A Cognitive Computational Approach to Understanding Theory of Mind and Its Impairment in Autism Spectrum Disorder

Isabelle Rosenthal
irosent@wellesley.edu

Follow this and additional works at: https://repository.wellesley.edu/thesiscollection

Recommended Citation
https://repository.wellesley.edu/thesiscollection/403
A Cognitive Computational Approach to Understanding Theory of Mind and Its Impairment in Autism Spectrum Disorder

Isabelle A. Rosenthal

Submitted in Partial Fulfillment of the Prerequisite for Honors in Neuroscience

May 2016

© 2016 Isabelle Rosenthal
ACKNOWLEDGEMENTS

First and foremost, I want to thank Damian Stanley for going above and beyond as a mentor to me, for his patience in dealing with my myriad coding errors, and for constantly pushing me to accomplish things I thought were beyond my capabilities. Similarly, many thanks to Professor Ralph Adolphs for welcoming me into his lab at Caltech so enthusiastically and for giving me advice and establishing a sense of direction whenever it was needed.

I would also like to thank Marisol Espino for her work collecting data for this project, as well as Professor John O’Doherty for his help constructing the paradigm for neuroimaging, and Dan McNamee for guidance on building the computational models.

At Wellesley College, I want to thank Professor Bevil Conway for being exactly the kind of advisor I needed – always providing honest and incisive advice but also constantly encouraging me to take on new challenges. In addition, the other members of my thesis committee, Professor Margaret Keane and Professor Mike Wiest, have been invaluable sources of knowledge and guidance during my time at Wellesley, even before I began writing this thesis. Special thanks as well to Professor Barbara Beltz, who despite not being my advisor or on my committee always found time to ask me how I was doing and to help me work through my difficulties with an objective eye.

I must also thank my friends and peers – without whose constant support I would never have been able to stay afloat at Wellesley, let alone write a thesis. It would be impossible for me to name everyone who has had a meaningful impact, so I will just thank Monica Gates, Stela Petkova, and Blair Daniel for making my path through the neuro major as pleasant as possible, as well as Natalie Solomon, Allie Mathews, and Kaitlyn Brady for being the absolute best emotional support infrastructure I could ever ask for.

Finally, I would like to acknowledge the Amgen Foundation, the Howard Hughes Medical Institute, and the Caltech Summer Undergraduate Research Fellowships (SURF) program for funding my time at Caltech from 2015-2016.
# TABLE OF CONTENTS

Abstract ........................................................................................................................................... 4

Chapter 1: Introduction .................................................................................................................... 5
  1.1 Overview
  1.2 Theory of Mind (ToM)
  1.3 ASD and the Link to ToM
  1.4 The Computational Approach
  1.5 A Novel Paradigm

Chapter 2: Methods ......................................................................................................................... 20
  2.1 Participants
  2.2 Behavioral Data Collection
  2.3 Agent Generation
  2.4 Models of Learning

Chapter 3: Results .......................................................................................................................... 29
  3.1 Behavioral Data Analysis
  3.2 Computational Models
  3.3 Cross-Fitting
  3.4 Correlations

Chapter 4: Discussion ..................................................................................................................... 40
  4.1 Behavioral Differences
  4.2 Implications of Modeling Results
  4.3 Neural Correlates of Theory of Mind
  4.4 Future Directions

Author Contributions ...................................................................................................................... 48

References ......................................................................................................................................... 49

Appendix A – Mentalizer Stimulus Presentation Instructions ......................................................... 57

Appendix B – Code for Computational Models ................................................................................ 59
ABSTRACT
Theory of Mind (ToM), the mental ability to represent other people’s beliefs and intentions, remains a faculty that is poorly understood on a computational level. While a “mentalizing network” of implicated brain structures exists, the interaction of these structures and their specific functions still requires further study. Furthermore, although it is known that individuals with Autism Spectrum Disorder (ASD) tend to exhibit impairments in ToM, the nature of these impairments is not known. In this thesis, ToM is investigated through a novel learning task which bears a similar structure to past ToM tasks, but requires participants to update their representations of another individual’s beliefs and intentions continuously over time. Participant behavior on this task was investigated using eight different computational models. Neurotypical individuals were able to complete the task, answering questions about beliefs and intentions with an accuracy significantly above chance. Although individuals with ASD had an accuracy on belief questions that was not significantly different from the control group, they were impaired in answering questions about intent. Similarly, a computational model based on the Rescorla-Wagner update rule was found that fit well to the control group, but this model fit much differently to the ASD population, and failed to perform as well in predicting ASD behavior. No model was found that predicted ASD behavior better than control behavior. This suggests heterogeneity in the way individuals with ASD compute ToM. While expanding current knowledge about the way neurotypical individuals generate ToM representations, these computational modeling results also provide an opportunity to better understand the neural underpinnings of these representations via neuroimaging.
CHAPTER 1: INTRODUCTION

Overview

Theory of Mind (ToM) is the ability to form mental representations of the beliefs and intentions of other people. Humans use ToM, also known as “mentalizing,” extensively in day-to-day life in order to navigate social situations appropriately and to form predictions about the actions of those around them. Past research on ToM has implicated several brain structures as being part of a “mentalizing system” (van Overwalle & Baetens, 2009), primarily the temporo-parietal junction, the medial prefrontal cortex, and the precuneus (Rushworth et al, 2013; Koster-Hale & Saxe, 2013; Sallet et al, 2012; Van Overwalle & Baetens, 2009; Hampton et al, 2008; Rilling et al, 2004; Frith & Frith, 2003; Frith, 2001). However, despite a broad knowledge of which structures in the brain might be involved in ToM computation, how these structures interact with one another and with other brain areas remains unknown. Furthermore, it has yet to be revealed if these brain areas are uniquely specialized for ToM representations, or if they implement broader functions that happen to factor into social cognition (such as attention). Impaired ToM has been implicated in social disorders such as autism spectrum disorders (ASD) (Moran et al, 2011; Baron-Cohen et al, 1985; Yoshida et al, 2010a). Understanding the neural computations responsible for ToM is key to elucidating how normally developing individuals form ToM representations as well as how these social representations in ASD are different. It is clear that ASD is a heterogeneous disorder, and the development of a more detailed understanding of the underlying mechanisms might help with identifying subgroups within the ASD population (Yoshida et al, 2010a; Adolphs et al, 2001).

The focus of this thesis is on investigating the computations which give rise to ToM, mathematically speaking, and how/whether these calculations differ in neurotypical and ASD populations. A sub-question within this deals with the variance within the ASD population. While it is clear that within the label of “autism” lies a highly heterogeneous population with impairments in representing the mental and emotional states of others optimally (Schuwerk, Vuori, & Sodian, 2015; Yoshida et al, 2010a; Moran et al, 2011; Baron-Cohen et al, 1997a, 1997b; 1985; Adolphs et al, 2001), this heterogeneity has yet to be explored in depth in the context of ToM. It is unknown how much variance lies in ToM calculations in ASD, or to what extend this variance can account for impaired ToM representation.
To better understand how ToM is enacted in neurotypical and ASD populations, it is necessary to examine the computations which build ToM representations in both groups in greater depth. Here, this objective is approached by isolating different components of ToM, determining how individuals perform when trying to represent each component accurately, and subsequently generating computational models whose fit to the behavioral data may indicate how the brain generates ToM representations. For this purpose, a novel learning task that requires participants to use ToM processes in order to make predictions about an independent third party’s behavior was designed. While the broad goal of participants in this task – to learn about and represent both the beliefs and intentions of other individuals – is the same as previous tasks in the literature that have investigated the neural basis of ToM (e.g. Young et al., 2007; Saxe & Kanwisher, 2003; Gallagher et al., 2000; Vogeley et al., 2001), this task builds on earlier efforts by requiring updates of these representations to occur on a continuous basis as new information is received over time (much like how ToM is used in real life). Additionally, this task has been structured so as to allow for learning and decision-making to take place with regards to these updates, which allows for the fitting of computational models of learning to participant behavior. These models representing the uncertainty of participants about the beliefs and intentions of other agents and their surprise (i.e. prediction error) when the agents behave unexpectedly (Yoshida et al., 2010b). However, unlike past research, our task does not concern a single vignette, but rather requires the participant (Mentalizer) to learn about a single individual (Agent) over the course of many trials, during each of which the Mentalizer must make judgments about the Agent’s state of mind and receives new information with which to update their representations of the Agent’s beliefs and intentions. The implementation of dynamic representations, as well as the fact that the Mentalizers must represent varying degrees of uncertainty over time, serves to mimic real life use of ToM more accurately than the single vignettes used in previous research. This is crucial, because if modeling of ToM is to be useful, the neural activity of Mentalizers must be as close to what it would be in real social situations as possible.

By examining the performance of neurotypical and ASD individuals on this task, it is possible to elucidate the nature of the impairment in representing ToM that is evident in ASD. Various reinforcement learning models, fit to both groups, can give information about how a representation of another’s state of mind is updated with new information, as well as how this
updating may be altered in ASD. Using these methods, a more complete image of how ToM is enacted behaviorally and mathematically can be realized. The task has also been designed with neuroimaging in mind, leaving open the possibility that examination of MRI data of participants will yield more precise information about which brain areas are responsible for specific computations.

Theory of Mind (ToM)

The first concept of ToM – although use of the term “Theory of Mind” to describe the ability to infer the beliefs and intentions of other individuals did not occur until nearly 50 years later – had its roots in theories of child development pioneered by Jean Piaget in the early 20th century, which divided cognitive development into four stages (Flavell, 1963). Based on a variety of psychological and clinical observations, Piaget’s sociological model suggested that children progressed during the third stage, the concrete-operational stage (during preadolescence), from a position of egocentrism to one of sociocentrism (Flavell, 1963; Piaget, 1999; Fisch, 2013). Piaget defined an egocentric child as one unable to conceptualize situations from the perspectives of others, while a sociocentric one as able to consider the viewpoints of others, causing a dramatic shift in thought patterns and in moral outlook (Piaget, 1999). Although Piaget’s theories of child development have failed to be well-supported by empirical research in a larger sense (Fisch, 2013), his identification of a “sociocentric” framework represents one of the first attempts at defining the ability to think about the mental states of others.

One of the earliest studies examining how individuals ascribe intentions to others – in other words, looking for a sociocentric framework – required human participants to watch short animations of geometric shapes (Heider & Simmel, 1944). In the animation, the shapes (a circle and two triangles) moved around a few connected lines. When participants described the animation as a story in which the shapes fought and chased each other, this was taken as evidence of being able to perceive the shapes as beings with their own mental states.

However, the first actual use of the term “Theory of Mind” is relatively recent – first occurring in a 1978 study by Premack and Woodruff examining whether chimpanzees had ToM. The question was examined through means of a task in which a chimpanzee was shown videos of humans struggling to solve a problem (for example, attempting to grasp bananas out of reach). The chimpanzee was then shown photographs, one of which depicted the relevant solution to the
problem. The ability of the chimpanzee to accurately select the correct photograph was taken as an indication that the chimpanzee was able to understand the goals of the human and to find a solution in line with what the human would want – in other words, that the chimpanzee could understand a human’s intentions.

The findings of this seminal experiment were later called into question by those who suggested that correct performance of the Premack and Woodruff task was not sufficient to establish ToM as it could have occurred through the use of non-mentalizing inferences or by chance (Heyes, 1998; Povinelli & Bering, 2002). Furthermore, it has been argued that even if it is true that the single chimpanzee subject of the Premack and Woodruff study was able to solve the problems presented, this was neither necessary nor sufficient to indicate ToM because in subsequent studies, chimpanzees failed to perform other tasks that required ToM (Burge, 1978, Fisch, 2013). Interestingly, although the verdict appears to still be out regarding ToM and chimpanzees, there has been evidence that a large amount of social cognition takes place in monkeys on the neural level (Sallet et al, 2012; 2011). However, the validity of ToM in nonhuman primates aside, since 1978, many studies regarding the representation of the mental states of others have employed the term “Theory of Mind” to describe their subject matter – for instance, the seminal “reading the mind in the eyes” task (Baron-Cohen et al., 1997b, 2001a). In this task, the ToM of participants was investigated by showing them closely cropped images of eyes and asking them to identify the emotion of the individual in the image.

