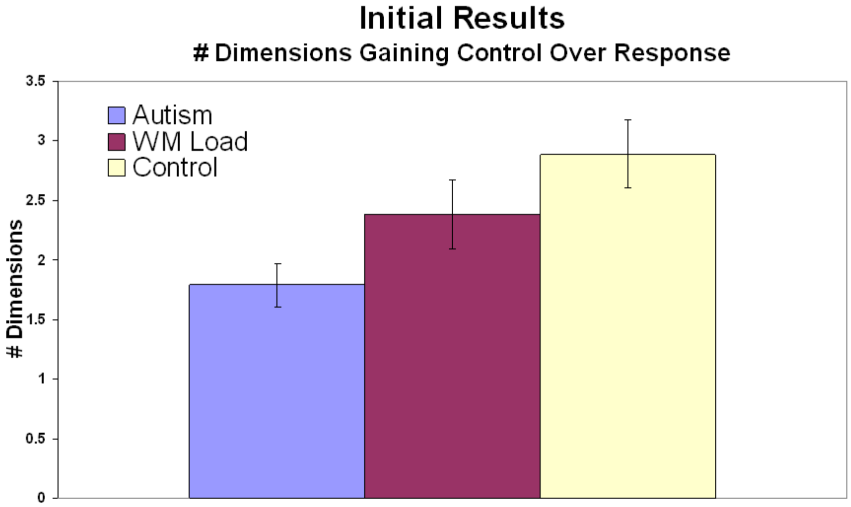


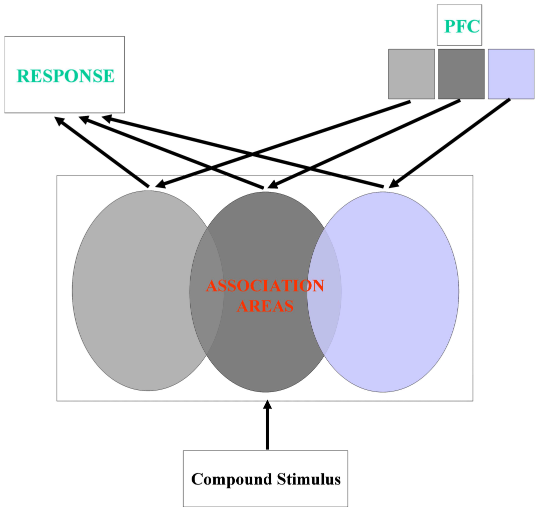


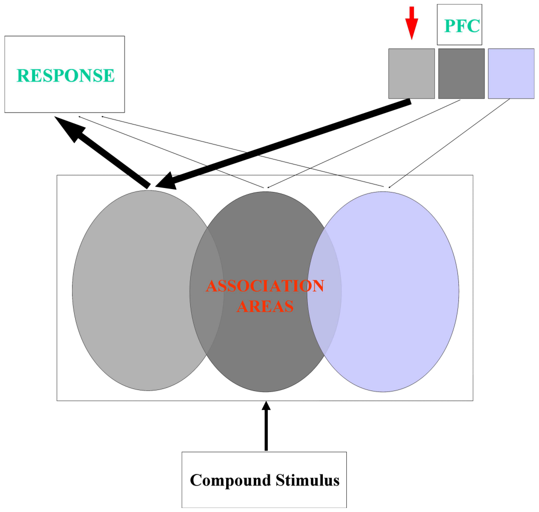

Stroop Effect Autisim vs. Control
Stroop Effect: (Cycles to Settle)
12
10
8
6
4
2
0
-2
-4
0
10
20
30
40
50
60
70
80
90
100
Training Epoch
CN Conflict : Control
CN Conflict : Autism

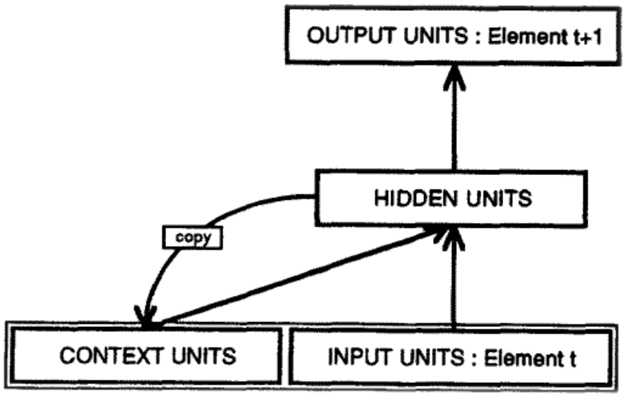

WCST Perseverative Error Results
Persev - Control
Persev - Autism
Number of Errors
Training Epoch
0
5
10
15
20
25
30
0
10
20
30
40
50
60
70
80
90
100

