Adjudicating the simulation theory/theory theory debate (with especial attention to the case of autism spectrum disorders)

Susan M. Parrillo

University at Albany, State University of New York, smparrillo@gmail.com
ADJUDICATING THE
THEORY THEORY / SIMULATION THEORY DEBATE
(WITH ESPECIAL ATTENTION TO THE CASE OF
AUTISM SPECTRUM DISORDERS)

by

Susan M. Parrillo

COPYRIGHT 2016
Abstract

Adjudicating the Simulation Theory/Theory Theory Debate
(With Especial Attention to the Case of Autism Spectrum Disorders)

Philosophers, cognitive scientists and developmental psychologists largely agree that we understand mental states and use them to explain and predict the behaviors of ourselves and of others (i.e. we ‘mindread’) by using a cognitive capacity known as the theory of mind (henceforth, ‘ToM’). However, a question remains as to what, exactly, underpins the ToM ability thereby allowing us such great accuracy in our first-person and third-person reports. My dissertation is an adjudication of the ongoing debate between two competing theories, each of which claims to have the best explanation of ToM. These two theories of ToM are known as the theory-theory (henceforth, the ‘TT’) and the simulation theory (henceforth, the ‘ST’). Because it is thought by cognitive developmental psychologists that autism spectrum disorders (henceforth, ‘ASD’) result from a ToM impairment or deficit, I pay especial attention to the case of ASD, using the features of the disorder, to adjudicate the debate. I suggest that the particular deficits and talents associated with ASD provide reasons to favor the ST over the TT in general and to favor an account of the ST that includes introspection over an account of the ST that excludes it.

Key words: theory of mind, autism, theory theory, simulation theory,
Dedication

For Nan

Acknowledgements

I owe a great debt to so many people who have read and commented on this dissertation in its many mutations during the last few years. First, and foremost, I owe the greatest debt to my Advisor, Bradley Armour-Garb, whose keen and critical eye saved me from so many errors – both in my prose and in my arguments. I have also been incredibly fortunate to benefit from the wisdom of the members of my dissertation committee, Ron McClamrock and Jason D’Cruz, whose insights on the front end of my project helped to mold and shape it. I owe a great debt to Leslie Halpern who worked tirelessly with me to help me understand what I needed to know about cognitive science, developmental psychology, theory of mind, empathy, and autism. Without her vast knowledge of the empirical data concerning autism spectrum disorders, this dissertation would have never got off the ground.

I also want to thank some great academics that have, perhaps unknowingly, had a hand in this dissertation and my academic career as a whole. Of course, my work has been deeply influenced by Robert Gordon and Alvin Goldman for their groundbreaking work on the Simulation Theory. I am grateful for the work that Simon Baron-Cohen has done on autism spectrum disorders: He has moved and inspired me every day since the inception of my project and made time in his busy schedule to correspond with me about my project. I must thank Bonnie Steinbock for the wisdom in her advice, for her constant support and unending friendship, all of which helped me to see the finish line and continue to strive toward it. I also owe a debt to Francisco Gonzalez, who advised my undergraduate thesis at Skidmore College, taught me how to argue for some
rather unorthodox views, and modeled for me what it is to be a truly great scholar. I also owe a
debt to Joel Smith at Skidmore College, who has been my professor, my mentor, my colleague,
and my friend.

I believe no dissertation can be completed without the help of friends and family, and I wish
to acknowledge mine here. My Mother and Father always believed that I could do this. It was
incredibly comforting to know they never stopped cheering for me. I am incredibly grateful for
my friend Ross Gard, who read, edited, and commented on every one of my undergraduate
papers so that I could learn to write academically, and offered me an unending supply of patience
and friendship. My beloved friend, Karen Birnie, to whom I owe so much! I must thank her for
encouraging me to go to college in the first place, for modeling for me what it is to be a woman,
and for the hours spent helping me to understand cognitive science. Finally, I must thank my
friend and fellow philosopher, Stephanie Teller, who listened to my arguments and helped me to
sharpen them.

Last, but by no means least, I owe the greatest debt by far to my children, Katie and Drew,
and my husband, Darren. My debt to the three of you is immeasurable and ranges from the
tangible to the intangible – for the way you move and inspire me, for the many hours you had to
spend in silence so that I could concentrate, for the times you felt as though the dissertation was
all I could think about (or care) about, for your patience, tolerance, and for your unconditional
love. I love you all and am forever grateful for you.
# TABLE OF CONTENTS

## CHAPTER 1: INTRODUCTION

1. 1. Theory of Mind: The Phenomena We Seek to Explain ........................................... 4  
1. 2. Taking Autism Spectrum Disorders into Consideration When Examining the Theory of Mind Debate ............................................................. 8  
1. 3. The Theory Theory View of Theory of Mind ............................................................ 13  
1. 4. The Simulation Theory View of Theory of Mind ....................................................... 16  
1. 5. Concluding Remarks ................................................................................................. 19

## CHAPTER 2: THE THEORY OF MIND

2. 1. A Brief History on The Emergence of Theory of Mind Studies .................................. 22  
2. 2. Theory of Mind Defined and Described ..................................................................... 23  
2. 3. Developmental Emergence and Proper Function of Theory of Mind ......................... 28  
2. 4. Absence of Theory of Mind and Potential Resultant Impairments ............................ 33  
2. 5. Correlation Between Theory of Mind Deficits and Autism Spectrum Disorder ........... 38  
2. 6. Conclusion ................................................................................................................ 43

## CHAPTER 3: AUTISM SPECTRUM DISORDER

3. 1. A Not-So-Brief History of The Diagnostic Criteria for Autism ................................. 46  
3. 2. The Current DSM: DSM-V and The Current Conception of Autism ....................... 52  
3. 3. The Impact of Different Conceptions of Autism on the Theory of Mind Debate ......... 55

### Appendices to Chapter 3

Appendix A DSM-I (1952) 000-x28 Schizophrenic Reaction, Childhood Type ............ 60  
Appendix B DSM-II (1968) 295.8 Schizophrenia, Childhood Type ............................... 60  
Appendix C DSM III (1980) Infantile Autism ................................................................. 60  
Appendix D DSM-III-R (1987) Autistic Disorder 299.0 ............................................... 60  
Appendix E DSM-IV (1994) Autistic Disorder 299.0 ..................................................... 62  
Appendix F DSM-IV-TR (2000) Rhett’s Disorder 299.80 ............................................ 63  
Appendix G DSM-IV-TR (2000) Childhood Disintegrative Disorder 299.10 .............. 64  
Appendix I DSM-IV-TR (2000) Autistic Disorder 299.0 ............................................. 65  
Appendix L World Health Organization (WHO) Definition of Autism (1987) ............... 69
5.3.3. Mirror Neurons: Important for Goldman, but Controversial ....................... 184
5.3.4. Introspection is Imperative for Simulation and Mindreading and is a Critical 
Component of Goldman’s Version of the Simulation Theory .......................... 189
5.4. The Theory Theorist’s Criticism of the Simulation Theory: Collapse ............. 193
5.5. Inaccurate or Failed Mindreading on the Simulation Theory Account ............. 209
5.5.1. Egocentrism as Some Evidence for Autistic Hyper-Introspectionism ......... 222
5.5.2. Does a Bad Copy Still “Count” as a Copy ........................................ 234
5.6. Concluding Remarks .............................................................................. 236

Chapter 6: AN ADJUDICATION OF THE DEBATE BETWEEN THEORY 
THEORISTS AND SIMULATION THEORISTS WITH 
CONCLUDING REMARKS ........................................................................ 238

6.1. A Brief Overview of the Argument .......................................................... 239

BIBLIOGRAPHY .......................................................................................... 242
Preface

A few words ought to be said about the impetus of this dissertation:

I have always had a keen interest and passion for philosophy. Even before I started my formal education in the discipline, I had a passion for argumentation and a never ending list of queries that began with “But why …? But why …? But why …?” I had the luxury of entering college as an adult in my thirties. I say this is a luxury because, by the time I began my higher education, my years on the planet afforded me the knowledge that I wanted to study and eventually teach philosophy. During my undergraduate years, my interests in philosophy were broad (by which I mean “all over the place”). I was, and still am, vastly interested in the philosophy of religion, existentialism, environmental ethics, aesthetics, and – for my senior thesis for my Bachelor’s degree, I argued for an unorthodox interpretation of Plato’s dialogues.

I took one year off from school after my undergraduate work at Skidmore College and before beginning graduate school at the University at Albany. It was during that year that I had to face the most significant, most complex, and most painful “But why…?” of my life. My daughter, Katie, was diagnosed with autism spectrum disorder. As any parent of a child with autism knows, hearing her diagnosis, quite literally, turned my world upside down. I was profoundly and irreversibly changed by Katie’s diagnosis.

As I entered graduate school, I wanted to learn everything there was to know about the autism spectrum and developmental disorders and, for a short period of time, contemplated changing disciplines to developmental cognitive psychology. Alas, my passion for philosophy prevailed, and I began to search for a way to blend my desire to learn philosophy and my need to learn about autism. I delved into topics such as the ethics of disability, the aesthetics of disability, and even contemplated following the lead of Dr. Eva Feder Kittay, whose book, Love’s Labor,
addresses the topic of *justice* as it pertains to people with disabilities. I was floundering in my search for a dissertation topic.

Finally, during my last year of graduate studies, I enrolled in a *Philosophy of Language* course, taught by Dr. Bradley Armour-Garb. I was required to read an article by Dr. H. Paul Grice called *Logic and Conversation* (1975). In this article, Grice describes his theory of conversational implicature and the cooperative principles and maxims of conversation. Briefly, these are: (1) A Maxim of Quality – that is, one ought not say what one believes to be false or for which one lacks adequate evidence; (2) A Maxim of Quantity – that is, one’s contributions to a conversation ought to be as informative as required without being overly informative; (3) A Maxim of Relation – that is, one’s contributions to a conversation ought to be relevant; and (4) A Maxim of Manner – that is, one’s conversational contributions ought to be brief and orderly, avoiding ambiguity and obscurity of expression. Importantly, Grice’s fourth maxim of conversation also requires a speaker to moderate the volume and pace of her speech patterns in order to adhere to the cooperative principles of conversation. Grice then goes on to discuss reasons that a speaker might fail to fulfill the maxims of conversation. One might violate a maxim by being intentional deceptive. One might opt out of the maxims by being unwilling to comply with a maxim. One might flout a maxim by blatantly failing to fulfill it when she is able to fulfill it.

Grice’s *Logic and Conversation* immediately brought Katie and autism to my mind. *Could it be that the social deficits and conversational awkwardness, which are both inherent and pervasive to autism, could be explained by an inability to adhere to Grice’s cooperative principles and maxims of conversation? Could it be that someone with autism can’t engage in the kind of cooperative conversation Grice describes?* I approached my professor with my
thoughts during his office hours. During this meeting, he explained theory of mind to me, introduced me to the work of Dr. Simon Baron-Cohen, and also gave me a copy of *Theory of Theories of Mind* (Eds. P. Carruthers and P. Smith, 1996).

I became engrossed in the topic of theory of mind and the debate between the two camps that seek to explain the mechanism that underpins it: the theory theorists and the simulation theorists. I came to understand that the most current data suggests that a theory of mind deficit may be responsible for the impairments we see in people with autism spectrum disorder. With great fervor, I began to read everything I could about the subject. Admittedly, my initial research was guided by my experiences as a parent of a child on the autism spectrum. When I read the positions of the theory-theorists, my instincts would say, “*But that’s just not Katie….***” or “*That’s not how Katie thinks…***” By contrast, when I read the arguments of the simulation theorists, my instincts would say, “*Yes! That’s Katie!***” Such thoughts were the seeds of my initial support of the simulation theory.

However, as my research continued, I had a more substantive basis for supporting the view that people with autism bear a theory of mind deficit (they are “mindblind”) and for supporting the views of simulation theorists. Moreover, I am now convinced that the simulation theory is the correct theory and that we can use the interesting case of autism to show this. So here, I pin my colors to the mast: I am a simulation theorist and this dissertation is meant to explain why this is the case.

Dr. Daniel Dennett once said, “I am a philosopher, not a scientist, and we philosophers are better at questions than answers…Finding better questions to ask, and breaking old habits and traditions of asking, is a very difficult part of the grand human project of understanding ourselves and our world” (1996). I am not a cognitive scientist or a cognitive developmental psychologist.
I am a philosopher and I am better at questions than answers. This dissertation is just one attempt of mine to raise new questions and bring them forth to the theorists and scientists who work on theory of mind. And I ask them, for Katie, “But why?”
CHAPTER 1: INTRODUCTION

For nearly three decades, a debate as raged between philosophers and cognitive developmental scientists (among others) regarding what skill, capacity, mechanism or other cognitive feature forms the basis for the human ability to understand one other’s mental states and to predict subsequent behaviors. Entire journal editions have been devoted to sketching out the opposing sides of the debate. Many books have been published, their pages filled with collections of works by various scholars who seek to support one side or the other on the issues. All parties to the debate concur that the Theory of Mind is what, cognitively speaking, affords us this ability with such great accuracy. The chasm of disagreement opens when discussions turn to defining what specific ability, skill, or mechanism, in particular, underwrites the Theory of Mind ability. In this dissertation, I not only sketch out the debate, but I also adjudicate the debate, and give reasons why one ought to prefer one theory to the other.

‘Theory of Mind’ (henceforth, ‘ToM’) is a term that is used interdisciplinarily to refer to the cognitive process of, and ability to, impute mental states both to oneself and to others. It is meant to encapsulate (and is often used synonymously with) terms such as ‘mindreading,’ ‘mentalizing,’ ‘intentionality,’ and ‘understanding other minds’.¹ Since the emergence of the original ToM framework in the mid-1970s,² a variety of competing versions of the theory have emerged in an attempt to explain this ability. The most prevalent of these theories have largely fallen into two categories.

¹ Simon Baron-Cohen, in his 2001 paper “Theory of Mind in Normal Development and Autism,” explicitly states that he uses the term ‘theory of mind’ synonymously with the mentalistic terms I mention here.
² See Premack and Woodruff, 1976
In the late 1970’s and early 1980’s, *The Theory Theory of Mind* (henceforth, ‘TT’), was the first theory to emerge and quickly became the orthodoxy on the matter. Later, in 1986, *The Simulation Theory of Mind* (henceforth, ‘ST’) \(^3\) emerged, which challenged the orthodoxy of the TT. There are many iterations of each of these theories of ToM. ‘TT’ serves well as an umbrella term for the group of theories of mind that claim that what underwrites the ToM capacity is some form of theoreticity or some theoretical accomplishment via observation and hypothesis formation by a perceiver. Also often included under the umbrella of the TT are modularity theories – those that aim to show that the capacity for theorizing is underpinned by a mechanism that is both modular and innately given. By contrast, ‘ST’ serves well as an umbrella term for group of theories of mind that claim the ToM ability is not underpinned by any theory employed by the perceiver (tacit or otherwise), but, rather, by the perceiver’s ability to imaginatively project herself into the position of the person being observed (i.e. ‘the target’), using her own mind as a model for the target’s mind (i.e. ‘to simulate’).

It is noteworthy that there exist other theories that seek to explain the mindreading phenomenon, such as, for example, the neurocomputational theory, and the rationalization theory. Despite the emergence of these competing theories, the ToM discussions and debates have remained largely centered on the TT and the ST, with a tacit assumption that it must be either the TT or the ST that explains mindreading. That being said, it is also noteworthy that there are some authors who attempt to correlate the TT with *knowing that* (hereafter, ‘KT’) and the ST with *knowing how* (hereafter, ‘KH) in the Rylean sense and some who even couch the

\(^3\) See Goldman, 1986.
TT/ST debate as a sort of KT/KH dispute. There is potentially a complex relationship between the TT/ST debate and the KT/KH distinction, the discussion of which is beyond the scope of this dissertation.

In this dissertation, I adjudicate between the two main competing theories (the TT and the ST), since they are the most widely assumed explanations of ToM and the most hotly debated. To accomplish this, I will focus primarily on the special case of autism spectrum disorders (henceforth, ‘ASD’) and will ask the following questions:

• *Can the special case of ASD bolster one of these competing theories, causing us to favor one over the other?*

• *Which theory of ToM can best explain ASD?*

What I aim to show is that, because the elements required for mindreading on the ST account are the very same well-known deficits and impairments of ASD, the case of ASD will provide reasons for favoring the ST over the TT and the ST will prove to be better equipped at explaining the social impairments of ASD than the TT. What follows in these first pages, is an overview of my dissertation and a brief description of what you can expect to find in its pages.

---

4 See Gilbert Ryle 1946 for insight as to the KT/KH distinction. See Short, 2015, who draws a strong correlation between the TT and KT and the ST and KH. See Currie, 1996, as an example of one who sets up the TT/ST debate as a sort of KT/KH dispute.

5 The KT/KH distinction and the corresponding intellectualist/anti-intellectualist view are as of yet not established. Thus one cannot claim that TT involves just KT and that ST involves just KH. If instances of KT rely on KH (or vice versa), then advocates of the TT or the ST could end up relying on both KT and KH. Since anti-intellectualism is not established, it seems that one cannot claim that either the TT or the ST involve just either KT or KH, so the issue of the relationship between the TT/ST and KT/KH ought to be postponed, at least pending some resolution to the dispute about the interrelation between KT and KH.
1.1. Theory of Mind: The Phenomena We Seek to Explain

Chapter two of this dissertation is, intentionally, rather non-philosophical. In this chapter, I seek only to give an expository account of what the ToM is, to give the history of the studies of ToM, and to define and describe the functioning of a both typically and atypically developed ToM, such as is assumed to be the case in people with ASD. In particular, I will discuss the functioning of ToM in cognitively typically developed individuals and I will contrast this with what impairments we should expect to find in individuals with a fractured, deficient, or altogether absent ToM. This chapter begins with the question: “To what do we refer when we talk about ToM?”

To begin with, at least *prima facie*, we are very reliable about our first-person reports – that is, our ability to attribute mental states to ourselves. For example, we are reliable about our first-person reports in the sense that when we are in pain, we know that we are, both directly and non-inferentially; and with great precision, these reports inform our predictions and explanations concerning our own behavior. We also make third-person reports in which we attribute mental states to others. These third-person reports concerning the mental states of others are, at least *prima facie*, also fairly reliable; and we use them to predict and explain the behavior of others with a fair amount of accuracy.

How do we reliably and effectively explain and predict the behavior of others and ourselves? What mechanism, module, skill, or capacity affords us the ability to attribute mental states to ourselves and to others? What allows us to know, understand, and evaluate what others believe, what they are thinking, and what they might do next? If mental states are in some sense private in nature, (I shall assume that they are), how are we able to impute intentionality and propositional attitudes to another person in the absence of direct access to their mind with any
sort of reliability? Divergence concerning the answers to these questions is at the root of the
debate between the TT and the ST, and I will explain, in this dissertation, how the two sets of
theorists answer these questions differently.

It seems that, because we have direct, non-inferential access to many of our own mental
states, we possess a sort of privileged access to particular aspects of the contents of our own
minds that we can know instantaneously and with great acuity. For example, should I
accidentally slice through my finger with a knife, then, provided that I am attending to my own
mental state, I will know immediately, directly, and non-inferentially that I am in pain.
Importantly, first-person reports and predictions of this nature are incredibly reliable and
authoritative. For now, I leave aside worries about “unconscious beliefs” and the like, hence, my
knowledge (or awareness) of my own mental state (e.g. pain) is virtually indisputable and will be
impervious to any attempt that another person might make to persuade me otherwise. This is not
a claim that my knowledge of my mental states is indubitable or infallible, but rather that it can
be incorrigible.

By contrast, it seems like we do not – apparently - have this same kind of direct, non-
inferential access to the mental states of others. I cannot know indubitably that another person is
“in pain” in the same way that I can know that “I am in pain.” I may need to (perhaps) infer from
her behavior that she might be in pain. But I may arrive at my assessment only through my
empirical observations of her behaviors, together with my inference, which, collectively, can
only indicate that she is “in pain.” However, if she were merely acting, I might incorrectly
attribute the mental state of “pain” to her.

---

6 Here, I use the term “contents of mind.” To be clear, by “contents” I simply mean “mental state” (i.e. the thought
that I am thinking). By “mind,” I mean only, “the container that holds this content.” I set aside, for now, the
philosophical question, “What is a mind?”
Indeed, on the assumption that we have knowledge (or awareness) of our own mental states in a way in which we do not know the mental states of others, it follows (i) that no other person possesses the same privileged access to our mental states that we possess; and (ii) that we do not possess any privileged access to any minds other than our own. This sounds like a lonely enterprise: I can know certain contents of my own mind; I cannot know the contents of any other minds with the same reliability; and no other mind can know the contents of my mind – at least not in the same way that I can know them.

Yet we clearly do not live in mental isolation of one another. That is, unless we exhibit a developmental or cognitive deficit, we are not oblivious to the mental lives of other people. Moreover, it seems unequivocal that we do at least sometimes possess some awareness of the contents of another’s mental states, although admittedly not with the same privileged direct access, or in precisely the same way in which we know our own. Nonetheless, it seems patently obvious that, despite the private nature of mental states, more often than not, we do possess a kind of awareness about the mental states of other people, and we can use that information to predict their behaviors.

Actually, provided we are neither developmentally nor cognitively deficient, more often than not, we can recognize the mental states of others and can then explain and predict their behavior brought about by those mental states with a surprising amount of reliability. We are able to attribute mental states, even to people unfamiliar to us, with a fair amount of accuracy. Ostensibly, the more we know the individual in question, the greater our acuity in attributing mental states and explaining and predicting their behavior. We would presumably be much more accurate in our guesses about our spouses, for example, than in our guesses concerning strangers. Nevertheless, it seems that we are able to accurately ascribe some mental states to persons
otherwise unknown to us. Here is an example illustrating the case of our ability to ascribe mental states to an unfamiliar person:

Your choice as to whether or not to accept information from a potential informant is often filtered through [mental state distinctions of know versus guess and truthfulness versus deceit]. For instance, you are lost in a city and seek directions. But something in the manner of your informant suggests that he is merely guessing, or that he is only pretending to know when he really does not, or that he knows very well, but is misleading you. You may also encounter the informant who relieves you of the burden of assessing his mental state, for he knows that he is guessing and tells you as much. In all cases you are likely to move on, even at some risk, and to look for the acceptable informant. The ideal informant is one whom essentially you could substitute for yourself without loss: he knows as much about the matter in question as you would know if you had the benefit of his experience and his account is impeccable.⁷

In this example, your third-person report of the informant’s mental state dictates how you will proceed. If you attribute truthfulness to him, you will listen to him and rely on his information. If you attribute guessing to him, you will move on, in search of a better informant. The assessment of the mental states and the intentions of the other person and the predictions of the other person’s behavior inform your responses.

We attribute mental states to others and then use them anticipatorily to explain and predict the behavior of others. We then use this information to determine our responses and to react adaptively. This is our everyday way of understanding each other.

Again, the questions are: Given that we can do this correctly and accurately, how do we do it in the absence of direct, non-inferential access to the mental states of others? How is it that, despite the apparently private nature of mental states, and in the absence of direct access to the minds of others, we are able to impute intentionality and propositional attitudes to one another with success? How are we able to understand and evaluate the thoughts, feelings, beliefs, desires,

and intentions (i.e. the mental states) of others with any precision whatsoever? And, perhaps more importantly, how are we able to do it in such a way so as to explain and predict each other’s behavior with even a modicum of accuracy? The intriguing answer and is that we do all of this by using a ToM, and the flint that sparks the debate is found in the questions concerning what specific ability or mechanism underpins the ToM capacity.

In this dissertation, I follow suit. I shall move the discourse forward, operating under the assumption that the question of how we mentalize has been answered: We do this by employing our ToM. Further, I attempt to identify its underlying nature by identifying which of the two theories, the TT or the ST, is the correct theory.

1.2. Taking Autism Spectrum Disorders into Consideration When Examining the Theory of Mind Debate

The third chapter of this dissertation, like Chapter 2, is also intentionally non-philosophical. My overarching argument herein depends upon the special case of ASD to show that the ST gives a more cogent explanation of the ToM capacity, than does the TT. Better stated, when reviewing the debate through the lens of ASD, we find actual reasons to think that the ST is the correct theory of ToM. Accordingly, the third chapter is, by necessity, a detailed description of ASD, its triad of impairments, the relevancy of ASD to ToM, and the relevancy of ASD to the TT and ST debate. The third chapter and its appendixes discuss the history of the autism diagnosis and gives in-depth descriptions of the various difficulties experienced, on some level, fairly ubiquitously by those with ASD. The chapter is also an in-depth reflection on the development of our understanding of autism, and the consequential ever-evolving changes to the diagnostic criterion.
Given what has been written thus far, a fair question might be, “What, if anything, does ASD have to do with ToM and the debate between theory theorists and simulationists?” In what follows, I will answer this question by showing how the correlation between an ill-functioning ToM and ASD was uncovered, and how this correlation has become the most current conventional wisdom concerning the root cause of ASD.

Since the ToM was presupposed in 1976, various theorists have used cognitive developmentalist research to explain the ToM capacity in developmental terms. Developmentalist accounts of the ToM ability often attempt to correlate the development of the ToM capacity with other capacities known to be modular such as the development of language acquisition, executive function, episodic memory, empathy, and perspective taking. Such empirical research is often an attempt to learn when (in what age range) we should expect to see the development of ToM in typically developing young children. The research generally uses preschool and early school-age children as its subjects. Most research compares groups of typically and atypically developing children (often comparing groups of non-autistic children with groups of autistic children). However, some studies also include children with mental retardation and Downs Syndrome as their subjects. Despite some quibbling over what other modules might be involved or correlated with the ToM development, the majority of cognitive scientists and cognitive psychologists now concur that the ToM is the primary capacity facilitating the human ability to mentalize, and that this ability emerges in typically developing children around age 3 – 4 years.

---

8 See Premack and Woodruff, 1976.
In the late 1970’s three philosophers separately but simultaneously suggested that one effective way to determine whether a creature possessed any belief concepts about other creature’s mental states would be to test its ability to impute false beliefs to others.\textsuperscript{11} The thought was that if a creature could anticipate the false beliefs of another creature and could predict the actions that would proceed as a result of those false beliefs, while simultaneously knowing the true circumstances and possessing a true belief about those circumstances, then the claim that the creature was mindreading could be empirically substantiated.

Developmental psychologists seized upon this idea and constructed a variety of false belief tasks that could be used to gather empirical data about the matter. Failure of these tasks would indicate a deficit in the mindreading capacity (i.e. a ToM deficit), while successful completion of them would demonstrate possession of the mindreading capacity (i.e. ToM proficiency).

For example, Wimmer and Perner (1983) devised a story in which a subject sees a child, Maxi, place chocolate in location A, and then goes out to play. While playing, Maxi’s mother moves the chocolate from location A to location B. The question asked of the subject is, “Where will Maxi look for the chocolate when she returns?” Successful completion of this task requires that the subject assert location A, on the grounds that the subject now knows that Maxi holds a false belief. Failure of this task would be marked by the subject asserting that Maxi will look for the chocolate in location B, which would indicate that the subject couldn’t impute a mental state to Maxi that is different than her own. She assumes that the contents of Maxi’s knowledge and beliefs mirror her own. This is but one example of one false belief task, of which there are many iterations.

\textsuperscript{11} See Dennett, 1978; Bennett, 1978; and Harman, 1978.
Wellman and others\textsuperscript{12} determined that the ability to pass the false belief task – thereby demonstrating the development of a ToM – is acquired in a series of developmental accomplishments including other modular accomplishments such as those previously mentioned. It was determined that the typically developing child will generally begin to pass the false belief tasks, demonstrating development of the ToM capacity between ages 3 and 4.

In 1985, cognitive developmentalists Baron-Cohen, Leslie, and Frith published a groundbreaking article asking, “Does the Autistic Child have a “Theory of Mind”?” They noticed that ASD is typically diagnosed in four year-olds, which is approximately the same age during which the ToM capacity (as a specific cognitive developmental capacity) \textit{ought to be} noticeably proficient in a typically developing child. Hence, with their article, they became the first to expose a link between ASD and a deficit in the mindreading capacity. After administering a variety of false belief task tests to autistic children, non-autistic children, and children with Downs Syndrome, they determined that the overwhelming majority of autistic children were incapable of passing false belief tasks (while children without ASD and those with Downs Syndrome were able to pass the same tests).\textsuperscript{13} Baron-Cohen thus asserted that children with ASD lack such a ‘theory’ – that is, they lack a ToM.\textsuperscript{14} They are, in Baron-Cohen’s terms, ‘mindblind’.

The vast majority of Baron-Cohen’s research during the last two decades has largely focused on this correlation between the ToM deficit and a diagnosis of ASD and, by and large, the scientific community has embraced the idea that ASD results primarily from a ToM deficit.


\textsuperscript{13} In the same study, the false belief task tests were likewise administered to children with Downs Syndrome. It was found that children with Downs Syndrome performed as well as typically developing children on these tests. This underscored the importance of the finding that the vast majority of children with ASD were unable to pass the false belief task tests (See Baron-Cohen, et al., 1985).

\textsuperscript{14} Baron-Cohen, Leslie, and Frith (1985).
The main argument of my dissertation rests primarily on one simple premise: If ASD results when one’s ToM is impaired, then we might expect that a person with ASD would likely also show impairments in the mechanism or skill that underpins the ToM capacity. At the risk of oversimplifying the matter, let me use an exaggerated example to illustrate what it is I am saying. Let’s pretend that ToM (qua the ability to mentalize or ‘mindread’) was underpinned by the ability to count to ten. We might then expect that anyone with an impaired ToM would be unable to count to ten. Conventional wisdom currently claims that ASD results from an impaired ToM. Hence, we should expect an individual with ASD to be unable to count to ten. Now, exchange the phrase “count to ten” with either “theorize” (as the TT asserts), or “simulate” (as the ST asserts). If the TT explanation of mindreading is correct, we should expect that individuals with ASD (who possess the impaired ToM that is associated with ASD) are unable to theorize - not only about mental states in particular and not only within the psychological domain, but also that they would experience difficulties with theorizing in general – even outside of the psychological domain. Likewise, if the ST explanation of mindreading is correct, we should expect persons with ASD to be utterly incapable of simulating, not just mental states relative to the psychological domain, but that they would experience difficulties with simulations in general – even outside of the psychological domain.

As stated, with this dissertation, I intend to show that the case of ASD will provide some reason for favoring the ST over the TT. The reason I offer is a relatively simple one: The most current research shows not just that people with ASD are able to theorize, but that the overwhelming majority of people with high-functioning ASD actually excel in their abilities to theorize in a wide-range of disciplines (especially STEM disciplines). Moreover, they often rely on theoreticity to sort of hack out a compensatory way to understand the mental states of others.
in social situations, thereby showing an ability to theorize even within the psychological domain. However, despite their ability to theorize in a variety of circumstances and situations, these same individuals with ASD are, nevertheless, mindblind. They cannot mentalize, and this impairment bleeds into every aspect of their social lives, giving rise to the features intrinsic to a diagnosis of ASD. These considerations strongly suggest that it is not the ability to theorize (that is, construct theories, whether tacit or otherwise underpins the ToM capacity: People with ASD can and theorize in multiple domains but cannot mentalize. Thus, it seems that some means apart from theorizing must underpin ToM. The remaining chapters of this dissertation aim undermine the arguments favoring the TT by elucidating the autistic ability to theorize both outside of and within the psychological domain, while simultaneously showing that individuals with ASD are intrinsically poor simulators for a variety of reasons and in multiple domains (thereby bolstering the arguments favoring the ST).

1.3. The Theory Theory View of Theory of Mind

The fourth chapter of this dissertation begins with a detailed expository account of the TT, including a brief discussion of its roots, history, and development. The chapter aims to cover, at length, the work of two distinct theory theorists. I discuss the work of Alison Gopnik who approaches the ToM question from the cognitive developmental psychologist standpoint and essentially pioneered the TT “child-scientist” view, which holds that children learn about the world and about other people in ways identical to the way that scientists learn about science – that is, by employment of successively more sophisticated theories via observation and

---

15 In fact, the most current research by Baron-Cohen suggests that people with ASD are so comfortable with theorizing and systemizing that they actually prefer activities involving this over activities that require any form of empathizing (such as, for instance, simulation).
hypothesis formation. After my review of Gopnik, I turn my attention to the work of Simon Baron-Cohen, who puts forth two theories. The first is a modularity approach to the TT and attempts to support the TT directly. The second theory is less about ToM or support for the TT, and more of a theory regarding the causes of ASD, which he frames as difficulties with mindreading owed to an autistic difficulties with empathizing.

Admittedly, I handle the works of Gopnik and Baron-Cohen slightly differently. I give a highly detailed exposition of Gopnik’s assertions, and then, as the dissertation finally waxes more philosophical, I work to raise objections to the most crucial elements of her theory. I do this both by elucidating the objections of others before me and also by raising a couple of my own objections. I think that, by then end of the section, Gopnik’s account is in a great deal of trouble and her positions seem untenable.