The presence of situations in which task participants are representing the false beliefs of others has come to be a gold standard for ToM studies. This is because an individual who thinks someone else has a false belief must be holding their representation of that person’s state of mind as separate from their representation of the actual, true, state of the world. Piaget’s notion of egocentrism would dictate that an individual without ToM would be unable to integrate information about the false beliefs of others, because they do not have the capability to represent any viewpoint other than their own. The Sally-Ann task designed by Wimmer & Perner (1983) represents one of the earliest and most famous versions of a task testing this ability (Fig. 1). In the task, a participant (usually a child) is shown two dolls (or in other versions of the task, real people) named Sally and Ann. Sally has a marble, the participant is told, and Sally is placing it in her basket. When Sally leaves the room, Ann moves the marble to her box. The participant is then asked where Sally will look for the marble when she returns. If they are able to model the
false beliefs of Sally accurately, they will answer that Sally will look in the basket. However, if they are unable to represent Sally’s beliefs as different from their own knowledge about the state of the world, they will answer that Sally will look in the actual location of the marble, in the box. It has been shown that neurotypical children age 4 and above are consistently able to pass the Sally-Ann task (Wimmer & Perner, 1983), while children with autism are unable to do so even past age 5 (Baron-Cohen et al, 1985), suggesting that they are unable to ascribe beliefs correctly to others. Many studies have been conducted since the 1980s aiming to better characterize how ToM develops in children (Flavell, 2004; Frith & Frith, 2003). By age 5, nearly all children are able to represent second-order beliefs (for example: “He thinks that she thinks that the ball is in the basket”) in other people (Sullivan et al, 1994). Interestingly, it appears that prior to this age (between ages 2-4), the implicit behavior of many children indicates that they possess ToM (for instance, glancing to the spot where a person mistakenly believes candy is stashed), while their explicit answers indicate no ToM abilities (Clements & Perner, 1994; Garnham & Perner, 2001). Children as young as 18 months old will ascribe intentions to adults attempting to perform a task, which they will not ascribe to a mechanical device performing the same actions (Meltzoff, 1995). Even infants couple words and objects by tracking a speaker’s gaze, signaling their recognition of the intention of the speaker to refer to a specific object (Baldwin et al, 1996; Bloom, 2000) – although how explicit this recognition is can be disputed.
While much behavioral work has been done examining how ToM develops with age, ToM is still poorly understood on the neural level. However, there are a handful of brain structures that have been identified as belonging to a “mentalizing network.” This network primarily contains the temporal parietal junction (TPJ), the superior temporal sulcus (STS), the precuneus, the anterior temporal lobe (ATL), and the medial prefrontal cortex (mPFC) (Rushworth et al, 2013; Koster-Hale & Saxe, 2013; Sallet et al, 2012; Van Overwalle & Baetens, 2009; Hampton et al, 2008; Rilling et al, 2004; Frith & Frith, 2003; Frith, 2001). The first studies to examine ToM neural correlates used paradigms in which participants were tasked with answering questions about stories or cartoons which required ToM to fully comprehend, while neural data was collected using positron emission tomography (PET) (Fletcher et al, 1995; Brunet et al, 2000) or functional magnetic resonance imaging (fMRI) (Saxe & Kanwisher, 2003; Gallagher et al, 2000; Vogeley et al, 2001). One seminal study used computer animations of geometric objects which participants either attributed mental states to or gave physical descriptions of while PET scanning was conducted (Castelli et al, 2000; Heider & Simmel, 1944). Increased activation was found in the mPFC, the TPJ, extrastriate cortex, and the ATL, even though the geometric shapes in no way resembled human figures, suggesting that the mentalizing network may be concerned more with intentions and predicting actions than with face processing or some other form of human-specific cognition. In another study which used fMRI, humans played games well-known in behavioral economics (the Ultimatum Game and the Prisoner’s Dilemma Game), requiring them to represent the intentions of a partner that was either human or a computer algorithm (Rilling et al, 2004). Activation in the anterior paracingulate cortex, the STS, and the precuneus, among other areas, exhibited stronger activation when participants were aware of having a human partner for both games, indicating some difference in the circuits employed to think about a human partner versus a computer. This was the case even though the optimal strategies for dealing with a human versus a computer algorithm were the same. However, it may well be that although optimal strategies for dealing with a human versus an algorithm are identical, individuals changed their strategies for the human, eliciting different activations despite not explicitly representing ToM.

The mentalizing network has also been implicated in determining moral responsibility (Young & Saxe, 2007; 2009; Koster-Hale & Saxe, 2013). Studies of cognitive development in children will often examine ToM in the context of moral responsibility, because assigning moral
weight to someone’s actions often depends on understanding that individual’s frame of mind (Cushman, 2008). For instance, Thomas Shultz and colleagues (1986) asked children (aged 5-11 years) to imagine that they had been harmed by another child (for instance, that the child had broken a toy of theirs) and to decide how much punishment this child should receive. All the children sought more punishment for deliberate harm than negligent or accidental harm – showing that they could incorporate information about the intentions of others into their assessments of a situation.

An fMRI study assessing the brain activation of participants as they read vignettes in which individuals caused either negative or neutral effects, while they believed that they were causing either negative or neutral effects (a 2x2 experimental paradigm), showed that the right TPJ responded the most to cases in which agents intended a negative effect but actually caused a neutral one (Young & Saxe, 2007). This suggests the mentalizing network and the right TPJ in particular uses information about intentions to compute ToM and representations of moral responsibility in others – perhaps generating some type of moral prediction error calculation that quantifies the difference between intentions and real-world results. More broadly, a study in which participants read “moral” or “nonmoral” facts about an action’s effect or a situation showed that the right TPJ, the precuneus, and mPFC are recruited during moral considerations selectively (Young & Saxe, 2009). Even more strikingly, a causal relationship between the right TPJ and moral judgment was established which it was shown that disruption of that area with the use of transcranial magnetic stimulation (TMS) causes individuals to rate events in which a person intends harm but fails to cause it as more morally permissible than otherwise (Young et al, 2010) – in other words, it appears that the TPJ is responsible at least in part for integrating beliefs held by others into calculations of moral responsibility. However, this relationship is far from completely accepted in the literature. For example, one recent study found that TPJ activity was modulated by the re-orienting of attention, even in non-social situations, suggesting that the role of the TPJ is more complex than at first glance (Mitchell, 2008).

While the establishment of ToM over the course of childhood has been well characterized, and key brain structures isolated as a part of the mentalizing network, the precise calculations underlying ToM remain poorly characterized. As the bulk of research surrounding ToM to date has used vignettes, whether in the form of stories, cartoons, or animations to examine how representations of the mental states of others are built up (e.g. Vogeley et al, 2001;
there is little information on how ToM updates over time as individuals learn new information about a person, or how different components of ToM may update differently. For instance, it is possible (although unlikely) that neurotypical humans update information about the beliefs of others in a way that is completely different from how they update information about the intentions of others, although beliefs and intentions are both taken to be parts of ToM. Furthermore, a learning paradigm affords more opportunities to model representations of thought that are uncertain or incomplete, and to examine the process by which they become more solidified.

**ASD and the Link to ToM**

Autism Spectrum Disorder (ASD) is a social developmental disorder, which occurs in around 1 in 68 children in the US at varying degrees of severity, according to the Center for Disease Control (Baio, 2014). Although symptoms of ASD, which include abnormal social interaction and communication deficits, have been well catalogued, the underlying causes remain poorly understood. In particular, it is understood that as a spectrum disorder, ASD remains a label for a population exhibiting a great deal of individual variance, although a difficulty with representing the mental states of others, especially beliefs, is known to exist (Schuwerk, Vuori, & Sodian, 2015; Yoshida et al, 2010b; Moran et al, 2011; Sigman et al, 2006; Baron-Cohen et al, 1997a, 1997b; 1985; Adolphs et al, 2001; Leslie, 1987). Many studies have shown a link between ASD and impairment in ToM. Children with ASD cannot perform the Sally–Ann task (Fig. 1) like their neurotypical peers (Baron-Cohen, 1985), even though they can perform a similar task where they are asked if a photograph accurately reflects a scene that is changed after the photo was taken, which requires no modeling of the mental states of others (Leslie & Thaiss, 1992). Even in adulthood, impairments exist that manifest themselves in performance on various ToM tasks – for instance, in a version of the geometric shapes animation ToM task (originally in Heider & Simmel, 1944), participants with ASD provided less accurate descriptions of the shapes with goals or intentions than neurotypical participants, and also exhibited less PET activation in the mPFC, the TPJ, and the temporal poles (Castelli et al., 2002). On the task in which individuals needed to judge how morally permissible an action was depending on whether the intentions were neutral or negative and the outcomes were neutral or negative, the ASD group was not able to differentiate accidental and attempted harms (Moran et al., 2011).
Likewise, in a social hunting game where two players choose to either hunt high-value “stags” together or split up to hunt low-value “rabbits,” individuals with ASD showed an impairment with representing the strategy of the other player – an impairment that correlated with ASD symptom severity (Yoshida et al., 2010b).

It has also been shown that individuals with ASD do not alter their choices about whether to donate to charity in the presence of an observer unlike neurotypical controls, indicating an insensitivity to social reputation that may be due to an inability to model the mental states of others (Izuma et al., 2011). When performing the “reading the mind in the eyes” task, where participants must ascribe emotions to closely-cropped photographs of eyes, individuals with ASD are consistently impaired to a degree that correlates with the severity of their symptoms, indicating an inability to extract social information about the mental states of others from the eyes (Baron-Cohen et al., 1997b, 2001a). A similar results was found in a study using photographs of entire faces (Baron-Cohen et al., 1997a).

However, one study showed that ASD performance is better for explicit ToM tasks than implicit ones, and that with training implicit ToM task performance can improve – leading to a hypothesis that perhaps explicit and implicit ToM are processed differently and that implicit ToM can be positively affected by experience (Schuwerk et al., 2015). In a similar vein, it has been suggested that because individuals with ASD perform better on a competitive ToM game than on the Sally-Ann task, which in non-competitive, deficits in ToM performance may be due to a lack of proper incentives (Peterson et al., 2013). Another study compared performance on the evaluation of emotions and trustworthiness in faces between individuals with ASD and individuals with amygdala lesions, and found similarities implying that amygdala damage may play a role in the social impairments symptomatic of ASD (Adolphs et al., 2001).

It is undisputed that incidence of ASD is correlated with impairments in social cognition. If there were a difference in how people with ASD compute ToM in comparison to neurotypical people, this would serve to explain a great deal of the findings of previous studies – findings that are for the most part behavioral and which are not conclusive (Fisch, 2013). However, the precise computations that the mentalizing network undertakes to represent ToM in neurotypical people are poorly understood, and it is unclear how much variance lies in the impairments exhibited in ASD (Yoshida et al., 2010a). Quantifying the variance in impairments in ASD, as well as investigating the places in which calculations fail to occur or different calculations take
place would significantly contribute to understanding how individuals with ASD think about the mental states of others and use that information in a social context.

**The Computational Approach**

In order to better understand ToM in both neurotypical and ASD populations, this thesis employs a computational approach by attempting to model how individuals update ToM over time given new information, using principles of model-based learning. Reinforcement learning is a method of decision making where the goal is to maximize the total reward gained based on probabilistic assessments about the environment (Sutton & Barto, 1998). To date, several subfields in neuroscience have made use of reinforcement learning to better understand how the brain makes predictions. The dopaminergic system, in particular, has been explored in depth using concepts from reinforcement learning such as prediction error (the amount by which expected reward diverges from actual reward) (Niv, 2009; Niv & Schoenbaum, 2008; Shultz et al., 1997).

Dopaminergic neurons in monkeys will fire when an unexpected reward occurs, but when the reward is expected the dopaminergic neurons will only fire for a predictive cue, not the actual reward (Shultz et al., 1997). When a reward fails to occur when expected, these neurons will exhibit inhibition. Thus dopaminergic neurons can be said to encode prediction error, and their behavior has been modeled accurately using a temporal difference (TD) model, which is a type of reinforcement learning model (Sutton & Barto, 1990; Shultz et al., 1997; Morris et al., 2006).

A TD model, like many reinforcement learning models, is based on the Rescorla-Wagner update rule (Rescorla & Wagner, 1972). The Rescorla-Wagner update rule is as follows:

$$ V_t = V_{t-1} + \lambda(\alpha - V_{t-1}) $$

where $$ V_x $$ refers to a representation (for instance, of expected reward) at a given time $$ x $$, $$ \alpha $$ refers to the actual outcome (here, reward received) and $$ \lambda $$ is a learning rate, which affects how deeply the information about the actual outcome affects the next prediction of reward. So a representation of expected reward at time $$ t $$ is equal to the last expected reward plus the discrepancy between the last expected reward and the actual reward at that time, modulated by a learning rate. This example is one of the simplest possible reinforcement learning models, which can be altered in a variety of ways to suit a given paradigm (O’Doherty, 2012). To return to the example of dopaminergic neurons, a TD model takes the basic framework of a Rescorla-Wagner model, but instead of computing the next predicted reward, calculates the predicted reward over
a longer span of time by summing across all future estimates. Therefore a TD model takes the form:

$$\hat{V}_t = \sum_i V_i$$

where $i$ goes from time $t$ to some time in the future. This type of model is good for encoding the activity of dopamine neurons because it is able to take into account situations in which there is a predictive cue, a time delay, and then a reward (Shultz et al., 1997).

While the behavior of dopaminergic neurons has provided some of the strongest evidence for the use of prediction error and other principles of reinforcement learning in the brain when learning about reward (Sutton & Barto, 1990), computational modeling is applicable to many other types of learning, including social learning. Examining the goodness of fit of models to behavioral data can help to advance hypotheses about what types of computations, sub-optimal or otherwise, are occurring neurally (for example, Wilson & Niv, 2011; Strauss et al., 2011; Wunderlich et al., 2009; Seo & Lee, 2008; Daw et al., 2006)

Some criticisms have been leveled at the practice of fitting reinforcement learning models to behavioral data. For instance, while computational neuroscience has gained a good deal of information about the functional organization of the brain and how it calculates prediction errors, historically it has struggled to cope with situations where uncertainty, regret or other such nuances have effects (Dayan & Niv, 2008). Furthermore, there exists the perennial problem of finding models which tread the line between being simple enough to be easily understood but failing to capture all the variance in the data, and being so complex that they are able to mimic all the data accurately but interpreting them becomes nearly impossible.