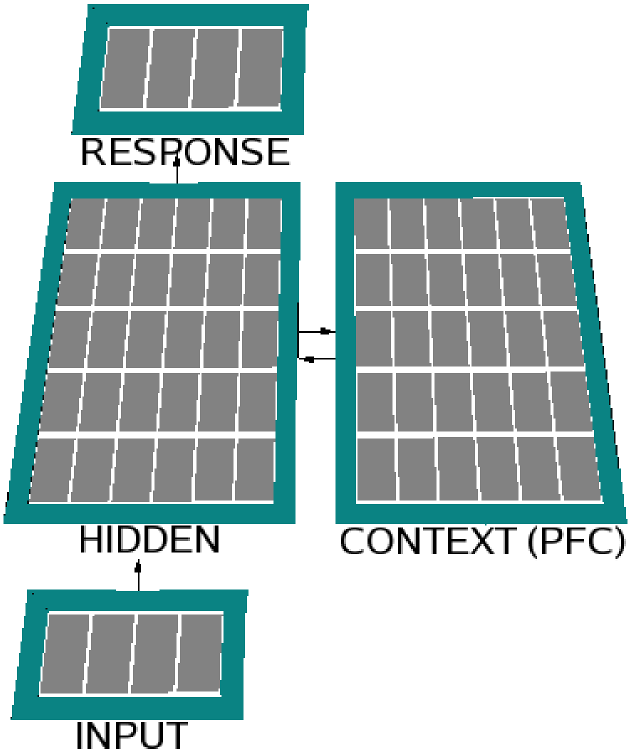

Response
PFC
Hidden Layer (Posterior Areas)
Stimulus Compound
No_Stimulus

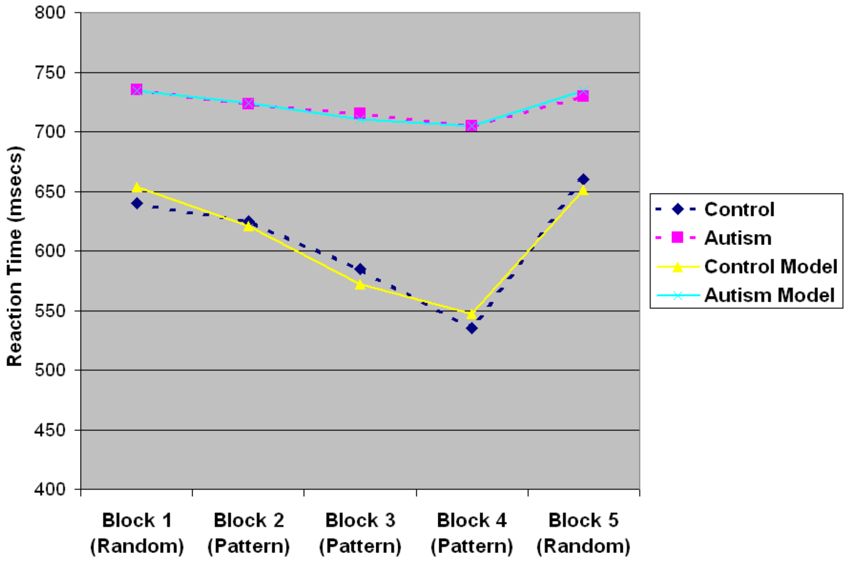

Initial Results
# Dimensions Gaining Control Over Response
# Dimensions
3.5
3
2.5
2
1.5
1
0.5
0
Autism
WM Load
Control