After completing the sections describing and undermining Gopnik, I turn my sights to the works of Simon Baron-Cohen. In 1986, Baron-Cohen was the first to publish studies that correlate the phenomenon of autism with the phenomenon of an impaired ToM mechanism. He has written an impressive corpus concerning his ‘mindblindness theory’ and has done much to move autism research forward. Baron-Cohen has two distinct theories concerning the absence of ToM in people with ASD, and I give space to both of them.

First, I describe Baron-Cohen’s Innate Minimalist Modularity Theory (henceforth, ‘IMMT’), in which he describes four discrete but interconnected subsystems, which comprise the ToM as a whole. Baron-Cohen, when working on the IMMT, is perhaps best characterized as a “modularist” – that is, a proponent of the modular-nativist theory or “the modularity approach” to mentalizing. Nevertheless, I discuss Baron-Cohen’s work in Chapter Four because most authors classify the modularity approach (along with the child-scientist approach) as forms of the
TT and the IMMT is best understood as a form of the TT (See Goldman, 2006, pp. 17, 93, and 100). There are some arguments against his IMMT, but these arguments are largely against the notion of modularity in general, and not so much about Baron-Cohen’s IMMT specifically. I discuss these objections in some detail.

Importantly, in 2003 Baron-Cohen extended his IMMT to a new theory reflecting his current theory of ASD: The Extreme Male Brain Theory of Autism (henceforth, ‘EMB’). This represents an apparent shift for Baron-Cohen, whose earlier work on the IMMT certainly seemed most appropriately classified as a theory theorist view. His more recent work on the EMB seems to represent a shift from the TT to the ST. Notably, Baron-Cohen never explicitly mentions the ST. However, with the EMB, he begins to describe ASD as an empathy disorder, claiming that those with ASD “have major difficulties with mindreading or putting themselves in someone else’s shoes, imagining the world through someone else’s eyes…” (Baron-Cohen, 2003, p. 137). Because of this apparent shift to the ST, I pay much more attention to the EMB than to the IMMT in the Baron-Cohen section of the dissertation. This comes up in the dissertation as a bit of a hiccup or a detour. However, I do this in the service of undermining Gopnik’s “child-scientist” account in particular, and the TT as a whole.

Perhaps most importantly – at least, for the purposes of my argument herein, Baron-Cohen’s EMB does much of the heavy lifting for my overarching view. Specifically the EMB uses empirical evidence to show that people with ASD, who are severely mindblind, are not only excellent systemizers and theorizers, but also that they are exceptionally poor empathizers. Hence, as I will show, EMB does a great deal to undermine the TT. Baron-Cohen’s EMB serves another role for this dissertation in that, EMB literally opens the door for a competing theory to

---

16 Empathy is viewed as a crucial component of simulation routines on the ST account.
offer a better explanation of the ToM phenomenon. I claim that Baron-Cohen opened the door to a competing theory, and the ST will prove to be a more cogent explanation of ToM, which does not face the difficulties that the TT views must face.

1.4. The Simulation Theory View of Theory of Mind

The fifth chapter herein, bears much the same structure as the TT chapter, which precedes it, and, like the TT chapter, is also substantially more philosophical in nature than chapters two and three.

I this chapter, I include a detailed account of ST, reflecting on its main tenets as discussed by the primary proponents of the theory. This chapter also includes a description of some challenges and objections that ST faces, some criticisms of ST by proponents of TT, as well as some possible responses that advocates of the ST might have to those challenges and objections.

First, I discuss the works of Robert Gordon, a philosopher who challenged the idea that the human capacity to impute mental states to others is theoretical in nature at all17 and it was the first theory of its kind to challenge the orthodox of the TT. Gordon asserts that mindreading is not, in fact, sponsored by any kind of theory – tacit or otherwise. Gordon introduces us to the ST as an alternative to the theory-laden TT. On Gordon’s account, we are not theorizers. We are simulators. This means that we do not employ some kind of theory to impute mental states to others. Instead, he claims, we use our own mental resources as a model for other minds, to sort of project ourselves (in imagination) into the position of another person. In running this sort of offline simulation, we are able to understand the mental states of ourselves and of others, and predict subsequent behaviors on that basis.

Again, the divide between the TT and the ST is not a disagreement about what we use to mentalize. All are in agreement that we use a ToM. Like all theory theorists, Gordon’s ST acknowledges the activity of, and the capacity for, understanding the mental states of others and uses that to predict and explain their behavior. But Gordon denies that either is achieved via the employment of a theory by a thinker. In fact, not only does his ST deny the central features of TT, it denies any sort of theoretical involvement at all – whether tacit or not, and even at the subpersonal or neural level.

Also in this chapter, I cover the works of Alvin Goldman, who is also a philosopher and a proponent and early defender of the ST. Like Gordon, Goldman maintains that any capacity we have to mind read is process-driven, denying theoretical involvement in the process. But Goldman rejects certain aspects of Gordon’s account of ST and, hence, extends Gordon’s ST in a couple of ways, which turn out to be crucial for my arguments in this dissertation.

First, Goldman (Goldman, 2006) distinguishes between kinds of mirroring processes. He asserts that there are different kinds, or levels of simulation. For example, he says there is a “low-level” simulation, which is an automatic and unconscious mirroring process that we engage in when observing the behaviors of another. And he contrasts this with “high-level” simulation, which occurs when we consciously and voluntarily engage in putting ourselves in the place of another.

Second, Goldman maintains that introspection and its connection to mental states are required for simulation - that is, the simulator must be aware of her own mental state concepts and have

---


19 Goldman, Alvin. Simulating Minds. 2006: New York: Oxford University Press. Goldman mentions the distinction between “low-level” and “high-level” simulation in his introductory chapter and then expands these notions greatly in chapters 6 and 7.
some introspective basis for applying them. By contrast, whereas Goldman asserts that introspection plays an important or necessary role in simulating, Gordon denies introspection. While Gordon’s ST does not require such introspective application of mental concepts, it does instead require that the simulator is able to express mental state concepts and extend these concepts to the simulated other.

The chapter elucidates the differences between the ST account by Gordon and the ST account by Goldman. It also elucidates some important (for my purposes) similarities. While Gordon holds that introspection is not necessary and Goldman holds that introspection is crucial, both philosophers hold that there is something that it is like to simulate accurately or “well,” and something else that it is like to simulate inaccurately or “poorly.” Both philosophers agree that if the simulator (qua perceiver) is unable to make an egocentric shift in order to suspend her own mental states, her resultant simulation of the target will be ‘contaminated’ – that is, inaccurate.

Goldman describes the importance of the simulator’s ability to ‘quarantine’ her own idiosyncratic beliefs and desires in order that she can successfully anticipate and explain those of her target. He claims that the failure to do this results in both egocentric biases and projection.20 However, Goldman asserts that projection is a critical part of mindreading and, hence, he allows for “misfires.” That is, Goldman claims that successful simulation and attempted (but ultimately failed) simulation both count equally as simulation for the purposes of mindreading.21 Each of Goldman’s objections to Gordon will be discussed at length in this dissertation.

Goldman’s version of the ST allows a certain tolerance for misfired simulations. I believe the tolerance found in Goldman’s version of ST accounts for the kind of simulation we see

---

20 Ibid. Pg. 29-30.
21 Ibid. Pgs. 36, 38, and 40 – 41.
performed by persons with ASD. Thus, as I will show, a version of ST which allows for misfired or non-quarantined (hence, contaminated) simulations will certainly prove more able to explain ASD than TT and, will even be more probative than other versions of ST (such as Gordon’s). In turn, the case of ASD, with its triad of impairments, will provide us with actual reasons for favoring ST over TT and, in particular, for favoring Goldman’s version of the ST.

Lastly, in the fifth chapter, I will adjudicate the debate between TT and ST. To accomplish this, I will draw on the arguments elucidated in chapters four and five concerning TT and ST and on various points made in chapter three concerning ASD. Here, I will aim to establish that, contrary to what has been claimed, the TT cannot be the correct theory because people with ASD can theorize in a wide range of areas – including those within the psychological domain. I will provide evidence-based reasons for thinking that people with ASD, in fact, have a great deal of difficulty with simulating. They can and do simulate. However, as I will show, they fail to quarantine their own idiosyncratic mental states, which results only in contaminated simulation, but produces simulation, nevertheless. I will support these claims by relying on certain empirical data. With the data in hand, I will provide and argument in which I show that, with the inclusion of the notion that poor or failed simulation counting as simulation, ST is more capable of explaining ASD than TT. On the basis of this, I will conclude that the case of ASD provides reasons for favoring ST over TT.

1.5. Concluding Remarks

In the sixth and final chapter of this dissertation, I summarize how the preceding chapters facilitate my adjudication of the debate between TT and ST and, ultimately, support my thesis, which is that – owed to the case of ASD - ST offers us a more favorable explanation of ToM
than does the TT. In an effort to bolster my argument, I will draw on some notions from the history of philosophy concerning the nature of copying and, in particular, when a “bad copy” still counts as a copy of the original. I also discuss that, philosophically speaking, sometimes a copy is so poor (and bears so little resemblance to the original) that it literally ceases to be a copy at all, and is just something different altogether.

In short, by the end of the dissertation, I will have undermined the TT on multiple grounds. I will have shown that the most important reason that the TT fails is because, if the TT were correct, we should expect that anyone without a fully functional ToM would be utterly incapacitated when it came to theorizing in general and especially incapacitated at theorizing in the psychological domain. But, alas, this is not the case. People with ASD do theorize. According to Baron-Cohen’s empirical data, people with ASD absolutely can and do theorize, and they do it particularly well, so much so that they prefer to join in activities that require proficiency at theorizing and systemizing. (I imagine that for the theory theorist, this is a rather unexpected result.) In the absence of the orthodoxy of the TT, we are left to explain, then, what might underpin the ToM ability. I assert it becomes immediately clear that the ST is a superior candidate. The ST can easily account for the triad of impairments that are hallmarks of ASD. Moreover, the special case of ASD shows that the TT cannot be the correct theory of ToM. And most importantly, the case of ASD bolsters arguments favoring the ST: People with ASD cannot mentalize, but they can and do theorize, so the ability to theorize cannot underpin the ToM capacity. However, when and if people with ASD do simulate, they utterly fail at the task, give way to idiosyncratic simulations, fail to make the appropriate egocentric shifts, and so can only simulate poorly, if at all. Thus, simulation seems to be a better explanation of the ToM capacity than its competition, the TT.
CHAPTER 2: THE THEORY OF MIND

In this chapter I discuss a topic of cognitive science referred to as ‘Theory of Mind’ (henceforth, ‘ToM’). This area of cognitive science investigates how we, as human beings, come to explain and predict the mental states and subsequent actions both of ourselves and of other persons. Cognitive scientists claim that the ToM ability (the ability to “mentalize”) is a skill possessed ubiquitously by human beings after reaching certain developmental milestones in early childhood.22 For the purpose of this dissertation, I shall assume that cognitive scientists are correct in their two assumptions (a) that ToM exists and (b) that ToM actually is the medium that affords a person the ability to mentalize.

I will first define and describe the ToM phenomena. Second, I will discuss the point at which, developmentally, we begin to see the emergence of ToM, and will briefly describe how ToM functions in the typical (especially the non-autistic) case. Third, I will discuss the results which occur in the cases of individuals whose ToM is either impaired or altogether absent, and how, interestingly, these results appear strikingly similar to the hallmark traits (or the ‘triad of impairments’) we find in individuals who are diagnosed with autism spectrum disorder (henceforth, ‘ASD’). Finally, I will conclude the chapter with a segue into the remainder of the dissertation by discussing how ToM is relevant to the debate between the two competing theories of ToM that I will be adjudicating in later chapters.

---

22 Alternatively, some cognitive scientists surmise that the ToM ability is a mechanism that is developed ubiquitously by human beings after reaching certain developmental milestones in early childhood.
2.1. A Brief History on The Emergence of Theory of Mind Studies.

Premack and Woodruff (1978) coined the term ‘Theory of Mind’ in their influential and generative article, “Does the Chimpanzee Have a Theory of Mind?” So, the term ‘Theory of Mind’ and its first concise definition – that is, “the original framework” – were both initially derived from research seeking to learn about intentionality in non-human primates. They wrote,

"An individual has a theory of mind if he imputes mental states to himself and others. A system of inferences of this kind is properly viewed as a theory because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior or other organisms."

Premack and Woodruff identified this ability in humans, gave it a name, and were testing to see if the same ability was to be found in apes. Since Premack and Woodruff’s work in 1978, the consensus seems to be that ToM is a capacity that is unique to human beings and higher primate mammals. Indeed it is claimed that ToM is “a quintessential ability that makes us human” (Whitten, 1991).

Research since Premack and Woodruff has attempted to identify why some people are better at explaining and predicting another person’s behavior by appealing to mental explanations than others. ToM studies are now largely performed as a branch of cognitive science. However, cognitive scientists are not alone in their interest in understanding ToM. In fact, ToM lies at the intersection of several disciplines (some of which fall under the purview of cognitive science) including: philosophy of mind, primatology, linguistics, psychology, and developmental

25 While research into whether or not animals possess a Theory of Mind (and which animals possess it) is interesting and relevant to Theory of Mind study as a whole, it is not relevant to this dissertation, and so I do not take up this issue or adjudicate the matter here. That being said, however, I think it is important to state that Whitten’s claim that ToM is “unique to humans and higher primates” is overly strong given the amount of research that has been completed using animal subjects of a variety of species.
psychopathy. In short, ToM is how we understand other people and view them as intentional agents. “‘Theory of Mind’ has come to be shorthand for the capacity to attribute mental states to oneself and to others and to interpret and predict behavior in terms of mental states” (Baron-Cohen, 1995, p. 55). To be clear, ToM does not necessarily imply any actual theory of the mind. Rather, ‘ToM’ is merely the name used to denote the apparent ability to use mental ascriptions for the self and others.

This skill is not only referred to as ‘ToM’ or ‘mentalizing.’ Over time, many terms have been used to describe this ability. It has also been alternately referred to as ‘mindreading,’ ‘social insight,’ ‘cognitive role-taking,’ ‘accurate empathy,’ and ‘understanding other minds.’ It has also been referred to as ‘folk psychology’ by philosophers, and either ‘naïve psychology’ or ‘intuitive psychology’ by cognitive scientists. Most recently, some cognitive scientists have gravitated toward the term empathic accuracy, which is defined as one’s ability to accurately infer the specific content of another person’s covert thoughts and feelings. Yet, philosophers and others continue to use the term ‘ToM,’ ‘mentalizing,’ or ‘mindreading.’

2.2. Theory of Mind Defined and Described

When explaining and predicting the behavior of another, we do so by using mental state terms. This employment of ToM is, in some sense, mindreading. Mindreading (in this sense) is not spooky, mystical, or telepathic. Simon Baron-Cohen claims this is something we do constantly, effortlessly, reflexively, and almost unconsciously. To be clear, the term “to attribute mental states to others” does not only refer to the ability to tell what a person is thinking, but also the ability to infer a full-range of mental states, such as what they may be feeling, believing, wanting, intending, planning, agreeing with, fearing, promising, hoping, (and so on) that causes
their actions. According to cognitive scientists, possession of the ToM skill enables one to understand and evaluate the mental states of others, taking their perspective for both first-person and third-person mental state ascriptions, in multiple contexts, at a very high speed. In brief, “having a theory of mind is to be able to reflect on the contents of one’s own mind and other minds” (Baron-Cohen, 2001, p. 3).

Many cognitive scientists claim that ToM makes human behavior (and to a certain extent, animal behavior) comprehensible and quasi-predictable. It is the understanding of people as “cognitive beings with rich mental lives that are available to themselves and not to others” (Schaffer, 1996, p. 66). Schaffer’s claim is that people possess internal mental representations of the world that influence their behavior. ToM is used to treat other agents as actual bearers of psychological states and processes (which may or may not be observable or salient), and to anticipate and explain the agents’ behaviors in terms of such states and processes. This ability to adopt a ToM is to adopt an intentional stance, which characterizes human social interaction. This stance is useful in everyday life, under ordinary circumstances, and in social interactions. Not only is ToM useful, but also, in some sense, it is required for mere survival. Baron-Cohen suggests, “It is evident, I think, that if another organism’s next action is going to be to attack you, or share its food with you, or mate with you, you would do well to anticipate this quickly” (Baron-Cohen, 1995, p. 12).

For the sake of clarity, I wish to point out that there are, importantly, (at least) two kinds of occasions in which we employ ToM. (1) We generate first-person reports using ToM to understand and interpret our own mental states and the actions that are produced from them. (2) And, we generate third-person reports using ToM to understand and interpret the mental states of
other people, among other things, and to explain and predict the actions that their mental states will precipitate.

There is a long-standing tradition in the history of philosophy of mind which ponders whether we can know the minds (and the mental states) of others, and how we might even go about doing that. This has historically been dovetailed by the oft-made assertion that we can only truly know the contents of our own minds. Mental states were thought to be an entirely private affair and can only be fully known and understood by the person who is having them. Under this tradition, it would seem that ToM is “a diabolically difficult task, in large part because one never has access to the content of another person’s mind” (Myers and Hodges, 2009, p. 281).

According to the traditional Cartesian view, a person has direct knowledge of the best imaginable kind of the workings of his own mind. I “must be directly and authentically seized of the present state and workings of my own mind…. I, alone, can possess direct cognizance of the states and processes of my own mind” (Ryle, 1949, p. 11). Engaging my ToM, I can take a special kind of perception – introspection – to understand my own mental states and predict and explain my own behavior. These are, of course, pretty fairly reliable first-person reports.

The assumption that often follows from this is that people are irremediably blind and deaf to the workings of one another’s mind and inoperative upon them. The theory (originating with Descartes) has been simply that we know the contents of our own minds and can predict our own behavior to create highly accurate first-person reports, but that we cannot do this with others; that we cannot understand, explain and predict the mental states of others and their subsequent actions. Ryle, though contesting the traditional Cartesian view, summarizes the tradition

---

26 The ‘problem of other minds’ harks back to Descartes’ Meditations (1640). The question and was later taken up by Locke (1689), and Mill (1865). The tradition of discourse articulating the problem and its various solutions has continued in contemporary philosophy (see Ryle (1949), Wittgenstein (1953), Sartre (1958), Ayer (1963), and Nagel (1986).
succinctly: “A person has no direct access of any sort to the events of the inner life of another. He can do no better than make problematic inferences from the observed behavior of another person’s body to the states of mind which, by analogy of his own conduct, he supposes to be signalized by that behavior” (Ryle, 1949, p. 14).

As research on ToM has continued, this has turned out not to be the case. It turns out that using a ToM, we use mental states to interpret, explain, and even predict the behavior of other people (in addition to ourselves), and we are seemingly able to generate these third-person reports with roughly the same acuity as we find in our first-person reports. According to Robert Gordon (1986), we can fairly accurately know the mental states of others, and we express this knowledge as we learn to utter sentences that are construed as statements about our own future behavior – that is, our immediate intentions (e.g. I feel hungry. I now intend to drive to the store. I hope to find something I will want to eat there.). We can then apply this knowledge of ourselves to knowledge of others.

We assume that people’s actions are motivated by their desires, in light of their beliefs and, equipped with ToM, we are quite often able to determine what these things are. ToM allows us to view others as individual agents with self-propulsion, goal-directedness, intentionality, language, and emotive expression. Furthermore, we take these to be honest signals of a whole host of inner mental states, all of which are further indicative of future actions, which we can predict by reflecting upon these honest signals and mental states.

Similarly to Gordon, and based on her extensive research in the field, Uta Frith claims that we are able to say a great deal about the inner lives – that is, the mental states – of others.27 We are able to make accurate third-person reports. We can infer, by putting together some certain facts and

---

observations, that others are experiencing certain mental states (such as knowing, believing, feeling, or desiring.) Frith thinks that once we put these facts and observations together, our ToM affords us the ability to attribute mental states to others not just by some handicapped, tenuous, or vague speculation, but rather with logic and precision. For Frith, there is no question that we possess a ToM, the question is, rather, how does the ToM work and what underwrites the ability to mentalize. She says, “We indulge in a kind of unconscious mind reading. We freely assume that we can tell what people are thinking, what they know, and what they don’t know…. Our [tacit] automatic inferences even extend to what kind of emotional states might arise in others” (Frith, 2003, p. 78). I would add to this that our tacit, automatic inferences of others also extend to what kinds of actions might be taken next by others.

As I mentioned earlier, mindreading or mentalizing is not to be erroneously construed as spooky, mystical, or telepathic. Frith also speaks to this point when saying, “To be able to mentalize does not mean that we have fanciful ideas about what might be in the mind of someone else, but that we know for certain what one can surmise about another’s thoughts and what one cannot surmise” (Frith, 2003, p. 214). The point is: an ordinarily developed individual can not only draw inferences about their own mental states, but also about the mental states of others. A typically developing person can explain and predict their own behavior based on the attribution and awareness of mental states (creating accurate first-person reports), and they can explain and predict the behavior of others based on the inferred mental states of others (creating accurate third-person reports). We do this (at least in part) using some measure of discernment.

---

28 Here, the kind of inference that we employ seems to be roughly inductive in nature.
29 The adjudication of the question ‘what underwrites ToM abilities?’ is the very topic of this dissertation.
about what we can and cannot infer, and this ability – the ToM skill – can be thought of as a ubiquitously human characteristic.

2.3. Developmental Emergence and Proper Function of Theory of Mind

My assessment of the current literature on the subject suggests that most researchers in cognitive science and developmental psychology seem to agree that humans are not born with a ToM – rather, this is a skill that is acquired in a series of developmental achievements after a specific amount of brain development approximately by the age three or four.\textsuperscript{30} The question is what supports this activity in the normal case, and what “makes it impossible or difficult to attain (for some individuals)?” (Frith, 2003, p. 80).

Around age three, the developing mind of a human being begins to reach several milestones in rapid succession – the explosion of language, engaging in imaginative play and pretense, development of episodic memory and empathy, development of executive functioning, and the beginnings of interpersonal relationships with peers and adults. One hypothesis of some theorists claims that one of these milestones is the development of the ability to mentalize – that is, the development of a ToM. It is thought that at least a portion of the ToM development comes in conjunction with or as a feature of the attainment of language. This is because by age four, most typically developing children can pick out mental state words from a word list (such as “think”, “feel”, “dream”, “pretend”, “hope”, “wish”, “imagine”) and distinguish these from other kinds of non-mental words (such as “jump”, “walk”, “eat”, “sit”, “move”). By age five, Uta Frith says, the child will already have the beginnings of a fully-fledged ToM (Frith, 2003, p. 81).

\textsuperscript{30} Although it is widely agreed upon that humans are not born with ToM, there is great disagreement among researchers as to both what it is that is acquired and what process(es) or mechanism(s) underpin it – again, adjudication of this debate is the very topic of this dissertation.
While it would be difficult to say which of these developmental milestones is “most important,” it certainly is not far fetched to claim, as Alan Leslie highlights, that the ability to engage in pretense is a major developmental milestone (See Leslie, 1987). “The ability to pretend presupposes the capacity to form and process internal memory representations of mental states, and keep them separate from internal memory representations of physical states” (Frith, 1991, p. 18). It seems the ability to engage in pretense, as with all of these multifarious developmental accomplishments, interactively benefits from the simultaneous development of a functional ToM.

Owed to the development of a fully-fledged ToM, children become able to attribute thoughts and wishes to others and to themselves. As their capacity for language simultaneously develops, they begin to learn what words mean and to distinguish mental state terms from physical state terms and other terms, enabling them to share a make-believe world of pretense play with others. According to Uta Frith, with the joint development of ToM and language, children’s social interactions gain force as they realize that what they wish and what others wish may not always be the same. Making mental state attributions and predictions is, apparently, universal in normal humans older than four years of age.31 Moreover, according to Daniel Dennett, “quite unselfconsciously, we spend a substantial portion of our waking lives formulating the world – not excluding ourselves – in these terms”.32 In time, this implicit insight will give children the

impetus for learning about social negotiation and manipulation and, given sufficient learning experience, the mentalizing mechanism enables the child to learn surprisingly fast about beliefs and deception (Frith, 2003, p. 102-103).

Understanding what ToM, in particular, is and how it functions is pivotal to understanding cognitive development in general. The underlying means of ToM remain unclear. Equally unclear is the relationship of ToM to the other developmental acquisitions of early childhood such as executive functioning, language development, empathy, and episodic memory. Nevertheless, it seems that the interplay between each of the various developmental achievements and ToM, I think, underscores just how very important the development of a functional ToM is. It would be difficult to overstate the importance of ToM, as it appears to be at the very core of how we successfully interact with others. “It is thought that the ability to understand other people’s beliefs, motivations, desires, and goals (mentalizing) is crucial for social interaction” (Frith and Frith, 1999).

ToM allows for complex mappings of intentionality. “Taking the intentional stance, we perceive, interpret, predict, and respond appropriately” (Dennett, 1996, p. 125). Employing the intentional stance, we can move away from mere physical or goal-directed descriptions of our observations of what someone is doing toward descriptions of behavior, which include intention. Baron-Cohen underscores the meaning of intentionality and its application to ToM in the following example:

A full-fledged theory of mind, then, requires a representational system. This permits the representational mapping of others' emotional states in a manner that is different from picking up their emotions directly. For instance, an intention can be mapped onto a representational emotional topology, going from "the fox is chasing the
chicken" (goal-directed) through "the fox is trying to catch the chicken" (intentionality) through "the fox wants to eat the chicken" (motivational) to "the fox is chasing the chicken and trying to catch it because it is hungry and wants to eat it" (emotional). Similarly for the chicken: it is running (goal directed) away from the fox (intentionality) because it is afraid (emotional) of being eaten (motivational) (Baron-Cohen-1995).

Notice, that ToM also requires a representational system that is different and more complex than just picking up on emotions directly. It requires a highly complex system of observation, the ability to take an intentional stance, and explanation/prediction of behavior (ToM). Further, all of this must be conjoined to knowledge of a language that employs mental state terminology. Picking out and stating “The fox who is running is hungry” would never be sufficient to explain and predict the behaviors represented by the agents in this example (provided, of course, that we consider the fox and the chicken to be ‘agents’).

According to H.P. Grice and others, we use mindreading to help us understand what a speaker’s communicative intentions might be. “What does he mean?” is akin to asking “What does he intend me to understand?” We search for relevance. When the police yell, “drop it” the robber knows what the police want him to do and what the “it” is. We use words, facial expressions, body language, context, introspection, and much more to understand another’s communicative meaning. Further, “mindreading is even more essential to decoding figurative speech (such as irony, sarcasm, metaphor, or humor)” (Baron-Cohen, 1995, p. 27). Without communicative intentionality, dialogue seems disconnected and random. We need communicative intentionality to find relevance in any given communicative exchange.

Why do we need a ToM? There are many occasions (every moment of interaction, actually) in which we want or need to know someone’s thoughts, desires, and intentions. “Such knowledge helps us predict that person’s behavior, which is especially important if that behavior
has consequences for us or if we need to adjust our behavior to it – as when we cooperate or compete. Such knowledge may also help us communicate. It provides the context for interpreting what people say to us and for framing what we say back” (Batson, 2009, p. 276). ToM is at the very heart of our social communication and interaction with others.

ToM is required in order to understand most conversational implicatures and for adhering to the kinds of ‘speech rules’ that Grice refers to – things such as controlling one’s volume, tone, inflection, and content; staying on task and being relevant; knowing when it is okay to interrupt; being able to tell when your listener is listening; avoiding ambiguity and obscurity; presenting your information briefly and concisely – these, Grice says, are the cooperative principle (CP) of conversational maxims relevant to conventional implicature. Lacking proficiency in ToM skills may result in such a dearth of social skills that it becomes impossible to engage in Grice’s CP.

To summarize, skillful employment of ToM allows us to engage in linguistic pragmatics including:

- Tailoring one’s speech to a particular listener
- Adapting the content of one’s speech to what your listener already knows or needs to know
- Respecting (Grice’s) conversational maxims: be truthful, relevant, concise, and polite
- Taking appropriate turns in conversation so that there is room for both participants to engage in the dialogue
- Being sensitive to the other person’s contribution to the conversation
- Recognizing the right or wrong things to say in a particular context
- Staying on topic
- Appropriately helping your listener to follow when a topic change is occurring.
- Understanding a speaker’s literal, non-literal, and hidden meanings of their utterances.

Literally every aspect of pragmatics involves a sensitivity to the mental states of both the speaker and the listener, and, hence, mindreading.

In short, as Wellman says, “Why is achievement of ToM important?... It is our framework theory of persons…It dictates our basic ontology…. It dictates our causal-explanatory
infrastructure, our basic grasp of how to go about making sense of ourselves and others”
(Wellman, 1990, p. 328).

2.4. Absence of Theory of Mind and Potential Resultant Impairments

The foregoing sections of this chapter say what a ToM is, how various researchers from various disciplines denote, define, and describe ToM, and the emergence and applications of a properly (i.e. typically) developing ToM. I think it is not enough to only discuss what it looks like to possess and employ a typically developed ToM, but that it is equally important to discuss what it would look like to have an atypically developed ToM – that is, either an impaired or altogether absent ToM.

Explaining and predicting the behavior of oneself and of others, and doing so in terms of mental states, is a highly complex cognitive task that can rightfully be seen as a great feat of mental gymnastics. The absence or erroneous operation of a typically functioning ToM, can be the source of inaccuracy and miscalibration in perspective taking, which is absolutely key to performing ToM tasks. Without ToM, there can be no meaningful psychological inferences about one’s own behavior or the behavior of others and there can be no understanding that a person’s beliefs affect their actions.

ToM is so automatic, such a deeply unconscious activity, and our use of it so constant and reflexive, that perhaps its features are more salient when they fail than when they do not fail. “Abilities are of practical importance only when people use them, and there is not more immediate barrier to accurate perspective taking than failing to use it in the first place” (Epley and Caruso, 2009, p. 297). One’s failure to successfully engage ToM and the resultant impaired
perspective taking would result in taking an incredibly literal interpretation of the world. The quintessential case of failure of communication due to an impaired or absent ToM will “occur precisely where nothing but the intention to communicate is conveyed” (Happé, 1993, p. 103).

Accordingly, ToM is also required to decipher and comprehend conversational implicature because often what is implied, suggested, or meant, is radically distinct from what is being said. This is the case with all non-literal speech forms. At a minimum, a deficient, impaired, or absent ToM would present as an inability to understand irony, pretense, humor, sarcasm, figurative or idiomatic speech, hyperbole, false belief awareness, ambiguity and deception - all of which, on some level, require ToM. The ability to understand non-literal speech - that is, being able to detect that a person is speaking non-literally and to know the intention that underlies the non-literal speech – is reliant upon the ToM ability.

Some examples are in order. In absence of ToM, understanding sarcasm and, in particular, the mental states that underpin it would be nearly impossible. “Sarcasm involves interpreting a message in precisely the opposite tone of its literal content and is therefore communicated by the

---

33 Baron-Cohen (1997) offers an example of what literal interpretations of the world might be like for someone who is aware of physical things but not mental things. Baron-Cohen’s example is that of a simple human act, “John walked into the bedroom, walked around, and walked out.” He claims that a mindreader (using ToM) will make sense of John’s behavior by using mental state attributions such as “John wanted something and thought it was in the bedroom, but realized that he forgot he left it in another room.” Without ToM, Baron-Cohen claims, we would have to make sense of John’s behavior without the use of mental-state terms or attributions. For example we might say, “Maybe John just does this every day at this time: he wake into the bedroom, walks around, and walks out again.”

Gopnik (199) offers a much more frightening (indeed nightmarish) example to describe the chasm between mindreaders and those who are mindblind. She says, “This is what it’s like to sit round the dinner table. At the top of my field of vision is a blurry edge of my nose, in front are waving hands…Around me bags of skin are draped over chairs, and stuffed into pieces of cloth, they shift and protrude in unexpected ways…Two dark spots near the top of them swivel restlessly back and forth. A hole beneath the spots fill with food and from it comes a stream of noises. Imagine that the noisy skin-bags suddenly moved toward you, and their noises grew loud, and you had no idea why, no way of explaining them or predicting what they would do next.”

Importantly, here, Gopnik’s example makes a whole host of (silly) assumptions that the mindblind person also is incapable of understanding language, or that the “skin-bags” are in fact other people who are eating. Presumably one does not rely on ToM to discern that people surround one, or to describe their actions. Nevertheless, I think the point that Gopnik is attempting to make salient is how terrifying it must be to be mindblind (i.e. to be without the ability to read and understand behavior in terms of mental states.)
tone in one’s voice or other paralinguistic cues rather than by the literal content of the message” (Epley and Caruso, 2009, p. 300).