In terms of inherent logistical flaws, there has also been push-back against “double dipping” during model fitting, which occurs when the same data set is used both to select the best models and for analysis of those models, resulting in artificial effects (Kriegeskorte et al, 2009). Luckily, double dipping can be avoided by simply dividing data sets into two, allowing for separate selection and analysis groups. There have also been concerns of “model-hacking,” or the over-fitting of models, which results in extremely good fits to the data but fails to preserve any generalizability. However, these concerns with applications of computational modeling to neuroscience do not represent irreparable problems. Rather they are simply issues of which to be aware when conducting a study that uses such an approach.
Social learning, like other types of learning, has been modeled in a variety of ways using these techniques. For instance, a computational model of the behavioral and neural data of a task in which participants took or ignored the advice of a partner when making choices has suggested that humans create representations of social value using reward-based associative learning techniques (Behrens et al., 2008). Another interesting study examined observational learning (the process by which an individual learns about an environment by watching the fate of a confederate) and suggested that two types of prediction error may be in use in such a situation: the “observational action prediction error” (actual outcome – predicted choice of others) and “observational outcome prediction error” (actual outcome – predicted outcome of others) (Burke et al, 2010). Such studies provide compelling evidence that computational modeling can contribute relevant information to our understanding of how social representations are formed and updated.

ToM specifically has also been investigated in a computational context. Computational modeling has been paired with ToM to examine how individuals represent the mental states of others in competitive scenarios requiring game theory (Yoshida et al., 2008), as well as to investigate how children understand the choices of an agent through the lens of the costs and benefits of the possible actions available to them (Jara-Ettinger et al., 2015), and to explore how individuals will rationalize the actions of others through backwards-induction of their planning process (Baker at al., 2009), to mention only a few examples. To unravel the neural correlates behind the calculations that result in real-world behavior, behavioral data is first collected, candidate models are fit to the behavioral data, and then neuroimaging data is analyzed through the lens of the best models to attempt to isolate which brain structures may be responsible for specific components of the calculations conducted (Behrens et al, 2009; O’Doherty et al., 2007). By taking this approach, it is possible to pair specific components of a cognitive process with specific brain areas (Boorman et al, 2013; Nicolle et al., 2012; Burke et al, 2010; Wunderlich et al., 2009; Hampton et al., 2008; 2006; Behrens et al., 2008; O’Doherty & Bossaerts, 2008; Daw et al., 2006)

As mentioned above, there have been many studies examining ToM behaviorally and neurally. However, there have only been a few attempts using the computational approach in conjunction with neuroimaging to attempt to better understand the roles of individual brain structures within the mentalizing network. In one such study, Alan Hampton and colleagues
(2008) scanned participants in using fMRI while they played a competitive strategy game with an opponent, finding that mPFC activity corresponded to predicted reward for a given choice, while posterior temporal sulcus activity correlated with prediction error. This suggests that the mPFC may be responsible for modeling the complete representation of another’s state of mind while other areas compute the components which form this representation. However, another study found evidence for representations of the prediction error of others in the mPFC (Suzuki et al., 2012). A third study which asked participants to make choices on the part of a partner added an important caveat to this finding: while the study corroborated the finding that when performing a competitive task a specific area in the mPFC reflects the prediction error of the opponent, when an individual is asked to consider the interests of another individual and make choices for them, the neural areas which model the prediction error of the self and which model the prediction error of the other individual switch places (Nicolle et al., 2012). So it may be that the ToM area of the mPFC actually serves as a location to model prediction error for an intelligent opposing agenda, rather than for other minds themselves. Yoshida et al. (2010b), using the competitive stag/rabbit hunting game described above, also found a correlation with uncertainty about an opponent’s mental state in the mPFC – interestingly, they found that activity in the dorsolateral prefrontal cortex correlated to the level of recursion a participant was employing about their opponent’s strategy; in other words, “I think he thinks that I think that he thinks that…” saw more activation than “I think that he thinks that…” Therefore it seems that the more sophisticated the ToM processes being undertaken, the more the dorsolateral prefrontal cortex is recruited. Yet despite these few studies, the precise function of all the structures implicated in the mentalizing network remains nebulous.

A Novel Paradigm

While a few studies, discussed above, have used a computational approach to study ToM (Suzuki et al., 2012; Nicolle et al., 2012; Yoshida et al., 2010b; Hampton et al., 2008), these studies have all relied on competitive or cooperative games, usually borrowed from behavioral economics, to study how individuals represent the thought processes of others. These games provide powerful incentives for individuals to accurately predict the actions of their opponents or partners, but they have the potential to present a variety of confounds associated with earning reward. They also are functionally unable to isolate different components of the calculations that enter into ToM –
each trial is simply a guess about the end-product in the form of a prediction about the behavior of the other individual. For these reasons, among others, this thesis uses a novel task, the Mentalizer task, which is designed to eliminate extraneous confounds while providing isolated components of ToM which can each in turn be modelled.

Past ToM studies have used vignettes to great effect, requiring participants to represent the beliefs and intentions of independent third-party individuals (Agents) in order to think about the choices the Agents are making (Young et al., 2007; Wimmer & Perner, 1983; Fletcher et al, 1995; Brunet et al, 2000; Saxe & Kanwisher, 2003; Gallagher et al, 2000; Vogeley et al, 2001; Castelli et al, 2000; Hieder & Simmel, 1944). The Mentalizer task used here echoes these studies by giving participants the same goal of representing beliefs and intentions in an Agent to predict the Agent’s choices, but unlike past research, this task contains many trials as opposed to a single vignette, during each of which the participant (the Mentalizer) must make judgments about the Agent’s state of mind and receives new information with which to update their representations of the Agent’s beliefs and intentions. On each trial, the Mentalizer is explicitly asked about the belief and the intent of the Agent, and then what choice the Agent will make (a question which should be directly predicted by the answers to the belief and intent questions, thus enabling the assessment of internal consistency in Mentalizers). Importantly, it is possible for the Agent to have false beliefs, and as the Mentalizer knows the actual state of the world, the Mentalizer is capable of recognizing these beliefs as such. This means that the Mentalizer is forced to represent these beliefs as separate from the actual state of the world. It is for this reason that false belief has long been the gold standard for ToM research (Saxe & Kanwisher, 2003).

Therefore the Mentalizer task is a novel social learning task in which belief, intent and actual prediction are isolated, enabling them to be computationally modeled separately from one another. The implementation of dynamic representations, as well as the fact that the Mentalizers must represent varying degrees of uncertainty over time, serves to mimic real life use of ToM more accurately than the single vignettes used in previous research. This is crucial, because if modeling of ToM is to be useful, the neural behavior of Mentalizers must be as close to what it would be in real social situations as possible. However, unlike past attempts at computational modeling of ToM, the task does not have a competitive or cooperative element, eliminating any confounds that might be associated with earning reward in these situations. Therefore, the
Mentalizer task is well suited to better elucidating the mechanisms of ToM in both neurotypical and ASD individuals.

That being said, there are limitations to this paradigm. For instance, while the task is structured to be more ecologically valid than previous tasks which relied solely on vignettes without implementing learning, it still fails to perfectly mimic the experience of implementing ToM in the real world. However, it remains a closer approximation of the nuances of ToM than most of the previous literature. The complexity of the task also signifies that it is a challenge to ensure that all participants understand it completely before collecting behavioral data. This has been addressed by having Mentalizers complete a worksheet designed to assess understanding beforehand, but it is still possible that some confusion with the task design remains.

This thesis is also fairly liberal with its use of broad ToM terms – especially “belief” and “intent.” In order to avoid issues surrounding of what exactly constitutes “belief” or “intent,” when the task was explained, the use of these words was avoided (see Appendix A for exact wording). This was also intended to side step the question of whether or not an individual can truly “intend” an action if their knowledge of the state of the world is uncertain. In this thesis, these terms are used simply as shorthand to refer to the different questions posed to Mentalizers and should not be taken as references to absolute definitions.

While this thesis is solely concerned with behavioral data and the fit of models to this data, it is important to note that the Mentalizer task is fully viable for neuroimaging purposes, and when fMRI data is collected of participants, analysis of this data in conjunction with the computational models of best fit to the behavioral data will yield information about the specific roles of brain structures within the Mentalizing network. As currently little is known about how these structures enact ToM this thesis represents an opportunity to better understand how mental representations of the thoughts of others are constructed, as well as a chance to investigate how these representations are different in individuals with ASD.
CHAPTER 2: METHODS
All computations were performed in MATLAB (version 2010a & 2014a; Mathworks Inc., Natick, MA, USA), and behavioral data was collected using Psychtoolbox (Brainard, 1997; Pelli, 1997). (See Author Contributions section for more details on allocation of labor).

Participants

Participants were recruited from the established database of the Adolphs lab (California Institute of Technology). 53 neurotypical (40 male) individuals and 26 ASD (21 male) individuals each completed one Charity task (acting as the Agent) and one Mentalizer task (acting as the Mentalizer). For participants to be included in the ASD group, they were required to be over the age of 18, to be verbal, and to meet the criteria for ASD according to the ADOS test (Autism Diagnostic Observation Schedule, Module 4; Lord et al, 2000) as well as the criteria for Asperger’s Syndrome or Autism according to the Diagnostic & Statistical Manual of Mental Disorders, 4th edition (DSM-IV, American Psychiatric Association, 1994). All participants also possessed a full-scale IQ score (FSIQ) above 85, determined with one of the versions of the Wechsler Adult Intelligence Scale: the Wechsler Abbreviated Scale of Intelligence, 1st or 2nd edition (Wechsler, 1999, 2011); the Wechsler Adult Intelligence Scale, 3rd or revised edition (Wechsler, 1981, 1997); or the Wechsler Intelligence Scale for Children, 3rd edition (Wechsler, 1991; 1 participant). Controls were psychologically and neurologically healthy individuals matched to the ASD group on age, gender, years of education, and IQ (Table 1), with no family history of ASD. The control group was then split into two groups for model fitting, which were balanced across each other in terms of the same metrics, as well as which version of the task they performed (old vs new) (Table 1). The purpose behind the splitting was to fit the models on one group and then assess to goodness of this fit on the other group in a technique known as cross fitting, to avoid concerns associated with double-dipping.

All participants had normal or corrected-to-normal eyesight and gave informed consent as approved by the California Institute of Technology Institutional Review Board. All participants successfully completed the task and were paid $20/hour as well as any bonus money earned over the course of the Charity or Mentalizer tasks.
Behavioral Data Collection

The main experimental protocol consists of a task-within-a-task. The inner task (the Charity task) takes place over 121 trials during each of which the participant (the Agent) chooses whether to give to a charity or to keep money for themselves. However, at any given time the program is in one of two modes, which alternate periodically without the knowledge of Agents (after the first 21 trials, in which the mode is revealed in order to allow the Agent to practice).

In the normal mode, the Agent’s choices are implemented 65% of the time, and reversed 35% of the time. In the reversal mode, the Agent’s choices are reversed 65% of the time and actually happen only 35% of the time.

Importantly, the current program state switches every so often, but is stable enough that the Agents had time to learn. In order to get the choices they want, Agents must keep track of the recent history of reversals versus normal outcomes so that they can continuously update their belief about what state they are in.

In the outer task (the Mentalizer task) participants (Mentalizers) watch an Agent perform the Charity task while mentally updating representations of the Agent’s beliefs and intentions from trial to trial. Because the Agent must infer which one of the two possible modes is in effect at any given time, whereas the Mentalizer is always kept informed about what the true mode is, there are opportunities for the Mentalizer to represent the agent’s beliefs as false. The Mentalizer must take into account both the Agent’s current belief about the mode

<table>
<thead>
<tr>
<th></th>
<th>MC1 (N=27)</th>
<th>MC2 (N=26)</th>
<th>ASD (N=26)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>mean</td>
<td>30.26</td>
<td>30.46</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>9.14</td>
<td>8.85</td>
</tr>
<tr>
<td>Years of education</td>
<td>mean</td>
<td>14.89</td>
<td>14.92</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.39</td>
<td>1.41</td>
</tr>
<tr>
<td>Male</td>
<td>#</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>PIQ</td>
<td>mean</td>
<td>105.26</td>
<td>105.62</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>10.67</td>
<td>11.28</td>
</tr>
<tr>
<td>VIQ</td>
<td>mean</td>
<td>108.11</td>
<td>108.12</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>10.47</td>
<td>9.88</td>
</tr>
<tr>
<td>FSIQ</td>
<td>mean</td>
<td>107.52</td>
<td>107.96</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>9.76</td>
<td>10.18</td>
</tr>
<tr>
<td>AQ</td>
<td>mean</td>
<td>20.33</td>
<td>24.01</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>16.62</td>
<td>21.68</td>
</tr>
<tr>
<td>EQ</td>
<td>mean</td>
<td>48.69</td>
<td>53.42</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>13.03</td>
<td>11.38</td>
</tr>
<tr>
<td>SQ</td>
<td>mean</td>
<td>71.15</td>
<td>75.00</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>21.49</td>
<td>25.56</td>
</tr>
<tr>
<td>EYES Score</td>
<td>mean</td>
<td>27.30</td>
<td>27.68</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>3.57</td>
<td>3.78</td>
</tr>
<tr>
<td>SNI Network</td>
<td>mean</td>
<td>4.72</td>
<td>4.80</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.62</td>
<td>1.89</td>
</tr>
<tr>
<td>SNI PPL</td>
<td>mean</td>
<td>19.44</td>
<td>26.00</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>10.26</td>
<td>28.52</td>
</tr>
<tr>
<td>SNI Embed</td>
<td>mean</td>
<td>2.12</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.51</td>
<td>1.94</td>
</tr>
</tbody>
</table>

Table 1. Participant Demographics. MC1, MC2 = Matched Controls groups 1 and 2; ASD = Autism Spectrum Disorder. Shaded cells indicate this measure was used to match groups.
and their hypothesis about the Agent’s intentions when predicting choice for maximum accuracy (for instance, it is possible that the Agent thinks that the state is reversal, and so they picked the charity intending for the decision to be reversed and for the money to actually go to themselves). On each trial before the Agent’s choice is revealed, the Mentalizer is asked to answer 3 questions: 1) does the Agent intend to donate to the charity or not, 2) what mode does the Agent believe to be in effect, and 3) what choice will the Agent make.