PFC
RESPONSE
ASSOCIATION
AREAS
Compound Stimulus

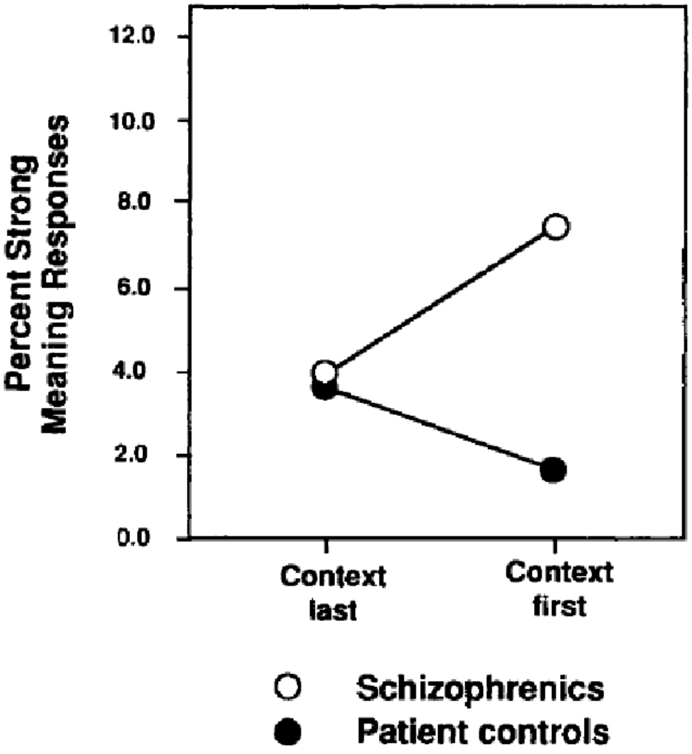

PFC
RESPONSE
ASSOCIATION
AREAS
Compound Stimulus

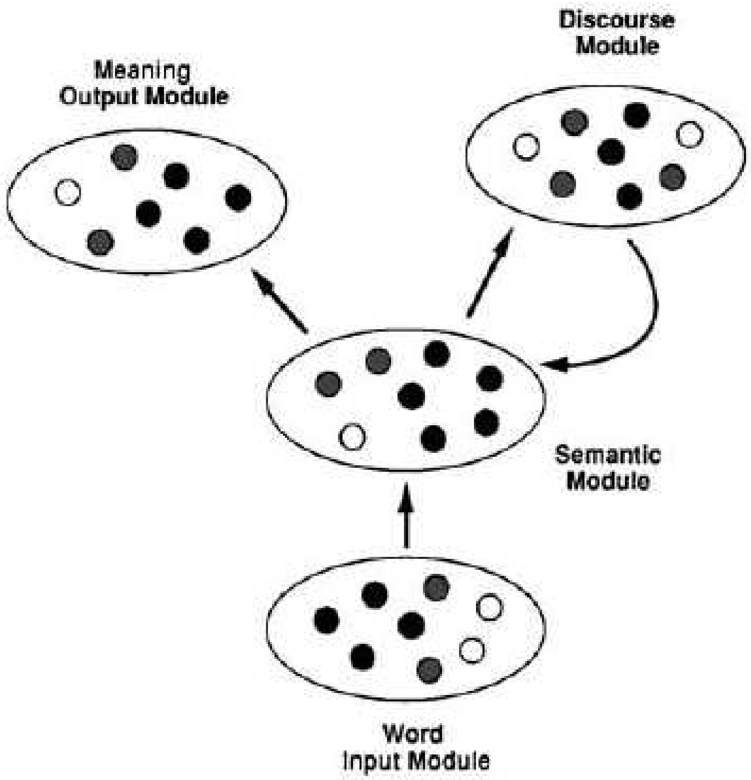

OUTPUT UNITS : Element t+1
HIDDEN UNITS
copy
CONTEXT UNITS
INPUT UNITS : Element t