Suppose that a teenage girl, who is attempting to be mean to another teenage girl, might sarcastically say, “Nice dress.” If the girl is in fact utilizing sarcasm as a platform for being mean, the literal words (according to Epley and Caruso) will be accompanied with paralinguistic cues indicating this to be the case. She may say the words ‘nice dress’ using a flat or even a nasty tone of voice – one in which the accent or tone applied to the word ‘dress’ will tend to fall couple of notes – and she may employ other paralinguistic cues: such as rolling her eyes, or making a smug or sour facial expression.

Conversely, the literal content of the words ‘nice dress’ is just that. The literal connotation says, “I think your dress is nice,” or something such as, “In my opinion, your dress is a nice dress,” or even just simply, “I like your dress.” When the paralinguistic cues are consistent with the literal content of the words being uttered, we take the speaker to truly think that the dress actually is nice. In this case, the speaker’s paralinguistic clues, by contrast, may include a smiling face, a bright and friendly disposition, and the accent or tone applied to the word ‘dress’ will tend to raise couple of notes. The listener must attend to the whole range of the paralinguistic cues utilized by the speaker to know what the speaker means by the utterance, ‘nice dress.’

Notice that attending to the speaker’s tone, body language, facial expressions, and other paralinguistic cues are necessary to detect the sincerity or insincerity of the speaker’s utterance and to determine whether sarcasm is at play. However, it is important to note that, while noticing the paralinguistic cues is necessary to detect the usage of non-literal speech, these cues alone are not sufficient to know the intention or the mental states that underpin the utterance. In other
words, while tone alone may alert the listener to the fact that the speaker is speaking non-literally, tone alone cannot aid the listener to further infer the mental states and intentions of the speaker. In this example, the mental states inferred by the targeted teenage girl might include: “She does not like my dress.” “She is being mean.” “She does not like me.” “She is not my friend.” “She is making fun of me.” Assuming these mental states would allow the listener to then predict her own behavior as well as the other girl’s behavior: “She will be mean to me the next time I see her.” “I should avoid her in the hallway.”

To draw such mental state inferences and to use them to explain and predict behavior takes much more than simply attending to tonal changes in a speaker’s voice or other paralinguistic cues. And, it takes more than simply recognizing that sarcasm is at play. To do this, - that is, to know the literal meaning and intention that are behind the literal speech, the listener must skillfully engage in the ToM process. In the case of an impaired or absent ToM, the listener may be aware that sarcasm is at play but will likely not be able to take the second (and arguably more important) step: to interpret the utterance in light of mental states and explain and predict behavior in accordance with those mental states. The same or similar difficulties will follow suit with examples of irony, pretense, metaphor, and virtually all other forms of non-literal speech.

Grice uses the example of someone stating idiomatic expression, “He was in the grip of a vice.” Here, knowing the English language is insufficient to know what the speaker means by the utterance. The literal interpretation of the words is vastly different than the meaning and intention of the speaker. Employment of a ToM, allows the listener the ability to determine whether a part of someone’s anatomy was actually caught in some kind of tool, or if that person was unable to rid himself of a bad habit. Similar difficulties arise with the use of many common idioms: “I’m between a rock and a hard spot.” “He has an axe to grind.” “Let’s not beat a dead
horse.” “She has a heart of gold.” None of these idiomatic utterances mean what they say in a literal context.

Notice it is not enough to know the literal meaning of such utterances. Nor is it enough to know the idiomatic colloquialism or the conventional use of the idiom. Of course, as any foreign person attempting to learn English knows, idioms tend to be odd speech utterances and the only way to know recognize them is via rote memorization, much like one might memorize a vocabulary list. Like sarcasm, recognizing the employment of an idiom is but the first step in what I take to be a two-step process. Recognizing the idiom from the list of idioms one has studied will ‘tip off’ the listener that non-literal idiomatic speech is at play. However, there is a second step necessary in order to understand the speaker’s meaning and intention. The listener must also be able to infer a full range of mental states, employing a ToM, to know that the speaker is not speaking literally and to infer what the speaker means. Great confusion arises when taking idioms and other non-literal speech forms as literal speech utterances. Such confusion is an example of how an impaired, deficient or altogether absent ToM will result in grave difficulties or total failures in social communication and social interaction.

Another, perhaps clearer, example is found in looking to metaphor as a non-literal speech form. Metaphor, much more than either sarcasm or idiomatic speech, absolutely requires some understanding of intentions because metaphor functions much more like a loose interpretation of a speaker’s thoughts. As such “metaphors cannot be fully understood or properly used without a first-order ToM” (Happé, 1993, p. 104). Using literalness as a default, or memorizing them (like a vocabulary list of idioms) as a compensatory mechanism, simply will not suffice to understand the metaphor. Interestingly, according to Happé, we can draw a parallel between failing to understand the intention behind metaphors and failing the false belief test since, in both cases,
“the actor’s mental states (belief) is crucial, and reality alone is no guide to actions, so in metaphor the speakers’ mental state (intention) is vital” (Happé, 1993, p. 104).

The cases of non-literal speech utterances such as sarcasm, idiom, and metaphor, (like false belief tasks) all “require the processing of literal content through a pretense” (Egan, 2008, p. 385). And it is widely accepted that engaging in pretense requires a ToM (see Wellman, Gopnik, Happé, Frith, and Baron-Cohen). In all of these cases, the paralinguistic cues may tip off the listener that non-literal speech is at play. However, something additional is required.

“Interpreting a metaphor or an idiom involves representing a particular (literal) content as pretended, and deriving the speaker’s intended meaning by reference to that pretense, i.e. by working out what would follow if one were so pretending” (Wearing, 2012, p. 501). What is additionally required, I maintain, is a typically functional ToM.

2.5. Correlation Between Theory of Mind Deficits and Autism Spectrum Disorder

In the previous section, I have noted that the features of ToM are, perhaps, most salient when they are absent, such as in various cases of non-literal speech utterances. I have also described some results and some potential difficulties that can be experienced by individuals in who the ToM skillset is impaired, deficient, or absent. As a matter of great interest, the most salient diagnostic characteristics described in individuals with autism spectrum disorder (henceforth, ‘ASD’) closely mirror the results we expect to find in the absence of a typically functioning ToM. “It is widely reported (Happé, 1991; Tantum, 1991) that even the most verbally able autistic people fail to understand non-literal speech such as irony, joking, and metaphorical expressions” (Happé, 1993, p. 103). In these instances, communication will noticeably
breakdown whenever the speaker’s attitude and mental states must be taken into account by the listener to modify the literal content and infer the non-literal content.

In fact, when studying ASDs, the social consequences that result from such conversational implicature follies, which can be justifiably viewed as the absence of ToM, are striking. Frith writes,

[A] normal child comes equipped with a mechanism for manipulating representations of mental states, and that, given normal developments, this mechanism causes us to understand mental states such as pretence and belief. We hypothesizes that autism results if this particular component of the mind is faulty. Above all, the fault would impeded development and learning of social imagination and communication skills (Frith, 1991, p. 19).

A cognitively typically developing person without ASD will view the world and couch it in terms of both physical and mental states and this ability will help her successfully navigate the social world. By contrast, a person with ASD will firmly grasp the world and couch it almost strictly in terms of physical states, but will seldom view the world in terms of mental state, nor will they describe the world or their views of the world using mental state terms. They will appear to know the physical functions of the body and brain but fail to mention any mental function of the brain (Baron-Cohen, 1989a). “[Individual on the autism Spectrum have] significantly less success in inferring mental states as compared with physical states” (Kaland, et. al., 2002, p. 524). According to Barron-Cohen, “the full range of mental-state terms is missing both in the language and the thought of people with autism” (Baron-Cohen, 1995, p. 84).

Here, an example will server to illustrate this point. Imagine that you are a patron in a coffee house. Presently, you observe another patron – a young woman. You watch as the young woman enters the coffee house, approaches the counter, and looks over the menu. She peers through the glass cases at the offerings of pastries and half sandwiches, and then abruptly departs. If you are
a typically developed person in possession of a robust ToM, you will describe what you’ve witnessed in terms of both mental states and physical states: She came in because she was looking for something in particular; She might have been craving something; She wanted something specific; When she didn’t find it perhaps she felt disappointed and decided to go elsewhere, and so she left. As a person without ASD, you would use all of your awareness of the young woman’s mental states to explain and predict her behavior. In stark contrast, if you were a person with an ASD and so not in possession of a full ToM, you would describe this same event strictly in physical terms and not at all in mental state terms: She came in the coffee house; She looked around; She left. Indeed, the more fractured your ToM, the more befuddling this event would be to you, and you might only be able to offer some temporal explanation for the young woman’s behavior, such as “Perhaps she does this every day.” Notably, your explanation of her behaviour would be devoid of mental state terms such as looking for, craving, wanted, felt, decided, and etc. Many theorists often assume that ToM – that is, the skill the ability to attribute mental states to others to explain and predict their behavior - is crucial for successful social interaction.

The view that ToM is necessary for every day social interactions is, perhaps, most strongly stated in the assertion by Baron-Cohen, Leslie, and Frith in the ToM deficit account of autism (1985), where they claim that failure to acquire and develop a ToM results in ASDs. In fact, they claim that the whole triad of impairments seen in ASDs can be reduced to the single cognitive deficit in ToM resulting in an inability to cognitively represent mental states. “The so-called “ToM deficit” hypothesis proposes that a fault in just one of the many components of the social brain can lead to an inability to understand certain basic aspects of communication…This can explain both aloofness and indiscriminate social approach” (Hill and Frith, 2006, p. 283) often
seen in people with ASD. “This theory, sometimes referred to as ‘mindblindness’ or ‘mentalizing failure’, has been tested extensively (see chapters in Baron-Cohen et al. (1993, 2000)) and proved fairly robust” (Hill and Frith, 2006, p. 283).

Mindblindness is Baron-Cohen’s way of conceptualizing the impaired ability of people with ASD to understand and respond to what other people are doing and thinking. People on the spectrum have an impaired or absent ToM. Mindblindness with atypical (dampened) responsiveness to others’ mental states conceptualizes the first diagnostic criterion for ASD: impaired social interaction. These deficits lead to diminished ability to interact with others in normal (i.e. socially expected) ways.

The ToM impairment or deficit is a part of the core feature of autism in which the person with ASD experiences a profound disorder in understanding and coping with the social environment. Being able to conceive of mental states and infer them is a crucial aspect of social skills competency and ToM is but one factor (albeit a crucial one factor) for social competence. “Difficulty in understanding other minds is a core cognitive feature of ASD’s. The ToM difficulties seem to be universal among such individuals” (Baron-Cohen, 2001, p. 3).

A ToM impairment or deficit, such as we see or believe we see in ASDs, would lead precisely to a literal interpretation of others’ behavior and utterances without regard for context. All of the aforementioned non-literal, idiomatic, and conversational implicatures are just largely lost on individuals with ASD. The end results are that ASDs will tend to choose physical state descriptions over mental state descriptions and the ‘social awkwardness’ we see so prevalently in people with ASD.

Indeed, some theorists claim that the whole triad of impairments that are seen as the behavioral and diagnostic hallmarks of ASD can be explained by an underlying impairment or
failure to understand and employ mental states as a way of understanding others. This failure would necessarily confine a person to excessively literal interpretations of speech, which is considered one key diagnostic criterion for ASD. “In addition, we can see how a failure to relate to people emotionally (as is the case with people on the spectrum) could also be explained as an inability to comprehend their belief states” (Mitchell, 1997, p. 89).

Other elements of the triad of impairments meeting the diagnostic criteria for ASD can also be viewed as shortcomings in the possession of a typically functioning ToM. This would include an ignorance of others’ beliefs and mental states resulting in treating people as machines, difficulty relating to and communicating with others, diminished ability or no ability at all to engage in pretend or imaginative play, a dearth of eye contact and social awareness, and one sided interactions coupled with autistic aloneness can all be understood as results of a ToM deficit. Baron-Cohen writes, “hence the three cardinal symptoms of autism – the abnormalities in social development, in communication development, and in pretend play – might be the results of a failure in the development of mindreading” (Baron-Cohen, 1995, p. 63).

The correlation between what we would expect to see in the absence or impairment of ToM and what we do in fact observe in ASD is striking and lends a great deal of credence to Baron-Cohen’s ToM deficit hypothesis of autism, or his Mindblindness as an account of autism. The ToM deficit sheds lights on every aspect of the triad of impairments that comprise the diagnostic criteria for ASD and all of “The conspicuous communicative impairments in autism could fall under the umbrella of “deficient ToM” (Mitchel, 1997, p. 86).
2.6. Conclusion

This chapter has explored a branch of cognitive science known as “theory of mind.” I have defined and described ToM at length, and given examples of how it functions when functioning properly or typically (especially in people without ASD). I have also provided examples of results we could reasonably expect in the breakdown case when there is an impairment, deficit or absence of ToM in an individual. I have drawn the correlation between a deficit in ToM and the diagnostic criteria for autism spectrum disorder. This correlation bolsters Simon Baron-Cohen’s assertion that the triad of impairments found in ASD can be reasonably viewed as mindblindness under his ToM deficit hypothesis.

Explaining ASD is of deep theoretical interest given its relation to the issue of how the mind (and the ToM) is normally comprehended. In the chapters that follow, I will discuss two of the main competing theories of ToM (the theory theory of mind and the simulation theory of mind) and some iterations of each of these theories. I will discuss the most plausible iterations of each of these competing theories. There is much debate and it remains thus far unclear as to which theory provides the most explanatory power of ToM. The adjudication of this matter, I shall take up in great detail throughout the remainder of this dissertation.
CHAPTER 3: AUTISM SPECTRUM DISORDER

What, if anything, does autism have to do with the Theory of Mind debate? The simple answer is that, according to Theory of Mind theorists, it appears that autistic individuals tend to bear impairments in the functions that are ordinarily explained by, or fall under the umbrella of, Theory of Mind (also called ‘mind-reading’ or ‘mentalizing’) capabilities – that is, the ability to explain and predict behaviour in terms of mental states. Both the theory theory (henceforth, ‘TT’) and the simulation theory (henceforth, ‘ST’) attempt to explain both “normal” Theory of Mind (henceforth ‘ToM’) acquisition, and what is acquired, in typically developing individuals.

I am interested in the ToM that is either fractured or altogether absent when acquired by certain atypically developing individuals, specifically developmentally delayed individuals who fall on the autism spectrum. “Even though quite difficult to grasp and prone to misunderstandings, the notion of a lack of an everyday “Theory of Mind” in autism has become widely known and accepted…and a large body of empirical work now supports this hypothesis” (Frith, 2003, p. 79). Because these individuals have ToM deficiencies, by seeing why they cannot mentalize, we can then get a sort of confirmation for one or another of the competing accounts of ToM – the TT or the ST. The theory theorist and the simulation theorist will have different answers as to how the ToM, which is responsible for mentalizing, works. Hence, looking at the case of autism (in which there is a lack of mentalizing) can assist in settling the debate between the competing theories.

---

34 Frith makes two assumptions in her work: (1) that the ToM exists and exists as a mechanism in the brain; (2) that people with autism have a ToM deficit, which explains their autistic impairment(s).
In her article *What Could Explain Autism?* Jill Boucher (1996) suggests multiple criteria that ought to be met in any discussion concerning autism at the theoretical level. As an initial criterion, she contends that, to be considered satisfactory, any theory about autism must necessarily include both what definition of autism is being used and what concept of autism is being assumed (Boucher, *Ibid.*). Boucher calls for this specifically in light of the fact that autism has proven, over time, to be remarkably difficult to describe and explain. This difficulty arises because autism is always defined by behavioral traits, and the diagnostic criteria for autism has repeatedly changed over the last 50 years as our understanding of autism has evolved. In fact, even the most current science of autism remains dynamic. While we know more about what autism is *not* more than ever before (i.e., it is not schizophrenia, psychopathy, retardation, or as a result of “refrigerator mothers” or vaccinations), we still cannot claim to know what autism *is* or, even less, what causes it. I thus concur with Boucher’s suggestion of an initial criterion that is highly descriptive of the autistic phenomenon we seek to explain and just what we are explaining, and henceforth this chapter will be structured in accordance with Boucher’s initial criterion.

To set the stage for the debate over which mind-reading theory best explains autism, I will begin by discussing autism in a general way. I will pinpoint the impairments and propensities that are most common in autistic individuals, and will do this in a way that is accessible to those who are not specialists in autism but who are interested in the debate between theory theorists and simulationists. To accomplish this, I will first give an overview of the various descriptions and diagnostic criteria for autism and how they have evolved over time. I will, of course, include the psychiatric diagnostic categories and criteria from the primary diagnostic reference used in the United States by mental health professionals: the *Diagnostic and Statistical Manual of*
Mental Disorders (henceforth, ‘DSM’) published by the American Psychiatric Association (henceforth, ‘APA’). There have been several revisions to the diagnostic criteria for autism in the various DSM publications over the last several decades, and reviewing these will help to underscore both the evolution in our understanding of autism as well as the difficulty experienced in attempting to describe and explain autism. My discussion will secondarily include the somewhat less rigid and more varied ways in which autism is frequently viewed or described by scholars in relevant fields such as developmental psychology and cognitive science. Finally, in accordance with Boucher’s initial criterion, I will discuss which definition and concept of autism I deem relevant to the task of adjudicating the debate between mind-reading theories.

Before elucidating the rich history of the diagnostic criteria for autism, let me say why the history is important and relevant for my project. Admittedly, the history is lengthy. Its inclusion is, however, necessary for several reasons. First, the history of the diagnostic criteria for autism displays a wide variety of attempts to explain and characterize autism. It underscores the great difficulty we have had at understanding and conceptualizing autism, which has eluded and puzzled us for many years. It seems that if I am to argue for either one or another of the competing ToM theories to support an explanation of how we mentalize, I ought to first display that I have taken into account the lengthy history of autism and thus fully grasp the implications of my suppositions in confirming or denying any account of ToM.

3.1. A Not-So-Brief History of The Diagnostic Criteria for Autism

I concur with Boucher in that, “[The long history] should be taken into account by anyone… using the case of autism to argue for one or other particular theory concerning mentalizing and the development of mind” (Boucher, 1996). Hence, while charting the history and evolution of
the diagnostic criteria is a lengthy inclusion, it is, nevertheless, relevant to my task of adjudicating the debate between competing theories of mind. Thus, I include it here.

The term autism was first coined in 1908 by a Swiss psychiatrist, Dr. Paul Eugen Bleuler. Bleuler derived the term from the Greek *autos* (self) with the suffix *ismos* (an action or state of). The term was introduced as *autismus* and was meant as a descriptor to reflect the morbid self-absorption that Bleuler noticed in a specific set of his schizophrenic patients.

*Autism*, as a behavioral description, was first introduced by Dr. Leo Kanner of John’s Hopkins Hospital in 1943. Kanner had noticed several striking behavioral similarities in a group of children he was studying. These similarities included what he called “autistic aloneness” and the “insistence on sameness” of routines. Kanner dubbed the term “early infantile autism,” which is now often referred to as “Kanner’s Syndrome” (henceforth, ‘KS’) or “Classic Autism.” Many of the behaviors Kanner identified in his 1943 paper, *Autistic Disturbances and Affective Contact* (Kanner, 1943), continue to be considered the primary behavioral traits common to individuals who meet the modern diagnostic criteria for autism. However, since KS was identified, contemporary diagnostic criteria for autism has broadened considerably to include many more traits, symptoms, and behaviors.

The APA published the first DSM in 1952. Despite the fact that Kanner had identified and introduced “early infantile autism” nine years prior to its publication, the DSM did not include “autism” as a separate or discrete diagnosable condition. Instead, perhaps as reflective of Bleuler’s findings, it included autism under the diagnosis for schizophrenia in children. Autism
was considered a primary manifestation of the psychotic reactions found in children in whom schizophrenia was diagnosable prior to puberty.\textsuperscript{35}

Similarly, when the APA published the DSM-II in 1968, autism, again, was not mentioned as a separate discrete diagnosis and, in fact, the word, ‘autism,’ does not appear anywhere in the DSM-II except, again, under the diagnostic criteria for schizophrenia in children. Here, the APA still considered autistic behavior to be one of several behavioral characteristics that were manifestations of schizophrenia in children whom were diagnosable prior to puberty.\textsuperscript{36}

It was not until the DSM-III was published in 1980 that the APA briefly described autism a separate diagnosis, apart from, and unrelated to, schizophrenia. Using a name similar to that coined by Kanner in 1943, it was introduced into the DSM-III as “infantile autism.” Importantly, the DSM-III at last distinguishes individuals with autism from individuals suffering from schizophrenia and, in fact, specifically calls for an absence of schizophrenic like symptoms such as delusions, hallucinations, loosening of associations and incoherence. Instead, the DSM-III Diagnostic Criteria for infantile autism begins to reflect similarities to KS or “Classic Autism” from Kanner’s descriptions, which are still prevalent in the modern conception of autism, e.g., both a lack of responsiveness to other people (i.e., Kanner’s “autistic aloneness”) and a resistance to change (i.e., Kanner’s “insistence on sameness”). We also see the addition of the specific communication difficulties affecting pragmatics and the rigidity that have come to be thought of as hallmarks of the modern conception of autism.\textsuperscript{37}

\textsuperscript{35} For the full text from the DSM-I (1952) Diagnostic Criteria for 000-x28 Schizophrenic Reaction, Childhood Type, please see the Appendices to this chapter, Appendix A.
\textsuperscript{36} For the full text from the DSM-II (1968) Diagnostic Criteria for 295.8 Schizophrenia, Childhood Type, please see the Appendices to this chapter, Appendix B.
\textsuperscript{37} For the full text from the DSM-III (1980) Diagnostic Criteria for Infantile Autism, please see the Appendices to this chapter, Appendix C.
Just seven years later, in 1987, the APA published a revision to the DSM-III, the DSM-III-R. The revised version deleted the diagnostic criteria for “Infantile Autism” and replaced it with the diagnostic criteria for “Autistic Disorder” (henceforth, ‘AD’). The new diagnostic criteria are representative of a substantial shift in the historical understanding and explication of autism. Specifically, the diagnostic criteria changed from five general characteristics to sixteen more specific characteristics, and it allowed for the presence of some symptoms with the absence of other symptoms.

The new criteria detailed many behavioral characteristics that were exhibited as qualitative impairments in reciprocal social interaction, verbal and nonverbal communication, and imaginative activity. It asserted the presence of a marked restricted repertoire of activities and interests. And it specified that the onset of diagnostically significant behaviors and symptoms be during infancy or early childhood. By so doing, the DSM-III-R begins to reflect our modern understanding and explanation of autism, and is the starting place for what has come to be called the “triad of impairments” thought to be inherent in autism.\(^{38}\)

Though not explicitly employed in the DSM-III-R, the changes brought about the term “autism spectrum” into the vernacular of both developmental psychologists and laypeople alike, because this edition of the DSM defines autism by behavioral criteria. This term was meant to capture the apparent fact that not every autistic individual bears precisely the same behavior characteristics, and they do not necessarily share the same level of impairments – an individual might present with mild autism or severe autism. Thus the concepts of “high-functioning autistic” and “low-functioning autistic” were also introduced into the vernacular, furthering the

\(^{38}\) Notice that in the lengthier, more descriptive diagnostic criteria found in the DSM–III-R the symptoms originally noted by Kanner for KS are still present. For the full text from the DSM-III-R (1987) Diagnostic Criteria for Autistic Disorder 299.0, please see the Appendices to this chapter, Appendix D.
notion of a spectrum disorder. Moreover, the changes to the diagnostic criteria indubitably widened the diagnostic boundaries, cast a larger net and, as a result, many more individuals were then diagnosable as autistic or fell under the autistic spectrum. This trend would continue as our understanding of autism evolved and the APA introduced more revisions to the DSM that would include both AD and variations of AD classified under the umbrella of Pervasive Developmental Disorders (henceforth, ‘PDD’).

In 1994, the APA released the DSM-IV with some substantial and significant revisions related to AD. First, while still predominantly adhering to the descriptive language for KS, as developed by Kanner, the diagnostic criteria for AD became substantially streamlined, containing more descriptive specificities from what was published in the DSM-III-R. The DSM-IV furthers the modern concept of autism as possession of a triad of impairments. Specifically, the DSM-IV requires for diagnosis, some symptoms in each of the following categories: qualitative impairments in social interaction and communication; restricted and repetitive patterns of behavior and interests; delays or abnormal function in social interaction, language used in social communication, and symbolic or imaginative play; and that the disturbance is not better accounted for by Rhett’s Disorder or Childhood Disintegrative Disorder.

39 For the full text from the DSM-IV (1994) Diagnostic Criteria for Autistic Disorder, please see the Appendices to this chapter, Appendix E.
40 Rhett’s Disorder is also an old diagnostic term for a PDD that is no longer found in the DSM-V (2013), and that is now under the umbrella of autism spectrum disorder (ASD). Rhett’s Disorder was found almost exclusively in females, and was characterized by normal development until 6–18 months of age followed by a severe regression in motor, social, and language skills and a deceleration in head growth.
41 For the full text from the DSM-IV-TR (2000) Diagnostic Criteria for Rhett’s Syndrome and Childhood Disintegrative Disorder, please see the Appendices to this chapter, Appendix F and Appendix G respectively.
A second, and perhaps more interesting, change to the DSM-IV, is the addition of diagnostic criteria for Asperger’s Disorder (henceforth, ‘AS’).\textsuperscript{42} Asperger’s Syndrome was first described in the 1940’s by a pediatrician in Vienna, Dr. Hans Asperger. Though on separate continents and with no known correspondence between Dr. Asperger and Dr. Kanner, Asperger noticed particularly autistic traits in (primarily male) children similar to those that Kanner was noticing in his male patients. The primary difference between KS and AS was that, unlike subjects studied by Kanner, Asperger’s subjects displayed normal or above average intelligence and did not display impaired or delayed language development.

Unlike KS, or “Classic Autism”, however, AS did not receive any attention as a separate diagnosis until Uta Frith characterized Asperger’s as a “syndrome” in her book, \textit{Autism and Asperger’s Syndrome}. Here, Frith describes individuals with AS as having “just a touch of autism” (Frith, 1991). In the years between Frith’s description of AS in 1991 and the publication of the DSM-IV in 1994, AS came to be described by mental health professionals as ‘high-functioning autism’ or ‘mild autism.’ In 1994, the DSM-IV includes AS as a separate diagnosis from AD, and includes both AD and AS (as well as Rhett’s Disorder and Childhood Disintegrative Disorder) under the category of “Pervasive Developmental Disorders” (henceforth, ‘PDD’).

Third, another change found in the DSM-IV is the addition of new diagnostic criteria for a new category of PDD: Pervasive Developmental Disorder - Not Otherwise Specified (henceforth, ‘PDD-NOS’).\textsuperscript{43} PDD-NOS functions as a sort of ‘catch-all’ for individuals who may

\textsuperscript{42} Since the DSM-IV specifically states that Asperger’s Disorder may also be called \textit{Asperger’s Syndrome}, I will, hereafter, refer to it as ‘AS’ so as not to confuse it with \textit{Autistic Disorder}, which I have already designated henceforth as ‘AD’.

\textsuperscript{43} For the full text from the DSM-IV-TR (2000) Diagnostic Criteria for Pervasive Developmental Disorder – Not Otherwise Specified, please see the Appendices to this chapter, Appendix H.
display multiple traits typical of AD or AS, but who may not meet the criteria precisely due to a variety of reasons including (but not limited to) late age onset or atypical and/or sub-threshold symptomatology, and who do not meet the criteria for other PDDs such as Rhett’s Syndrome or Childhood Disintegrative Disorder. The addition of PDD-NOS is important to the modern conception of autism as a spectrum because the diagnosis allowed for individuals with atypical autism or even extremely mild autism to be included on the autism spectrum, thereby assisting them in getting the services they need (such as speech therapy, social skills training, and so on).\textsuperscript{44}

Six years later, in 2000, the APA published a revision to the DSM-IV: the DSM-IV-TR, with updated diagnostic criteria for all PDDs. In the DSM-IV-TR, the discussion of AD is lengthier and more detailed, but the diagnostic criteria remains as it was written in the DSM-IV.\textsuperscript{46} However, this edition gives both a much lengthier discussion of AS, and more specific diagnostic criteria for AS.\textsuperscript{47}

\textbf{3.2. The Current DSM: DSM-V and The Current Conception of Autism}

The DSM has undergone further revisions. The revisions to the diagnostic criteria for all PDDs including AD have been open to public critique since some time in 2011. After much debate, the revision were finalized and published in the spring of 2013. There are significant

\textsuperscript{44} PDD-NOS is an old diagnostic term that is no longer a legitimate diagnosis in the DSM-V (2013), and is now simply under the umbrella of autism spectrum disorder (ASD).

\textsuperscript{45} PDD-NOS was used as a diagnosis for severe and pervasive impairment in the development of social interaction and communication skills but when the criteria for PDD and ASD are not met. It is often considered the mildest form of ASD, although this is not always true: One may have quite severe impairments that are associated with ASD, while other ASD-associated impairments may be mild or non-existent.

\textsuperscript{46} For the full text from the DSM-IV-TR (2000) Diagnostic Criteria Autistic Disorder, please see the Appendices to this chapter, Appendix I.

\textsuperscript{47} For the full text from the DSM-IV-TR (2000) Diagnostic Criteria for Asperger’s Syndrome, please see the Appendices to this chapter, Appendix J.
changes to the diagnostic criteria and categories for all PDD’s. According to the APA, it is hoped that the revision reflects our most current understanding of autism and does away with the needlessly overly complicated distinctions between the various disorders that previously either erroneously fell under the criteria for childhood type schizophrenia (in DSM-I through DSM-II), or were parsed apart by small categorical distinctions (e.g. as in the distinctions between Rhett’s Disorder and Childhood Disintegrative Disorder, or the distinctions between AD, AS, and PDD-NOS) (in DSM-III through DSM-IV-TR).

The DSM-V specifically changed the way all PDDs are diagnosed and coded by mental health professionals. The revisions include: (1) complete removal of all previously separate diagnoses under the category of PDD, \(^{48}\) and (2) a complete overhaul in the diagnostic criteria for AD. In fact, the DSM-V deletes “AD” from the manual and replaced it a unitary diagnosis of “Autism Spectrum Disorder” (henceforth, ‘ASD’), an umbrella term that now incorporates the several previously separate diagnoses (including AD, AS, Rhett’s Disorder, Childhood Disintegrative Disorder, and PDD-NOS). Under the new diagnostic umbrella of ASD, further diagnostic distinctions are to be made based on severity levels which are based on the amount of support required by the individual to face their challenges with social communication and restricted interests and repetitive behaviors. \(^{49}\)

There has been much research and debate over the last twelve years (since the release of the DSM-IV-TR) as to whether autism is best seen as a (a) syndrome, (b) a distinct type with several subtypes, or better seen as (c) a continuum or spectrum. The DSM-V asserts that the symptoms of the PDDs are best represented by a continuum from mild to severe, rather than a simple ‘yes’

\(^{48}\) (i.e. Rhett’s Disorder, Childhood Disintegrative Disorder, PDD-NOS, and Asperger’s Syndrome),

\(^{49}\) For the full text from the DSM-V (2013) Diagnostic Criteria for Autism Spectrum Disorder, please see the Appendices to this chapter, Appendix K.
or ‘no’ diagnosis to a specific disorder, type, or subtype. It reflects the current intuition that the phenomenon of autism is best defined by set of behaviors and impairments that are common, in varying degrees, to autistic individuals. Thus, these should be denoted by a single name, ASD, and then further categorized according to severity. The revisions to the criteria for a diagnosis of ASD are notably more thorough and strict than the previous criteria for just AD, but it is clear that a diagnosis of ASD continues to reflect Kanner’s initial findings as well as the well-known development of the triad of impairments, which have come to be known and accepted as the very hallmarks of autism.

To reiterate, the current DSM-V asserts that the best conception and understanding of autism is that of a continuum or spectrum on which autism is defined by a common set of behaviors and it should be characterized by the single name: ASD. The DSM-V revisions to the diagnostic criteria for ASD have been a subject of much debate. At issue is how the revisions will impact patients on the spectrum, and whether “high functioning autistics,” who were previously diagnosable with AS or PDD-NOS, will, as a result of the revisions, lose their diagnoses, and thus lose the valuable services and support they have been receiving. Though these are clearly important issues (and, indubitably, in the cases of Rhett’s Disorder and Childhood Disintegrative Disorder, these issues are absolutely critical), given my aims in this chapter, I do not take up these issues here.

Before moving on, I should say a bit about why it is important to chart the history and evolution of the diagnostic criteria for autism and autism spectrum disorders. First, for decades, autism was originally folded into the diagnosis for childhood type schizophrenia, which is very far from how we currently describe, understand, and explain autism. Secondly, before 1987,
autism was not developed into its own discrete diagnosis, as AD. Third, since its first elaboration as Infantile Autism, in 1980, the diagnostic criteria have gone on to include an increasing number of traits and characteristics, including some subtypes such as AS and PDD-NOS. Finally, the exposition underscores the profound difficulty that mental health professionals traditionally have had in defining, conceptualizing, and characterizing the phenomena we call “autism.”