This task structure mimics traditional ToM tasks, ensuring that Mentalizers must represent the true and false beliefs of Agents, and integrate their representations of Agent belief with that of Agent intent in order to predict the Agent’s actions (Fig. 1, Young et al., 2007). During each
trial, the Agent is offered a choice, the Mentalizer is asked about what they think the Agent’s beliefs about the program mode and intentions toward the charity are, and then Agent’s actual choice and the trial outcome is revealed. The task as a whole consists of either 80 (old version) or 84 trials (new version), plus 21 initial practice trials. However, only the first 80 trials were used for analysis (excluding practice trials), so the old and new versions of the task are functionally equivalent. 33 control participants performed the old version and 20 performed the new version, and splitting of matched controls for the purposes of analysis was balanced in terms of new/old versions of the task in the two groups. 11 out of 26 ASD participants performed the new version.

In addition to performing the Mentalizer task, all participants completed the “reading the mind in the eyes” task (Eyes task; Baron-Cohen et al., 2001a) as well as a standard IQ test yielding a PIQ (performance intelligence quotient; nonverbal intelligence) score and a VIQ (verbal intelligence quotient) score, as well as a FSIQ (full scale intelligence quotient) score (Table 1). In addition, all participants completed tests for AQ, EQ and SQ (Autism Spectrum Quotient, Empathy Quotient, Systemizing Quotient) scores (Wheelwright et al., 2006; Goldenfeld et al., 2006; Lawrence et al., 2004; Baron-Cohen et al., 2003; 2001b). They also were scored on the Social Network Index (SNI) for the diversity of their social networks (SNI Network), the number of individuals in their networks (SNI PPL) and the number of embedded networks (SNI Embed) (Table 1; Cohen et al., 1997).

On the day of the experiment, participants first completed a consent form, followed by a demographics questionnaire and finally a “Day of Visit” screening questionnaire containing questions on sleep, drugs or medications taken, and history of psychiatric diagnoses, concussions and seizures to assess their state of mind. Participants then received an extensive powerpoint briefing to familiarize them with the structure and rules of the Agent task. They next completed a quiz with the experimenter comprised of 4 questions to assess comprehension. For any questions that were answered incorrectly, the experimenter went through the logic of the task once more with the participant. Participants then completed the Charity task. Following the Charity task, participants received a second powerpoint briefing describing the Mentalizer task (for exact wording, see Appendix A), and then completed that task, which contained a different set of trials.

Upon completion of the Mentalizer task, participants filled out a questionnaire in which they indicated the preferences of the Agent they were observing for each charity (i.e. donate or take),
and then provided information about any strategies they used to learn about the agent. Once this was completed, the experimenter randomly selected one trial from the Charity task and one trial from the Mentalizer task and, depending on the task, paid the agent the true outcome (Charity task) or a reward of 5$ per correct answer (3 possible) and $0.50 for each correct probe trial (Mentalizer task). Finally, Participants completed on any of the following surveys if they had not done so on a previous visit: AQ, EQ, SNI.

Agent Generation

Agents presented during the Mentalizer task were generated by combining beliefs from real participants on the Charity task with engineered preferences for certain charities. Initial novel Agents were brute force generated by creating 10,000 sets of paired modes and outcomes, with two set average lengths of mode duration (4 trials and 10 trials) with a range of 3 and 5 respectively. The total number of trials was set to 84 (28 for each charity). Block (set of consecutive trials with the same mode) lengths were randomly generated in such a way that each mode was in effect for 42 trials overall. The outcomes of each trial (whether or not the Agent’s decision was executed or reversed) were generated by setting the total number of executed trials per mode to reflect the mode’s probability (either .65 or .35) as closely as possible over all trials with that mode, then randomly distributing those trials within all the blocks with that mode. The outcome data was then fitted to a basic model of belief (for more, see the details on the RW belief model in Reinforcement Learning Models). This model assumed belief before the first outcome was revealed on the first trial to be equally biased towards both modes, with a learning rate of 0.3.

This yielded 10,000 sets of modes, outcomes, and agent beliefs. The bottom 95% of these sets with the least stable agent beliefs (in other words, with the most switching of beliefs from one mode to another over the 84 trials) were cast aside. In order to calculate this, agent beliefs were treated as binary. Out of the remaining 500 sets, those with agents who possessed false beliefs over 30% to 40% of the trials were selected. These sets were then assigned charity information. For each set, 100 permutations of the order in which the three types of charities were presented and the amount of money the agent is offered for themselves ($7-13) were randomly generated. In none of these permutations was there more than 2 consecutive repeats of the amount the agent was offered or of the charity type, and the exact same trial was never
repeated consecutively. The amount of money offered for the charity was kept constant across all trials at $10, and the side of the screen on which the charity would be presented was randomly assigned for each trial.

Each of the 100 permutations of charity information for each of the remaining sets were then randomly assigned to have agents prefer one charity over themselves 100% of the time and vice versa for the other two charities, or to prefer two charities over themselves 100% of the time and vice versa for the remaining one. They were then fitted to a model for agent intentions and expected outcomes. Since charity preferences were set to either 100% or 0%, for all trials involving a specific charity the agent was given the same intent. The agent’s choice relied on the interaction of these intents with the beliefs garnered from the Rescorla-Wagner belief model (on the first trial the agent was set to choose randomly). The agent’s expected outcome of each trial was a function of what mode the agent believed was in effect and what choice they made. Finally, the RW/RW model of a Mentalizer estimating the agent’s beliefs and intents was fitted to the agents and permutations in order to select out the most learnable sets. The mean of the model’s intent predictions in the 4th quarter of total trials, compared to the agent’s actual intent, was used as an indicator of the agent’s “learnability.”

The top 20 sets with the highest learnability were selected out and pared down by using each for the Charity task on pilot subjects. The final 10 Agents selected were those for which participants exhibited the greatest accuracy. Due to an algorithmic error, the fact that beliefs are collected from Charity task participants after the trial in which they used that belief was not corrected for, so the constructed Agents using these beliefs access them one trial ahead of time. Although this resulted in Agents who technically possess one trial’s worth of precognition for the entire duration of the task, it is unlikely that this drastically affect Mentalizer performance as the task already contained much uncertainty to begin with, and the mental states of others are to a certain extent stochastic. Furthermore, Mentalizers did not receive explicit feedback during the task about their performance on belief.

Models of Learning

Computational models simulating the updating of belief and intent over time were used to investigate the updating process of Mentalizer’s ToM representations of the Agents. These models were structured so as to generate values for the signals that will later be investigated in
neuroimaging data – these signs being primarily uncertainty and prediction error. Uncertainty here refers to how sure the Mentalizer is of the accuracy of any of their representations of Agent belief or intent, and prediction error refers to the difference on each trial between what the Mentalizer predicted and what actually happened (in other words, how surprised the Mentalizer was by the outcome). Models used a Rescorla-Wagner update rule of reinforcement learning (Rescorla & Wagner, 1972) for belief and intent calculations, which was altered in a variety of ways to reflect different ways that these values may be computed.

The belief model focused on the program state aspect of the Mentalizing task – in which Mentalizers, watching the Agents, attempt to ascertain what the Agent’s current belief about program states is and to use this prediction to inform their interpretations of other information about the Agent. To mimic this process, the belief model was designed as a flexible model that used prediction error (calculated as the difference between trial outcome at time t (OCt) and representation of belief prior to that trial (Bt-1), in essence assuming that the more surprised by an outcome the Mentalizer is, the more likely they are to update their representation of belief. This model was able to accommodate rapid switches of belief between one program state and another, and was fitted to individual data using a learning rate $\lambda_{Bel}$.

**Basic Belief Model**

$$B_t = B_{t-1} + (\lambda_{Bel}) \times (OC_t - B_{t-1})$$

The same mathematical principles were used to represent intent, simply computing prediction error for intent and updating the representation of charity preference on each trial accordingly.

**Basic Intent Model**

$$I_t = I_{t-1} + (\lambda_{Int}) \times (OC_t - I_{t-1})$$

All models had a single parameter to fit (learning rate), whose value was determined by minimizing the negative log likelihood generated from the difference between the model’s predictions and the actual answers given by the Mentalizer participant. Thus for each set of data it was possible to arrive at two optimized learning rates: one for belief and one for intent.

Ultimately, eight models were selected for analysis: RW/RW, RW/RW_overT, RW/REAL_overT, RW/NONE_overT, RW_overT/RW, REAL/REAL_overT, normal/NONE_overT and reversal/NONE_overT. RW/RW is the model implementing the basic Rescorla-Wagner update rule for both belief and intent described above. RW/RW_overT uses the
same basic belief model, but the intent model gives earlier trials more weight than later trials. This is achieved by dividing the learning rate by the number of times a given charity has been presented. RW/REAL_overT uses the same scaled intent as RW/RW_overT and the same belief
model, but the belief calculations that go into the intent calculations are not those from the belief model, but the real state in effect at any given trial, which the Mentalizer is fully aware of. REAL/REAL_overT is nearly identical to this model except that it does not implement a belief model at all – instead using the real state for both belief and intent. RW/NONE_overT calculates intent without using belief at all – in other words, it takes the choices of Agents at face value rather than considering that if the Agent believes reversal mode is in effect, they will pick the opposite of what they truly want. Both normal/NONE_overT and reversal/NONE_overT are the same as RW/NONE_overT except that instead of using a standard Rescorla-Wagner model for belief, they assume that all trials are on normal mode or on reversal mode, respectively. Finally, RW_overT/RW uses a basic intent model, but belief is calculated using a scale learning rate, so that the later a trial occurs in the task, the less weight it brings to the belief calculation.

Figure 3 depicts how learning rates were optimized on MC1, MC2, and ASD for the RW/RWOverT model as an example. The same process was applied to all the other models as well. Optimized belief and intent learning rates were arrived at using a grid search approach where first a broad search over the 0-1 belief and intent learning rate space was conducted at intervals of 0.1, and then the minimum log likelihood (taken to be the indicator of the model with the best fit to the data) was used to conduct a second, narrower gridsearch (intervals of 0.01).

Negative log likelihood was calculated using the following equation:

\[ \text{Negative log likelihood} = \sum \left[ - \log \left( \frac{e^{-|\text{mentVals} - \text{modVals}|}}{\sum (e^{-|\text{modDiff}|})} \right) \right] \]

where \( \text{mentVals} \) represents a matrix of the actual answers of the participant, \( \text{modVals} \) represents the values given by the model, and \( \text{modDiff} \) is a matrix with one row being model predictions and the second row being \((1 - \text{model predictions})\).

The Bayesian Information Criterion (BIC) was used as a metric to assess goodness of fit of the models. This value is based directly on the negative log likelihood:

\[ \text{BIC} = 2 \times (\text{negative log likelihood}) + \text{numParameters} \times \log(\text{numData}) - \log(2\pi) \]

where \( \text{numParameters} \) is the number of parameters going into the fitting of the model, and \( \text{numData} \) represents the number of data points being assessed (so in this case the number of trials multiplied by the number of questions being asked).

For the code used to run all computational models, see Appendix B.
RESULTS

*Behavioral Data Analysis*

The accuracy of an ASD group (N=26) and a control group (N=63) on the Mentalizer task was assessed in terms of belief, intent, and choice. These metrics were isolated through analysis of answers to three questions asked on every trial:

1) What program mode, normal or reversal, does the Agent **believe** to be in effect on this trial?
2) Does the Agent **intend** the money to go to themselves or to the charity?
3) Will the Agent **choose** themselves or the charity?

Importantly, it was necessary for Mentalizers to integrate their answers for the belief and intent questions in order to answer the choice question accurately. For example, a belief in reversal mode would indicate that the Agent would ultimately choose the opposite of what they wanted to occur, so the Mentalizer would need to flip their choice prediction answer.