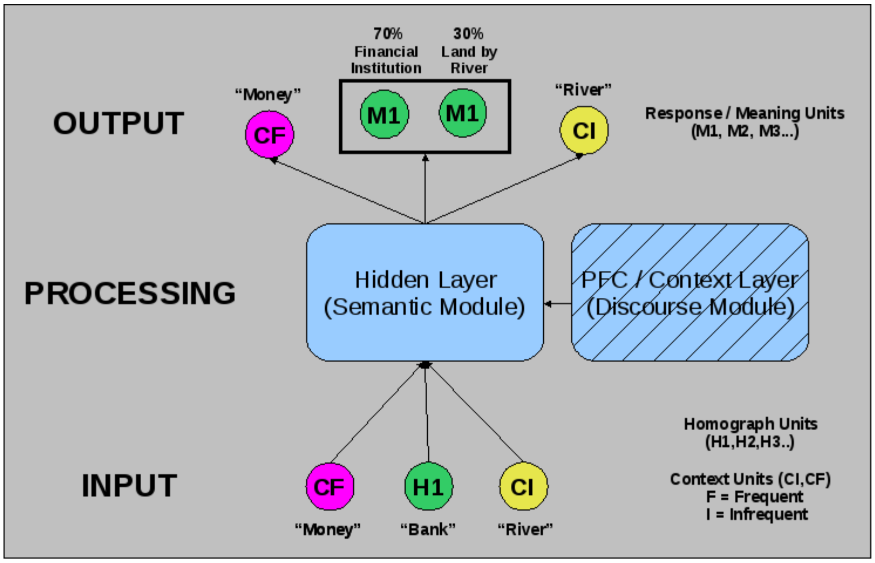

RESPONSE
HIDDEN
CONTEXT (PFC)
INPUT

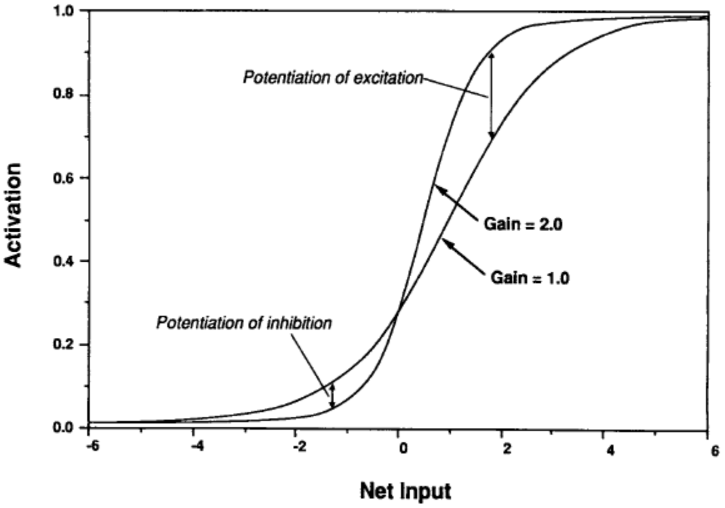

800
750
700
650
600
550
500
450
400
Reaction Time (msecs)
Block 1 (Random)
Block 2 (Pattern)
Block 3 (Pattern)
Block 4 (Pattern)
Block 5 (Random)
Control
Autism
Control Model
Autism Model

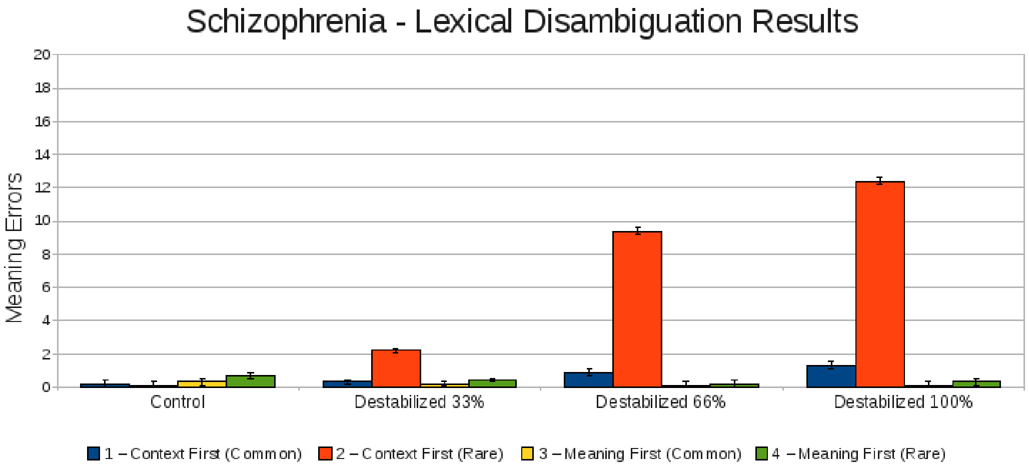
Participant Group		Common Pronunciations		Rare Pronunciations	
(n = 17)		Before	After	Before	After
Normal	Mean	4.88	5.00	4.88	5.00
	sd	0.33	0.00	0.33	0.00
	Range	(4 - 5)	(5 - 5)	(4 - 5)	(5 - 5)
Autism	Mean	4.77	4.88	3.82	4.06
	sd	0.44	0.33	1.19	0.97
	Range	(2 - 5)	(4 - 5)	(2 - 5)	(2 - 5)
Asperger	Mean	4.77	5.00	4.29	4.59
	sd	0.44	0.00	0.99	0.71
	Range	(4 - 5)	(5 - 5)	(2 - 5)	(3 - 5)