3.3. The Impact of Different Conceptions of Autism on the Theory of Mind Debate

Enormous amounts of time, research, and revisions have contributed to the evolution of the diagnostic criteria for what we now call ASD. Two common threads found throughout the history of ASD diagnostics are Kanner’s original observations, and, since 1987, that the “triad of impairments”50 (also sometimes called the “triad of deficits”). These are specific to autism and are noted throughout the historical evolution of how we understand autism. Though these are articulated in various ways, and elaborated upon with slight variances, and are refined most currently in the DSM-V.

In reviewing the diagnostic criteria for AD (now ASD), since 1987, one finds lengthy descriptions and multiple behavioral characteristics or possible behavioral characteristics. Each of these characteristic traits easily falls under the umbrella of the triad of impairments. Many scholars, such as Simon Baron-Cohen, tend to focus on the triad of impairments as that which needs explaining. Often, the triad of impairments is noted and discussed as a concise enough definition of autism, and seldom is the full text of the diagnostic criteria for ASD given. For

50 “Triad of impairments” is a term which is meant to denote the three broad categories in which autistic individuals experience one or more variety of deficit: 1) Specific abnormalities in social behavior and reciprocity, 2) Verbal and/or nonverbal communication difficulties, and 3) Lack of creativity, imaginative play, or pretense. Each of the three categories in the DSMs is comprised of lengthy lists of potential expressions of the difficulty, of which an individual must exhibit some, but not all of the behavioral traits in each category. Thus, “triad of impairments” functions as a sort of “short hand” descriptor for summarizing autistic traits.
example, Baron-Cohen (2004) concisely describes autism as a “triad of deficits.” He states the “triad consists of deficits in social, communication, and the imagination of other’s minds.” He then focuses on the “cognitive mind-blindness” in autistic patients to explain the triad of deficits. The World Health Organization (1987) also proposes a concise definition of ‘autism’ purely via a brief reflection on the triad of impairments. Similarly, Lorna Wing (1988) claims in her article, The Continuum of Autistic Characteristics, “autism is best described as a continuum or spectrum of developmental disorders in which three core social abnormalities (the triad of impairments) plus or minus other disabilities all occur and vary independently of each other”.

Following the lead of Baron-Cohen and others, I shall assume that autistic behaviors are best summarized by just noting the broad categories of the triad of impairments that are – in some form or other, with varying degrees – universally expressed in persons with ASD, rather than citing every trait that may express each impairment (although, the full text including all of the peripheral and concomitant impairments is required for the purposes of thorough diagnosis in each case, these are not relevant to my project).

Throughout the historical evolution of the concept of autism, we have struggled to understand, explain, and conceptualize the triad of impairments that comprises autism. Researchers have in some instances conceptualized autism as a “syndrome.” In other instances, they have conceptualized autism as a “distinct type with multiple subtypes.” Lastly, they have conceptualized autism as a kind of continuum or “spectrum” of deficits that might be expressed in a patient to varying degrees. According to Jill Boucher, there are certain “logical constraints” that follow from the various conceptions of autism – some of which have been well-recognized.

---

51 For the full text definition from the World Health Organization (1987) for autism, please see the Appendices to this chapter, Appendix L.
and discussed with relevance to autism, and others that have not (Boucher, 223). By “logical
constraints,” I take it she means that if we conceptualize autism in some particular way, then we
are limited to what kind of explanation we can justifiably offer for it. For example, she says, if
autism is viewed as a “syndrome,” by definition, this is to say that autism (AD) stems from a
single physical cause and is conceptualized as a unitary condition. If, however, autism is
viewed as a distinct type (AD) with multiple subtypes (AS, Rhett’s Disorder, Childhood
Disintegrative Disorder, and PDD-NOS), this is – on Boucher’s account - to say that it stems
from two or more psychological causes. Thirdly, and finally, if autism is understood as a
continuum or spectrum, then the claim that it is best characterized by “at least three independent
impairments, usually accompanied by other dissociative impairments with primary deficits that
underlie the continuum” (Boucher, Ibid, pp. 225-226) may be a fair claim.

According to Boucher, “the distinction between the concept of autism as a syndrome and the
concept of autism as a continuum rests on whether or not the triad of impairments are dissociable
from one another: If they are not, we have a syndrome; if they are, we have a continuum”
(Boucher, Ibid, p. 226). I will sidestep this issue and not argue for one conception or other of
autism. Rather, I will work with the most current diagnostic criteria for ASD in current DSM-V,
in which the APA clearly conceptualizes autism as a spectrum. Given the ‘logical constraints,’
we can assume the APA is now prepared to contend that the triad of impairments is, in fact,

---

52 I think that Boucher takes some liberty here with her definition of ‘syndrome.’ In particular, a syndrome is
generally seen as a group of signs and symptoms that occur together and characterize a particular abnormality.
Contrary to Boucher’s claim, it does not follow that the group of signs and symptoms must come from a single,
underlying physical cause.

53 Here, I understand a ‘spectrum’ as describing a single disorder with a wide range of symptoms and a range of
severity of the various deficits and impairments that are linked to the underlying disorder. Despite their disparity, the
range of symptoms is, nevertheless, linked conditions of a single disorder.
dissociable from one another; hence the advent of the umbrella diagnosis of ASD and consequential deletion of all other PDD subtypes.

We should expect that, in accordance with a second criterion given by Boucher, “for a well articulated theory of autism, there should be an appropriate fit between the concept of autism assumed and the explanatory model offered” (Boucher, Ibid, p. 231) Given this criterion and the above logical constraints, we can expect that our explanatory model will be constrained by the conception of autism we assert. Also given that the most current description or conception of autism is that of a continuum or spectrum (ASD), the logical constraints so imposed will prove to have some bearing on the Theory of Mind debate.

It seems that the TT, if subject to the proposed logical constraints, will exhibit an inability to explain all three of the triad of impairments. Moreover, TT will prove to work best with a single primary deficit model, such as Baron-Cohen’s mind-blindness, for example, which would be to assert a single underlying physical cause of autism, thus asserting that autism is a syndrome. However, this is in stark contrast to our most current intuitions about autism – that is, if we take the DSM-V to be an authority that reflects the most current research and intuition on autism. After decades of research and much refinement, our most current conception of autism is that of a spectrum (ASD).

By contrast, the ST will be better equipped to explain all three impairments in the triad. If we take, as I do, ‘spectrum’ and ‘continuum’ to mean the same thing, then our most current intuitions of autism logically require a description of autism that is best characterized by at least three independent impairments, usually accompanied by other dissociative impairments with primary deficits that underlie the continuum. Given the current conception of autism as a spectrum, I will argue that, if applicable, ST will prove a better explanatory model for ASD than
the TT. I will argue this to be the case for two reasons: First, I will argue that simulationists can better account for all three impairments in the triad of impairments; and second, I will argue that (unlike TT) ST can work well with concurrent explanations of other deficits leading to autistic behaviors and impairments, such as a deficit to executive functioning or empathy. Moreover, I will argue that, taking ASD into consideration, the TT is likely not the right theory of ToM because, given the assumption of ASD as an impaired ToM, theorizing cannot underpin ToM. So, as I will further argue (again: given the assumption of ASD as an impaired ToM), the ST is a more desirable theory of ToM, as it offers a more plausible account of that which might underpin ToM – that is, simulation.

The most current conception of autism as Autism Spectrum Disorder (DSM-V, 2013) will help to show that ST has more explanatory power than TT. I will argue that, given the special case of autism, we have reason to favor ST over TT as the most preferable Theory of Mind.
Appendices to Chapter 3: Autism Spectrum Disorder

Appendix A

DSM-I (1952)
000-x28 Schizophrenic Reaction, Childhood Type

Here will be classified those schizophrenic reactions occurring before puberty. The clinical picture may differ from schizophrenic reactions occurring in other age periods because of the immaturity and plasticity of the patient at the time of onset of the reaction. Psychotic reactions in children, manifesting primarily autism, will be classified here.

Appendix B

DSM-II (1968)
295.8 Schizophrenia, Childhood Type

This category is for cases in which schizophrenic symptoms appear before puberty. The condition may be manifested by autistic, atypical and withdrawn behavior; failure to develop identity separate from the mother's; and general unevenness, gross immaturity and inadequacy of development. These developmental defects may result in mental retardation, which should also be diagnosed.

Appendix C

DSM-III (1980)
Diagnostic Criteria for Infantile Autism

A. Onset before 30 months of age   B. Pervasive lack of responsiveness to other people (autism)   C. Gross deficits in language development   D. If speech is present, peculiar speech patterns such as immediate and delayed echolalia, metaphorical language, pronominal reversal.   E. Bizarre responses to various aspects of the environment, e.g., resistance to change, peculiar interest in or attachments to animate or inanimate objects.   F. Absence of delusions, hallucinations, loosening of associations, and incoherence as in Schizophrenia.

Appendix D

DSM-III-R (1987)
Diagnostic Criteria for Autistic Disorder 299.0

At least eight of the following sixteen items are present, these to include at least two
items from A, one from B, and one from C.

A. Qualitative impairment in reciprocal social interaction (the examples within parentheses are arranged so that those first listed are more likely to apply to younger or more disabled, and the later ones, to older or less disabled) as manifested by the following:

1. marked lack of awareness of the existence or feelings of others (for example, treats a person as if that person were a piece of furniture; does not notice another person's distress; apparently has no concept of the need of others for privacy);

2. no or abnormal seeking of comfort at times of distress (for example, does not come for comfort even when ill, hurt, or tired; seeks comfort in a stereotyped way, for example, says "cheese, cheese, cheese" whenever hurt);

3. no or impaired imitation (for example, does not wave bye-bye; does not copy parent's domestic activities; mechanical imitation of others' actions out of context);

4. no or abnormal social play (for example, does not actively participate in simple games; refers solitary play activities; involves other children in play only as mechanical aids); and

5. gross impairment in ability to make peer friendships (for example, no interest in making peer friendships despite interest in making friends, demonstrates lack of understanding of conventions of social interaction, for example, reads phone book to uninterested peer).

B. Qualitative impairment in verbal and nonverbal communication and in imaginative activity; (the numbered items are arranged so that those first listed are more likely to apply to younger or more disabled, and the later ones, to older or less disabled) as manifested by the following:

1. no mode of communication, such as: communicative babbling, facial expression, gesture, mime, or spoken language;

2. markedly abnormal nonverbal communication, as in the use of eye-to-eye gaze, facial expression, body posture, or gestures to initiate or modulate social interaction (for example, does not anticipate being held, stiffens when held, does not look at the person or smile when making a social approach, does not greet parents or visitors, has a fixed stare in social situations);

3. absence of imaginative activity, such as play-acting of adult roles, fantasy character or animals; lack of interest in stories about imaginary events;

4. marked abnormalities in the production of speech, including volume, pitch, stress, rate, rhythm, and intonation (for example, monotonous tone, question-like melody, or high pitch);
5. marked abnormalities in the form or content of speech, including stereotyped and repetitive use of speech (for example, immediate echolalia or mechanical repetition of a television commercial); use of "you" when "I" is meant (for example, using "You want cookie?" to mean "I want a cookie"); idiosyncratic use of words or phrases (for example, "Go on green riding" to mean "I want to go on the swing"); or frequent irrelevant remarks (for example, starts talking about train schedules during a conversation about ports); and

6. marked impairment in the ability to initiate or sustain a conversation with others, despite adequate speech (for example, indulging in lengthy monologues on one subject regardless of interjections from others);

C. Markedly restricted repertoire of activities and interests as manifested by the following:

1. stereotyped body movements (for example, hand flicking or twisting, spinning, head-banging, complex whole-body movements);

2. persistent preoccupation with parts of objects (for example, sniffing or smelling objects, repetitive feeling of texture of materials, spinning wheels of toy cars) or attachment to unusual objects (for example, insists on carrying around a piece of string);

3. marked distress over changes in trivial aspects of environment (for example, when a vase is moved from usual position);

4. unreasonable insistence on following routines in precise detail (for example, insisting that exactly the same route always be followed when shopping);

5. markedly restricted range of interests and a preoccupation with one narrow interest, e.g., interested only in lining up objects, in amassing facts about meteorology, or in pretending to be a fantasy character.

D. Onset during infancy or early childhood


Appendix E

DSM-IV (1994)
Diagnostic Criteria for Autistic Disorder 299.0

A. A total of six (or more) items from (1), (2), and (3), with at least two from (1), and one each from (2) and (3):
(1) **Qualitative impairment in social interaction, as manifested by at least two of the following:**
   (a) marked impairment in the use of multiple nonverbal behaviors such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction.
   (b) failure to develop peer relationships appropriate to developmental level
   (c) a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest)
   (d) lack of social or emotional reciprocity

(2) **Qualitative impairments in communication as manifested by at least one of the following:**
   (a) delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gestures or mime)
   (b) in individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others
   (c) stereotyped and repetitive use of language or idiosyncratic language
   (d) lack of varied, spontaneous make-believe play or social imitative play appropriate to developmental level

(3) **Restricted repetitive and stereotyped patterns of behavior, interests, and activities, as manifested by at least one of the following:**
   (a) encompassing preoccupation with one or more stereotyped patterns of interest that is abnormal either in intensity or focus
   (b) apparently inflexible adherence to specific, nonfunctional routines or rituals
   (c) stereotyped and repetitive motor mannerisms (e.g., hand or finger flapping or twisting, or complex whole-body movements)
   (d) persistent preoccupation with parts of objects

B. Delays or abnormal functioning in at least one of the following areas, with onset prior to age 3 years: (1) social interaction, (2) language as used in social communication, or (3) symbolic or imaginative play

C. The disturbance is not better accounted for by Rhett’s Disorder or Childhood Disintegrative Disorder (DSM-IV, 1994 and DSM-IV-TR, 2000).

**Appendix F**


**Diagnostic Criteria for Rhett’s Disorder 299.80**

A. All of the following:
(1) apparently normal prenatal and perinatal development
(2) apparently normal psychomotor development through the first 5 months after birth
(3) normal head circumference at birth

B. Onset of all of the following after period of normal development:

(1) deceleration of head growth between ages 5 and 48 months
(2) loss of previously acquired purposeful hand skills between ages 5 and 30 months with the subsequent development of stereotyped hand movements (e.g. hand-wringing or hand washing)
(3) loss of social engagement early in the course (although often social interaction develops later)
(4) appearance of poorly coordinated gait or trunk movements
(5) severely impaired expressive and receptive language development with severe psychomotor retardation.

Appendix G

DSM-IV-TR (2000)
Diagnostic Criteria for Childhood Disintegrative Disorder 299.10

A. Apparently normal development for at least the first 2 years after birth as manifested by the presence of age-appropriate verbal and non-verbal communication, social relationships, play and adaptive behavior.

B. Clinically significant loss of previously acquired skills (before age 10 years) in at least two of the following areas:

(1) expressive or receptive language
(2) social skills or adaptive behavior
(3) bowel or bladder control
(4) play
(5) motor skills

C. Abnormalities of functioning in at least two of the following areas:

(1) qualitative impairment in social interaction (e.g., impairment in nonverbal behaviors, failure to develop peer relationships, lack of social or emotional reciprocity)
(2) qualitative impairments in communication (e.g., delay or lack of spoken language, inability to initiate or sustain a conversation, stereotyped and repetitive use of language, lack of varied make-believe play)
(3) restricted, repetitive, and stereotyped patterns of behavior, interests, and activities, including motor stereotypies and mannerisms

D. The disturbance is not better accounted for by another specific Pervasive Developmental Disorder or Schizophrenia.

Appendix H

DSM-IV-TR (2000)
Diagnostic Criteria for Pervasive Developmental Disorder Not Otherwise Specified (Including Atypical Autism)

This category should be used when there is a severe and pervasive impairment in the development of reciprocal social interaction associated with impairment in either verbal or nonverbal communication skills or with the presence of stereotyped behavior, interests, and activities, but the criteria are not met for a specific Pervasive Developmental Disorder, Schizophrenia, Schizotypal Personality Disorder, or Avoidant Personality Disorder. For example, this category includes “atypical autism” – presentations that do not meet the criteria for Autistic Disorder because of late age at onset, atypical symptomology, or subthreshold symptomology, or all of these.

Appendix I

DSM-IV-TR (2000)
Diagnostic Criteria for Autistic Disorder 299.0

A. A total of six (or more) items from (1), (2), and (3), with at least two from (1), and one each from (2) and (3):

(1) Qualitative impairment in social interaction, as manifested by at least two of the following:
   (a) marked impairment in the use of multiple nonverbal behaviors such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction.
   (e) failure to develop peer relationships appropriate to developmental level
   (f) a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest)
   (g) lack of social or emotional reciprocity

(2) Qualitative impairments in communication as manifested by at least one of the
following:
(a) delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gestures or mime)
(e) in individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others
(f) stereotyped and repetitive use of language or idiosyncratic language
(g) lack of varied, spontaneous make-believe play or social imitative play appropriate to developmental level

(3) Restricted repetitive and stereotyped patterns of behavior, interests, and activities, as manifested by at least one of the following:
(a) encompassing preoccupation with one or more stereotyped patterns of interest that is abnormal either in intensity or focus
(e) apparently inflexible adherence to specific, nonfunctional routines or rituals
(f) stereotyped and repetitive motor mannerisms (e.g., hand or finger flapping or twisting, or complex whole-body movements)
(g) persistent preoccupation with parts of objects

B. Delays or abnormal functioning in at least one of the following areas, with onset prior to age 3 years: (1) social interaction, (2) language as used in social communication, or (3) symbolic or imaginative play

C. The disturbance is not better accounted for by Rhett’s Disorder or Childhood Disintegrative Disorder (DSM-IV, 1994 and DSM-IV-TR, 2000).

Appendix J

DSM-IV-TR (2000)
Diagnostic Criteria for Asperger’s Disorder 299.80

The essential features of Asperger's Disorder are severe and sustained impairment in social interaction and the development of restricted, repetitive patterns of behavior, interest, and activity. The disturbance must clinically show significant impairment in social, occupational, and other important areas of functioning. In contrast to Autistic Disorder, there are no clinically significant delays in language. In addition there are no clinically significant delays in cognitive development or in the development of age-appropriate self-help skills, adaptive behavior, and curiosity about the environment in childhood.

A. Qualitative impairment in social interaction, as manifested by at least two of the following:

(1) marked impairment in the use of multiple nonverbal behaviors such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction
(2) failure to develop peer relationships appropriate to developmental level
(3) a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest to other people)
(4) lack of social or emotional reciprocity

B. Restricted repetitive and stereotyped patterns of behavior, interests, and activities, as manifested by at least one of the following:

(1) encompassing preoccupation with one or more stereotyped and restricted patterns of interest that is abnormal either in intensity or focus
(2) apparently inflexible adherence to specific, non-functional routines or rituals
(3) stereotyped and repetitive motor mannerisms (e.g., hand or finger flapping or twisting, or complex whole-body movements)
(4) persistent preoccupation with parts of objects

C. The disturbance causes clinically significant impairment in social, occupational, or other important areas of functioning.

D. There is no clinically significant general delay in language (e.g., single words used by age 2 years, communicative phrases used by age 3 years)

E. There is no clinically significant delay in cognitive development or in the development of age-appropriate self-help skills, adaptive behavior (other than in social interaction), and curiosity about the environment in childhood.

F. Criteria are not met for another specific Pervasive Developmental Disorder or Schizophrenia (DSM-IV, 1994 and DSM-IV-TR, 2000).

Appendix K

DSM-V (2013)

Autism Spectrum Disorder: Must meet criteria A, B, C, and D:

A. Persistent deficits in social communication and social interaction across contexts, not accounted for by general developmental delays, and manifest by all three of the following:

(1) Deficits in social-emotional reciprocity; ranging from abnormal social approach and failure of normal back and forth conversation through reduced sharing of
interests, emotions, and affect and response to total lack of initiation of social interaction,

(2) Deficits in nonverbal communicative behaviors used for social interaction; ranging from poorly integrated- verbal and nonverbal communication, through abnormalities in eye contact and body-language, or deficits in understanding and use of nonverbal communication, to total lack of facial expression or gestures.

(3) Deficits in developing and maintaining relationships, appropriate to developmental level (beyond those with caregivers); ranging from difficulties adjusting behavior to suit different social contexts through difficulties in sharing imaginative play and in making friends to an apparent absence of interest in people

B. Restricted, repetitive patterns of behavior, interests, or activities as manifested by at least two of the following:

(1) Stereotyped or repetitive speech, motor movements, or use of objects; (such as simple motor stereotypies, echolalia, repetitive use of objects, or idiosyncratic phrases).

(2) Excessive adherence to routines, ritualized patterns of verbal or nonverbal behavior, or excessive resistance to change; (such as motoric rituals, insistence on same route or food, repetitive questioning or extreme distress at small changes).

(3) Highly restricted, fixated interests that are abnormal in intensity or focus; (such as strong attachment to or preoccupation with unusual objects, excessively circumscribed or perseverative interests).

(4) Hyper- or hypo-reactivity to sensory input or unusual interest in sensory aspects of environment; (such as apparent indifference to pain/heat/cold, adverse response to specific sounds or textures, excessive smelling or touching of objects, fascination with lights or spinning objects).

C. Symptoms must be present in early childhood (but may not become fully manifest until social demands exceed limited capacities)

D. Symptoms together limit and impair everyday functioning.

The APA goes on to describe the severity levels for the social communication challenges, restricted interests, and repetitive behaviors for individuals with ASD. Severity levels include:

- Level 1: Requiring Support – Without supports in place deficits cause noticeable impairments.

- Level 2: Requiring Substantial Support – Marked deficits are apparent even with supports in place.
• Level 3: Requiring Very Substantial Support – Severe deficits cause severe impairments in functioning.

Appendix L


(1) Specific abnormalities of social behaviour, affecting in particular reciprocal relating and empathy;

(2) Communication difficulties affecting non-verbal communication, conversational skills (pragmatics) and prosody;

(3) Lack of creativity and imagination as is evident in, for example, a paucity of creative pretend play and an inability to role-play, this lack of creativity being accompanied by a characteristic rigidity and repetitiveness of behaviour.

They go on to state that

…other types of impairments may be present including (4) Sensory and perceptual abnormalities of various kinds; (5) Nonmodality-specific language learning difficulties; (6) Ritualistic, obsessional, and compulsive behaviour and marked resistance to change; (7) Generalised learning difficulties (mental retardation); and (8) Spoken language disorders
CHAPTER 4: THE THEORY THEORY ACCOUNT OF THEORY OF MIND

This chapter explores what has come to be known as The Theory Theory of Mind (henceforth, the ‘TT’). After giving a brief description of the TT in the general sense, I will give a more detailed expository account of the TT, by reflecting on its main tenets as discussed by the primary proponents of the theory. While there are many such proponents of the TT, (i.e. ‘theory theorists’), I will focus primarily on the accounts of the two theory theorists whom I think give the most interesting and plausible accounts of the TT that are relevant to this dissertation.

Specifically, I will first discuss the TT as “the child-scientist” view formulated, at length, by one of the earliest theory theorists, Alison Gopnik, whose work has tended to lead and dominate the field. I will then discuss the TT as “innate minimalist modularity theory” put forth by Simon Baron-Cohen et.al. This particular modularity theory is often characterized as falling within the TT camp, and is relevant to my task of adjudicating the debate between competing theories of ToM, as it was Baron-Cohen’s work that first drew the connection between ASD and mindreading deficits which are discoverable via false belief tasks and, according to Baron-Cohen, appear to be a result of a dearth of ToM. Importantly for this dissertation, reflections on Baron-Cohen’s theories concerning ASD serve to undermine Gopnik’s account of the TT (as well as the TT en toto) by indicating that some other ability (besides theorizing) must underwrite the ToM capacity.

After my expositions of Gopnik and Baron-Cohen’s accounts of the TT, I will review, in turn, the main arguments that favor their iteration of the TT as the most plausible explanation for the capacity that supports the ToM capacity. Along the way, I will review the main arguments against the TT and will mention some of the challenges and objections that the TT faces, some of
which are launched by proponents of a competing ToM: *The Simulation Theory* (henceforth, ‘ST’). I will discuss some possible responses that advocates of the TT might have to those challenges and objections. Moreover, I will show why it is that, when we focus on Baron-Cohen’s account of ASD, it seems especially clear that the TT is *not* the most plausible theory of ToM.

4.1. A General Discussion of the Theory Theory of Theory of Mind

ToM studies have quickly become one of the fastest growing bodies of empirical research in psychology over the last 30 years (Leudar and Costall, 2009, p. 1). Since the groundbreaking work of Premack and Woodruff in the 1970’s, in which they developed the original framework for ToM studies, a wide range of disciplines including primatology, developmental psychology, cognitive science, neuropsychiatry, and philosophy has had a stake in the interdisciplinary and collaborative inquiry into ToM.\(^5^4\) Initially, all research in the ToM field started with the same assumption as Premack and Woodruff: that some theory or other underwrites the ToM ability – that is, all of the early theories concerning ToM operated under the presupposition that the development and employment of ToM is, in fact, theory laden in some way or other, whether tacit or overt. In fact, until the mid-1980’s,\(^5^5\) any theoretical iteration concerning ToM was a variant of the original assertion: that the achievement of understanding the mental states of

---

\(^{54}\) For a detailed description regarding Premack and Woodruff's "*Does the Chimpanzee Have a Theory of Mind,*" see Chapter 2: The Theory of Mind, this dissertation.

\(^{55}\) In 1986 Robert Gordon and, independently, Jane Heal proposed an alternative theory concerning ToM called The Simulation Theory (ST). There are also many iterations, variants, and gradations of ST that have evolved over the last twenty-five years. See the chapter Simulation Theory chapter of this dissertation for further discourse regarding ST.
others, and the social understanding inherently involved in such understanding, come as a result of some kind of theoretical and intellectual accomplishment on the part of the perceiver.56

Competing theories regarding the nature, the development, and the underlying source of ToM began to emerge during the 1980’s. The emergence of competing theories of mind in which ToM is executed in different ways (simulation rather than via theorization) has necessitated a name change to the original ToM framework proposed by Premack and Woodruff in the 1970’s. This name change is meant to help distinguish between the set of theories which claim ToM is generated by theorizing from the set of theories that claim ToM is generated by a process of simulating. The theories which viewed mentalizing as a theoretical activity and was once simply referred to as the ‘Theory of Mind,’ is now referred to with the rather quite awkward expression coined by Morton, the ‘Theory Theory of Mind’ (TT) (Morton, 1980).

“The TT is one general theory of how we understand the minds and mental states of other people. According to the TT, I know the contents of other minds by way of either a) observation and hypothesis formation, or b) some theory innately given” (Currie, 1996, p. 242). The theory theorist considers how children come to understand one particularly important aspect of the world – the fact that other people have emotions, desires, and beliefs and that those mental states cause their behavior, and questions how we come to know other minds. They say, “Infants seem to be born believing that people are special and there are links between their own internal feelings and the internal feelings of others” (Gopnik, 2004, p. 22). As noted, it is widely believed

56 In a very general sense, a theory theorist holds the view that if and when we are able to predict the mental states of others we do this by combining our observations with what we know about ourselves and others in general and that, from this as well as other factors, we can draw a sort of inference that is inductive in nature – a theory – about the mental state of others and ourselves. Importantly, for the theory theorist, these theories are in a constant state of revision. Moreover, and importantly, the theories derived ought to be contrasted with other ways in which one might draw inferences concerning the mental states and resultant behavior predictions of oneself and of others such as strict observation, rationalization governed by a set of principles, guessing, and the like.
that possession of a ToM sponsors this activity, and the bulk of the developmental literature is devoted to determining how the ToM is acquired and developed.\footnote{Importantly, the term ‘Theory of Mind’ has become so much a part of the furniture of developmental literature that there is scarcely any question raised as to whether the ‘ToM mechanism’ indeed exists as a separate function of the mind, rather than assuming that humans might know the mental states of themselves and others as an application of a more general mechanism of knowledge. Leudar and Costall (2009) argue extensively against the very existence of a ToM or a ToM mechanism.}

Generally speaking, the TT asserts that, “third-person [mental state] attribution proceeds wholly \textit{via} inferences that are guided by the causal principles of folk psychology” (Goldman, 2006, p. 69). People hold a ‘naïve theory’ or ‘\textit{folk} psychology’ about the mental states of others. That is, to mentalize ordinary people construct, or are endowed with, a naïve psychological theory that guides their assignment of mental states to other people (Goldman, 2006, p. 4). Notice that, on this account, one would not simply \textit{notice, observe, or perceive} the apparent mental states of oneself or of others. On the TT account, ‘something’ happens beyond the mere observation or perception. For example, you might notice, observe, or perceive that someone near you is angry. You might notice their contorted facial expression, observe the fact that they are throwing the dishes, and perceive the tone of their voice with its raised decibels as they are yelling. These observations must be combined with other actual and virtual experiences you have had with anger and angry people for you to take the next step – that is, to actually hold a \textit{theory} (or, in Gopnik’s terms, ‘draw a theoretical inference’) about the mental state of the person you are observing. According to the TT, people then use the information gained through mentalizing to make sense of the actions of others and predict the future behavior of others. Importantly though, mentalizing does not only make sense of the actions of others \textit{via} ascribing mental states to them; also, we also ascribe mental states to ourselves to explain and predict our own behavior.
For the theory theorist, “self-ascription, like other-person ascription proceeds by theoretical inference” (Gopnik, 1993).

The TT is thus perhaps best seen as a family of theories that may be construed in various ways (and in various strengths), but share certain common elements. For example, all TT iterations maintain, in accordance with its roots and original principles, that theoretical activity (as theoretical and intellectual prowess on the part of the perceiver) underpins our understanding of our own mental states as well as the mental states of others, and makes possible the acquisition of the social understanding that supposedly results from such understanding.

Many scholars have contributed to the web of iterations that comprise the family of TT. The purest (i.e. strongest) form of the TT adopts the following positions: (1) Both third-person and first-person mental state ascription are executed by theoretical reasoning the premises of which are observed features of the target’s environment and behavior, plus causal laws or principles of ‘folk psychology’. (2) Mental-state concepts are understood in terms of theoretically specified causal relations between behavior, environment, and other mental states. (3) Mentalizing (ToM) skills develop in step with, and as a consequence of, development of a mentalizer’s folk-psychological theory. (Goldman, 2006, p. 69).

Other scholars (see Carey, 1985; Gopnik, 1990 – 2012 all; Wellman, 1990) have argued for a much stronger theoretical model of ToM that also, in part, comprises the family of theories called the TT. On these strong accounts, ToM is not seen as mere mental machinery but rather an actual theory bearing all of the features of any other (e.g. scientific) theory with rules of inferences. So, from their perspective, it is not possible to pry apart our concepts of mental states from the set of inferences that individuate them within our theory of the mental world. One advantage of this position is that it better explains the articulation of the development of ToM in
conjunction with children’s other abilities as mind readers, such as, for example, detection of
desires (see Gopnik et al., 1994), pretense (see Perner, 1991), and emotions (see Harris, 1989).

The TT has become popular with developmental psychologists who use the TT iteration of
ToM to explain the development of human social understanding (Carruthers, 1996). These
proponents tend to focus on the development of ToM throughout early childhood, and seek to
examine and explain “breakdown cases” – that is, they use empirical data and longitudinal
studies of children on the autism spectrum who, they claim, apparently ubiquitously exhibit a
theoretical view assists in “fitting in” the development of ToM via the TT with other cognitive
achievements such as development of language, episodic memory, and executive functioning.

Unfortunately, developmental psychologists focus less on the philosophical arguments
regarding the correctness or incorrectness of the TT, and tend to focus more primarily on how
the TT fits with the aforementioned working theories in cognitive developmental science and
developmental psychology. In fact, nowhere in the literature does any theory theorist offer any
concise argument to support the suppositions of the TT. Goldman, however, attempts to glean
the TT argument from the literature and couches it like this:

(i) Young children’s performance on mentalizing tasks changes over time as a
function of changes in their grasp, or understanding, of mental concepts.
(ii) These changes in concepts, or conceptual understanding, reflect successive stages
in children’s theories of the mental.
(iii) Therefore, mental concepts must be theoretical concepts.
(iv) Hence, all determinations of the instantiation of mental concepts, in both self and
others, must be inferential in character.58 (Goldman, 2006, p. 70).

58 One might notice that premise (ii) does not seem entail the conclusion (iv) of the argument as Goldman states it. I
think that Goldman is trying to illustrate that the theory theorist seems to make a leap from the TT assumption that
(a) young people get better and better at understanding others, and the assumption (b) that this is done via
One advantage of TT is that it can explain how the theory develops and changes over time. This is consistent with other kinds of theories that experience radical changes over time such that a theory at a given stage may be incommensurable with its earlier stages.