Both control and ASD populations performed well above chance on the belief question but not significantly differently from one another (Fig. 4a). Bootstrapped 95% confidence intervals

![Figure 4. Behavioral Accuracy](image)

**Figure 4.** Behavioral Accuracy a) Percent accuracy by group (N<sub>CTL</sub>=53, N<sub>ASD</sub>=26), computed on a trial-for-trial basis. For conditions in which learning was expected to occur across the duration of the experiment (i.e. intent and choice) accuracy is shown for the first and last thirty trials (10 from each charity). b) Consistency across trials by group. Trials qualified as consistent if the participant’s answer for choice was logically predicted by their answers for belief and intent. Asterisks within columns refer to bootstrapped 95% confidence intervals as significantly different from 50% behavioral accuracy (values depicted with error bars). Asterisks between columns (within population and question type only) refer to bootstrapped 95% confidence intervals not overlapping one another.
were used to determine significance from chance (50% accuracy) and significance between groups.

Since the mode switches back and forth many times over the course of the task, Mentalizers could only rely information from previous trials to a limited extent. Therefore, they were not expected to perform better in later trials on belief than in early trials (and on average the data echoed this expectation). However, the intent of Agents vis-à-vis the charities remained constant throughout the task, and therefore it was possible to build up an understanding of which charities the Agents preferred over time and use this information to gradually increase performance on the intent and choice questions over the total 80 trials. To determine if this was the case, the intent and choice performance were analyzed in two sections: the first 30 trials, where Mentalizers were operating on relatively little data, and the last 30 trials, where Mentalizers had at least 50 trials worth of information on which to base their answers.

Performance of the control population on the intent question was significantly different between the first 30 trials and the last 30 trials, indicating a net improvement in accuracy on this question over time (Fig. 4a). In contrast, the ASD population showed no difference in performance between the first 30 trials and the last 30 trials for intent. Accuracies for both the first 30 trials and the last 30 trials for controls were also significantly different from chance, while for ASD neither value was significantly different from chance.

In terms of the choice question, neither group showed a significant difference between the first 30 trials and the last 30 trials. However, the control group’s accuracies were above chance for both instances, whereas the ASD population’s accuracies were at not above chance in either instance.

To assess the degree to which the groups were using a logical policy to perform the Mentalizer task, the percentage of trials in which their answers for belief and intent directly predicted their answer for choice was quantified as a “consistency” metric (Fig. 4b). For example, a trial where a participant answered “reversal” for belief, “charity” for intent, and “self” for choice was classified as consistent. The levels of consistency of both control and ASD population were between 70-80%, which was significantly different from 50% of trials. However, the consistency levels of controls and ASD were not significantly different from one another.
Computational Models

Eight different computational models were fit to behavioral data from three different participant groups – two Matched Control groups (MC1; MC2) and one group with autism (ASD), where N = 27, 26, and 26 respectively. The basic model off which all other models were based, RW/RW, used a basic Rescorla-Wagner update rule for both belief and intent, each of which had a learning rate which could be fitted to individual or group data (for more, see Methods). The best performing model on control data was RW/RWoverT, a model that was identical to RW/RW except that the intent model learning rate was scaled over the course of trials. This had the effect of giving more weight to interpretations of earlier trials and gradually tapering off the ability of the model to update as time went on.

The other six models were purposely designed to update in a logically suboptimal way, in the hopes of capturing possible errors present in Mentalizer calculations. Thus RW_overT/RW used a basic intent model but a weighted learning rate for belief, giving more credence to the early trials, despite the fact that the mode switched back and forth intermittently, rendering the information given in early trials useful for only a limited amount of time.

Likewise, RW/REAL_overT used the standard belief model for calculating belief, but the real mode of the program to calculate intent rather than calculated belief. REAL/REAL_overT functioned in the same way except that belief was never modelled in the first place – rather, the real mode was used both the answer the belief question and to calculate intent. Similarly, RW/NONE_overT used the basic belief model but disregarded the mode entirely for intent computations – instead interpreting the choices of Agents as direct indications of their intentions. normal/NONE_overT and reversal/NONE_overT functioned along the same lines, except that they assumed for belief that the Mentalizer constantly believed it was normal or reversal mode, respectively, and used this constant belief to construct intent.

Models were fitted to the data to find optimized learning rates, which dictated how much weight each update to the models’ predictions would have (in other words, the speed at which the model learned). All models were fitted for a belief learning rate and an intent learning rate, except for REAL/REAL_overT, normal/NONE_overT, and reversal/NONE_overT which only needed an intent learning rate as belief was not learned. Learning rates were optimized via a gridsearch method (see Methods, Fig. 3).
Belief learning rates were relatively similar among all groups within each model (Fig. 5a). RW/RW and RW/RW\_overT received identical belief learning rates to each other within each group, as did RW/REAL\_overT and RW/NONE\_overT. For all groups, belief learning rates went to 1 for RW\_overT/RW. For no model were the belief learning rates significantly different between participant groups (bootstrapped 95% confidence intervals).

The fitting of optimized intent learning rates showed nearly identical learning rates amongst all groups and for all models that were small in magnitude, except for the RW/RW\_overT model (Fig. 5b). This model showed a striking dissociation between the fits to both MC groups and the ASD group, where MC1 and MC2 had intent learning rates of 0.58 and 0.69, respectively, while the intent learning rate of ASD was 0.09. This model was also the only one to achieve any kind of significant dissociation between groups. The intent learning rate of
MC1 was significantly different from that of ASD, and this was nearly the case between MC2 and ASD as well (lower bound MC2 = 0.21 and upper bound ASD = 0.22). Significance values were determined in the same manner as for belief learning rates, with bootstrapped 95% confidence intervals.

To explore the degree of accuracy the models achieved in mimicking the answers given by Mentalizers, the percent of trials in which the models directly agreed with the Mentalizers in terms of belief, intent, and choice were quantified (Fig. 6). In terms of belief, (Fig. 6 top) RW/RW, RW/RW_overT, RW/REAL_overT, and RW/NONE_overT showed nearly identical accuracies (between 67-72%) for all groups. RW_overT/RW, normal_NONE_overT and reversal/NONW_overT were the worst performing models in terms of belief – failing to perform significantly above chance (50% accuracy). No models had belief accuracies that were significantly different between groups (bootstrapped 95% confidence intervals).

Intent accuracies (Fig. 6 bottom left) showed more variation across groups in RW/RW and RW/RW_overT, which were more accurate for matched control groups than they were for the ASD group. However, RW/RW_overT was the only model where this difference in accuracy

![Figure 6](image)

**Figure 6.** Percent trials in which models agree with Mentalizers. On a trial-by-trial basis, the percentage of trials in which the models gave the same answers as the Mentalizers in terms of belief (top), intent (bottom left) and choice (bottom right). Error bars represent bootstrapped 95% confidence intervals (5000 iterations). Dotted lines represent 50% accuracy (chance). *ASD accuracy is significantly different from both MC1 and MC2 accuracies. N_{MC1}=27, N_{MC2}=26, N_{ASD}=26.
between both matched control groups and ASD was significantly different. The other models all showed intent accuracies that were similar to one another, and where at least one matched control group was not significantly different than ASD. For all of these remaining models, accuracies were not significantly different from 50% (chance), or below chance, on at least one group of matched controls, with the exception of RW_overT/RW which showed higher accuracies.

Choice accuracy (Fig. 6 bottom right) was poor across the board, at no point rising higher than 56% of trials correct. All models and all groups showed accuracies around 50% (chance), with the exception of RW/REAL_overT and the MC1 group of RW/NONE_overT, which had very low accuracies (32-35% correct). RW/RW and RW/RW_overT were the only models for whom both MC1 and MC2 accuracies were significantly above 50% (bootstrapped 95% confidence intervals), while no model took ASD accuracies above 50%.

To further investigate the goodness of fit of the models to behavioral data, the Bayesian Information Criterion (BIC) was examined (Fig. 7). The BIC uses the negative log likelihood as well as the number of parameters entering into a model and the number of data points under

Figure 7. Goodness of fit of each model as described by the Bayesian Information Criterion (BIC). The more elevated the BIC value, the worse fit a model is to a set of data. Combined BIC refers to fit as judged by a sum of belief, intent and choice modelling. Belief BIC, Intent, BIC, and Choice BIC refer to the fits on each of these questions individually. Error bars represent bootstrapped 95% confidence intervals (5000 iterations). Dotted lines represent BIC values for a random model. N_{MC1}=27, N_{MC2}=26, N_{ASD}=26.
consideration to calculate how well a model fits to the data, and the lower the BIC value, the better the fit.

Four different BICs were calculated: a combined BIC value incorporating belief, intent and choice model predictions, and separate belief BIC, intent BIC, and choice BIC values to assess fits to those specific questions (Fig. 7). For no models were ASD BIC values below the BIC value of a random model (dotted line) – and this was true for overall BIC as well as the individual BICs for belief, intent, and choice.

Combined BICs were lowest for all participant groups for RW/RW, RW/RW_overT, RW/REAL_overT, and RW/NONE_overT (Fig. 7 top left) and were significantly below the BIC value for a random model for MC1 and MC2 (bootstrapped 95% confidence intervals). Notably, RW/RW and RW/RW_overT both showed markedly lower combined BIC values for both matched control groups than for ASD (although this did not achieve significance), whereas all other models showed no major differences between groups. The same pattern was true for the first four models with regards to the belief BIC values (Fig. 7 top right). These first four models also had drastically lower belief BIC values than the second four models.

RW/RW and RW/RW_overT stood out with regards to the intent BICs as the only two models with any difference between participant groups (Fig. 7 bottom left), although this difference was not significant (bootstrapped 95% confidence intervals). While ASD showed BICs as high as any other model (significantly above the random model BIC value), both matched control groups on these models showed lower values.

Choice BICs, in contrast, showed very little difference between models and little to no difference between groups. Only the last three models, which did not fit a belief learning rate, showed across the board slightly lower values.

In general, in terms of comparing BIC values, the models were best able to fit belief, while some were able to better grasp intent than others, and no model excelled in terms of choice.

Cross-Fitting
The generalizability of the optimized models was tested by assessing the fit of these models across groups (Fig. 8). More specifically, the learning rates found by fitting the models to the MC1 group were tested on the MC2 group and vice versa. The percent accuracies and BIC
values from these two tests were averaged together to yield measures of how well the models generalized across different groups of neurotypical (control) individuals. Similarly, the MC1 and MC2 learning rates were tested on the ASD group, and these two sets of values were averaged together to yield a measure of how well the ASD behavioral data could be described by models using control learning rates.

The overall BIC values for the models cross fit across MC groups (Fig. 8a, white bars) were only significantly below the random model BIC value for RW/RW and RW/RW_overT (bootstrapping 95% confidence intervals). The BIC values for the ASD data fit with MC (Fig. 8a, gray bars) were significantly higher than the random model BIC value for all models.

The accuracies of the cross fit models were also assessed (Fig. 8b). For the models cross fit across MC groups, the first four models performed identically well in terms of belief, and the
RW/RW and RW/RW_overT also performed well in terms of intent and above chance in terms of choice. All other models had at least one metric perform at chance or worse.

For the models with ASD data fit to MC learning rates, the same pattern of accuracies is evident. For the RW/RW model and the RW/RW_overT model, the ASD cross fit had significantly lower accuracies for belief, intent and choice than the MC cross fit (bootstrapping 95% confidence intervals).

**Correlations**

In order to examine the relationship between performance on the Mentalizer task and a variety of other traits and behavioral measures, correlations between these values were assessed for matched controls and for ASD using Pearson’s rho values (Tables 2 and 3).

<table>
<thead>
<tr>
<th></th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>FSIQ</th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>SNI Network</th>
<th>SNI PPL</th>
<th>SNI Embed</th>
<th>EQ</th>
<th>SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8517</td>
<td>0.8307</td>
<td>0.3576</td>
<td>0.0172</td>
<td>0.0855</td>
<td>0.1768</td>
<td>0.1456</td>
<td>-0.0823</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4403</td>
<td>0.3737</td>
<td>0.2170</td>
<td>-0.0616</td>
<td>0.0633</td>
<td>0.0756</td>
<td>0.0723</td>
<td>-0.0597</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2170</td>
<td></td>
<td></td>
<td>0.1188</td>
<td>0.1211</td>
<td>0.2761</td>
<td>0.1304</td>
<td>-0.1066</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.1891</td>
<td>-0.0334</td>
<td>-0.1459</td>
<td>0.4339</td>
<td></td>
<td></td>
<td></td>
<td>-0.1429</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7152</td>
<td></td>
<td>0.7096</td>
<td>0.1205</td>
<td>0.0809</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8527</td>
<td></td>
<td>0.8241</td>
<td>0.2168</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1501</td>
<td></td>
<td>0.1897</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Control group correlations in terms of Pearson’s rho values. FSIQ, PIQ, EYES Score, SNI Network, SNI PPL, EQ, SQ, and AQ represent behavioral metrics as described in Methods. OVERALL belief/consistency represents these measures over all trials in the Mentalizer task. CHANGE intent/choice/consistency represents the magnitude of the difference of these measures between the first 30 trials and the last 30 trials. Green shading indicates p-values < 0.01; yellow = indicates p-values < 0.05.
Control subjects (Table 2) showed unsurprisingly strong positive correlations between FSIQ and PIQ, FSIQ and VIQ, and PIQ and VIQ. There was also a strong correlation between PIQ and the EYES Score, as well as a correlation between FSIQ and the EYES Score. Similarly, there were strong correlations between SNI Network and SNI PPL, SNI Network and SNI Embed, and SNI PPL and SNI Embed. Additionally, there was a strong correlation between the EYES Score and EQ as well as EQ and SQ, and a strong negative correlation between EQ and AQ. Age exhibited a strong positive correlation with AQ. Gender was not correlated with anything.