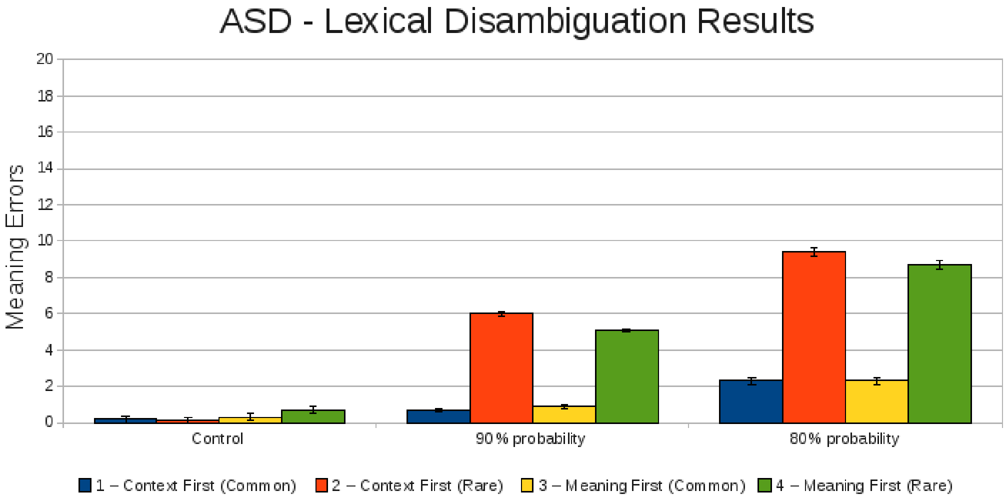

Percent Strong Meaning Responses
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Context last
Context first
○ Schizophrenics
● Patient controls

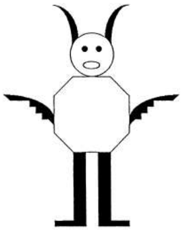

Discourse
Module
Meaning
Output Module
Semantic
Module
Word
Input Module


70%
Financial
Institution
30%
Land by
River
"Money"
"River"
OUTPUT
CF
M1
M1
CI
Response / Meaning Units
(M1, M2, M3...)
PROCESSING
Hidden Layer
(Semantic Module)
PFC / Context Layer
(Discourse Module)
Homograph Units
(H1,H2,H3..)
INPUT
CF
H1
CI
"Money"
"Bank"
"River"
Context Units (CI,CF)
F = Frequent
I = Infrequent

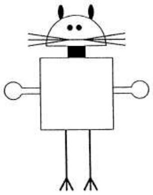

1.0
0.8
0.6
0.4
0.2
0.0
Activation
-6
-4
-2
0
2
4
6
Net Input
Potentiation of excitation
Gain = 2.0
Gain = 1.0
Potentiation of inhibition


Schizophrenia - Lexical Disambiguation Results
Meaning Errors
20
18
16
14
12
10
8
6
4
2
0
Control
Destabilized 33%
Destabilized 66%
Destabilized 100%
1 – Context First (Common)
2 – Context First (Rare)
3 – Meaning First (Common)
4 – Meaning First (Rare)

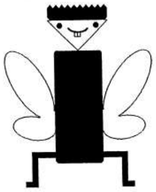

ASD - Lexical Disambiguation Results
Meaning Errors
20
18
16
14
12
10
8
6
4
2
0
Control
90% probability
80% probability
1 – Context First (Common)
2 – Context First (Rare)
3 – Meaning First (Common)
4 – Meaning First (Rare)