Developmental psychologists and some philosophers, who support TT (see Churchland, 1981, 1995; and Fodor, 1992), share the assumptions that naïve psychology, at bottom, is driven by a *science-like* theory, where a *theory* is understood as a set of law-like generalizations. Naïve psychology would include generalizations that link (1) observable inputs to certain mental states, (2) certain mental states to other mental states, and (3) mental states to observable outputs (behavior). The first type of law might be illustrated by “Persons who have been physically active without drinking fluids tend to feel thirst.” An example of the second might be “Persons in pain tend to want to relieve that pain.” An example of the third might be “People who are angry tend to frown.” The business of attributing mental states to others consists of drawing law-guided inferences from their observed behavior, stimulus conditions and previously determined antecedent mental states. For example, if one knows that Darren has been engaged in vigorous exercise without drinking water, one might, by employing a sort of loose inductive reasoning, infer that he is thirsty.

4.1.1. Some Elementary Notions of the Theory Theory

At this point, and before advancing further in the chapter, it seems important to say more concisely what is meant by certain terms employed by theory theorists and used repeatedly throughout the remainder of the chapter. Specifically, at the base of the TT are the notions of

successively improved theories to the apparent TT assertion that knowing the mental states of others must be inferential in nature. It seems clear that the argument, as Goldman states it, does not seem to follow and one might rightly wonder whether Goldman sets up a bit of a straw man argument.
theorizing, theory, and like scientific theory, (or, alternatively, like science, or like scientists). I shall briefly discuss each in turn.

Throughout the literature, and thus throughout the dissertation, the term theorizing is used repeatedly. I take it that the theory theorist use of the word ‘theorizing’ is just the same as the conventional use and lexical definition of the word, which means to form a theory or a set of theories about something, thereby creating a theoretical premise or framework of something. When the theory theorist claims that ‘a person is theorizing when mentalizing,’ she means that a mental state attributor forms (or has formed) a theory or a set of theories about the mental states of others and that, by deploying theories, the mental state attributor creates a theoretical framework for the understanding of other minds, and that the ToM ability is underpinned by such theories and theoretical framework.

What, then, does the theory theorist mean by ‘theory’? When Premack and Woodruff coined the phrase ‘Theory of Mind,’ what they meant by ‘theory’ was a system of inferences about unobservable states (i.e. the mental states of others) that can be used to make predictions (specifically about the behaviors of others) (See Premack and Woodruff, 1976). Following Premack and Woodruff, Alison Gopnik employs the term ‘theory’ as well. When Gopnik (and theory theorists who came after her) says ‘theory,’ she is referring to “abstract, coherent, systems of entities and rules that provide predictions, explanations, and interpretations, and that change in characteristic ways in response to counterevidence” (Gopnik, 1996). These entities and rules are, she says, “a succession of related intuitive causal theories of the world – including the physical, biological, psychological worlds – theories that expand, elaborate, modify, and revise in light of new evidence (Gopnik, 2004). Importantly, not all theory theorists think that the kinds of theories that underpin the ToM ability are causal – that claim is fairly unique to Gopnik.
However, all theory theorists maintain that the theory referred to by the TT consists of law-like generalizations with propositional attitudes playing a central role in these generalizations.

This brings me to the third elementary notion of relied on by a theory theorist (such as Gopnik): that the kinds of theories that underpin the ToM ability are like scientific theories, (or, alternatively, that mentalizers are like scientists whose theory development is like science). As you will see further in the chapter, the analogy that is drawn between theorizing for mentalizing and scientific theorizing is an extremely strong one.\(^59\) She says that the children learn about other minds using the very same cognitive devices that are used in science – the very same learning mechanisms are used for both ToM and the sciences (See Gopnik, 2003, 2004). Gopnik also says that the kind of knowledge that drives mindreading has the same structure of scientific learning and that it is acquired, stored, and used in the same way that scientific theories are used (See Gopnik and Meltzoff, 1997). In short, she says, “the body of knowledge underpinning ToM has all of the structural, functional, and dynamic features that characterize scientific theories” (Ibid).

Perhaps what is most important (at least, for my argument against the TT) is not just the fact that the theory theorist claims that ToM is underpinned by theories that are strongly analogous to science, but that the theory theorist claims that mindreading is the product of a scientific-like exercise of a domain-general theorizing capacity. The claim is that this is a domain-general learning mechanism (or, in some places, a domain-general psychological structure) that can be used across many different context domains – including and especially areas that require theorizing, such as the sciences. To be clear, the theory theorist does not hold that the theorizing that underpins the ToM ability is domain-specific psychological structure that is solely dedicated to resolving problems in the psychological domain. Again, the claim is that this is a domain-

\(^{59}\) This is elaborated more thoroughly in sections 4.2 – 4.2.3 of this chapter.
general learning mechanism used in multiple contexts, including the sciences. Because of this, I shall rest my argument against the TT on this premise: that if the ToM ability is underpinned by a domain-general psychological structure that is also thought to be used for scientific theorizing, then (if this were right) we should expect one who has difficulty with ToM to also experience difficulty with using the same structure in other contexts – namely, the sciences.

In the remainder of this chapter, I will give expositions of the two primary articulations of the TT. The first exposition will be of the child-scientist account of the TT, as put forward by Alison Gopnik. According to Gopnik, much of the work suggesting that children learn about the causal structure of the world has been done in the context of the TT: the idea that children have intuitive theories of the world analogous to scientific theories, and that these theories change in ways that are similar to scientific theory change. The second exposition will be of two discrete theories put forth by Simon Baron-Cohen: the Innate Minimalist Modularity Theory and the Extreme Male Brain Theory. I will examine each of these variations of the TT for both their validity and plausibility. I will then discuss these primary arguments, which favor the TT as the best explanation for the ToM. I will end the chapter with an exposition of my key argument against the TT (as the best explanation of the ToM), in which I use Baron-Cohen’s theories to undermine Gopnik’s theory as well as the TT in toto.

4.2. Gopnik’s Child-Scientist Account of the Theory Theory

Cognitive psychologists who support a general version of TT claim that children are born with innate theories about the mental states of others. They claim that, even in infancy, children observe others and the world, gather data, and continuously revise their theories to include these new observations and to make new predictions regarding the mental states and behaviors of
others. These observations of others are what the theory theorist calls the ‘naïve psychological theories’ or the ‘folk psychology’ that people develop. They assert that even very young children learn about the mental states of others via a process of formulating and revising a succession of intuitive theories (Carey, 1985; Gopnik, 1988; Heil, 1989; Wellman, 1990; Gopnik and Meltzoff, 1997; Wellman and Gelman, 1997). To varying degrees, they say this process of theory revision somewhat resembles the manner in which scientists propose and revise scientific theories. They claim that theory changes in infants and children are generated by the same learning or acquisition methods used by adult scientists – that is, children go about forming and changing their theories of mind in the same way that scientists go about forming and changing their theories: by general-purpose causal reasoning (Goldman, 2006, p. 83). This is known as the ‘child (as little) scientist’ version of the TT.60

Alison Gopnik is a professor of psychology and philosophy who is well known for her work in the areas of cognitive development and the development of ToM. Gopnik is one of the earliest proponents of the TT and was the first author of the child-scientist version of TT (See Gopnik, 1990, 1993, 1997, 2003; Gopnik & Meltzoff, 1997, 1998; Gopnik & Wellman, 1992, 1994; and Gopnik & Astington, 1991). Gopnik studies ToM development from the perspective of cognitive psychology, and her child-scientist view exemplifies a particularly strong (i.e. “purist”) version of the TT.

There are, however, important reasons that Gopnik adopted such a strong, purist version of the TT. Gopnik developed her child-scientist account as a sort of response to the work of Jean Piaget, and in an honest attempt to better elucidate explanations for children’s behavior. Perhaps

---

60 Notably, Gopnik’s ‘child (as little) scientist’ account of the TT invokes the quite similar work of Susan Carey (1987), who worked to show that the acquisition of concepts (roughly the equivalent of a single word such as object, animal, alive, heat, weight, and matter) consequently allows young developing minds to understand and use successively more mature theories using these concepts in the realm of science and, particularly, biology.
it was as a result of her response to Piaget that she adopted such a firm and unyielding position as a theory theorist.\textsuperscript{61} Let me briefly explain.

More specifically, Piaget developed a developmental stage theory that divided children’s cognitive development into five distinct stages (with some sub-stages), each of which are characterized by measurable differences in behavior and are especially colored by egocentrism in the earlier stages. Piaget interpreted certain developmental data in light of this theory. Whilst engaged in her ToM research, Gopnik and other ToM researchers reviewed the same developmental data but arrived at quite different interpretations of it. Gopnik thus balked against the Piagetian developmental stage theory on several grounds, primarily because of the differences in their interpretations of certain developmental data. I will briefly mention three of these here.

First, Piaget claimed that before the age of 6, children have no appreciation for mental life at all – they are adualists and are completely unable to distinguish between mental and real phenomena (Piaget, 1926/1929, pg. 55). However, early developmental research on ToM showed that children under 6 were not the adualists that Piaget claimed them to be (See Wellman, 1986). Second, although Piaget posited that children under 6 had no appreciation for mental life, he did discuss representation in general and interpreted certain data to mean that children as young as 1 year, 3 months of age could understand representing one object as another object in pretense. This stands in stark contrast to the ToM literature, which gave rise to the view that children can and do understand pretense before age 2, but that they do so without understanding

\textsuperscript{61} It is important to acknowledge the kind of dominance of Piagetian models at the beginning of Gopnik’s career. Indeed, Piaget was seen as the “Great Pioneer” of the Constructivist theory of knowing. His ideas were popularized in the 1960’s and his work led to the study of development as a major sub-discipline of psychology. By the end of the 20\textsuperscript{th} century, Piaget was second only to B.F. Skinner as the most cited psychologist of that era (Haggbloom, et.al., 2002). Perhaps Gopnik wrote her child-scientist account and adhered to it as strongly as she did because her findings competed with those of Piaget, the preeminent developmental psychologist of the time.
representation and, further, that children under age 4 cannot understand the mind as representing or misrepresenting (as with a false belief task) at all until roughly age 4. Third, there is a tacit Cartesian presupposition underlying Piaget’s notion of egocentrism as it relates to the representation of mental state concepts in that it claims that we know each mental state by our own introspective (i.e. egocentric) experience (See Perner, 1991). This Piagetian egocentric way of acquiring mental state concepts is in direct opposition to Gopnik’s TT view of ToM. Gopnik held that the developmental data ought to be interpreted as incompatible with the Cartesian position because mental state terms do not obtain meaning through introspection, but through acquisition and development of a theory – a theory in which mental states (1) are theoretical concepts, (2) have the status of hypothetical constructs, (3) are not directly observable but need to be inferred, and (4) provide advantages for predicting and explaining the behavior of others (Belini and Pufall, 1992, pg. 151).

Given the foregoing, Gopnik likely maintained her unyielding position as a theory theorist and adhered to her child-scientist view of the TT (often even in the face of evidence to the contrary) because her interpretations of the developmental data for her ToM research came into basic conflict with Piagetian interpretations of the same data. In short, ToM research stands in direct contrast to Piagetian developmental stage theory and “is anchored in perspectives on Philosophy of Mind that antedate Piaget’s work and put the [ToM] field firmly on anti-Piagetian footing” (Belini and Pufall, 1992, pg. 146). As such, Gopnik strongly maintains her view in the face of opposition. I explain her theory in detail in the following.

---

Like other (indeed, most) theory theorists, Gopnik claims that children develop their everyday knowledge of the world using the same cognitive devices that adults use in science (Gopnik, 1997). These devices include “static features of theories, such as their abstract, coherent, causal, counter-factual-supporting character; functional features of theories such as their ability to provide predictions, interpretations, and explanations; and dynamic features such as theory changes in the light of new evidence” (Gopnik and Glymour, 2002). Also, like most theory theorists, Gopnik argues for a theoretical model of ToM that sees ToM as a mental capacity which underwrites our understanding of both our own and of others’ mental states. The theory theorist holds that a mentalizer exercising the ToM ability (mentalizing) is similar a scientist employing a scientific theory, and it is acquired, stored, and used in a similar way that scientific theories are: A hypothesis is put forth as some kind of explanation or for making a prediction. Then, some kind of experiment is carried out, and the theory gets revised and re-revised (hopefully) reflecting new and improved explanation and understanding.

However, in contrast to other theory theorists (e.g. Meltzoff, Wellman, and Astington), Gopnik asserts a version of TT that is particularly strong when compared to the general account of TT. First, she conceives of ToM as not merely a mental faculty, but an actual naïve theory complete with posits, axioms, and rules of inferences. Mental states (such as beliefs, for example) are considered theoretical entities or posits of this theory. Second, Gopnik’s child-scientist account is more radical than other theory theorists in that, just like in science, the body of knowledge that underpins mentalizing is open to both revision and invalidation (defeasibility and annulment) if and when counterevidence occurs. Hence, Gopnik’s child-scientist account of TT relies heavily (if not entirely) upon the TT/child-scientist analogy between the development
of mental state attribution theory and the development of theories in the so-called “hard sciences” (e.g. biology, physics, physiology, and chemistry).

Gopnik claims that there are powerful similarities between the cognitive development of individuals and scientific theory change. Both, she says, “involve abstract, coherent systems of entities and rules - theories. In both cases, theories provide predictions, explanations, and interpretations. In both, theories change in characteristic ways in response to counter-evidence” (Gopnik, 1997). She further claims that children experiment with and explore the world, test the predictions of their theories, gather relevant evidence, and constantly revise their theories. She suggests that infants are born with innate theories and that they begin revising those theories, even in infancy.\(^{63}\) In short, according to Gopnik, children use the same psychological learning and acquisition methods that scientists use when constructing scientific theories. Thus, she says, children are “little scientists” or, more to the point, “scientists are big children” (Gopnik, 1997). Her point is, I think, that children learn in precisely the same way that scientists learn because both kinds of learning rely on the same domain-general psychological structure. Gopnik’s extensive corpus is literally replete with this claim, which she reiterates and re-reiterates.

Often, she will begin by making the general claim that children learn like scientists. Human minds, she says, “derive abstract, complex, highly structured, veridical, representations and rules, namely theories, from limited input, namely evidence. These are just the mechanisms we use in science” (Gopnik, 2003). Relying heavily on the assumption that both kinds of knowledge and

\(^{63}\) Here, note the influence of Noam Chomsky (1990), who (similarly) asserted that language acquisition is an innate structure or function of the brain. To support this claim, Chomsky argued: (1) there is an optimal learning age in which a child will grasp a language and attain fluency; (2) language acquisition occurs spontaneously and without an impetus; and (3) children go through stages of language development and do so at roughly the same age and in roughly the same order. As you will see, theory theorists use these same arguments to justify their claim that understanding mental states (i.e. mentalizing) is also an innate structure or function of the brain. Thus, Chomsky’s influence in the debate is to be noted.
both kinds of learning are supported by the same psychological structure, she restates this
comparison to general science elsewhere. She says, “To a striking extent, children use data to
formulate and test hypotheses and theories in much the same way that scientists do. Scientists
learn about the world in three ways: They analyze statistical patterns in the data, they do
experiments, and they learn from the data and ideas of other scientists. The recent studies show
that children also learn in these ways…” (Gopnik, 2012). And again:

The basic idea is that children develop their everyday knowledge of the world by
using the same cognitive devices that adults use in science. In particular, children
develop abstract, coherent, systems of entities and rules, particularly causal
entities and rules. That is, they develop theories. These theories enable children to
make predictions about new evidence, to interpret evidence, and to explain
evidence. Children actively experiment with and explore the world, testing the
predictions of the theory and gathering relevant evidence. Some counter-evidence
to the theory is simply reinterpreted in terms of the theory. Eventually, however,
when many predictions of the theory are falsified, the child begins to seek
alternative theories. If the alternative does a better job of predicting and
explaining the evidence it replaces the existing theory (Gopnik 2010).

Although her account of scientific theory change is somewhat naïve, her general claim is that
children are innately endowed with the rudimentary beginnings of a theory and that they develop
successively more sophisticated theories as they acquire mental-state concepts. They do this via
continual theory construction and revision - a process that maps onto the construction and
revision of scientific theories in general.

Next, Gopnik extends her child-scientist account. She changes her initial claim that the way
children learn is just the way that scientific theories in general are generated, and moves her
analogy forward by claiming that the way children learn is akin to specific kinds of scientific
theories, including physics, biology, and chemistry. She claims that children develop “causal
knowledge of everyday physical phenomena, like gravity and movement, and everyday
biological phenomena, like illness and growth…Like scientists, children seem to develop a
succession of related intuitive causal theories of the world, theories that they expand, elaborate, modify, and revise in light of new evidence” (Gopnik, 2004).

Next, Gopnik further extends her child-scientist account by moving her analogy forward to also include psychology. For Gopnik, the child-scientist view proposes that human beings use abstract models of how minds work that are analogous to the model we have of how the physical world works. She claims “studies suggest that children learn about the world much in the same way that scientists do … by forming intuitive theories of the physical, biological, and psychological realms” (See Gopnik, 1996, 2003, 2004, 2005, and 2010). And, again, these theories, like scientific theories, are complex, coherent (i.e. logical and consistent) representations of the causal structure of the world that change and evolve in light of new evidence. She states her view particularly strongly here:

Human beings, unlike other animals, develop everyday theories of the world around them. Two decades of research have shown that children construct and revise an everyday physics and biology and, above all, an everyday psychology. These everyday theories are much like the formal, explicit theories of science. Theorizing lets children understand the world and other people more accurately… Theories are profoundly powerful and adaptive…. Theories explain the world we see and lets us imagine other worlds… children are devoted intuitive scientists (Gopnik, 2005).

This quote illustrates precisely how strong the analogy to science is for Gopnik. She holds that children utilize theories to understand the physical and biological worlds and, importantly, extends this to include the child’s understanding of the psychological world (Gopnik, 2003). All of these ideas, she says, are best illustrated by an account of the child developing an understanding of the mind, employing a ToM, or ‘mentalizing.’ So, for Gopnik, a child’s capacity to learn and develop theories is not found just in how a child explores the world, but a child’s capacity to learn about the world in general is best illustrated by the way the child learns
about other people through ‘naïve folk psychology,’ through ‘mentalizing,’ and through employing a theory ToM.

The strong analogy to science serves the theory theorist well with its explanatory power. Human beings require a timeframe from infancy to full-fledged adulthood than any other species that we know of (Gopnik, 1998). Developmental psychologists have long studied the arc of development that occurs during these formative years. The notion that learning about the world and learning about the mental states of others is underpinned by science-like theoreticity is compelling to the theory theorist because the processes of the scientific method could serve to nicely explain many aspects of child development including domains such as language, empathy, executive functioning, episodic memory, and – yes – ToM. The notion that humans theorize, explore, revise theories, and retest them could potentially explain the stages of childhood development, growth, and learning. It could explain the shifts we see from the psychological profiles of infants to toddlers, toddlers to preschoolers, preschoolers to children, and so on.

To illustrate her point, Gopnik borrows from Neurath likening theory formation and development view to a boat that is being perpetually rebuilt as we sail in it. At every stage of development, our work is necessarily limited by the kind of boat we pushed off from the pier in, the tools we have at our disposal, and the constraints of keeping our vessel seaworthy. Eventually, the boat we sail in may not have one single plank or rivet from the original structure.64 Like Neurath’s boat, our theory development (or, perhaps, knowledge development) need not depend on some fixed point or foundation.

---

64 Otto Neurath (1882-1945) captured the main features of his doctrine of scientific theory and knowledge using the image of a boat: “Imagine sailors, who, far out at sea, transform the shape of their clumsy vessel from a more circular to a more fishlike one. They make use of some drifting timber, besides the timber of the old structure, to modify the skeleton and the hull of their vessel. But they cannot put the ship in dock in order to start from scratch. During their work they stay on the old structure and deal with heavy gales and thundering waves. In transforming
If theory use and revision is as Gopnik claims it is, this metaphor fits well. Ongoing revisability, as an explanation for ToM, allows the theory theorist to explain how it is that children of different ages acquire and improve upon their ToM skills over time. Gopnik focuses primarily on preschool age children, claiming that there is a critical, observable shift that occurs around age 3 ½ - 4 that indicates the development and honing of ToM skills. According to TT, it is the development of ToM that changes the toddler from “a little observer of human behavior to a folk psychologist, capable of making complex mental state attributions” (Slaughter & Repacholi, 2003, p. 1).

Empirical studies have documented this developmental arc. “By [the age of] two, children seem to understand that their own desires may differ from the desires of others,” by thirty months, children can note that their own visual perceptions may differ from another’s perception based on viewpoint, and “By [the age of] four, children can understand that beliefs, as well as desires and perceptions, may differ, and that beliefs may be false” (Gopnik, 2004, p. 22-23). “By the age of five, children understand some of the basic causal principles of everyday physics, biology, and psychology” (Gopnik, et.al., 2004, p. 4).

Of particular interest to the ToM debate is the claim that children around the age of four begin to have a concept of beliefs, some of which may be false. To measure this, Perner and Wimmer (1985) developed a ‘false belief task’ (henceforth, ‘FB’) to test whether children of varying ages were not only able to understand that others have beliefs that are separate from their own, but also that the beliefs of others might be false. There are various iterations of the FB. Generally, the tasks require the subjects to understand the true nature of a given situation, and to
simultaneously understand that a different person, who is not privy to the details of the situation, will have a different thought or belief (mental state) and hence exhibit different behavior. Success on this task is said to detect the presence of the ToM capacity and ToM skill-levels in individuals.

After completing empirical studies with children of varying ages, Gopnik concludes that preschool children (age 3 ½ - 4) experience some sudden development of their ToM skill. She justifies this claim with data which shows that the overwhelming majority of children under 3 ½ fail FB tasks, and that the overwhelming majority of these same children can pass the FB task after the age of four. This, she says, is owed to a shift in the child’s ability to grasp mental state concepts (a shift which she equates with the development of a revised theory, ToM). According to Gopnik, the radical change that occurs for a preschool age child (from not understanding FB to understanding FB) is explainable by the TT version of ToM. She claims that only the constant process of making, changing, and revising theories in light of new evidence is likely responsible for this radical developmental change.

Gopnik takes this evidence and asserts that its progression mirrors that of the progression of a bonafide scientific theory. She then draws the strong analogy between the development of a mature ToM and the development of science. The progression of ToM occurs as a result of theory revision, and the progression of science occurs as a result of theory revision. Relying on this analogy, she claims, “Young children are little scientists who form and revise their thinking about various domains in the same way scientists do (Gopnik and Wellman, 1992; Gopnik and Meltzoff, 1997). They collect evidence, make observations, and change their theories in a highly science-like fashion. They generate theories not only about physical phenomena but also about unobservable mental states like belief and desire. As in formal science, children make transitions
from simple theories of the phenomena to more complex ones” (Goldman summarizing Gopnik, 2012, p. 3).

Repeatedly, Gopnik asserts that very young children have abstract, structured, logically consistent, causal representations of the world around them – representations that are similar to scientific theories. The learning and thinking of young children as it pertains to various worlds – even the social world - is strikingly similar to the experimentation and learning process of science. She says, “[Children] use those representations to make wide-ranging new predictions… it is particularly clear that preschoolers have intuitive theories of the physical, biological, psychological, and (she now adds) social world… Theoretical and empirical research has begun to show that children’s learning mechanisms do indeed resemble the basic inductive processes of science…” (Gopnik, 2012). In fact, Gopnik’s version of TT as child-scientist rests entirely on the veracity of the strength of her analogy to science. Moreover, her evidence for this relies on the notion that children experience a conceptual shift between 3 ½ - 4 years of age – a conceptual shift which changes them from being people with incompetent ToM skills and incapable of understanding false beliefs and mentalizing to becoming people with competent ToM skills who are quite capable of understanding false beliefs and mentalizing.

To summarize Gopnik’s position: Gopnik is a developmental cognitive psychologist whose view is a very strong version of the TT. She asserts that the ability to mentalize requires possession of ToM, and that this ToM is theory-driven. She claims that theory-driven ToM functions in much the same way that science does: through an unending process of hypotheses and experiments, which produce revised theories. This, she calls the “child-scientist approach” to TT. She first asserts that children learn the way scientists do in general. She later extends this to specific sciences such as physics, biology, and chemistry. When this analogy appears to go
through, she further extends the analogy to include psychology and eventually extends it to their understanding of the social world. Her account is wholly reliant upon the “ToM development to scientific theory development” analogy. She supports this analogy with evidence from FB tasks that show a developmental improvement in a typically developing child’s capacity to grasp mental state concepts, which Gopnik equates to the child’s ability to theorize more proficiently.

4.2.1. Some Objections to the Child-Scientist Account of the Theory Theory

Gopnik’s account raises several questions. I will raise several objections here, saving my final and strongest argument for last (which must be presented following the Baron-Cohen section). First, is the parallel between scientific theory development and the development of a child’s ToM an accurate or appropriate analogy? Or, in any case, is the analogy strong enough to warrant Gopnik’s reliance upon it in order to support her child-scientist version of TT? Second, does the empirical data surrounding FB tasks support Gopnik’s premise that children undergo a (rather dramatic) conceptual shift at 3 ½ - 4 years of age? Third, and finally, does the evidence that Gopnik provides actually indicate that children learn and theorize about minds the way she says they do? That is, does the data from her empirical experiments support her contention that children learn and theorize about “psychological and social worlds” via the same mechanisms and in the same manner by which they learn and theorize about physical and biological worlds? I will now address each of these three questions.

The first question inquires whether the child-scientist account of the TT, which conceptualizes mentalizing as the progression of a theory-based ToM, is as strongly analogous to science, scientific learning, and the progression of scientific theories as is claimed. We should ask this because Meltzoff and Gopnik say, “Our understanding of the mind, at least in large part,
involves abstract causal structures from evidence. Hence, the analogy to theory change in science – involving initial hypothesis, tests, and conceptual revision – is a strong analogy” (Meltzoff & Gopnik, 2013, p. 19). There are several features of the mentalizing domain and development of ToM that are, in fact, rather unlike science and the scientific method, which I will now elucidate four of these.

First, according to the child-scientist account (unless there is some kind of developmental delay or developmental difficulty), children reach basically the same theories, including the same theoretical posits and axioms, about their own mental states and the mental states of others. This is very different from the way science progresses. Rarely if ever, do individual scientists posit the same theories, perform the same experiments, and revise their theories in the same way as other scientists. Goldman (2006) notes this disanalogy. He says, “A notable feature of professional science is the diversity of theories that are endorsed by different practitioners” (Goldman, 2006).

Second, according to Gopnik, children arrive at the same theories in roughly the same time frame, which is a relatively short period of (3 ½ - 4 years), and employ them in roughly the same manner (again, unless there is some kind of developmental delay or difficulty). By contrast, as Goldman notes, when scientists do concur, their agreement comes after lengthy periods of disagreement, debate, further research, and so on. Sometimes these debates can continue for decades after initial research data is published, and an individual scientist may devote her entire career to determining which theory is the correct theory. “This pattern of [on-going] controversy [in science] contrasts sharply with what is ascribed to young children in the mentalizing domain. They are said to converge on one and the same theory, [in the same order] all within the same narrow time-course” (Goldman, 2006).
Third, according to cognitive psychologists, not only do children arrive at the same theories in the same brief timeframe, but also, they arrive at them in the same order. Goldman highlights this point when he says, “What seems particularly miraculous is that normal children acquire competence with respect to the same assortment of mental-state concepts and in exactly the same order. They acquire the same concept of belief …, and they acquire it after they acquire the concept of desire” (Goldman, 2006, p. 86). This is, again, rather different from the sciences. In “real adult science,” it would seem to be a miracle if scientists (who are working independently) converged upon not just the same theories, but did so in exactly the same order as other scientists (who also just happen to be working on the same thing, at the same time, independently).

Finally, I add, according to the child-scientist account of TT, children perform all of this without collaboration with each other and without any self-conscious reflection. At one point, even Gopnik concedes something like this point. She says, “Of course, formal scientific thinking involves a level of self-conscious reflection, including reflection on the very process of science itself. We do not see this reflection in very young children: The preschoolers see probabilistic evidence and revise hypotheses, but they don’t necessarily know that that is what they are doing… “(Gopnik, 2012). It would seem to be some kind of miracle if every scientist working on the Higgs boson derived the same theory and explanation, in the same short timeframe, in the same order, and with no collaboration or self-reflection. Unlike children who are learning to mentalize, scientists are keenly aware of what it is they are doing.

I think that these four disanalogyes, when taken together, add up to a fairly serious objection to the child-scientist account of TT. When fully fleshed out, glaring differences exist between the way children use data to formulate and test hypotheses and theories and the way that scientists do. I have discussed four of these. It seems that the way children form and revise their thinking
about various domains bears little resemblance to professional science. Given these differences, it is evident that the “striking analogy to science” (upon which the child-scientist account hinges) turns out to be a pretty weak analogy. But Gopnik’s account relies heavily (if not entirely) on this analogy. The fact that this analogy appears to be faulty gives rise to skepticism about Gopnik’s claim that children should be viewed as little scientists.\(^{65}\)

The second question about the “child as scientist” account concerns the validity of Gopnik’s claim that typically developing children experience a “conceptual shift” that endows them with the ToM skills to mentalize, understand that others might hold false beliefs, and consequently predict the behaviors of others. Her support for these claims are (i) that this conceptual shift is illustrated by children’s performance on ToM tasks (e.g. FB tasks) – performance that improves with time because of a development in their understanding of mental concepts; and (ii) that these changes in conceptual understanding reflect successive stages in children’s theories about the psychological and social worlds. I explore these two premises in greater detail here.

Regarding (i), Gopnik (1993) cites empirical data, from her own research and others (See Perner and Wimmer), that there is a noteworthy developmental change in children’s performance on FB tasks. She provides data showing that children under the age of 3 consistently fail FB task whereas. Conversely, children consistently pass the same FB tasks after they attained an average age of 3 ½ - 4 years. Gopnik uses this data to support a principle claim of her child-scientist approach, viz., that a conceptual deficit explains the younger child’s (e.g. younger than 4 years) poor performance on FB tasks and acquisition of mental state concept explains later proficiency on the same tasks. That is, she claims that improved performance on FB tasks is a result of a

\(^{65}\) As stated, you will see later in this chapter, after the Baron-Cohen theories are elucidated, there is another (stronger) reason for not accepting Gopnik’s theory or her conclusions.
developmental shift from a mental *concept* that doesn’t make room for the possibility of a FB to a mental *concept* that admits of the possibility of a FB.

Regarding (ii), Gopnik asserts that the developing ToM allows for a change in concept, which amounts to a change in *theory* explained by successive *theory development*. She says, “normal children’s transition from failure to success on mentalizing tasks hinges on a transition from a nonrepresentational to a representational conception of belief” (Gopnik, 1993). Gopnik thus claims that younger children have a conceptual deficit that older children do not share. She then further concludes, “all attributions, including self-attributions, are executed by theoretical reasoning” (Gopnik, 1993). Thus, as Goldman points out, “the argument from conceptual change is one of the primary arguments in the child-scientist theory’s arsenal” (Goldman, 2006, p. 80).

We should ask some questions to ascertain whether these two premises of this argument support its conclusion. First, since Gopnik says that a young child’s newfound ability to succeed on FB tasks is as a result of a conceptual change, it is worth asking: Is a *conceptual change* just the same thing as a *theoretical change*? Second, does development and acquisition of a ToM sufficiently explain newfound success on FB tasks? Or could some other module explain this success either in conjunction with ToM or in lieu of ToM? And, third, does the supposed conceptual shift occur in children at the age she claims it does? I shall address each of these questions in turn.

Regarding the first question, there is a difference between a *conceptual change* and a *theoretical change*. Importantly, these are not necessarily synonymous in the sense that is relevant to this discussion. Conceptual change is the process whereby concepts and relationships between them change over the course of an individual person's lifetime. A theoretical change, in the sense that is relevant to this discussion, involves change to a scientific-like theory – one that
assumes as fact a relationship between observable and unobservable states. The child-scientist account of TT asserts that children develop theories in the same manner as adult scientists. So, the kind of theory they are talking about is scientific or “science-like” theory. Scientific theories, by definition, “posit causal relationships between observable and unobservable states” (Goldman, 2006, p. 71). But, as Goldman points out, not all concepts or conceptual changes involve the kind of theory or theoretical changes employed by scientists. Further, I would add that thus far there is little if any evidence that changes in mental state concepts involves the kind of theory or theoretical changes employed by scientists.

Indeed, many conceptual changes do not involve any kind of theoretical change, let alone a theoretical change that is science-like. And, he says, “even if a concept is a theoretical one, in the sense relevant here, it doesn’t follow that determinations of its instances must proceed wholly by theoretical inference” (Goldman, 2006, p. 71). On Goldman’s account of concepts vs. theories, it seems that possession of a concept is neither a necessary nor a sufficient condition for the possession of a theory. The fact is that ‘concepts’ and ‘theories’ are not as synonymous as Gopnik portrays. So the child-scientist claim that the shift from a set of mental state concepts that excludes the possibility of a FB, to one that admits of the possibility of a FB reflects successive stages in children’s theories about the psychological world seems dubious. After all, it is entirely plausible that the younger child would do poorly on FB tasks because of a language deficiency or information processing deficiency rather than a conceptual deficiency. Removing the premise from conceptual change from the child-scientist’s arsenal substantially weakens their argument.