Overall belief performance of control Mentalizers was correlated with EQ, but not strongly. The same was true for overall consistency and VIQ, and overall consistency and SNI Network. The change in intent (signifying difference in intent question accuracy from the first 30 trials and from the last 30 trials and thus functioning as a marker of learning) was weakly

<table>
<thead>
<tr>
<th></th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>SNI Network</th>
<th>SNI PPL</th>
<th>SNI Embed</th>
<th>EQ</th>
<th>SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7702</td>
<td>0.9050</td>
<td>0.8052</td>
<td>-0.0249</td>
<td>-0.1114</td>
<td>-0.0101</td>
<td>-0.3460</td>
<td>0.1098</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FSIQ</th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>SNI Network</th>
<th>SNI PPL</th>
<th>SNI Embed</th>
<th>EQ</th>
<th>SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1826</td>
<td>0.4891</td>
<td>0.8035</td>
<td>0.0938</td>
<td>-0.1075</td>
<td>0.0165</td>
<td>-0.3098</td>
<td>0.1355</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FSIQ</th>
<th>0.1235</th>
<th>-0.1911</th>
<th>-0.0923</th>
<th>-0.1322</th>
<th>-0.2013</th>
<th>0.3928</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1159</td>
<td>0.2787</td>
<td>0.3667</td>
<td>0.0637</td>
<td>0.0677</td>
<td>-0.0521</td>
<td>0.0043</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AQ</th>
<th>Age</th>
<th>Male</th>
<th>OVERALL</th>
<th>Belief</th>
<th>Consistency</th>
<th>Intent</th>
<th>Choice</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1021</td>
<td>0.2440</td>
<td>-0.2761</td>
<td>0.2256</td>
<td>0.0932</td>
<td>-0.0976</td>
<td>-0.3245</td>
<td>-0.0496</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FSIQ</th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>SNI Network</th>
<th>SNI PPL</th>
<th>SNI Embed</th>
<th>EQ</th>
<th>SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2787</td>
<td>0.1159</td>
<td>0.2787</td>
<td>0.3667</td>
<td>-0.0923</td>
<td>-0.1322</td>
<td>-0.2013</td>
<td>0.3928</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FSIQ</th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>SNI Network</th>
<th>SNI PPL</th>
<th>SNI Embed</th>
<th>EQ</th>
<th>SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2787</td>
<td>0.1159</td>
<td>0.2787</td>
<td>0.3667</td>
<td>-0.0923</td>
<td>-0.1322</td>
<td>-0.2013</td>
<td>0.3928</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FSIQ</th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>SNI Network</th>
<th>SNI PPL</th>
<th>SNI Embed</th>
<th>EQ</th>
<th>SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2787</td>
<td>0.1159</td>
<td>0.2787</td>
<td>0.3667</td>
<td>-0.0923</td>
<td>-0.1322</td>
<td>-0.2013</td>
<td>0.3928</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FSIQ</th>
<th>PIQ</th>
<th>VIQ</th>
<th>EYES Score</th>
<th>SNI Network</th>
<th>SNI PPL</th>
<th>SNI Embed</th>
<th>EQ</th>
<th>SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2787</td>
<td>0.1159</td>
<td>0.2787</td>
<td>0.3667</td>
<td>-0.0923</td>
<td>-0.1322</td>
<td>-0.2013</td>
<td>0.3928</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** ASD group correlations in terms of Pearson’s rho values. FSIQ, PIQ, EYES Score, SNI Network, SNI PPL, EQ, SQ, and AQ represent behavioral metrics as described in Methods. OVERALL belief/consistency represents these measures over all trials in the Mentalizer task. CHANGE intent/choice/consistency represents the magnitude of the difference of these measures between the first 30 trials and the last 30 trials. Green shading indicates p-values < 0.01; yellow = indicates p-values < 0.05.
correlated with FSIQ, and strongly correlated with PIQ and overall belief performance. Change in choice was similarly weakly correlated with both FSIQ, VIQ, and change in intent, as well as strongly correlated with overall consistency. Change in consistency was not correlated significantly with anything.

The ASD group (Table 3) showed many similar patterns. For instance, there were strong correlations between FSIQ and PIQ and FSIQ and VIQ, with a weak correlation between PIQ and VIQ. The EYES Score was strongly correlated with FSIQ, PIQ, and VIQ. SNI Network was strongly correlated with SNI PPL, as were SNI Network and SNI Embed, and SNI PPL and SNI Embed. SQ and AQ were weakly correlated, and age and SQ had a weak negative correlation. Gender again exhibited no significant correlations with anything.

For ASD, there was a strong negative correlation between EQ and overall belief performance, and a weak correlation between SNI Network and overall consistency. Change in intent and change in consistency were not correlated with anything, but change in choice was weakly negatively correlated with VIQ.
Behavioral Differences

Performance of two participant groups, matched controls and ASD, was assessed on the Mentalizer task via accuracies on answering three questions about third-party Agents: belief, intent, and choice. Analysis of these accuracies showed that the performance of control and ASD groups were not significantly different from one another on belief (Fig. 4a). However, an examination of intent performance by accuracy in the first 30 trials and last 30 trials showed that control participants exhibit a significant improvement in performance over the course of the task, whereas ASD participants do not. Furthermore, controls perform above chance in both instances while individuals with ASD are never significantly above chance. This indicates that individuals with ASD have some impairment performing the calculations necessary to answer the intent question that does not prevent them from computing belief accurately.

It is likely that an individual who is performing the Mentalizer task begins each trial with two priors: the belief that the Agent has about the mode and the intention that the Agent has towards the relevant charity (Fig. 9). Based on these priors, the Mentalizer makes a prediction about what choice the Agent will make on that trial. The Mentalizer then sees the outcome of the trial, which can be filtered into two bits of information: whether or not the decision that the Agent made was reversed or not, and whether the Agent chose to donate or take money.

Following principles of reinforcement learning, the more surprising this outcome was (in other words the more it deviated from the priors), the more significant the update to the priors will be. However, in terms of intent, the choice outcome information must be filtered through the existing representation of belief in order for it to account for the possibility of reversal mode being in effect.

Once the current belief has been accounted for, the prediction error resulting from the outcome can then go

**Figure 9.** The Mentalizer task updating process. Prior beliefs and intents go to forming a prediction of choice. The Decision Outcome and the Choice Outcome of the trial then go to updating these priors. Intent must be updated by filtering choice through the current representation of belief in order to account for the possibility of the mode being in reversal.
towards updating the intent prior, which paired with the belief prior can be used to make the prediction about the Agent’s choice on the next trial.

The behavioral results found here suggest that there is some issue with the updating of the intent prior that ASD individuals are not able to do. Whether they simply fail to update the prior, or do so with a sub-optimal calculation is still an unanswered question. However, it is unlikely that the lower performance for the intent question can be simply ascribed to the cognitive load placed on working memory in this task. It has been shown that individuals with ASD have unimpaired verbal working memory using a standard N-back letter task (Williams et al, 2005), which requires a very similar cognitive load to the one required to store preferences of Agents for different charities. Therefore it seems that the social component of the task is what creates the dissociation in behavioral performance between controls and ASD.

It may also be that individuals with ASD have difficulty integrating their intent and belief priors into a choice prediction. However, this study finds no evidence for this as it is possible that the fact that choice performance is never statistically different from chance for ASD (unlike controls) is due to their impairments with intent. Since a correct intent calculation is necessary for a correct choice prediction, a failure to calculate intent correctly could be entirely responsible for never answering the choice question correctly. Furthermore, the high percentage of consistent trials in the ASD group, which is not significantly different from the percentage of consistent trials in the control group, indicates that individuals with ASD are able to integrate their belief and intent priors in a consistent way to give a logical (if incorrect) answer for choice on a regular basis (Fig. 4b).

This finding supports a more nuanced understanding of how ToM is represented in individuals with ASD. While the link between ASD and impaired ToM has been established (Schuwerk, Vuori, & Sodian, 2015; Yoshida et al, 2010b; Moran et al, 2011; Sigman et al, 2006; Baron-Cohen et al, 1997a, 1997b; 1985; Adolphs et al, 2001; Leslie, 1987), there has been little research to date investigating the exact nature of this impairment. The behavioral data suggests that not all components of ToM are equally affected – rather, the type of ToM that governs “intent” in this paradigm appears to be dissociable from the type of ToM that governs “belief.” The challenge now is to determine how these labels can be generalized to other ToM tasks and ultimately to representation of others in the outside world.
Additionally, the high percentage of consistent trials in both groups (Fig. 4b) suggests that these populations are using the internally consistent model described in Figure 9, and that any systemic errors are due to problems updating priors in an appropriate way when a surprising outcome occurs. In other words, the logic behind using belief and intent to update choice is present in all groups, so the error must lie in the representations of belief or intent or choice if there is low accuracy in answering those respective questions. The presence of this internally consistent model suggests that this task is a viable one for the purposes of applying the computational approach in order to unearth what type of updating of priors is occurring.

While it is possible that the differences the task is measuring that are evident both behaviorally and in the modelling are due to some other factor than ASD, the results of the correlation tests (Table 2, 3) suggest that this is not the case. Most of the correlations found were expected because the tests are known to be linked (for example, the positive correlations between VSIQ, PIQ and VIQ). For Matched Controls, weak positive correlations (0.01 < p < 0.05) were found between intent change over time, choice change over time, and FSIQ, as well as between choice change over time and VIQ. A strong positive correlation (p<0.01) was found between intent change over time and PIQ (Table 2). However none of these correlations were found in the ASD population, and in fact there was found a weak negative correlation between choice change over time and VIQ (Table 3). Thus it is likely that differences between MC and ASD learning over time cannot be ascribed to these factors. In fact, there was no correlation between some behavioral metric and some other test that was in the same direction for both ASD and MC and significant.

Implications of Modeling Results

Eight models were fit to behavioral data in order to investigate the computational processes underlying ToM learning in controls and in ASD. Two of these models were designed to provide accurate answers: RW/RW, which uses a basic Rescorla-Wagner update rule for both belief and intent representations, and RW/RW_overT, which uses the same belief model but weights the intent learning rate so that more importance is assigned to updates occurring earlier on in the task. The other six models were designed with logical flaws in the hopes of determining what impairment might be causing lower intent accuracy in ASD (Fig. 4a).
The group of matched controls (N = 53) was split into MC1 (N = 27) and MC2 (N = 26) in order to allow for cross fitting to occur and to avoid double-dipping into the data when examining goodness of fit.

Although for all the models belief learning rates were relatively similar between groups, there was some variation across models (Fig. 5a) – most notably, all belief learning rates were under 0.8 except for RW_overT/RW, where every group had a learning rate of 1. This model’s logical flaw (weighting earlier updates of belief more than later ones, when in fact the mode switches constantly so there is no benefit to this) is most alleviated when the learning rate is highest because it raises the degree to which any updating can occur, so it makes sense that if all groups are able to predict belief with a relatively high level of accuracy (Fig. 4a), the belief learning rate here would go to 1. Since behavior accuracies across groups were not significantly different (Fig. 4a), the fact that belief learning rates across groups are not significantly different is to be expected.

The fitted intent learning rates were also remarkably similar across all groups and models, and were all values under 0.3, except for one model: RW/RW_overT. This model shows a clear dissociation where the matched control groups have a higher learning rate than the ASD group. This means the degree to which this model updates intent is much faster for the fit to MC1 and MC2 than it is for the fit to ASD data. This model is one with some of the best accuracies for matched controls on all metrics as well as some of the lowest BIC values for matched controls (neck-and-neck with the RW/RW model) (Fig. 6-7). The dissociations this model shows in learning rates, accuracies, and BIC values in comparing both matched control groups with the ASD group indicates that this model is most likely the best suited to unraveling the difference in calculations occurring between the two.

In terms of accuracy of all models (Fig. 6), it is notable that all models predict belief answers relatively equally across groups. Since there is no significant difference between matched control and ASD accuracies on the belief question (Fig. 4a), this is to be expected. However, there is more variation in intent accuracies. While no model predicts ASD behavior better than behavior of both matched control groups, the two models without logical flaws (RW/RW and RW/RW_overT) are the ones which most markedly predict matched control behavior better than ASD behavior, suggesting that they may be closer to approximating neurotypical calculations of ToM than ASD calculations of ToM.
No model excels at predicting choice (Fig. 6), which may perhaps be attributed to the fact that although consistency is high in both groups behaviorally (Fig. 4b), there still remained around 30% of trials which were not consistent, and the task of predicting choice which relies on integrating two separate models correctly may be too challenging for the simple models used here. Interestingly, RW/REAL\_overT seems to have fit to nearly an inverse model of behavior, achieving an accuracy for choice that does not rise about 35% for any group. However it is worth mentioning that MC1 and MC2 were the only two groups to have model choice accuracy significantly above 50%, which occurred for both RW/RW and RW/RW\_overT, whereas no model was able to reach choice accuracy significantly above chance for ASD.