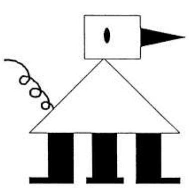
Mip



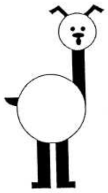
Dak



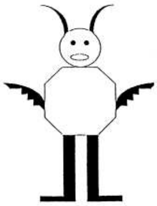
Sop



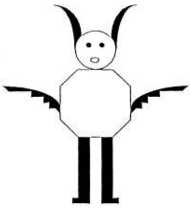
Pev



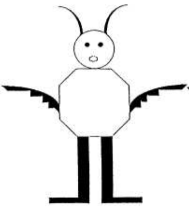
Tuz



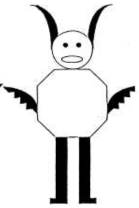
2244



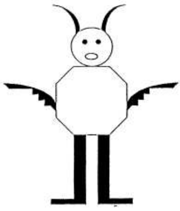
5511



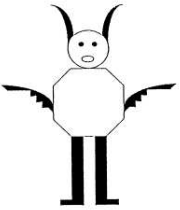
1515



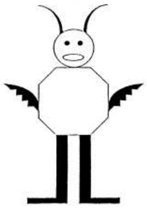
5151



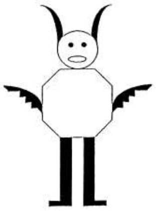
2424



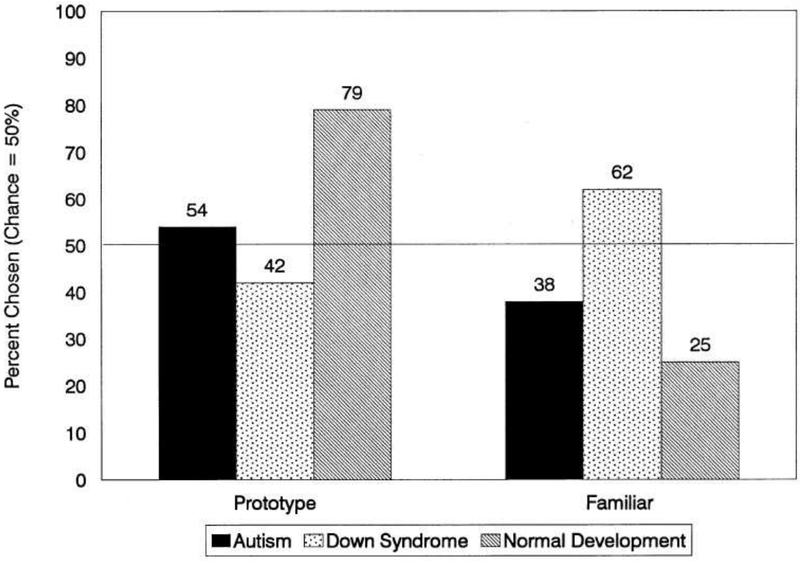
4422

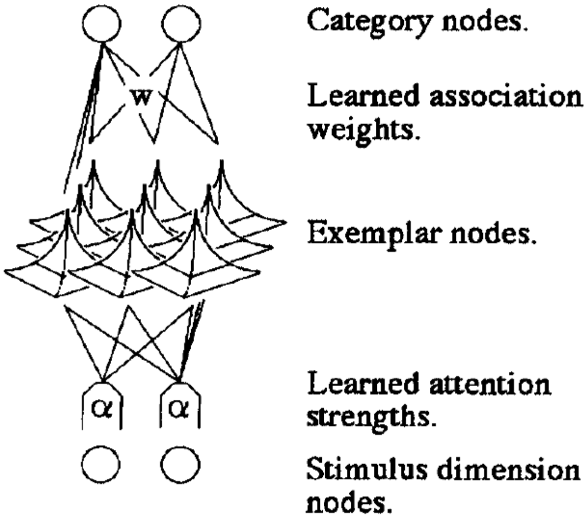


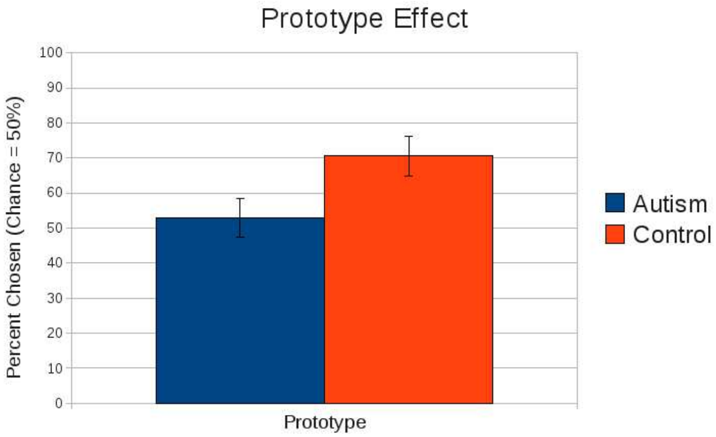
1155



4242


Percent Chosen (Chance = 50%)
100
90
80
70
60
50
40
30
20
10
0
54
42
79
38
62
25
Prototype
Familiar
■ Autism
▒ Down Syndrome
▨ Normal Development


Category nodes.
w
Learned association weights.
Exemplar nodes.
α
α
Learned attention strengths.
Stimulus dimension nodes.


Prototype Effect
100
90
80
70
60
50
40
30
20
10
0
Percent Chosen (Chance = 50%)
Prototype
Autism
Control