Regarding the second question, the child-scientist approach explains the failure of younger children on FB tasks by claiming that younger children performed poorly because they exhibited
a conceptual deficit. And that, further, there is a shift in the child’s ability to conceptualize that Gopnik takes to be a shift in and development of a more sophisticated theory of mind comprised of successively more sophisticate theories. However, the timeframe in question - 3 ½ - 4 years of age – is one in which typically developing children experience a veritable explosion of modular acquisitions responsible for a wide variety of new developmental achievements. So, it warrants inquiry as to whether this developmental shift represents solely a shift in concept acquisition, or whether some other developmental achievement could either partially or wholly explain FB task proficiency.

There are many hypothesized contenders (other than ToM) that could explain the low proficiency rates on FB tasks seen in younger children. I have previously mentioned language and information processing speed deficiencies as possible candidates. Additional candidates include immature episodic memory, executive functioning, or empathy. Acquisition of language skills is just one of many developmental milestones reached during this timeframe, so I will use language acquisition as an example of such a candidate to illustrate my point. It is entirely possible that children under the age of 4 are actually in possession of ToM skills, but simply lack the linguistic skills to articulate their understanding of mental states and FBs; that children under 3 years of age simply do not carry mental-state terms in their limited vocabularies, whereas children over 4 years of age are more apt to know and articulate such terms.

Gopnik does little to contemplate other unattained developmental modules that would produce processing deficiencies, which might explain the supposed failure on FB tasks by children under age 4. Moreover, if an alternative module, such as language acquisition, could offer a better explanation of 4 year-olds successes on FB tasks, and this could be supported by empirical data, then this would seriously undermine this crucial part of the child-scientist account
of TT. Of course, at that juncture, Gopnik could simply claim that the development and acquisition of ToM contributes to the explanation of this success. But in so doing, she would be making a substantially weaker claim about ToM. In any case, as it turns out, empirical data does suggest that children exhibit ToM skills long before the age at which Gopnik claims they do.

Regarding the third question, there is a fair amount of empirical evidence to support the idea that a different processing deficiency could explain the failure that 3-year-olds on false-belief tasks. I find the most compelling evidence to be empirical data that shows that 3-year-olds actually are quite capable of recognizing false beliefs (Wellman and Bartsch, 1988). In fact, Clements and Perner (1994) were able to show that children as young as 2 years, 11 months can demonstrate an understanding of false belief and, more specifically, that the target may possess a false belief that is not shared by the child. The most recent data provides evidence that 15-month-old infants demonstrate an understanding of the false-beliefs of others when the task demands are reduced in such a way so as not to necessitate linguistic skills as a prerequisite for passing the false-belief task (Onishi and Baillargeon, 2005). As Goldman states, these findings together with the confirmation of these findings “disconfirms a core thesis of the child-scientist approach, that young children’s mental-state theory either omits belief entirely or construes is as a nonrepresentational state (which cannot be false)” (Goldman, 2006, p. 78).

It is now evident that, at the very least, there are some problems with the claim that typically developing children experience a “conceptual shift” that endows them with the ToM skills to mentalize, understand that others might hold false beliefs, and consequently predict the behaviors of others. The literature suggests: (a) that even if the shift from failure to success on FB tasks could be explained as a conceptual change, it is not at all clear that this is also a theoretical change which reflects the successive development of more sophisticated theories; (b) that it is
entirely plausible that development of some other module, such as language acquisition, could explain the sudden change in a child’s ability to pass FB tasks; and (c) that there is enough empirical data to show that ToM skills are in place well before the target age of 3 ½ - 4, indicating that, in fact, some other module besides ToM development can explain the conceptual shift that Gopnik proposes. “Their conceptual deficit diagnosis of early performance on false-belief tasks has crumbled in the face of accumulating evidence, and new studies suggest quite a different time line for the understanding of false belief [as well as quite different methods for understanding and articulating false beliefs]” (Goldman, 2006, p. 92-93). If this is right, then Gopnik’s points about conceptual development in young children cannot be supported.

Following Goldman, my third and final question regarding Gopnik’s child-scientist account of TT queries whether the evidence (i.e. the empirical data) that she provides actually indicates that children learn and theorize about minds in the way that she says that they do. Moreover, I question whether the data from her empirical experiments supports her conclusions that children learn and theorize about “psychological and social worlds” via the same means and in the same manner by which they learn and theorize about physical and biological worlds.

In several articles, Gopnik describes her claim that children are intuitive scientists and that they learn about the world – especially the physical, biological, and physiological worlds – in precisely the same way that real adult scientists do. Scientists start with a hypothesis; they perform experiments and gather data; they analyze their results; and they then revise their theories as necessary. They posit causal relationships between observable and unobservable states. This method results in successively more sophisticated theories about the world. Gopnik’s child-scientist learns about physics, biology, and chemistry in this same way, but through a
process of “naïve theory formation” (“naïve” ostensibly because they are children and not trained scientists).

It is plain that scientists work in this way and that scientific theories progress in this way. Like Goldman (2006, 2012), I question whether children learn about the world in the manner that Gopnik describes. Goldman (2006) claims that because different kinds of problems require different kinds of problem solving and solutions, it is dubious that children employ one single universal learning method for different kinds of problems (i.e. solve the problems of physics in one way and solve the problems of algebra in a different way). Goldman says,

Even if Gopnik and colleagues could persuade us that children use their favored causal learning procedure for some domains, it is highly speculative to infer that the same procedure is used in learning mentalizing skills. This is especially pertinent because Gopnik et al. (2004) report no experiments with specifically psychological, that is, mentalizing, subject matters. One cannot assume that if children have the capacity to use learning mechanisms like the Bayes net, they will surely apply them to the problem of mentalizing (2006, p. 85).

Similarly, but more specifically, I question whether children solve the problems of psychology, linguistics, and sociology (i.e. the so-called “soft sciences”) using the same tools by creating naïve theories, in the same manner, as they solve the problems that Gopnik cites in physics, biology, and chemistry (i.e. the so-called “hard sciences”).

If we can question the validity of the claim that children learn about all of the different sciences in the same way, we can surely question the validity of the claim that children learn about mentalizing and develop naïve psychological theories in the same way that children learn about gravity and develop naïve physical theories. Surely, it is at least plausible that children learn about physics and/or biology using one means of learning mechanism, but learn about minds and mentalizing via a different means altogether. This plausibility leaves open the possibility that children do not learn about or theorize about mental states the way that they do
about science, or for that matter whether scientists theorize about science the way that the child-
scientist account would have us believe.

This raises the second question of this objection: Does her empirical data support her
conclusions that children learn and theorize about psychological and social worlds via the same
process and in the same manner by which they learn and theorize about physical and biological
worlds?

First, I think it bears repeating that Gopnik and her colleagues report not one single
experiment gathering information from specifically psychological - that is mentalizing, subjects.
Gopnik and her colleagues create experiments to test children’s ability to theorize. They do this
by using physical objects such as “gearboxes” and “blicket detectors,” (See Gopnik and Schulz,
2004 and Gopnik, 2012). They also test for predictions based on mental states using observables
such as goldfish crackers vs. broccoli and ducks vs. frogs combined with the dramatic facial
expressions of the experimenter (Gopnik, 2010). All of her data is collected using physical
objects: mechanical objects and toys that work in specific ways causing the child to assert a
causal theory of how the observable material objects work.

Gopnik says, “We have shown that by age four, children can use information about the
interventions of others appropriately to make new causal inferences. Consider the gear toy
experiment described above…” (Gopnik, 2004, p. 27). Here, again, Gopnik uses gearboxes and
blicket detectors, etcetera to prove what children might infer about mental states. However, it
seems plausible to me that children might, as the child-scientist account claims, employ theories
to understand causal structures of the physical world, but that they use some other process or
ability (such as simulation and/or introspection) to understand psychological and social worlds.
If so, this would dramatically undermine Gopnik’s version of TT.
Goldman (2012) also notes that Gopnik does not use one ounce of empirical data derived from experiments revolving around mentalizing to formulate her hypothesis regarding how children metalize. This seems *prima facie* to be problematic. There is a significant difference between understanding gearbox toys or blicket detectors, and understanding the mental states of others. Obviously, the former are physical, tangible, things that can be manipulated to ascertain their underlying mechanism(s) and inner workings. Conversely, mental states are not immediately tangible physical things in the same way that blicket detectors and gearbox toys are, and they are not manipulatable (at least, not in the same way). That these are so clearly two different *kinds* of things renders them disanalagous and this should, at the very least, not be ignored.

We should note not just the distinction between physical test objects versus non-physical mental states, but also the second (equally important) distinction between observable and unobservable test objects. Goldman reminds us “that the problem facing children in the mentalizing domain is a problem with *unobservable* states (according to theory theorists)” (2006). Gopnik and her colleagues provide very little evidence or empirical data to support her claims that children can and do theorize about unobservable things. They do cite one experiment that involves unobservable entities, which involves a puppeteer manipulating a puppet from behind a puppeteer box. Goldman notes, “Even in this experiment, the entities are unobserved but they are observable – that is, they are both known and discoverable…these are not *unobservables*. These have been observed on many previous occasions” (Goldman, 2006). By contrast, the TT claims that mental states such as desires and beliefs are *unobservable states*.

Gopnik et al. offer little if any direct experimental evidence of children learning to postulate such states or learning to theorize about immaterial or unobservable entities. It seems that their
argument launches from their account of development in general, and not from an account of mentalizing specifically. This, again, seems *prima facie* to be problematic for the child-scientist version of TT, since they are seeking to prove an account of mentalizing which is derived from a successively sophisticated ability to theorize about unobservable entities.

### 4.2.2. Discussion

For now, I have elucidated what I think are three pretty serious objections to the child-scientist account of TT. First, the parallelism to science is neither an appropriate nor an accurate analogy and, in any case, the analogy is not strong enough to support Gopnik’s reliance upon it to support her position. Second, the empirical data surrounding FB tasks does not support Gopnik’s premises that (i) children experience a conceptual shift regarding mental states – a capacity which improves suddenly and dramatically when children attain 3 ½ - 4 years of age; and (ii) that these changes in conceptual understanding of mental states reflect successive stages in kid’s theories about the psychological and social worlds. Third, the evidence that Gopnik provides doesn’t suggest that children learn and theorize about *minds* the way she says they do. That is, her conclusions that children learn and theorize about psychological and social worlds *via* the same process and means, and in the same manner, by which they learn and theorize about physical and biological worlds utterly outstrips her evidence. Each of these objections reveals a different sort of weakness to the account.

While I think that each of these objections are valid, I think the most damaging of this set of objections is the first objection. Gopnik claims that the way in which children learn to mentalize and develop a ‘folk psychology’ is *strongly* analogous to the way scientists learn and perform the research functions of science – that is, *via* the development of progressively more sophisticated
I concur with Goldman and others (see Stone and Davies, 1996) that Gopnik’s *strong analogy* to science – upon which her entire account depends – *fails* in several ways.

It would have to be by some kind of miracle if every scientist, in every location, arrived at the same theories, and developed them in the same order and in the same time frame without any collaboration with the scientific community, and, moreover, without any self-reflection or introspection as to which theories the individual scientist is working on. But, as we have seen, this is precisely how Gopnik claims that young children come to possession of ToM skills.

When I reviewed Gopnik’s descriptions of how it is that typically developing children develop a full-fledged ToM, I had a knee-jerk (reflexive) reaction and thought, “But wait… that’s just not how scientists *work*… That’s just not how scientific theories develop and progress!” Very little research into secondary and tertiary literature reveals that this is a common objection to Gopnik’s child-scientist account of TT. Goldman (2012) notes that, indeed, scientists develop theories that progress in sophistication and accuracy *via* ongoing discussion and debate within the scientific community regarding which version of the theory has the most veracity. Often different scientists will contribute to existing theories, or revise theories that were hypothesized by someone else. Seldom, if ever, is *just one* scientist doing *all* of the work on a given theory. And never is the scientist doing the work in isolation. Scientific theories, unlike the theories of naïve folk psychology, are “written in learned journals and textbooks, are subject to rigorous investigation, and must be actively taught (Stone & Davies, 1996). The disanalogies between Gopnik’s child-scientist account of TT and “real adult science” are well documented in the literature.

It occurs to me however that there is a further objection that dovetails with the well-known disanalogy objection. I have taken note that Gopnik *says* that children learn about and develop
the ability to mentalize in precisely the same way as a scientist does. Thus, she says, children are little scientists. When we consider “real adult scientists doing real adult science,” we see that scientific theories become successively more sophisticated over long periods of time via the contributions and collaborations of the collective scientific community. Thus, what Gopnik implies is that the typically developing child does the same theoretical work as an entire scientific community completely on their own and with no instruction or collaboration. Part of what makes Gopnik’s assertion so weak is that she says:

\(a\) The child = a scientist.

However, my interpretation is that her claims imply:

\(b\) The child = an entire scientific community.

So, if we take Gopnik’s analogy seriously, we can extrapolate that, on her account, an individual child is doing the same kind of theoretical work in the domain of mentalizing – the same heavy lifting – as an entire team of scientists would do in the development of any given theory over a lengthy period of time. Taken to its logical ends like this, not only is her account not analogous to science, but also it even seems a bit absurd. Even if we are charitable to Gopnik’s account, granting that she really does mean the lesser claim, \(a\), she still becomes entangled in the disanalogy between children doing folk psychology and “real adult scientists” engaging in professional science (Gopnik, 1997, 2004, 2010, 2012). Such an entanglement is a real worry for Gopnik’s child-scientist version of TT because breaking this analogy strips her of one of the primary pillars of her theory of ToM.

Gopnik could, of course, respond to this. Oddly, if Gopnik could concede part of my third objection (i.e. *Children do not learn and theorize about psychological and social worlds via the same process and in the same manner by which they learn and theorize about physical and*
biological worlds), she could hold onto a part of what she loses because of the first objection. Perhaps here is how:

My third objection emerged from my sense that something seems faulty with Gopnik’s application of her empirical data. It seems to me that Gopnik has strained her data with a hasty overgeneralization by extrapolating from her suppositions about observables and applying them to unobservables. Let me explain. In my review of the literature, I had noticed that all of Gopnik’s experiments with children in the laboratory utilized test materials that are both material and observable (e.g. gearboxes, blicket detectors, goldfish crackers, frogs, puppets, and etc.). From the data she collects, she draws conclusions about the way that children learn, theorize, and make predictions about the physical and biological worlds – each of which are also both material and observable. She claims, perhaps rightfully, that children develop a sort of “folk physics” and “folk biology.” She then relies on these various experiments to draw conclusions pertaining to the way children learn, theorize, and make predictions about psychological and social worlds neither of which is material or observable in the same (concrete) sense as the physical world or the biological world.

Although he doesn’t expand upon the point, Goldman also briefly notes, “Gopnik et al. report no experiments with specifically psychological, that is, mentalizing, subject matters” (Goldman, 2006). I continue to maintain that there is an important difference between the observable and material (physical objects), versus the unobservable and immaterial (mental states). It is hard to understand how we can really know anything about mental states without conducting experiments with mental state subject matters.

Gopnik draws another analogy, apart from her child as scientist analogy. She claims that the way that children learn about physical material things is analogous to how they learn about
mental, immaterial things. Of course, an analogy compares two things, and, on the basis of their similarities, allows us to draw conclusions about the objects. The more closely each thing resembles the other, the more accurate the conclusion. The objects under scrutiny must be similar enough to allow for a meaningful analogy. I am simply arguing that physics (for example) and psychology are too dissimilar to allow for analogy.

Gopnik could respond by conceding this point, conceding that there is a notable distinction between kinds, and conceding that, “in many respects, the practices of folk psychology and even scientific psychology do not bear the marks of quantum physics, or inorganic chemistry, or molecular biology, or neurophysiology” (Stone & Davies, 1996). Moreover, in conceding this third point, she would get back part of what she loses by the first objection. She could say, “Look, maybe there is no continuity between folk psychology and the aforementioned sciences, but, certainly, there is continuity between ‘folk psychology’ and “real adult” scientific psychology. So, even if the way a child learns about mentalizing does not map onto the way a child learns about physics or biology, surely the way a child learns about mentalizing (‘folk psychology’) is analogous to how real adult scientists do real adult psychology!” Such concessions would allow Gopnik to hold onto an analogy – perhaps even a strong analogy, but one that is limited to cognitive psychology.

The question is: Would Gopnik’s supposed response be sufficient to save an intact version of her child-scientist account of TT? It seems to me that even this response would not be sufficient to accomplish this. Even if Gopnik gives up part of the third objection, she would only recover a small part of what she lost by the first objection. That is, she would only be able to assert a strong analogy as mentalizing relates to cognitive psychology. However, really, she could only argue a weak analogy to science in general. Accordingly, her child-scientist account would be
radically reduced and might not resemble her initial account substantially enough to save her position.

4.2.3. Additional and Concluding Remarks Regarding Gopnik’s Child-Scientist Account of the Theory Theory

I think that the objections and questions raised against the child-scientist version of TT thus far in this chapter are real worries for Gopnik’s child-scientist account. First, is the parallelism to science an accurate or appropriate analogy? Second, does the empirical data surrounding FB tasks support Gopnik’s premise that children undergo a (rather dramatic) conceptual shift in the way Gopnik claims and in the timeframe in which she says it occurs? Third, and finally, does Gopnik’s empirical data provide evidence to support her conclusions that children learn and theorize about minds, psychological worlds, and social worlds the way she claims they do?

Gopnik herself says,

The real question for developmental cognitive science is not so much what children know and when they know it, but how children’s theories develop and change and why children’s theories converge towards accurate descriptions of the world. It is all very well to suggest that children’s learning mechanisms are analogous to scientific theory-formation. However, what we would really like is a more precise specification of the mechanisms that underlie learning in both scientists and children (Gopnik and Schulz, 2004, p. 371).

A more thorough examination of Gopnik’s account shows that we really do not have a precise specification of the mechanisms that underlie learning. Gopnik concedes that we do not know how children are learning about mentalizing. She, in fact, says we are very far from knowing the story of how children (and scientists) learn and grow their bodies of knowledge. “There is no precise and convincing explanation of learning” (Gopnik, 2004, p. 24). Certainly, her research provides compelling evidence that, developmentally, children’s ToM skills increase throughout
childhood. However, her data leaves unclear that the TT or the child-scientist version of TT is the underlying medium of ToM.

There is an additional objection to Gopnik’s version of TT that is important, and I can now note it at this juncture. Recall that in the first section of this chapter, I mentioned the fact that nowhere does any theory theorist write a concise argument in favor of the TT. So, in the same section, I state this argument the way that Goldman couches the general argument in favor of TT. Goldman (and I, following his lead,) then set about undermining the premises of that argument (the TT argument as Goldman understands it).

Like other theory theorists, Gopnik does not actually offer any clear and concise argument favoring her child-scientist approach to the TT. Her work could be interpreted as an inductive argument by analogy. Alternatively, her work could also be understood as an argument of the form: \( P1 \ldots \) Children are like scientists with respect to learning. \( P2 \ldots \) Specifically, children are like scientists in these ways: \( a, b, \) and \( c. \) Therefore, children are miniature scientists. The problem is, again, that nowhere does Gopnik actually make an argument in favor of the TT in general or for her child-scientist version of the TT in particular. Instead, Gopnik just boldly asserts her claims without really making any argument whatsoever.

Thus far in this chapter, I have elucidated how Goldman is more or less working to undermine Gopnik’s conclusions. I have followed suit. I agree with Goldman on the points that undermine Gopnik’s conclusions. And I have added to them.

I have shown that Gopnik begins with the claim that kids learn the way scientists (in general) learn. She then extrapolates from science in general to some specific “hard” sciences: biology, physics, and chemistry. This weakens her analogy. She then further extends (and I would say,  

---

66 A & B are each alike with respect to property Q. A has quality P. Therefore, B probably also has quality P.
“strains”) her analogy with her claim that kids learn about specific “soft sciences” such as psychology and the social world in the way that scientists do. I hold that the further Gopnik moves away from her analogy (that “kids are exactly like scientists because kids learn the same way that scientists work”), the weaker her analogy becomes until it just breaks down. It is not appropriate to assume that kids have only one dedicated learning mechanism, that they learn about the physical worlds in the same way that they learn about mental worlds, or that they learn in the same way that scientists learn about both the hard sciences and the soft sciences.

Restated, it is not the case that Gopnik’s premises do not support her conclusions, because there really are no premises. She offers only boldly stated conclusions. Gopnik’s conclusions simply outstrip her evidence and I intend to show that the considerations in favor of Gopnik’s conclusions are weak because they are based on a weak analogy. It turns out that kids are not very much like scientists at all, and the ways that kids are like scientists are not in the ways that are relevant to mindreading.

This brings forth a novel and compelling argument against Gopnik’s child-scientist account of TT. To elucidate this argument, I first need to begin with a detour away from Gopnik and her work, and look to the work of another theory theorist, Simon Baron-Cohen. It may seem off-topic to offer an exposition of Baron-Cohen’s two theories at this juncture, but I do so in the service of undermining both Gopnik’s child-scientist version of TT and the TT as a whole. As I will show, Baron-Cohen’s work and findings serve an important role in a novel argument against Gopnik’s child-scientist account and the TT in toto.
4.3. Baron-Cohen’s Innate Minimalist Modularity Theory and Empathizing – Systemizing Theory/Extreme Male Brain Theory

As stated in the previous section, there are many theory theorists and a variety of iterations of the TT. There are some views that theory theorists share. For example, I think it is fair to say that most theory theorists assert that there is at least some innate basis to ToM (although some assert this claim more strongly than others). I think that most theory theorists would also deny that when we mention or allude to mental states (both our own mental states and the mental states of others) these are derived from our own first-person experiences or personal reflections. Also, I think most theory theorists represent mental states as concepts (abstract ideas) that serve as self-evident postulates to a theory, which have been developed through a process of abstract theorizing (much like Gopnik describes). However, theory theorists often disagree about just how and when a ToM is acquired and just how the ToM skill is honed. In particular, there is disagreement among theory theorists about whether the theory in question is acquired via the way that Gopnik suggests (comparable to the acquisition and development of scientific theory), or if the theory is innate, modularized, or starting-state nativistic.

Simon Baron-Cohen is a Professor of Developmental Psychopathology at the University of Cambridge and Director of the Autism Research Centre in Cambridge. In his early work, Baron-Cohen aligns himself with theory theorists. He says, “In this sense, ToM can be thought of as a theory: It explains and predicts other’s behavior” (Baron-Cohen, 2008, p. 57). Unlike Gopnik, who asserts that ToM acquisition is comparable to the acquisition and development of scientific theories, Baron-Cohen’s early work falls into the latter category, asserting that the basis of ToM is both innate and modularized in some sense. This modularity approach to ToM studies is most often conceived of as a form of the TT. I cover his work in this section of the dissertation.
because he was among the first researchers to link a ToM deficiency or impairment to ASD, a link that he uncovers in his book, *Mindblindness* (1995) and again in his book, *Autism and Asperger Syndrome: The Facts* (2008). I also cover his work because his most recent theory links ASD to the practices of science. An exposition of the links between the ToM (both as ability and as mechanism), ASD, and science will work to undermine Gopnik’s child-scientist account and the TT as a whole. But first, I must delve into the intricacies of Baron-Cohen’s two theories.

During the course of his work, Baron-Cohen has developed (at least) two separate theories, both of which reflect his background in evolutionary biology and developmental psychology, and both of which extend in an attempt to explain the phenomenon of autism. First, Baron-Cohen proposed a modular form of the TT, the *Innate Minimalist Modularity Theory* (henceforth, ‘IMMT’) (1995). This is a correlative theoretical model, through which he asserts four discrete sub-systems (or mechanisms) that work together to comprise the human mindreading capacity. These four sub-systems, correspond roughly to components in the world: volition, perception, shared attention, and epistemic states. This theory argues that individuals with autism spectrum disorders exhibit a delay, a deficiency, or an impairment of the ToM mechanism. Baron-Cohen extends this theory to his *Mindblindness Theory of Autism*.

Later, Baron-Cohen introduced another theory: the *Empathizing – Systemizing Theory*. This theory tests individuals using the Empathy Quotient and Systemizing Quotient (which Baron-Cohen developed) to determine whether that individual is more proficient at empathizing or systemizing. The quotients turned out to be a fairly reliable predictor of a difference between the sexes in that, according to Baron-Cohen, men tend to be better systemizers and women tend to be better empathizers. The Empathizing – Systemizing Theory also turned out to be a more reliable predictor than is gender as to who would be more likely to enter professions in science,
technology, engineering, and mathematics (henceforth, ‘STEM’). In an effort to explain certain
diagnostic features of autism that could not be explained by his IMMT, Baron-Cohen extended
his Empathizing – Systemizing Theory to his Extreme Male Brain Theory of Autism (henceforth,
‘EMB’) (2003, 2008). The development of the EMB seems to represent a shift for Baron-Cohen.
Whereas his earlier modularity approach to ToM placed him in the TT camp, his more recent
work on the EMB seems to place him in the ST camp because he begins to frame mindreading as
requiring empathy and the ability to imagine what it is like to be someone else.67

I will now offer a brief exposition of these theories by Baron-Cohen, after which I will
provide a discussion of some objections, comments, and criticisms of these theories. Lastly, I
will show how Baron-Cohen’s theories, when taken together, serve to undermine Gopnik’s child-
scientist view and the TT.

4.3.1. Baron-Cohen’s Innate Minimalist Modularity Theory

Baron-Cohen elaborated the modularity approach to ToM when he proposed IMMT (1995).
This is correlative theoretical model that postulates four discrete sub-systems (or lower level
perceptual mechanisms) that work together to extract relevant social information and so
comprise the human mindreading capacity (that is, the ToM).

These mechanisms include an Intentionality Detector, an Eye Direction Detector, a Shared
Attention Mechanism, and a Theory of Mind Mechanism. According to Baron-Cohen, these four
sub-systems or mechanisms, while they do not seem to form a natural class, work in conjunction
with each other to inform the overall ToM, providing the critical inputs to developing a ToM,

67 Note: Baron-Cohen does not explicitly mention the labels “theory theory” or “simulation theory” but, in his
efforts to explain ASD, his early work seems to support the TT and his later work seems to support the ST.
which allows us to impute mental states to others. A description of the four sub-systems follows.

The first mechanism that Baron-Cohen describes is the Intentionality Detector (henceforth, ‘ID’). The ID, which works through vision, audition, and sensation, is proposed to be a perceptual device that interprets ‘primitive volitional’ mental states such as goal and desire. These are seen as basic mental states required for making basic sense of the movements of all organisms in the environment. For example, an infant seeing an adult move across the room would (through the ID mechanism) interpret that as ‘the adult wants to go there’ or ‘the adult does not want to stay here.’ The ID is considered to be an innate mechanism that infants possess for reading mental states. The ID “grabs even an infant’s attention to animate actions, providing opportunities for the infant to learn about goal-directedness” (Baron-Cohen, 2006).

The second mechanism that Baron-Cohen describes is the Eye Direction Detector (henceforth, ‘EDD’). Baron Cohen suggests that EDD has three basic functions of detecting the presence of eye-like stimuli, computing whether eyes are directed towards it or towards another direction, and inferring of something else that it actually sees something. In other words, it interprets ‘gaze’ as ‘seeing.’ The EDD “grabs the infant’s attention to the eye region of faces and thus provides opportunities for the infant to learn the significance of gaze as a clue to a person’s mental states” (Baron-Cohen, 2006). The EDD is considered especially important to mindreading as it allows the infant to attribute a specific perceptual state to the organism (“the monkey sees the banana” or “Mommy sees me”). Both of these mechanisms allow the infant to interpret observed behavior as a small number of mental states.

---

The third mechanism Baron-Cohen describes is the Shared Attention Mechanism (henceforth, ‘SAM’). The SAM is considered to be a higher-order skill that allows the individual to form what is called ‘triadic representation.’ Triadic representations conceptualize relations between an Agent, the Self, and an Object (which may or may not be another agent) (Baron-Cohen, 1995, p. 44). The SAM builds triadic representations by perceiving the perceptual state of another agent and computes shared attention by comparing another agent’s perceptual state with the self’s current perceptual state. SAM takes inputs from ID and EDD, which “enables the infant to work out if s/he and another person are attending to the same thing, thus ensuring that shared foci or common topics are a central experience for the developing infant” (Baron-Cohen, 2006). The SAM is critical to Baron-Cohen’s IMMT of ToM because, without some kind of a SAM, ToM cannot even get started, (Baron-Cohen, 1995, p. 55). Importantly, Baron-Cohen assert that the SAM, which allows us to build a triadic representation such as ‘you and I are attending to the same x’ is massively impaired in individuals with ASD (Baron-Cohen, 1995, p. 64).

The fourth and final mechanism that Baron-Cohen describes in this model is the Theory of Mind Mechanism (henceforth, ‘ToMM’). “In this model, ToMM is conceived of as either being a more mature development of SAM, or simply triggered by SAM” (Baron-Cohen, 2006).

Further, Baron-Cohen says that, although the other three mechanisms are constitutive of ToM, we need the ToMM over and above the other three mechanisms. The ToMM is considered to be a system for inferring the complete range of mental states observed from human and animal behavior – that is, for actually employing a ToM. The ToMM achieves this by representing the set of epistemic mental states (imagining, dreaming, knowing, deceiving, believing, etc.) and combining the different mental state concepts (volitional, perceptual, and epistemic) in a coherent understanding of how mental states and actions are related. ToMM has the dual
function of representing the set of epistemic mental states and turning all of this mentalistic knowledge into a useful theory” (Baron-Cohen, 1995, p. 51). This theory, we call the ToM (which both Gopnik and Baron-Cohen refer to as the TT).

After his in depth descriptions of these four discrete sub-systems (ID, EDD, SAM, and ToMM) that comprise ToM, Baron-Cohen extends his IMMT in an attempt to answer the question ‘What could be the cause of autism?’ Equipped with a theory of how a fully-fledged ToM operates in individuals who do not have ASD, combined with the diagnostic criteria for persons with ASD, and some evidence that individuals with autism fail false belief tasks (indicating a diminished ToM), he asserts his Mindblindness Theory of Autism (1995)

4.3.2. Baron-Cohen’s IMMT and the Mindblindness Theory of Autism

There is something that it is like to have a typically developed and normal functioning ToM. Human beings make observations, put facts together, and make inferences regarding what others (and what we ourselves) know and believe, feel and desire. Ostensibly, although we cannot see mental states, we attribute them with logic and precision, and not by tenuous and vague speculation. This understanding is based on a “powerful mental tool” (ToM) that every typically developing child develops in early childhood and every typically developed adult possesses and uses with varying degrees of skill. ToM provides us with the ability to predict relationships between external states of affairs and internal states of mind. Typically developed individuals will accumulate tacit knowledge and continually update it as they seek to share

---

69 Please see the Theory of Mind chapter of this dissertation, for a detailed description.
mental states with their peers and elders. This ability is called *mentalizing* or *mindreading* (Frith, 2003, p. 77).  

The Mindblindness Theory of Autism (henceforth, ‘MTA’) essentially links ToM to ASD by asserting that individuals with ASDs are delayed in developing a ToM, and that the ToM is somehow deficient or impaired. It has been observed that these individuals are lacking the ability to take another’s perspective, imagine their thoughts, feelings, and beliefs, and then predict their behavior based on that understanding. The difficulties found in ASD are not those of making inferences in general, but it is very specifically a difficulty with thinking about mental states. “Here, autism is a lack of knowledge or belief. The TT claims that there are just some propositions that the autistic does not know or cannot formulate” (Currie, 1996, p. 242). Hence, it is supposed that people with ASD, those who lack a ToM, are (in Baron-Cohen’s terms) ‘mindblind.’

Baron-Cohen believes that people with ASD have a fully functional ID – that is, they can distinguish between organisms with intentionality and organisms without intentionality. He also asserts that those with ASD have a functional EDD – which is a sort of geometric exercise in detecting which way a person is looking.