Model accuracies are for the most part mirrored in the pattern of their BIC values (Fig. 7). Clearly RW/RW and RW/RW\_overT have the best fits to the matched control data overall (Fig. 7 top left), while ASD fits overall is equally poor across the first four models and all groups have highly elevated BIC values indicating poor fits on the last four models. Breaking down the BIC values by belief, intent, and choice also echoes the story told by the accuracies of the model in Figure 6. The first four models reliably predict match controls better than ASD but predict all groups significantly better than the last four models. In terms of intent, the first two models show the same dissociation between matched controls and ASD, while all the other models seem to fit equally poorly.

For choice, no model has BIC values that are particularly low, but the last three models, REAL/REAL\_overT, normal/NONE\_overT, and reversal/NONE\_overT have lower BIC values that are lower than the previous five. This is most likely due to the fact that BIC values take the number of parameters fit to the model into account. The more parameters, the higher the value—a feature intended to discourage model-hacking, or the overfitting of models to data with many parameters, which sacrifices generalizability beyond a specific training set. Thus the last three models, which only fit an intent learning rate and no belief learning rate, have a slight advantage in this regard.

The fact that the RW/RW and RW/RW\_overT models, which are the best performing for matched controls on all metrics, only have one parameter to fit for belief and one parameter to fit for intent renders the possibility that the accuracies are due to model-hacking unlikely. It seems that RW/RW\_overT in particular is able represent a calculation that occurs in neurotypical individuals when performing the task, which does not occur or occurs erroneously in ASD, since
it is the only model that exhibits a significant difference in percent behavioral accuracy between MC groups and ASD. Interestingly, it appears than no model stands out as a good fit for ASD in particular – even amongst the ones with have purposeful logical flaws. It is possible that there does exist a single model that would capture ASD behavior like RW/RW_overT does for neurotypical behavior, and that it simply wasn’t among the ones considered here. However, it is also possible that ASD behavior is not homogenous enough to be represented by one model, fit to the data as a group. If individuals with ASD differ enough in the ways that they calculate ToM to answer the intent question, it is plausible that no one model would be able to capture all of that variation.

Another concern that may be raised is the coding error resulting in Agents having one trial of precognition in terms of viewing the pattern of trial reversals or executions for building their models of belief (for more, see Agent Generation in Methods). However, no participants mentioned noticing this irregularity, and it is likely that this discrepancy had a small enough impact as Mentalizers did not receive explicit feedback on their performance during the task. Furthermore, the fact that the dissociations between MC and ASD still occurred despite this suggests that it is possible to perform well on this task as a neurotypical individual even with this error. However, the error has been rectified in more recent versions of the task and is not present in the version that will be used for neuroimaging.

Neural Correlates of Theory of Mind

While the information the computational models and the behavioral data provide gives more details as to how neurotypical individuals might go about computing ToM, mathematically speaking, and where individuals with ASD have an impairment in doing so, neuroimaging techniques have the potential to link these behavioral findings with the “mentalizing system” of neural structures that has already been outlined in previous work (van Overwalle & Baetens, 2009; Rushworth et al, 2013; Koster-Hale & Saxe, 2013; Sallet et al, 2012; Hampton et al, 2008; Rilling et al, 2004; Frith & Frith, 2003; Frith, 2001). Because representing belief switches back and forth of the course of the experiment and representing intent is a process which builds slowly with the addition of more information with minimal switching, these signals are in a manner of speaking orthogonal to each other – in other words, it may be possible to isolate neural activation corresponding to belief or intent from each other.
The Mentalizer task has already been optimized for fMRI data collection, and it is anticipated that data will be for the most part collected by mid-June 2016 (target: 50 controls; 30 ASD). When examining this neuroimaging data, we may find activation in the medial prefrontal cortex and temporo-parietal junction as those are areas that have been directly implicated in ToM computations (Saxe & Kanwisher, 2003), although it is possible to think about others while avoiding explicitly recognizing their internal state, leading to alternate activation patterns. A few other locations are also potential candidates. The amygdala and temporal pole are known to mediate learning affective traits, so it is possible that activity there will correlate with intent computations, as learning what charity an Agent prefers is a type of affective trait (Todorov & Olson, 2008). Additionally, the posterior cingulate cortex and amygdala are known to be involved in forming first impressions of others (Schiller et al, 2009) and the prefrontal cortex and superior temporal sulcus in updating those impressions (Mende-Siedlecki et al, 2013), so it is highly probable we will see activity in those locations.

It is also likely that activity for ASD will be different than controls, especially for intent. It is my hypothesis that this activity will be highly heterogeneous within the ASD group, due to the fact that no one model was found that fit well to all ASD data – instead, it appeared to be the case that each ASD subject was performing computations in their own way, different from the others. The uniting factor is that these intent computations were all impaired in comparison to controls, but it seems that ASD individuals may all be taking different routes to arrive at the same destination, and thus there may be widely diverging patterns of activation among them.

**Future Directions**

In the immediate future, this task will be run using fMRI to image activation of neurotypical and ASD neural structures during the task. This may yield clues about the source of the dissociation in behavior between these two populations on the intent question of the Mentalizer task. As it appears that the RW/RW_overT model best captures the differences between MC and ASD behavior on the Mentalizer task, this model in particular will be used when examining the patterns of neural activation occurring while participants perform the task. Hopefully it will be possible to isolate certain components of the mentalizing network which perform specific mathematical roles, by correlating their activation with the calculations in the model. It will also be useful to investigate how activation differs between ASD and MC – and how these
differences may be modulated by which question they are answering. This may provide clues as where the differences in calculations are occurring between the two groups, even if the signals are too weak to be paired with the models for analysis.

Another approach will be to try to isolate subgroups within the ASD population by fitting the models to individuals rather than group data. If some individuals fit particularly well to certain models, then it may be that this model represents a subgroup within the broader label of ASD. For this, a larger population of ASD participants would be ideal to achieve maximum statistical validity.

It will also be useful to try to duplicate these results on a slightly different task. While it is clear that the ASD group performs differently when answering questions about belief or intent, it is still unclear what larger difference this dissociation is a reflection of. Finding another similar task where this dissociation is evident would be a good method of generalizing these findings further. Belief and intent are well defined in terms of this paradigm, but it will take further study to better understand what the differences in performance truly reflect. It may be that intent simply refers to the ability to understand the preferences and the likes or dislikes of another individual and this impairment is simply one of the many symptoms of ASD but cannot be generalized further. Alternatively, it may be that lowered intent performance is reflective of a larger impairment that is captured in this task, which can also be uncovered in other tasks that require similar but non-identical calculations. Therefore it will be important to investigate the “components” of ToM from as many perspectives as possible before making sweeping statements about the nature of how ASD individuals are impaired in representing the mental states of others.
AUTHOR CONTRIBUTIONS
This project was a joint effort between Damian Stanley of the Adolphs lab at Caltech and me. Dr. Stanley designed the task and ran behavioral subjects. I wrote the code to generate trials for the Charity and Mentalizer tasks, and ran pilot participants on the Charity task. I constructed Agents for the Mentalizer task using this data and code that I wrote. Dr. Stanley wrote the original code for the RW/RW model, which I heavily adapted to optimize it, assess fits, and to build all the other models. I also optimized the behavioral paradigm in Pyschtoolbox to be viable for neuroimaging using fMRI, and I piloted participants on this paradigm. I also wrote code to assess behavioral and modeling accuracies, as well as the code to assess bootstrapping (which was performed by Dr. Stanley) and other statistical fits. I also wrote the code to conduct and assess crossfitting. I made all figures and tables except for figure 3, which Dr. Stanley designed.
REFERENCES


APPENDIX A – Mentalizer Stimulus Presentation Instructions

After the participant completed the Agent Charity Task, the experimenter reviewed the instructions to the Mentalizer Task. Similar to the Charity Task instructions, the instructions for the Mentalizer task are explained via a PowerPoint presentation. The experimenter read each slide aloud to the participant.

The instructions on the slides are listed as follows:

1. In this second task, you will be watching some other people make the same decisions that you just did. We want you to try to learn what mode they think the program is in (‘Normal’ or ‘Reversal’). We will show you the actual mode the program is in, but remember that they don’t have that information and are guessing – just like you did in the last task.

2. We would also like you to try to learn what type of person they are (for example which charities do they want to donate to and when do they want take the money for themselves). Remember – if they believe the program is in ‘Reversal’ mode, they will reverse their choices to get what they want.

3. On each trial, you will first answer 2 questions in random order:
   “Which outcome does Jane want?”
   Press “1” for Definitely Jane
   Press “2” for Maybe Jane
   Press “9” for Maybe Charity
   Press “0” for Definitely Charity

4. “What does Jane think is the mode now?”
   Press “1” for Definitely ‘Normal’
   Press “2” for Maybe ‘Normal’
   Press “9” for Maybe ‘Reversal’
   Press “0” for Definitely ‘Reversal’

5. You will then guess whether you think they chose the charity or themselves.

6. Next you will see what they actually chose.

7. You will be notified when the mode has changed to either “Normal” or “Reversal”. Remember that they do not know that the mode has changed, only you do.
8. Go through the next trial with the experimenter to make sure that everything is clear and answer any questions you have.

9. To make sure that you are keeping the actual mode (‘Normal’ or ‘Reversal’) in mind, every once and a while there will be a test trial like this one.

10. When you see this test trial, indicate the current mode. For every test trial that you respond correctly on, you will get an extra $0.50. In addition, at the end of the experiment, we will randomly select one regular trial and you will receive $5 for each of the three questions that you answered correctly.
    So it pays to be right!

11. REMEMBER:
    The person you’re watching does not know what the actual mode is, only you do. Your goal is to figure out what they think the mode is, as well as what their preferences are.
    If you have any questions, please ask the experimenter now.
APPENDIX B – Code for Computational Models

The function `belief_and_intent_current_models` is the main function called to run a set of data (either a group or an individual set of trials) on a given model. It takes \( X \) which is a vector containing the belief and intent learning rates, `dataVals` which is a structure containing all the behavioral data, `modelType` which is a scalar indicating the model to run the data on, `totalSubj` which is a scalar indicating the total number of participants in `dataVals`, and `initValMat` which is a vector of the starting values for belief and intent models.

It returns \( e \) which is the negative log likelihood of the model fit to the data, `BIC` which is a matrix containing [overall `BIC`, belief `BIC`, intent `BIC`, choice `BIC`] for each individual participant, and `dataCell` which is a structure containing all of the original behavioral data as well as model accuracies and other values computed in the function.

```matlab
function [e,BIC,dataCell]=belief_and_intent_current_models(X,dataVals,modelType, totalSubj, initValMat);

%X contains learning rates
% belEst = Mentalizers' belief estimates [numTrials x numSubjs] normal=1, reversal=0
% belOut = choice execution outcomes [numTrials X numSubjs] executed=1, reversed=0
% intEst = Mentalizers' predictions about agent intent [numTrials X numSubjs] donate=1, take=0
% ag_choice = agents' choices [numTrials X numSubjs] donate=1, take=0
% chType = charity type (number) [numTrials X numSubjs] 1 or 2 or 3
% ment_choice = mentalizer's prediction of choice [numTrials X numSubjs] 1 = charity, 0 = self
% ag_belief = agent's belief about mode [numTrials X numSubjs] 1 =normal, 0 = reversal
% ag_intent = agent's belief about mode [numTrials X numSubjs] 1 =normal, 0 = reversal
% initValMat = if agent wanted to give to charity or self [numTrials X numSubjs] 1 = charity, 0 = self
% initValMat = starting value for models - .5 for now.
% ment_bel_unbinarized = unbinarized versions (1-4) of belief estimates by mentalizers
% ment_belUncertainty = 1 if mentalizer was certain, 0 if not

%versions of model encoded in modelType
%1= rw/rw
%2=rw/rwoverT
%3=rw/REAL_overT
%4=rw/NONE_overT
%5=rwoverT/RW
%6=REAL/REAL_overT
%7=normal/NONE_overT
%8=reversal/NONE_overT
```

59
switch modelType
  case {1,2,3,4,5}
    numParameters = 2;
  case {6,7,8}
    numParameters = 1;
end

%pull out learning rates from X
bel_lr=X(1);
int_lr=X(2);
if numel(X)<3
  sm = 1;
else
  sm=X(3);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%begin for loop iterating over trials
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for knd = 1:numel(dataVals)   %get matrix of values out of structure
  belOut = dataVals{knd}.bel_oc(1:80,:);   
  chType = dataVals{knd}.char_type(1:80,:);   
  ag_belief = dataVals{knd}.ag_belief(1:79,:);   
  belEst = dataVals{knd}.bel_est(1:80,:);   
  intEst = dataVals{knd}.int_est(1:80,:);   
  ag_choice = dataVals{knd}.ag_choice(1:80,:);   
  ment_choice=dataVals{knd}.ment_choice(1:80,:);   
  ag_intent=dataVals{knd}.ag_intent(1:80,:);   
  state=dataVals{knd}.state(1:80,:);   
  clear belMod intMod_ch chVals vals2 vals1 vals;   
  numTrials=size(belEst,1);   
  numSubj=size(belEst,2);   
  dataCell{knd}.bel_lr=X(1);   
  dataCell{knd}.int_lr=X(2);   
  dataCell{knd}.sm=sm;   
  %limit bel_lr to under 1
  if (any(X<0 | bel_lr >1));
    e=10000000;
    return;
  end
  if nargin<5
    initValMat=0.5; %default start value for belief and intent
  end

  % % Model Belief
  belMod=nan(numTrials+1,numSubj);   
  belMod(1,:)=initValMat;
  switch modelType
    case {1,2,3,4}
      for ind=2:size(belMod,1)
        belMod(ind,:)=belMod(ind-1,:)+bel_lr.*(belOut(ind-1,:) -
        belMod(ind-1,:));
      end
    case {5} %rwoverT/rw
      for ind=2:size(belMod,1)
        belMod(ind,:)=belMod(ind-1,:)+bel_lr.*(belOut(ind-1,:) -
        belMod(ind-1,:));
      end
  end
\[
\text{belMod}(\text{ind},:) = \text{belMod}(\text{ind}-1,:) + \left(\frac{\text{bel}_lr}{\text{ind}-1}\right) \cdot (\text{bel}_\text{Out}(\text{ind}-1,:)) - \text{belMod}(\text{ind}-1,:));
\]

end

\textbf{case} \{6\}
\begin{align*}
\text{for~} \text{ind} = 1:\text{size(belMod,1)}-1 \\
\text{belMod}(\text{ind},:) = \text{state}(\text{ind},:);
\end{align*}
end
\textbf{case} \{7\}
\text{belMod}(\text{size(belMod,1),:}) = \text{state}((\text{size(belMod,1)}-1,:),:);
\textbf{case} \{8\}
\text{belMod}(\cdot,:) = 0; \quad \% \text{assume~normal}
\textbf{case} \{8\}
\text{belMod}(\cdot,:) = 0; \quad \% \text{assume~reversal}
end
\textbf{dataCell\{knd\}.belMod=belMod; \% not binary}\n
% Model Intent
\text{numChar=numel(unique(chType(:)))); \% charities are labeled 1,2 and 3}
\text{intMod_ch=nan(size(intEst,1)+1,size(intEst,2), numChar); \% initial separate trackers for charity and agent preferences}
\text{intMod_ch(1,:,:)=initValMat; \% initial values for charity beliefs}
\textbf{for~} \text{ind} = 2:\text{size(intMod_ch, 1)} \% for 2 to number of trials
\text{intMod_ch(\text{ind},:,:}) = \text{intMod}\_ch(\text{ind-1,:,:}); \% copy the previous charity slice into the current slice
\text{int}\_\text{ind} = \text{sub2ind(size(intMod_ch), \text{ind}^*\text{ones(numSubj, 1)}, \[1:\text{numSubj}]', \text{chType(\text{ind-1,:})}'); \% indices to update in intMod_ch (ind trial, all subjects, charity index to update)
\text{intOld} = \text{intMod}\_ch(\text{int}\_\text{ind}); \% get out the charity preference values to update
\textbf{switch} \text{modelType}
\textbf{case} \{1,5\}
\text{intOutcome} = \text{abs}(-\text{round(belMod(\text{ind-1,:})'} - \text{ag}\_\text{choice}(\text{ind-1, :}')'); \% get interpretation of outcome of last trial
\text{intNew} = \text{intOld} + \text{int}_lr^*\text{(intOutcome} - \text{intOld}); \% get new prediction of intent ORIGINAL
\textbf{case} \{2\}
\text{intOutcome} = \text{abs}(-\text{round(belMod(\text{ind-1,:})'} - \text{ag}\_\text{choice}(\text{ind-1, :}')'); \% get interpretation of outcome of last trial
\text{intNew} = \text{intOld} + \text{int}_lr^*\text{(intOutcome} - \text{intOld}); \% get new prediction of intent ORIGINAL
\textbf{case} \{3,6\}
\text{intOutcome} = \text{abs(state(\text{ind-1,:}') - ag}\_\text{choice}(\text{ind-1, :}')'); \% get interpretation of outcome of last trial
\text{intNew} = \text{intOld} + \text{int}_lr^*\text{(intOutcome} - \text{intOld}); \% get new prediction of intent ORIGINAL
\textbf{case} \{4,7\}
\text{intOutcome} = \text{ag}\_\text{choice}(\text{ind-1, :}')'; \% get interpretation of outcome of last trial
\text{intNew} = \text{intOld} + \text{int}_lr^*\text{(intOutcome} - \text{intOld}); \% get new prediction of intent ORIGINAL
\textbf{case} \{8\}
\text{intOutcome} = (\text{ag}\_\text{choice}(\text{ind-1, :}')<0.5); \% get interpretation of outcome of last trial (reversing it because we assume reversal)
intNew = intOld + int_lr./((sum(chType(1:ind-1,:)==repmat(chType(ind-1,:),ind-1,1)))).*(intOutcome - intOld); % same as TwExp but without exponent
end
intMod_ch(int_ind) = intNew; %stick new intent predictions back in the model
end

intMod_ch(intMod_ch>1)=1;
dataCell{knd}.intMod_ch=intMod_ch;
dataCell{knd}.ag_choice = ag_choice;

%Model CHOICE
%pick out charity values, generates probabilistic estimates
for ind = 1:size(chType, 1) %for each trial
    for jnd = 1:size(chType,2) %for each subject
        chVals(ind,jnd) = intMod_ch(ind, jnd, chType(ind,jnd)); %for 1 trial, for each subject, chVals = that charity intent
    end
end
%translate into choices
choiceMod = nan(numTrials-1, numSubj); %empty model for choice numTrials-1 X numSubjs - we do not include the first trial
belNormSelect = belMod(2:end-1,:)>0.5;
chSelect = chVals(2:end,:);
choiceMod(belMod(2:end-1,:)>0.5) = chSelect(belNormSelect); %choice model = intents for the charity on that trial
belRevSelect = ~(belMod(2:end-1,:)>0.5);
choiceMod(~(belMod(2:end-1,:)>0.5)) = 1-chSelect(belRevSelect); %choice model = intents for the charity on that trial

dataCell{knd}.choiceMod =choiceMod;
%calculate percentage match with mentalizer
dataCell{knd}.belModMent = sum(sum(belEst(2:end,:) == (belMod(2:end-1,:)>0.5)))./numel(belMod(2:end-1,:));
dataCell{knd}.intModMent = sum(sum(intEst(2:end,:) == (chVals(2:end,:)>0.5)))./numel(intEst(2:end,:));
dataCell{knd}.choiceModMent = sum(sum(ment_choice(2:end,:) == (choiceMod>0.5)))./numel(choiceMod);

%calculate percentage match with agent
dataCell{knd}.belModAg = sum(sum(ag_belief(1:end,:) == (belMod(2:end-1,:)>0.5)))./numel(belMod(2:end-1,:));
dataCell{knd}.intModAg = sum(sum(ag_intent(2:end,:)==(chVals(2:end,:)>0.5)))./numel(chVals(2:end,:));
dataCell{knd}.choiceModAg = sum(sum(ag_choice(2:end,:) == (choiceMod>0.5)))./numel(choiceMod);

%%%individual accuracies
for jnd = 1:size(chType,2) %for each subject
    dataCell{knd}.indivBelMent(jnd) = sum(sum(belEst(2:end,jnd) == (belMod(2:end-1,jnd)>0.5)))./numel(belMod(2:end-1,jnd));
dataCell{knd}.indivIntMent(jnd) = sum(sum(intEst(2:end,jnd) == (chVals(2:end,jnd)>0.5)))./numel(intEst(2:end,jnd));
dataCell{knd}.indivChMent(jnd) = sum(sum(ment_choice(2:end,jnd) == (choiceMod(:,jnd)>0.5)))./numel(choiceMod(:,jnd));
end
%calculate belief fit
pl = dataVals{knd}.bel_est(1:80,:);  
vals = exp(-sm.*abs(pl-belMod(1:end-1,:)));  
vals1(1:size(belMod(1:end-1,:),1), 1:numSubj, 1) = 0;  
vals2 = sum(exp(-sm.*abs(vals1-repmat(belMod(1:end-1,:),[1,1,2]))),3);  
bel_pch=vals./vals2;

%calculate intent fit
all_int_pch = zeros(numTrials, numSubj);  
for ind = 1:size(intMod_ch, 3)  
  %for each charity
  char_vals = squeeze(intMod_ch(1:end-1,:,ind));  
  chInds = find(chType == ind);  
  pl = dataVals{knd}.int_est(1:80,:);  
  int_pch = pl(chInds);  
  int_qch = char_vals(chInds);  
  vals = exp(-sm.*abs(int_pch-int_qch));  
  clear vals1  
  vals1(1:size(int_qch,1), 1) = 0;  
  vals1(1:size(int_qch,1), 2) = 1;  
  vals2 = sum(exp(-sm.*abs(vals1-repmat(int_qch,1,2))),2);  
  int_pch=vals./vals2;  
  all_int_pch(chInds) = int_pch;  
end

%calculate choice fit
pl = dataVals{knd}.ment_choice(1:80,:);  
vals = exp(-sm.*abs(pl(2:end,:)-choiceMod));  
clear vals1  
vals1(1:size(choiceMod,1), 1:numSubj, 1) = 0;  
vals2 = sum(exp(-sm.*abs(vals1-repmat(choiceMod,[1,1,2]))),3);  
ch_pch=vals./vals2;

%this is for other BIC calculation - needed because it needs to be 80xnumSubj
%ordinarily ch_pch generates79xnumSubj
clear ch_qch2  
ch_qch2(:,1:numSubj) = 0.5;  
ch_qch2(2:numTrials,:) = choiceMod;  
choiceMod2 = ch_qch2;  
pl = dataVals{knd}.ment_choice(1:80,:);  
vals = exp(-sm.*abs(pl(1:end,:)-choiceMod2));  
clear vals1  
vals1(1:(size(choiceMod2,1)+1), 1:numSubj, 1) = 0;  
vals2 = sum(exp(-sm.*abs(vals1-repmat(choiceMod2,[1,1,2]))),3);  
ch_pch2=vals./vals2;

e=sum(-log([bel_pch(:);all_int_pch(:);ch_pch(:)]));  
%summarizing negative log likelihood for belief, intent and choice
dataCell{knd}.ch_pch2 = ch_pch2;  
dataCell{knd}.ch_pch = ch_pch;  
dataCell{knd}.bel_pch=bel_pch;  
dataCell{knd}.int_pch=all_int_pch;  
dataCell{knd}.e=e;
end

numOne = size(dataCell{1}.belMod,2);
numTwo = size(dataCell{2}.belMod,2);

e = 0;
for ind=1:numel(dataCell)
    e=e+dataCell{ind}.e;
    dataCell{ind}.belModMentALL = (dataCell{1}.belModMent*numOne +
dataCell{2}.belModMent*numTwo)/(numOne+numTwo);
dataCell{ind}.intModMentALL = (dataCell{1}.intModMent*numOne +
dataCell{2}.intModMent*numTwo)/(numOne+numTwo);
dataCell{ind}.choiceModMentALL = (dataCell{1}.choiceModMent*numOne +
dataCell{2}.choiceModMent*numTwo)/(numOne+numTwo);
dataCell{ind}.belModAgALL = (dataCell{1}.belModAg*numOne +
dataCell{2}.belModAg*numTwo)/(numOne+numTwo);
dataCell{ind}.intModAgALL = (dataCell{1}.intModAg*numOne +
dataCell{2}.intModAg*numTwo)/(numOne+numTwo);
dataCell{ind}.choiceModAgALL = (dataCell{1}.choiceModAg*numOne +
dataCell{2}.choiceModAg*numTwo)/(numOne+numTwo);
end

%construct BIC for belief, intent, choice, combined, by subject

lastCount = 0;
for ind=1:numel(dataCell)
    count =size(dataCell{ind}.belMod,2);
    %overall
    BIC(1, lastCount + 1: lastCount + count)= sum(-
    log([dataCell{ind}.bel_pch(1:80,:);dataCell{ind}.int_pch(1:80,:);
dataCell{ind}.ch_pch(1:79,:)])) *2+(numParameters*log(80+80+79)-log(2*pi));
    %belief
    BIC(2, lastCount + 1: lastCount + count)= sum(-
    log([dataCell{ind}.bel_pch(1:80,:)])) *2+(numParameters*log(80)-log(2*pi));
    %intent
    BIC(3, lastCount + 1: lastCount + count)= sum(-
    log([dataCell{ind}.int_pch(1:80,:)])) *2+(numParameters*log(80)-log(2*pi));
    %choice
    BIC(4, lastCount + 1: lastCount + count)= sum(-
    log([dataCell{ind}.ch_pch(1:79,:)])) *2+(numParameters*log(79)-log(2*pi));
    lastCount = lastCount + count;
end

end