However, he asserts that SAM, which allows us to discern that we are attending to the ‘same $x$’ as another agent, is massively impaired in those with ASDs (Baron-Cohen, 1995, pp. 62-64). In these individuals, SAM is not just a deficit of joint visual attention. In most people with ASDs, SAM does not appear to be working in any modality – vision, touch, or audition. So, for example, a deficit in SAM would logically lead to one typical trait of autism: inability to

---

70 Note Frith’s use of the term “powerful mental tool” when referring to ToM. Like Gopnik (and indeed *most* theory theorists), Frith seems to write under the assumption that ToM exists and exists as a modular mechanism, which, I think, is far from definitive.
modulate intonation to make speech audible and interesting to another person (i.e. a listening audience). According to Baron-Cohen, an autistic person either lacks or fails to employ the concept of another person as an interested (or disinterested) listener. Alternatively, a person with ASD may simply believe that the contents of the listener’s mind are the same as the contents of her own mind and, thus, the listener’s interests must be the same as hers. So, for example, the person with an ASD would simply assume that her listener shared the same interest in the memorization and recitation of 9,000 license plate numbers, thus continuing her recitation of the numbers while failing to notice the disinterest of her listener. Finally, without SAM to trigger ToMM, virtually all aspects of ToM are impaired in someone with an ASD (Baron-Cohen, 1995, p. 69). “People with ASDs may be puzzled by other people’s actions, or anxious because other people’s behavior seems unpredictable, precisely because they cannot use a ToM to interpret or anticipate what others are doing or are going to do” (Baron-Cohen, 2008, p. 57).

The MTA asserts some very specific deficiencies in ToM. It assumes that: (1) People with an impaired ToM are mindblind. (2) All people with ASD possess an impaired or deficient ToM. (3) So, those with ASD are mindblind. Uta Frith says, “The mind-blindness theory is appealing because many problems in social interaction and in communication can be understood as a consequence of the inability to realize fully what it means to have a mind and to think, know, believe, and feel differently from others” (Frith, 2003, p. 207) I wish to add that this theory is appealing because it accounts for and makes sense of many of the social impairments and communication impairments that are hallmarks of the diagnosis for anyone with an ASD according to the DSM-V (2013). However, the MTA does not account for all of the diagnosable aspects of ASD, and Baron-Cohen continues his work to explain ASD via his more recent theories of ToM.
4.3.3. Baron-Cohen’s Empathizing–Systemizing Theory and Its Extension to the Extreme Male Brain Theory of Autism

Baron-Cohen’s Empathizing–Systemizing Theory (henceforth, ‘E-S’) tests individuals for two distinct capacities: their interest in and ability to empathize (henceforth, ‘E’) and their interest in and ability to systemize (henceforth, ‘S’). To accomplish these assessments, Baron-Cohen et al. developed two tests: the Empathy Quotient (henceforth, “EQ”) and the Sympathy Quotient (henceforth, ‘SQ’). The EQ measures the relative strength of an individual’s empathy – that is, their ability (and desire) to identify, understand, and respond appropriately to the thoughts and feelings (i.e. the mental states) of other people. By contrast, the SQ measures the individual’s relative strength of interest in systems and desire to analyze or construct them. Systems, Baron-Cohen claims, include “anything that follows rules or lawful patterns. Key classes of systems include mechanical systems, natural systems, abstract systems, and collectible systems” (Baron-Cohen, 2009).

According to Baron-Cohen, when taken together, the EQ and the SQ reveal five discrete brain types (or profiles), which are determined by the presence or absence of discrepancies between the individual’s capacities for E and S relative to each other, as assessed by the individual’s scores on EQ and SQ. These five profiles are as follows:

- **Type E**, in which the individual’s ability or interest to empathize is at a significantly higher level than their interest or ability to systemize (E>S).

- **Type S**, in which the individual’s ability or interest to systemize is at a significantly higher level than their interest or ability to empathize (S>E).

- **Type B**, in which an individual’s abilities or interests for empathizing is balanced at the same levels as their ability or interest for systemizing (E=S).

- **Extreme Type E**, in which an individual’s ability or interest to empathize is above average, but whose ability or interest to systemize is below average (E>>S).
• **Extreme Type S**, in which an individual’s ability or interest to systemize is above average, but whose ability or interest to empathize is below average (S>>E).

Baron-Cohen claims that these E-S profiles are exhibited throughout the general population and are illustrative of a difference between genders. Consistently, more female than male test participants showed an E>S profile. By contrast, consistently, more male than female test participants showed an S>E profile (Baron-Cohen, Knickmeyer, & Belmonte, 2005). Baron-Cohen additionally claims that the E-S is a better predictor than simple gender as to who will enter the STEM disciplines and workforce (Billington, Baron-Cohen, & Wheelwright, 2007). That is, regardless of gender, those who select STEM disciplines and enter the STEM workforce are overwhelmingly profiled as S>E.

It is concurrently widely observed that individuals diagnosed with high-functioning ASDs (and those formerly diagnosed with Asperger’s Syndrome)\(^{71}\) also tend toward STEM disciplines and professions. Moreover, data suggests that more males are diagnosed with ASD than females.\(^{72}\) It is widely accepted that people with ASD exhibit an above average interest or ability to systemize while simultaneously exhibiting a below average interest or ability to empathize (Baron-Cohen, 2009). Baron-Cohen also found that the overwhelming majority of individuals with ASD bear the S>>E brain type profile, regardless of gender. Thus, Baron-Cohen extended his E-S theory to the Extreme Male Brain Theory of Autism (henceforth, ‘EMB’) and, in so doing, now links ASD to proficiency in STEM.

Baron-Cohen outlines EMB in detail in his book, *The Essential Difference* (2003). Here, he makes a somewhat bold move with his assertion that there are differences in the brains of males

---

\(^{71}\) Please see the Chapter 3 of this dissertation for a detailed discussion concerning guidelines for diagnosing ASDs.  
\(^{72}\) Baron-Cohen’s research shows ASDs are strongly biased toward males with ratios of 4:1 (males: females) for classic autism (Kanner Syndrome) and as high as 11:1 in individuals with high functioning autism (and those formerly diagnosed with Asperger’s Syndrome) (Baron-Cohen, et al. 2011).
versus females, as well as differences in the brains of people with ASD versus those without ASD. In particular, Baron-Cohen et al. have correlated increase fetal testosterone (henceforth, ‘FT’) with specific groups: males, those with an S>E brain type profile, and those with ASD who exhibit the S>>E brain type profile. EMB is essentially a shorthand explanation for the apparent phenomenon that more males than females show brain type profiles of S>E, and more males than females are afflicted with ASD, showing a brain type profile of S>>E (Baron-Cohen, 2005).

In the following section I will discuss some challenges and objections to that Baron-Cohen’s theories, and how he might answer them. Moreover, and most importantly for this dissertation, I will discuss how it is that Baron-Cohen’s EMB (in conjunction with his IMMT and MTA theories) makes explicit that a link exists between ToM, ASD, and a strong preference for systemizing and STEM disciplines.

4.3.4. Discussion of Baron-Cohen’s Theories: IMMT, MTA, E-S, and EMB

IMMT:

The precursor to Baron-Cohen’s IMMT account of ToM was the ToMM theory by Alan Leslie. Also a theory theorist and a proponent of the modularity approach, Leslie was the first to hypothesize that ToM impairment or deficiency could be responsible for the triad of impairments inherent in ASDs (See Leslie, 1987, 1994). Leslie’s work was the initial basis of a neuropsychological perspective on ASDs and various researchers, including Baron-Cohen, have since attempted to prove the neural basis of ASDs. The idea was born that the ToM capacity
might be modular in a Fodorian sense (see Fodor, 1983). A decade after introducing his modularity of mind, Fodor also asserted that some kind of modularity underlies our ToM ability (see Fodor, 1992). These assertions have (at least) this in common: that these modules are biological structures that are innate.

Baron-Cohen somewhat follows Leslie with his introduction of IMMT. Many philosophers (e.g. Goldman, Gordon, Harris, Heal,) who work on ToM questions will line up to argue against models that assert the concept of ToM as an innate modularity. However Baron-Cohen thinks he can escape these arguments. His theory is “minimalist” because it presupposes only a “general learning mechanism” that is biologically hardwired at birth (Baron-Cohen, 1989). He does not argue for strong nativist modularity, but for “minimalist innate modularity.” He says, “Rather than having to postulate ToM coming fully prepackaged as an innate module, this minimalist alternative specifies less that is innate – but still specifies some innate social-information gathering mechanisms” (Baron-Cohen, 2006). The initial structures are biologically present, but there is, for Baron-Cohen, nothing innate in cognition itself. Thus, Baron-Cohen claims that we come hardwired with four discrete lower level perceptual mechanisms (ID, EDD, SAM, and ToMM), which provide critical inputs to developing a ToM, but the actual ability to mindread is derived from the individual functioning in and learning from her environment (Baron-Cohen, 1998).

As a brief aside, the fact that Baron-Cohen hypothesizes these four perceptual mechanisms as part and parcel of his IMMT, and that he surmises these to be modular in nature, raises a very interesting problem for the debate between the theory theorists and the simulation theorists with

—

73 Fodor favors a dedicated cognitive mechanism or module - a single, innate, domain-specific, informationally encapsulated cognitive tool with fixed neural architecture - that is marshaled for each cognitive task such as perception, language, and so on.
respect to the question of that which underpins the ToM ability. Recall that the debate is over a rigid dichotomy between the theory theorist claim that the ToM ability is underpinned by some form of theoreticity versus the simulation theorist claim that it is underpinned by some simulative capacity. Because the paradigmatic cases of modules are cases of perception, Baron-Cohen’s IMMT perhaps exposes the TT/ST debate as a false dichotomy by raising the interesting possibility of a third option: that perhaps the ToM ability is underpinned by neither theoreticity nor simulative capacity but, rather, by perception. For the purposes of this dissertation, I do not introduce or address this problem or third option. I acknowledge this as a potential problem and an interesting possibility but, instead, I seek only to adjudicate the TT/ST debate, since many authors say that the TT and the ST are “the only two games in town,” and set up the debate in such a way that the TT and the ST exhaust the possibilities (See Stitch and Nichols, 1992; Goldman, 1995, 2002; and Davies and Stone, 2001). Thus, for now I will continue the discussion of the child scientist view of the TT and how Baron-Cohen’s IMMT bears on it, saving for later the discussion of the perceptual view of the mechanism underlying ToM abilities.

Earlier, following Goldman, I argued against Gopnik’s child-scientist account of TT, stating that, because children are said to converge on one and the same theory in the same order and all within the same narrow time-course, they cannot be learning in the way that scientists learn. For, it would be something like a miracle for scientist to converge on the same theory, in the same order, conducting the same experiments, in the same narrow time frame with no interaction or introspection.

In a sense, if correct, Baron-Cohen undermines this objection to Gopnik. If Baron-Cohen’s claim (that “ToM acts as a collection of minimally innate independent modules of the mind - and
so has a specific neural basis somewhere in the brain”) is correct, then this might explain how (or why) children come to the same theory, in the same time frame, and in the same order: He can simply say that the four components that comprise ToM are hardwired, setting the stage for the individual to process information from the environment in order to develop a ToM. So, minimally innate lower level cognitive perceptual mechanisms combined with environmental data collection should result in a mature ToM in the typically developing individual. Baron-Cohen clearly says, “Mindreading may have some innate component, … but this in no way excludes the important role of experience being necessary in the typical development of mindreading” (Baron-Cohen, 2006).

On the other hand, even if Baron-Cohen is right in his claims about these modules, he does not escape a different objection that I made against Gopnik’s account. Specifically, even on Baron-Cohen’s IMMT account, it remains entirely plausible that some other cognitive ability or mechanism (such as language development, empathy, episodic memory, executive function) are at play or at interplay in the development of a full-fledged mature ToM.

Finally, it seems that there is a problem with claiming that ID, EDD, SAM, and ToMM are even minimally modular – at least, if we are to subscribe to Fodor’s definition of modularity. Fodor claims that a modular system must - “at least to some interesting extent” - fulfill at least some of the following certain properties:

- Domain-specificity, modules only operate on certain kinds of inputs—they are specialized
- Informational encapsulation, modules need not refer to other psychological systems in order to operate
- Obligatory firing, modules process in a mandatory manner

The claim that localization follows from innateness is suspect and therefore debatable.
• Fast speed, probably due to the fact that they are encapsulated (thereby needing only to consult a restricted database) and mandatory (time need not be wasted in determining whether or not to process incoming input)

• Shallow outputs, the output of modules is very simple

• Limited accessibility

• Characteristic ontogeny there is a regularity of development

• Fixed neural architecture (Fodor, 1983).

Baron-Cohen’s IMMT hinges, in part, on a conception of ToM that is modular so modularity is crucial to his account. However, it seems to me that ToM meets some, but not all, of the criteria for modularity. First, there is a question as to whether ToM functions are domain-specific or domain-general and prima facie, it seems that, on Baron-Cohen’s description, ToM operates on multiple kinds of inputs (as does language processing one of Fodor’s paradigmatic cases of modules). Secondly, clearly ToM cannot be informationally encapsulated. ToM must marshal and refer to other cognitive abilities and psychological mechanisms to operate. At a minimum, ToM must call upon other cognitive developmental capacities such as language, memory, and empathy. Thirdly, the outputs of ToM are in no way “shallow” Rather, mentalizing is comprised of highly complex first-person and third-person inferences, largely based on inductive reasoning and entailing multiple levels of observation, perception, and (in some loose sense) inference. Finally, it seems that ToM does not fit the description of being a single, highly specific cognitive tool that is marshaled for a single cognitive task and serves no other function. “If this constraint cannot be met, then the appearance of modularity as an underlying cause of a given pathology is merely virtual. In both autism and schizophrenia, the apparent ToM deficits are a consequence of a disinfused array of cognitive and (possibly) non-cognitive malfunctions” (Gerrans & McGeer, 2003, p. 273).
Despite these objections, IMMT – as Baron-Cohen has cast it – has an appealing advantage over Gopnik’s child-scientist account. IMMT can explain many – though not all, of the common diagnosable deficits that we see in individuals with ASD. This gives rise to Baron-Cohen’s MTA account.

**MTA:**

As stated, the MTA has the appeal of accounting for many of the diagnosable aspects of ASD. IMMT and its extension to MTA can account for the social, imaginative, and communicative impairments seen in ASDs. Deficits in social-emotional reciprocity, failure to engage in back and forth conversation, abnormalities in eye contact, body language and use of non-verbal communication, deficits in developing peer relationships, difficulties in adapting behavior to suit different social contexts, and difficulties in sharing imaginative play can *all* appeal to IMMT and MTA for an explanation.

It is fairly straightforward to see how a fractured or impaired ToM might negatively impact development in all of these areas. For, obviously, if one has an impaired ability to read other people, impute mental states to them – that is, if one has an impaired ability to mentalize or mindread - one would exhibit great difficulties in engaging in typical conversation and normative social reciprocity. Additionally, it is observed that those with ASD tend toward literal-mindedness – that is, they struggle to understand and show insensitivity toward metaphor, irony, sarcasm, and even idioms qua idioms. To someone with ASD, “She has a heart of gold” means, literally, “the composition of her heart is a yellow precious metal called gold (Au)”. There would be little or no understanding without ToM that others intend to convey by their words something other than or more than just what their words mean (Gerrans & McGeer, 2003, p. 281).
As appealing as Baron-Cohen’s IMMT and its extension to the MTA are, these theories are not without their difficulties and shortcomings. One difficulty in particular for the IMMT and MTA is that they cannot account for the non-social diagnostic features of ASDs, such as stereotyped body movements, restricted range of interests, preoccupations with limited classes of objects or parts of objects (e.g. engine parts), or highly restricted, fixated interests that are abnormal in intensity or focus (e.g. as amassing facts about meteorology). It seems that ToM deficits cannot be solely responsible for ASD pathology since ToM deficits seem unrelated to these features of ASDs. Thus, ToM must either operate in conjunction with some other module of the mind (as the language module does for Fodor), or the presence of ASD cannot be attributed solely to a ToM deficit or impairment. There must be an additional deficit in an additional module – whether language, memory, or executive function – thereby debunking the theory that ToM is modular at all – or at least, it is not modular by the criteria set forth by Fodor. Included in the very definition of modularity, according to Fodor, is that to be a module it must be informationally encapsulated. But, clearly, ToM is not informationally encapsulated and must interplay with other modules of the mind to operate.

Another objection to and limitation of the MTA is that there exists a whole range of clinically diagnosable conditions show some level of mindblindness. Schizophrenia, narcissistic personality disorder, and borderline personality disorder all bare elements of mindblindness. Accordingly, mindblindness deficits can be acquired later in life. For example, schizophrenia (which was once called the “late onset of autism”) is seen as a form of mindblindness. The schizophrenic’s paranoid and ill-formed intuitions that “someone is out to get me” are indicative that she has lost the ability to accurately mindread and the ability to understand and predict the mental states of others. Also in the case of traumatic brain injury (henceforth, ‘TBI’), depending
on the region of the brain that was injured, the patient may exhibit mindblindness indicative of a loss of the ToM skill. In these examples (schizophrenia and TBI), the individuals had a typically developed ToM, and the ToM deficit was acquired. These various pathologies show that mindblindness may not be specific to ASDs. Baron-Cohen concedes this point and acknowledges, “obviously, we want a mechanism that specific to ASDs, and that we do not have that yet” (Baron-Cohen, 2008 p. 61).

Moreover, some argue that the ToM theory that can account for cases of ASD, schizophrenia, and TBI just fail. “There is no single ToM module responsible for successful social reasoning and behavior in normal subjects; hence, there is no dedicated module, realized in neural substrate, that fails to develop in ASD or breaks down in schizophrenia. In other words, we argue that the ToM theory fails at the first stage” (Gerrans & McGeer, 2003, p. 273).

It may just be the case that developmental disorders, such as ASD, may not even help us to understand and answer questions about modularity or vice versa. Complex systems (such as minds) break in complex ways and there are many ways to break a complex system, any of which could lead to distorted mindreading. Baron-Cohen appears to address these limitations of IMMT and MTA via his more recent theory, the Empathizing–Systemizing Theory and its extension to the Extreme Male Brain Theory of Autism.

**E-S and its Extension to EMB:**

EMB, as a supplement to IMMT and MTA, has some theoretical appeal. In particular, EMB works to answer some more of the questions that we have about individuals with ASD, as well as those with interests and talents in STEM disciplines. Examples of such questions might be:

---

75 Baron-Cohen has not addressed issues of late onset ToM deficit comorbid to schizophrenia or TBI.
“Why is ASD epidemiologically more prevalent in boys than it is in girls?” and “Why is that typically more boys and men naturally migrate toward STEM disciplines and professions, while girls and women tend not to do so?” and “Why does it seem like high-functioning individuals with ASD also tend to migrate toward STEM disciplines and professions?” Another virtue of EMB is that it can account for the social impairments associated with ASD and some (though not all) of the non-social features that are typical of ASD such as: unusually narrow interests and highly repetitive behaviors, also called 'resistance to change or need for sameness.'

However, as a criticism, EMB does no better than its predecessors (IMMT and MTA) in accounting for some other (e.g. non-social) aspects of ASD. Helen Tager-Flusberg notes this claiming, for example, that EMB does not address or explain “the many neurological features of the disorder, like the motor symptoms [such as repetitive movements and clumsiness], the sleep problems or the seizures” (See McGough Wall Street Journal. 16 Jul 2003, p. B1). To this, I add that E-S and EMB also cannot account for some behaviors described in the DSM-V, which are specific to many individuals with ASD such as: head banging, extreme sensory sensitivity to certain sounds, lights, and textures, or extreme food aversions.76

One aspect of Baron-Cohen’s E-S and its extension to EMB that is relevant (for my purposes) is that Baron-Cohen, in recent years and while asserting and defending these theories, seems to be moving away from his allegiance to the TT, and moving toward becoming an empathy

76 Additionally, EMB has been widely criticized by feminists for some of its implications, which are noteworthy, though not relevant to this dissertation. These are: (1) EMB implies neural biological differences between genders. (2) EMB may have negative consequences: First, that girls with ASD might be misdiagnosed because EMB asks us to revert to a time when autism was construed as a disorder found in males alone. Second, E-S and EMB might endorse more males entering STEM disciplines and professions than females, further exacerbating already male-biased fields of study and work, creating further inequality for females. Third, and finally, Tager-Flusberg worries that the EMB will be misconstrued in such a way that the ‘maleness’ of EMB is confused with and misinterpreted as ‘male aggression’. It is dangerous, she says, that people will draw an erroneous inference and, as a result, believe that people with ASD are overly aggressive, which (generally speaking) does not seem to be the case (McGough, 2003).
theorist of mindreading – that is, a simulation theorist (Goldman, 2006, pg. 201). He most recently says that there are components of ToM that are attributable to experience, and even to simulation. He now says, “mindreading involves putting oneself in someone else’s shoes, imagining the world through someone else’s eyes” (Baron-Cohen, 2003, pg. 137). In his most recent writing, he strenuously reiterates the need for experience in conjunction with the innate neural structures to feed into a ToM. He says that minimalist modules are in place, later allowing for the development of ToM but require the relevant kind of input for ToM to function normally (Baron-Cohen, 2006). 77

Another virtue of Baron-Cohen’s EMB theory relevant to this dissertation is that it dovetails nicely with his MTA theory. At long last, (and after a long detour) I shall now make good on my debt to you, as I am now finally in a position to elucidate how it is that these two theories, when taken together, serve to drastically undermine Gopnik’s child-scientist version of the TT and, ultimately, to undermine the TT in toto.

4.4. End Detour: How Baron-Cohen’s Work Undermines Gopnik’s Child-Scientist Theory

Baron-Cohen’s ES theory and its extension to EMB are indubitably interesting and relevant for any discourse pertaining to causes of ASD. Prima facie, these theories have seemingly little to do with the ToM capacity or the debate between the TT and the ST. To the contrary, Baron-Cohen’s work is significant for this dissertation because he provides the material for my novel argument against Gopnik. In this section of this chapter I intend to show that, when taken

---

77 I will later argue that the relevant kind of input best comes from simulation and, in some sense, non-idiosyncratic introspection.
together, Baron-Cohen’s MTA and EMB serve to undermine Gopnik’s child-scientist version of the TT and that, as a consequence, the TT collapses altogether. Let me explain.

**Assume as Premise 1:** Recall two essential claims in support of the TT by theory theorists: (1) ToM is the mechanism through which we are able to mindread (i.e. mentalize, or attribute mental states to others and to ourselves); and (2) this ToM activity is only accomplished *via* theorizing on the part of the perceiver (i.e. proficient ToM is an act of theoreticity or theoretical prowess on the part of the perceiver). It is strictly the ability to theorize that underwrites the ToM capability.

**Assume as Premise 2:** Recall that Gopnik employs the two essential claims of the TT (assumed as Premise 1 above) to ‘prove’ her case that “children just are miniature scientists.” Though, as stated, she never launches any concise argument to support this claim, I can glean an argument from her many publications that would proceed such as this:

1. Children learn about the world and mental states *via* the same process that scientists use to learn about science (i.e. theories).
2. Scientists theorize to learn about science.
3. Children theorize to learn about the world and mental states.
4. Both 2 & 3 are supported by the same *domain-general* psychological structure. Therefore, children are miniature-scientists (Or, Children = Scientists).

**Assume as Premise 3:** Baron-Cohen’s IMMT and its extension to MTA make additional assertions, which – since aligns himself with theory theorists - I add to the claims of proponents of the TT and to Gopnik’s account. Here, Baron-Cohen (following Leslie, 1993; and Frith, 1991, 1994, 1999, 2003) says that when ASD is present, ToM is absent or impaired. With this theory,

---

78 See this chapter, Section 4.1.
79 See this chapter, Section 4.2.
80 I must be careful here. Baron-Cohen does not claim that every time ToM is impaired or absent that ASD results. (There are many conditions that result in a late-onset impaired ToM, but do not result in ASD.) Baron-Cohen only claims that when ASD is present, ToM is absent or impaired.
Baron-Cohen has created a link between ToM and ASD, which is crucial to my argument against TT in this dissertation. 81

Assume as Premise 4: Baron-Cohen’s most recent E-S theory and its extension to EMB make clear his assertion that ASD should also be cast as EMB (ASD = EMB). He accomplishes this by first asserting that males typically have a greater ability and interest in systemizing (than in empathizing) than females (Brain Type S>E). Concurrently, females have a greater ability and interest in empathizing (than in systemizing) than males (Brain Type E>S). ASD, he says, is the equivalent of an Extreme Type S Brain (S>>E), which results when an individual’s ability or interest to systemize is above average, but whose ability or interest to empathize is below average (e.g. as is found in individuals with ASD). 82

Assume as Premise 5: Baron-Cohen’s says his E-S theory is a much better predictor, than is gender, of who will select STEM disciplines and professions. He claims an individual with a Brain Type S>E is more likely to select STEM than either a Type B Brain (which is balanced between systemizing and empathizing) or a Type E>S Brain (which prefers to empathize over systemize). Moreover, with Baron-Cohen’s extension of E-S to EMB, he claims that people with ASD are far more likely to select and be proficient in the STEM disciplines and professions because they are Brain Type S>>E. That is, individuals with ASD should experience success in STEM disciplines and professions that require systemizing in the ordinary course of business, and they should prove to be poor empathizers (as would anyone with an impaired, deficient, or

81 See this chapter, Section 4.3.4.
82 See this chapter, Section 4.3.3.
altogether absent ToM). Here, Baron-Cohen has created a link between ASD and proficiency in STEM.

This is a rather unexpected result. According to the TT and Gopnik, ToM is accomplished via theorizing, and theorizing is also the business of a scientist. Furthermore, theorizing for mindreading and scientific theorizing are each is supported by the same domain-general psychological structure. So, on Gopnik’s child-scientist account and the TT, I would expect the conclusion to say something like, ‘Individuals with ASD have an impaired ToM; so, individuals with ASD cannot mindread (i.e. mentalize); so (because theorizing underwrites the ToM and people with ASD have an impaired or absent ToM) it must be the case that individuals with ASD cannot theorize well in other domains that are also supported by the same underlying psychological structure.’ In other words, on Gopnik’s child-scientist account, we should expect that people with ASD (who cannot theorize or mindread) would be utterly incapable of performing the duties of science, which very much include theorizing, because they are purportedly supported by the same domain-general mechanism.

Quite to the contrary, on Baron-Cohen’s account, individuals with ASD have an impaired or absent ToM and cannot mindread and nevertheless they migrate toward and succeed at STEM disciplines and professions. That is, people with ASD excel in areas – such as science – which absolutely require the ability to theorize. And, yet, people with ASD cannot mindread (are “mindblind” in Baron-Cohen’s terms). Furthermore, we have plenty of anecdotal examples of people with ASD who are excellent theorizers. I mention here only one (well-known) example. Dr. Temple Grandin is autistic and she is great at theorizing about certain things (especially

83 Note: this claim is born out in statistical evidence, in which we see an unusually high percentage of persons with high functioning ASD in STEM disciplines and professions.
science) and yet she cannot mindread. She reports that she is *cognizant of* (and can recite succinctly) complicated psychological theories and yet she reports extreme difficulty explaining and predicting the behavior of others. This suggests that having such (presumably psychological) theories is not sufficient for mentalizing.

Given all of the above, it is clear to me that Gopnik’s child-scientist version of the TT cannot go through because – all of the above – at least suggests that there must be some other way (*apart from theorizing*) that underwrites our ability to execute mental state attributions.84

### 4.5. Concluding Remarks

In this chapter, I have outlined the TT in general and discussed two major proponents of it: Alison Gopnik and Simon Baron-Cohen. Also, I have given an exposition of their various theories including Gopnik’s child-scientist theory and Baron-Cohen’s IMMT which he extends into MTA and his more recent E-S which he extends to his EMB theory of autism. All of these theories, except perhaps the EMB theory, have been shown to face some fairly serious objections.

I have also discussed the fact that many theory theorists (including Baron-Cohen) have, in an attempt to support the TT, turned their attention to work concerning ASD. The claims are these:

1. People with ASD are mindblind and, as such, cannot understand, explain, or predict mental states. ToM is the instrument used in understanding both other’s mental states and our own. There is no difference between understanding the mental states of other and understanding our own mental states. Theoreticity underwrites both first-person and third-person mental state

---

84 Here, as everywhere throughout this dissertation, I take the terms ‘mindread,’ ‘mentalize,’ and ‘impute mental states’ as synonymous terms.
attributions. (2) ToM is a discrete domain specific module with a neural basis in the brain and is dedicated solely to the functions of ToM without reliance on or interplay with other modules. (3) The psychological structure that underwrites the ToM ability is a domain-general learning mechanism that is used across many different content domains including and especially areas that require theorizing (such as the sciences). (4) Since understanding mental states requires ToM and, according to theory theorists, this requires the ability to theorize, it should be the case that people with ASD, who are mindblind, are not able to theorize in other capacities that are also supported by the same underlying domain-general learning mechanism or psychological structure.

However, in my view, work on ASD has only proved to undermine, rather than bolster, the claims of the TT and the child-scientist account of TT. If we take seriously Gopnik’s account that theorizing underpins ToM and Baron-Cohen’s assertions concerning the relevancy of ASD to the ToM debate, then on Gopnik’s account, we would expect that people with ASD would be incapable of systemizing and theorizing in the scientific sense because they are presumed to be supported by the same domain-general learning mechanism. Here emerges a difficulty for Gopnik’s account because the evidence is clear that people with ASD can theorize and theorize exceptionally well. In fact, according to Baron-Cohen, ASD is a result of excessive fetal testosterone - which results in an autistic mind: a mind that is works exceptionally well with systems, but fails to be empathetic. He then cites multiple studies of people who have entered STEM disciplines and professions who could be considered autistic – that is, people with ASD do best in environments where they work with systems and so must theorize.

Specifically, if the TT and Gopnik have the right account (that mentalizing requires the employment of successively well-developed theories and that those who cannot mentalize cannot
do so because they are incapable of theorizing), then it would seem to follow that people with ASD (who possess a ToM deficit) would also be unable to formulate other, more science-like, theories. But taking cases such as Dr. Grandin and others into consideration, this is clearly not the case. If this is all correct, then the TT must be wrong. These considerations, taken together, strongly suggest that it is not theorizing– whether implicit or tacit – that underpins ToM.

Taking Baron-Cohen’s MTA together with his E-S and EMB theories, results in an objection that completely undermines Gopnik’s child-scientist account. Plainly stated, children can’t be miniature scientists. Also, Baron-Cohen’s theories overwhelm the TT with evidence that serve to undermine the TT in toto since his work shows that some other ability must underwrite ToM proficiency. It is my assertion in this dissertation that the work on ASD has served to undermine the TT. Work on the special case of ASD serves as a sort of confirmation for a competing theory: The Simulation Theory of Theory of Mind.
CHAPTER 5: THE SIMULATION THEORY ACCOUNT OF THEORY OF MIND

This chapter will explore and argue in favor *The Simulation Theory of Mind* (henceforth, the ‘ST’). In the first section, I will give a brief general description of the ST, including its history and development, and I will mention some views that are common to all variants of the ST. I then will give a more detailed account of the ST, by reflecting on its main tenets as discussed by the two simulationists whom I think give the most interesting and plausible accounts of the theory. Specifically, in the second section of the chapter, I will discuss the ST as formulated by its originator, Robert Gordon. In the third section of the chapter, I will discuss a formulation of the ST by simulationist Alvin Goldman, has recast and modified the ST in various important ways that cause me to support his account of the ST over Gordon’s account of the ST. After the discussions of Gordon and Goldman’s respective accounts of the ST, in the fourth section of the chapter, I will first discuss a common criticism of the ST by theory theorists: that the ST simply ‘collapses’ into the TT. Here I will elucidate three different versions of the collapse argument and the responses that simulationists have to each. In the fifth section of the chapter, and before concluding the chapter, I will discuss the potential for failed or ‘misfired’ attempts at mindreading *via* failure to quarantine one’s own idiosyncratic or egocentric mental states and the crucial role that hyper-introspectionism plays in quarantine proficiency. Lastly, I will discuss egocentrism as evidence for autistic hyper-introspectionism, and whether making a bad copy or using a bad model can still be construed as *making* a copy or *using* a model. These sections, when taken together, work to establish my views that the ST is currently the most preferable explanation of ToM, and that Goldman’s version of the ST is currently the most preferable version of the ST.
5.1. A General Discussion of the Simulation Theory of Theory of Mind

In the fields of cognitive science, cognitive psychology, developmental psychology, and philosophy, the TT was the reigning orthodoxy as an explanation of mindreading for approximately two decades. As seen in the preceding chapter, many scholars and scientists in these fields continue to assert that the TT still occupies this lofty status (See Gopnik; Meltzoff; and Perner). However, in the mid 1980s, the ST was developed (See Gordon, 1986; Heal, 1986; and Goldman, 1989), which offered an alternative to the TT orthodoxy. The TT and the ST are best viewed as competing families of theories, each such theory offering a different account of that which underpins our everyday folk psychology – in particular, our ability to mentalize. The two camps clearly disagree on what process or ability underlies and supports the ToM capacity. However, they do concur regarding the function of the ToM in general: that it allows us to mindread.

As seen in the TT chapter of this dissertation, theory-theorists claim that we mindread via employment of a theory, which grows successively more sophisticated with cognitive development. The ST is usually, but not always, taken to present a serious challenge to the assumption by theory-theorists that theoretical achievement underlies our everyday folk psychological competence regarding the mental states of others. Instead, the ST provides a different answer to the question, “How do we mindread?” Where the theory-theorists claim that we mindread via the employment of a theory, simulation theorists claim that we accomplish this by engaging in a form of role taking, which they call simulating.

As I explain at the end of the chapter, many proponents of the TT do not believe that the ST presents any serious challenge to the assumptions of theory theorists. In fact, many theory theorists claim that the ST simply “collapses” into the TT – a claim that others and I refute.
Like the TT, there are many versions of the ST put forth by various proponents. Also, like the TT, the ST is best seen as a *family* of theories (or an umbrella term) used to refer to different theories, which are tied together by sharing at least one common view. In this case, the shared view is one that is contrary to the assertions of the TT - that ascribing mental states to others does not proceed by the development of successively more sophisticated theoretical inferences, but rather that the mentalizing capacity is a *capacity to simulate* the mental processes of others. Generally speaking, simulationists deny that our understanding of one another primarily proceeds by the development of a theory, whether folk or otherwise. Instead, simulation theorists claim that we accomplish the mindreading task by using our own mental apparatus and (in some cases) our own experience of introspection as a simulation of another’s mental states. They claim that human beings are able to predict and explain each other’s actions by using the resources of their own minds to simulate the mental states that cause the actions of others. So instead of being theorizers, as the theory theorists assert, the simulation theorists assert that we are simulators. “We are mental simulators, not in the sense that we merely simulate mentation, but in the sense that we understand others by using our own mentation in a process of simulation” (Davies, 1994, p. 194).

5.1.2. Views Shared by Simulationists

In the remainder of this section, I will summarize what I take to be the five common threads found ubiquitously throughout the various accounts and descriptions of what the ST is and how simulation works to underpin accurate mindreading.

---

86 Simulation theorists will differ in their opinions regarding the necessity for introspection. For example, as I will show, Robert Gordon claims that no introspection is required for simulation. Conversely, Alvin Goldman claims that introspection is required for simulation (See Gordon 1995a, 1996a, 2007; Goldman 2006; and Goldman & Shanton 2010).
First, all simulationists share the view that mindreading is an ability or skill which absolutely does not derive from theoretical activity and is not, primarily, a kind of knowledge, although the ability or skill may yield knowledge in some cases. Simulation is seen as a kind of *ability or skill set* rather than a kind of *knowledge*. So, whereas theory theorists explain ToM development as theory-reduction and refinement (i.e. changes in acquisitions of and knowledge regarding mental state concepts) over a period of time, simulation theorists explain development as a refinement or honing of an actual *skill*. Developmentally speaking, children become increasingly better at mentalizing as they “gradually become more adept at imaginatively identifying with other people and at imagining counterfactual situations” (Davies and Stone, 1995). At this juncture, one might reasonably ask, “What kind of a skill? Skillful *at what*?” This leads to the second common thread among simulationists.

Most simulationists (such as Robert Gordon, Alvin Goldman, and Jane Heal) concur that to be a skillful simulator is to be adept at a kind of mental modeling in which the simulator uses her own mind as an analog model of another person. Simulationists call this other person “the simulated target” or just “the target.” According to the ST, I come to understand the minds of others by “imaginatively projecting” myself into their role. I use my mind as a model for theirs. My access to the minds of others is *via* empathetic contact with others. Rebecca Saxe says, “understanding someone else’s actions seems like a highly abstract – if not semi-miraculous entertainment-education-achievement. Simulation theories offer a demystification of the process: Knowledge of others is parasitic on our direct access to ourselves” (Saxe, 2009, p. 257). Thus,

---

87 Note: Different simulationists differ regarding the degree to which we do this. In particular, Gordon says that we can simulate without the analogical inference ‘from me to you’. See section 5.2.3 of this dissertation for a more detailed account of Gordon’s view.

88 I take it that, for the most part, I have direct (or ‘privileged’) access to my own mind – that is, I am in a better position to know the contents of my thoughts and mental states than anyone else. By “access to the minds of others,” I mean, “having or developing an understanding of the thoughts and mental states of others.”
the particular skill involved in proficient simulating is becoming adept at using one’s own mind as a model used for understanding the minds of others sufficiently enough to attribute mental states to them accurately. This may raise a second question, “But exactly how does one accomplish this adeptness?” This brings me to the third common thread among simulationists.

Thirdly, simulationists widely concur that effective simulation involves some kind of heuristic of one projecting oneself into the position of another person so as to attribute mental states to that person, predicting that person’s behavior, and planning an appropriate response. Daniel Dennett describes what he calls ‘the intentional stance.’ Although it would not be fair to consider Dennett a simulationist, his description of the intentional stance is particularly apt. He says, “When we decide to interpret an entity from the intentional stance (which is the key to unraveling the mysteries of other minds), it is as if we put ourselves in the role of its guardian, asking ourselves, in effect, “If I were in this organism’s predicament, what would I do?” (Dennett, 1996, p. 33). When we do this (viz., take the intentional stance), we are able to perceive, interpret, predict, and respond appropriately. Dennett’s claim is interesting because it is illustrative of how we (seemingly reflexively) tend to think about other people’s mental states: we ask ourselves, “What would I do, if I were that person?”

The ST suggests that we determine this (this: “what I would do if I were that person”) not through the use of some theory-based folk psychology, but, rather, we use our own mental apparatus to form predictions and explanations of someone by “putting ourselves in the shoes” of another person and simulating that person. When we try to figure out another person’s reaction to an event, we imagine ourselves in that person’s situation.89 We try to match up an experience

89 Simulation theorist Robert Gordon claims that one actually tries to imagine oneself as the other person so that one experiences a total transformation to the other person and not a mere transference of oneself (See Gordon 1995, 1995b, 1996).
that we have had to the experience that the other person is currently having. On this view, human beings are able to use their own cognitive resources to simulate the mental states underpinning the behavior of others, typically by making decisions within a “pretend” context such as role-taking or “putting oneself in the other’s place.” The ST asserts that we utilize our own motivational and emotional resources, together with our practical reasoning skills, so as to identify with another in imagination to replicate or reenact their mental life. We attempt to replicate their thinking, decision-making, beliefs, goal-directedness, emotional responses, desires, thoughts, and emotions.

Simulationists view this as a skill because our capacity for psychological understanding depends on our ability to run these cognitive simulations. It is possible to infer other people’s intentions and future actions by honing our ability to use our own mental states as models for theirs. Simulation doesn’t involve a complex theory of other people’s minds; it merely involves a capacity for pretense and for putting oneself in the other’s place and, the more skillfully we do this, the more accurately we are able to mindread.

Fourth, the main proponents of the ST share a standard explanation of approximately how simulational mindreading works. They claim that there are four essential steps involved in simulational mindreading. First, a mindreader constructs a pretend mental state intended to match, as closely as possible, the initial states of the target. Second, these pretend mental states (such as desires or beliefs), are fed into a cognitive system, such as a decision-making system, which operates on the pretend inputs in much the same way it would operate on any non-pretend (i.e. first-person) input. Third, the operation generates an output mental-state that is a token of the same general kind as that of the target (e.g. a decision). Fourth and finally, the mindreader “reads” this generated token state – that is, classifies it in terms of its type, content, strength, etc.
– and attributes it to the target. The first three stages together comprise the simulation process. The fourth stage exploits the output element of the simulation process and executes a mental-state attribution that is reasonably described as a kind of projection. In short, according to the ST, an attributor creates pretend states intended to correspond to those of her target, feeds them into her own cognitive equipment, and lets it produce an output state, e.g., a belief, decision, or emotion (Goldman, 2009a). If the output state closely resembles the target’s actual state, then the mental state attribution to the target is deemed accurate.

Related to this “output state,” there is a fifth thread that all simulation theorists hold in common, which is extremely important. Simulation theorists say that, in order to simulate, we must “run our own cognitive systems offline.” (No simulationist asserts that we do not need to do this, for failure to do this would not result in simulation at all but rather in actual behavioral output. When we run our thought process online, we run through a series of mental processes and states, allowing them to produce a given behavior. By contrast, when we run our cognitive systems offline, we run through the same series of mental processes and states, but we exercise restraint to prevent those mental processes and states from resulting in behavioral output.) We run our own mental states in an offline mode in order to simulate the mental processes of a given target. As a result of running our own systems offline, we are able to predict the target’s decisions or actions. Importantly, the “output state” is not actual, but imaginatively projected onto the target, so that there is no actual behavioral output at the tail end of running the processes. The mental mechanisms responsible for producing actions that result from mental states are disengaged in such a way so that no ‘decision-to-behave’ is translated to actual behavior. So, there is no output exhibited in any actual behavior of the simulator. Instead, simulation theorists say, because the simulator’s systems are being run offline, the only output of
the simulation process is a “pretend output.” Thus the only output is that of taking the result of
the thought process and imaginatively projecting it onto the target by assuming that (as a result
of the thought process) the simulator now has an understanding of the mental states of the target
and can accurately attribute mental states to the target.

Simulation is often conceived of as using one’s own behavior control system as a manipulable
model of other such systems. The system is first taken offline, so that the output is not actual
behavior. Inputs and system parameters are accordingly not limited to those that would regulate
one’s own behavior, but are extended to include those that would regulate the target’s behavior.90
We simply run our decision process offline and pretend to be in the shoes of our target to see
how our mind would resonate as if we were in the pretend context. We use ourselves as a model
to simulate their predicament and predict their behavior based on what we think we would do if
similarly situated as them. “An observer can understand someone else’s action using the same
cognitive and neural mechanisms that he or she uses to produce his or her own – that is, buy
running his or her action execution system in a “simulation” mode” (Saxe, 2009, p. 257).

The processing capacity involved here is taking a propositional attitude - that is, pretending or
believing something while temporarily suspending one’s own normally active processes that
check the truth and falsehood of states of affairs. “You can pretend something that is not actually
the case, and, likewise, you can believe something that is not true” (Frith, 1991). But, again, the
crucial element here is that systems of the simulator are run offline so as to avoid any actual
behavioral output and to result only in appropriate mental state attributions to the target.

90 In this case, the term “system parameters” refers to the parameters of the cognitive system of the simulator –
parameters that are largely dependent on the experiences of the simulator. For accurate and successful third-person
mental state attribution, these parameters must be adjusted (or altogether quarantined), so that the processes of the
simulator reflect the system parameters of the target and not the simulator.
To summarize, like the TT, there are many proponents of the ST, each with slightly different versions of the ST. (I will elucidate some of these differences in the remaining sections of this chapter.) There are some elements, however, that all simulationists hold in common. These are: (1) Mindreading is best seen as a skill and not as a kind of knowledge; (2) The skill in question is that of using one’s own mental states as models for those of a target; (3) This modeling is done via a heuristic of projecting oneself in imagination into the shoes of a target; (4) The projection of one’s self into the shoes of another is accomplished via a multi-step process such as described above; (5) Most importantly, simulation requires the capacity (i.e. skill) to run one’s own cognitive systems “offline” so as not to produce any actual behavior outputs on the part of the perceiver, but only to attribute mental states to the desired target.

Before delving into specific accounts of the ST, a word needs to be said regarding perception and simulation. Certainly, one could ask, “How is simulation different than perception?” or even, “Isn’t simulation just perception [or social perception]?” These are reasonable questions to ask since, prima facie, simulation appears to rely heavily on perception and because Baron-Cohen’s IMMT account seems to hint at the possibility of perception as underlying the ToM ability.

However, as you will see in the forthcoming sections, the ST is rightly treated as a theory of the personal and subpersonal processes that underlie an attributor’s direct perception of a target. Sometimes this is seen as a step-wise process that begins with perception and ends with running ones own processes in an offline mode. Indeed, simulation can be a process following perception. (As you will see, Goldman, for example, discusses high-level simulation as a controlled process following perception and low-level simulation as an automatic process at the neural level; and each of these are processes that occur at the subpersonal level following
perception.) But this is not always the case, as sometimes simulations can be run when there is no percept present to the simulator.

The tasks of mindreading and social perception are not tasks of purely perceiving. They are tasks of understanding and interpreting (and more so) explaining and predicting the behavior of others. Importantly, the ST aims to explain not only cases where the target is present to the attributor but also cases where the target is not present or available to direct perception, as in mental imagery.

Furthermore, simulation should not be considered as merely involving perception because the ST specifically posits pretend mental states and processes that resemble those of the target and reuse of the attributor’s own neural processes to simulate. It is this resemblance and reuse that are the essential features of the kind of simulation routine described by the ST. So a distinction exists between merely perceiving others (and the actions and facial expressions of others) and attributing mental states to others. It is resemblance and reuse that are the hallmarks of simulation, and not merely perception.

In the next two sections of this chapter, I will offer expositions of the two main proponents of the ST: Robert Gordon and Alvin Goldman. These expositions will also serve to make clear some different distinctions that each philosopher makes regarding kinds of simulation and their views on whether the mental states inferred and predicted in order to explain intentional behaviors are perceptual, volitional, or epistemic.
5.2. Robert Gordon’s Simulation Theory of Theory of Mind: An Alternative to the Theory Theory View of Theory of Mind

In 1986, philosopher Robert Gordon wrote an article entitled “Folk Psychology as Simulation,” criticizing the TT and introducing the notion of simulation, offering it as a better account of how human beings achieve psychological competence. Gordon observes that the TT “posits unobservable mental states such as belief, desires, intentions, and feelings, linked to each other and to observable behaviors by “law-like” principles, which are applied to observable situations via logical inferences that generate predictions and explanations of behavior” (Gordon, 1996b, p. 2). By contrast, on Gordon’s ST, one uses one’s own mental resources to simulate another person (i.e. a target) in order to arrive at an understanding of the target’s mental states and, so, to be able to explain and predict the target’s behavior on the basis of those mental states. Gordon was the first-person to put forth any sort of explanation of ToM that challenges, or is at variance with, the orthodoxy of the TT, and he is frequently considered a “radical simulationist.”

The ST he posited in his 1986 article attracted the attention of philosophers and developmental psychologists alike, with some writing in support of Gordon’s ST (e.g. Harris, Heal and Goldman) and others arguing against it to further the TT agenda (e.g. Gopnik, Perner, and Wellman). In his work, Gordon challenges the TT on several grounds and, in so doing, sketches out his version of the ST. Next, I will describe Gordon’s various challenges to the TT and the aspects of his ST.

First, Gordon challenges the TT position that all predictions of human behavior come from inferences that are underpinned by theories (about mental states such as beliefs, desires, and emotions) and that are in conjunction with a set of laws about behavior (such as “If A is having

---

91 Gordon’s 1996 article ‘Radical Simulationism’ elucidates his version of the ST. He claims his view is often viewed as ‘radical’ because he asserts a version of the ST that does not require introspection, does not require prior acquisition of mental state concepts, and does not admit of an analogical inference from the observer to the target.
mental states \(S_1, S_2, S_3, \text{ etc.}, \) and conditions \(C_1, C_2, \text{ and } C_3\) obtain, then \(A\) will (or probably will) exhibit behavior \(B.\)”) From the TT perspective, Gordon says, one has a deductive-nomological or an inductive-nomological basis for prediction, which, he says, is plainly not the case. Gordon says that first-person declarations of immediate intention are not a result of inferences drawn from theoretical premises. To illustrate the kinds of “first-person declarations of immediate intention” to which he refers, Gordon gives examples such as, “I shall now pour some coffee;” “I shall now pick up the cup;” and “I shall now drink the coffee.” Instead, he claims that all first-person predictions such as these are easily obtainable via practical reasoning as exemplified in simulation and, furthermore, they are incredibly reliable. He says, “[The success rates of these self-predictions] would be the envy of any behavioral or neurobehavioral scientist” (Gordon, 1995, p. 60). This is because our predictions of our own immediate and near immediate actions are usually a simple and accurate straightforward matter.

Second, Gordon challenges the TT on the grounds that, according to the TT view, only probable or typical effects of various mental states are posited because folk psychology does not specify a deterministic system. If this were true, he argues, then any predictions one makes about one’s own behavior would have to be qualified accordingly. So, for example, rather than saying, “I shall now pour some coffee,” I would have to say, “Typically, I would now pour some coffee.” Obviously, this is not the way we normally think of our self-predictions, nor is it the way we ordinarily speak of our self-predictions. More problematic for the TT account, Gordon says, “any action that is atypical, exceptional, or out of character for me… would defy prediction altogether, even seconds before I take the action” (Gordon, 1986, p. 161). This seems to be wrong, since it seems patently obvious that I am able to predict with great confidence and accuracy what I am about to do whether it is a typical action of mine and in-line with my character, or not. Gordon
says that we predict our imminent actions with great acuity, and that we are able to predict our atypical actions as well reliably as our any other.

Third, Gordon challenges the TT assumption that theoreticity underpins our mental state attributions to others or to ourselves. He argues, “we do not keep tabs on all of the relevant beliefs and attitudes and a fortiori we do not keep a reliable inventory of these” (Gordon, 1986, pp. 159-160). Seemingly, because any number of mental states would have to be combined with any number of conditions, inputs, and system parameters, such an inventory would be vast and its inventory unparsimonious. He further argues that even if we could know and keep tabs on all of the relative beliefs and attitudes, our resultant predictions would at best be “qualified and chancy.” One common assertion of theory theorists is that humans acquire a wide range of mental state concepts, which become increasingly sophisticated during early childhood development, and that these are referred to and used when one mentalizes. The benefit of the ST over the TT, in this case, is economy of storage. Gordon says that, on the ST account, because one uses one’s own mental resources to mentalize, one need not memorize, catalogue, and categorize all of the relevant factors involved in mentalizing.

Gordon’s ST opens the door to the very interesting possibility that predictions of behavior are not based on nomological reasoning but, rather, are based on a kind of practical reasoning (that provides the basis for the decision to do something and that the practical reasoning employed is based on certain salient facts and features). This is interesting because it allows for the idea of using simulation (qua simulated practical reasoning) as a predictive device. We can easily see that by simulating the situation-appropriate practical reasoning, we can certainly predict our own behavior in “what if” situations. If I can reason in this manner to explain and predict my behavior, it seems reasonable to expect that I could extend this capacity to reason in this manner
to also explain and predict the behavior of others (i.e. I could accurately mentalize for both first-person and third-person cases.) And, in fact, he says, we do use this sort of practical reasoning in simulation (Gordon, 1986, p. 160). For example, often when another person seeks our advice, we imaginatively put ourselves in his shoes, and we begin our advice by saying, “Well, if I were you…” Gordon says this is more than a mere figure of speech. Instead, he says,

We are actually imaginatively projecting ourselves into the other person’s problem situation. The reason, he says, is fairly obvious: We try to conceive the task as a practical one, one that calls for our action, so that we can marshal our own practical know-how or expertise, or at least our independent practical judgment. That is what we usually call on when giving practical advice (Gordon, 1995b, p. 740).

It is important to note here that although we are imaginatively projecting ourselves into the other’s situation, there are certain elements of ourselves that we must hold back (i.e. screen off) and for which we must make adjustments. Otherwise, we lose the very advantages that make our advice worthwhile in the first place: the special know-how and independent judgment.

Here are two examples of what is meant by ‘making adjustments.’ First, a friend might ask my advice what she ought to do to resolve a dispute with her partner. Suppose, in this case, I think that her partner is a bum and she ought to just dump him. However, my friend is not asking what I would do if I were she. She is asking my advice regarding what she should do. In this case, I must make the sort of adjustment that Gordon recommends. I must hold back this element of myself (my disdain for my friend’s partner) – that is, I must screen off or quarantine certain aspects of myself so that I can imagine myself as my friend and advise her as to what she ought to do – which may be to patch things up with her partner because she has feelings for him.

Here is a second example. Quite often, when we are golfing, my husband will seek my advice regarding which club he ought to use to best hit his ball closest to the pin. I must remember that my husband has a much more powerful swing than I do and can hit the ball further with a shorter
club than I could. Here again, I must quarantine the relevant or idiosyncratic details about my golf swing that would inform my decision about which club I would select for myself. In so making this adjustment, I can advise my husband as to which club he ought to select to make the best shot that he can, which will indubitably be a much shorter club than I would choose for myself. Screening off the relevant details about my golf swing allows me to adjust my advice to reflect his more powerful swing. I thus advise him to select the shorter club that is more appropriate for him.

In an effort to give a thorough exposition of Gordon’s ST, I need to explain some of the intricacies of his theory, which are found throughout his work. These intricacies consist of some distinctions and definitions that Gordon uses to clarify his view and support his version of the ST. These important notions include: 1) a distinction between (what Gordon calls) hot and cold methodology, 2) the importance of first-person versus third-person mental state attribution between which, he claims, there is little difference, 3) Gordon’s idea of transfer versus transformation and what it means to make relative adjustments and an egocentric shift, 4) the importance of making an egocentric shift to quarantine one’s own idiosyncratic mental states when simulating, and 5) the notion and importance of assent routines.

5.2.1. Hot versus Cold Methodology

In later articles, Gordon describes the capacity for practical reasoning as used in simulation as a “hot theory” or a “hot methodology,” which he contrasts with the TT (See Gordon, 1995, 1996, 1996b). He describes the TT as a “cold theory,” claiming that a “cold methodology” solely uses intellectual processes to make inferences from one set of beliefs to another, and “makes no essential use of our own capacities for emotion, motivation, and practical reasoning” (Gordon,
1996, p. 11). On the other hand, Gordon couches the ST as a “hot methodology” – that is, it does make use of one’s own motivational resources, emotional resources, and capacity for practical reasoning. It is my understanding that, according to Gordon, only a hot methodology adequately explains both first-person and third-person mental state attribution.

5.2.2. First-Person versus Third-Person Mental State Attributions

Gordon’s views regarding first-person versus third-person mental state attribution by simulating are deemed as the most radical of the simulation theorists (See Goldman, 2006).92 Most simulationist accounts depend on the ability to make inferences from oneself to others, utilizing a “like me” (or a “system like mine”) analogy. Because of this, these accounts also rely heavily on some form of introspection without which simulation fails. Alternate ST accounts hold that simulation is an analogical inference from the self to others that is largely premised on introspection and the prior acquisition of mental states concepts and require that mental state concepts be used in first-person self-ascriptions before being able to apply them during third-person mental state attribution (See Goldman, 2006). Gordon’s account of ST explicitly rejects the assumptions that simulation requires (1) the analogical inference from the self to others, (2) the need for introspection, and (3) the requirement for prior acquisition of mental state concepts for self-ascription of mental states and for subsequent third-person mental state attribution (Gordon, 1995, p. 53), because, importantly, he says that most of the arguments against the ST are dependent on the assumption that simulation requires the ability to recognize one’s own mental states via introspection. Gordon’s theory escapes the introspection and inference based

---

92 For more detail about first-person vs. third-person mental state attribution, please see the Theory of Mind chapter, this dissertation. As a brief reminder, ‘first-person mental state attribution’ refers to one attributing one's own particular mental states (including one's beliefs, desires, and sensations) to oneself. Conversely, ‘third-person mental state attribution’ refers to one attributing mental states (such as beliefs, desires, and sensations) to another person.
arguments against the ST because his version of the ST does not depend on introspection or the ability to draw inferences from the self to others.

An important question is whether it is reasonable to assume that we can extend the first-person based methodology to third-person cases. Gordon thinks we can and consistently, almost reflexively, do. In the case of first-person methodology, it is a simple matter of deciding, “what I shall do.” When I extend this to a third-person case, I must make adjustments for relevant differences [between myself and my target] (Gordon, 1986, p. 162). For Gordon, first-person statements, attributions, and predictions are not very different from those of the third-person case. He says the difference is only a matter of degree to which we employ our capacity for practical reasoning. That our first-person reports are so incredibly accurate (and seemingly rather easily attained) raises the important possibility that we might easily extend these to third-person cases. Davies and Stone query, whether “these psychological mechanisms that are used in making [first-person attributions] might be put at the service of more difficult predictive tasks [e.g. third-person mental state attributions and subsequent behavioral predictions]” (Davies & Stone, 1995, p. 16).

Gordon thinks that we can and do extend the psychological mechanisms used in our first-person attributions and predictions to third-person cases. Moreover, he indicates that we do this with regularity and accuracy, and that, again, there is scant little difference between our first-person attributions and our third-person attributions. Again, for Gordon, the difference between our first-person attributions and our third-person attributions is a matter of the degree to which I deploy my hypothetico-practical reasoning. He says,

As in the case of hypothetical self-prediction, the methodology essentially involves deciding what to do; but, extended to people with minds different than
one’s own, this is not the same as deciding *what I myself would do*. One tries to make *adjustments for the relevant differences* (Gordon, 1995, p. 63).

Gordon says many times, and I have reiterated in this section, the crucial need to make adjustments for relevant differences. It is reasonable, at this juncture, to wonder what Gordon means by this. They key to understanding Gordon’s meaning here is found in his explication of the primacy of transforming oneself into the target over the idea of merely transferring a mental state from oneself to the target, and, in so doing, running one’s own systems “offline.” I will explore this crucial difference in the following subsection.

5.2.3. Transference to the Target vs. Transformation into the Target

In a 1995 article, Gordon illustrated what he means by simulation. He calls this “the Tees and Crane illustration.” In this example, Gordon describes two men who are scheduled to depart from an airport at the same time but on different flights. The two men are sharing a taxi to the airport, which is delayed by traffic for thirty minutes. Because of their delay, they both miss their respective flights. It turns out, that Mr. Crane’s flight departed on time. So, Mr. Crane missed his flight by thirty minutes. However, upon arriving at the airport, Mr. Tees learns that his flight was substantially delayed. Mr. Tees still missed his flight, but only by a narrow margin. The question Gordon poses is, “who would be more upset about missing their flight? Mr. Crane or Mr. Tees?”

To discern which traveler will be more upset, Gordon says that he has two options - that is, there are two ways to answer this question. First, I can imagine *myself* (insofar as who I am, with all of my particular circumstances, experiences, and idiosyncrasies) missing *my* flight. I can imagine *my* reaction, and I can project my reaction onto Mr. Tees in an attempt to attribute mental states to Mr. Tees and predict his behavior. This, Gordon says, is an instance of me...
transferring myself, with my inputs and system parameters, by extrapolating the aforementioned elements of myself to Mr. Tees. This method is woefully insufficient, because my circumstances, inputs, and system parameters will likely be dramatically different than those of Mr. Tees. Hence, employment of these in simulation will not yield accurate mental state and behavior explanations and predictions that reflect those of Mr. Tees. Instead, they should only reflect my mental states, serving to explain and predict my behavior.

In the alternative, rather than transferring myself to Mr. Tees, Gordon says I can transform myself into being Mr. Tees. Gordon repeatedly asserts that his version of simulation does not involve the kind of transfer described above but, rather, a transformation of the simulator to the target (See Gordon 1995, 1995b, 1996). Gordon says that when he transforms himself to the target, he ceases to be Robert Gordon at all. In fact, even the referent of “I” is no longer Robert Gordon but, in this case, the referent of “I” would now be Mr. Tees. Gordon says that when a personal transformation into the target occurs, the simulator has no remaining tasks of mentally transferring various states from herself to the target and there are no further questions of the simulator needing to draw comparisons between herself and the target. He says,

For insofar as I have recentered my egocentric map on Mr. Tees, I am not considering what RMG would do, think, want, and feel in the situation. Within the context of the simulation, RMG is out of the picture altogether. In short, when I am simulating Mr. Tees missing his flight, I am already representing him as having been in a certain state of mind (Gordon, 1995, p. 57).

So simulation, for Gordon, does not entail pretending that one is in the situation of one’s target. Rather, the simulator simulates by taking on the traits of the target and, essentially, becomes the target in imaginative projection.

Importantly for Gordon, it is not enough to “put oneself in the target’s shoes,” as a perhaps less-radical simulationist would claim. Rather, one must simulate the other. Simulation is carried
out by taking on the target’s psychological traits. Let me illustrate this with a different example.

If I were my friend Greg, I would be at church on Sunday because Greg, being very religious, believes in the tenants of his church and, so, he generally wants and intends to go to church on Sunday. Even though I am not religious, I can predict that Greg will be in church by imagining myself, not as Susan, but as Greg. If I were Greg (qua Susan transferred into Greg’s shoes) I would not be in church because I would retain the elements of my mental states which are contrary to those of Greg’s mental states. But if I were Greg (Susan transformed as Greg), then I would take on Greg’s mental states as my own – including his beliefs, intentions, and desires with respect to his religious beliefs and practices, and I would be in church. I can thus predict that Greg will be in church on Sunday by taking on Greg’s psychological traits and mental states, and screening off my own beliefs, attitudes, intentions, and desires where religiosity is concerned. In this scenario, I simulate by imaginatively being Greg.

Gordon says that in order for this transformation to succeed, he would have to decide what mental state Mr. Tees is in or, in my example, I would have to decide Greg’s mental states to predict his behavior. Gordon would have to decide whether Mr. Tees would feel extremely upset, or perhaps he thought he missed his flight owed to some fault of the taxi driver. I would have to decide if Greg had, perhaps, some worries about pressing work that ought to be attended to, or perhaps the football game that he wanted to watch was on the television. (Knowing these would help me to know something about Greg’s mental states concerning attending church on that specific Sunday and to then predict his resultant actions.)

Here, critics can claim that Gordon’s ST requires introspection, knowledge of mental state concepts, and even intimate knowledge of the target. Gordon claims, to the contrary, that we can interpret the target’s behavior without using any kind of mental state concepts whatsoever.
Moreover, he says that simulating does not, in fact, require introspection and that the method that we ordinarily use for our own first-person mental state attributions are limited to identifying states in the first-person (Gordon, 1995a p. 53 and p. 57). However, he says, “Thanks to our capacity for imaginatively transforming ourselves into other “first-persons,” simulating is not exclusively a one-person method. Simulating is just as well suited for labeling another’s states as it is to labeling our own, provided we represent these states in the first-person and perform an egocentric shift” (Gordon 1995a, p. 57). So, for Gordon, simulation can occur without introspective access to one’s own mental states. One does not first assign a mental state to oneself and then transfer it to the target. Instead, via transformation, the simulator becomes the target, so introspective steps are unnecessary and no analogical inference from simulator to target is required. Perhaps, then, successful simulation primarily relies upon the specific element of (what Gordon calls) an egocentric shift – that is, an ability to make adjustments for the relevant differences between oneself and one’s target and, so, to “run one’s own systems offline.” I will explore these concepts in the next subsection.

5.2.4. Making Adjustments for Relevant Differences, the Egocentric Shift, and Running Our Systems Offline

There are many occasions while simulating in which one must make adjustments for relative differences between oneself and one’s target. Indeed, these adjustments must nearly always be made if simulation is to be at all successful. For example, Robert Gordon might be a relaxed and experienced traveler, while Mr. Tees might be a nervous, anxious, easily upset traveler. For Robert Gordon to accurately simulate Mr. Tees and successfully transform himself into Mr. Tees, he must first make adjustments to his simulation to allow for the fact that he and Mr. Tees
are entirely different kinds of travelers who would likely have entirely different kinds of responses and behaviors after missing a flight. Similarly, when I am simulating my friend Greg, I must adjust for the fact that I abhor religion while he has a strong religious faith and wholeheartedly enjoys religion. You see, if I do not make the requisite egocentric shift to adjust for differences between Susan and Greg, my simulation will be faulty. For example, I might predict that Greg might be agitated over going to church when he would rather be working or watching the football game on the television. However, the truth of the matter is that I (as Susan) would rather be doing such things than going to church. So it is imperative for me to screen off or quarantine my own idiosyncratic mental states when I am simulating Greg’s mental states regarding going to church.⁹³

In his 1996 article, Gordon gives two more examples adjusting for relative differences when transforming oneself to a target for the purpose of simulation. In one example, he describes viewing a list of award-winners. Viewing this list through my own eyes, I might be quite happy to find myself in the winning slot. However, if I view the list through the eyes of one of my competitors, and successfully quarantine my own views regarding the list so as to transform myself into my competitor, I would find myself rather upset to find the name Susan in the winning slot because “I” would now be the referent of the competitor, and “I” would be in a losing position.

Gordon gives another apt example in the same 1996 article. He describes what it is like for an actor to take on a certain role and to play it well. He claims that the actor must constantly be the person whom he is playing. He is pretending to be that character and, in so doing, he must take

⁹³ Note: This notion of quarantining (or ‘screening off’) one’s own idiosyncratic mental states will be crucial to my argument later in this chapter.
on the character’s psychological traits in order to become emotionally and conatively engaged in
the events occurring in the life of the character. “Many actors speak of transforming themselves
and of becoming the characters they play,” he says (Gordon, 1996b, p. 4). This involves an
imaginative shift in the reference of indexicals so that the referent of the pronoun “I” is the
character, not the actor. The actor is simulating the character and he is doing so by running his
own systems offline, and suspending his own idiosyncratic sets of thoughts, beliefs, desires, and
other mental states so as to become the character. The actor, while simulating, must block his
default mode by making adjustments to the role-play so as to reflect only the mental states of the
character. He must complete the egocentric shift by quarantining his own mental states to
essentially allow for role-play that is uncontaminated by his own mental states and only reflects
the mental states of the character. Gordon quotes one actor as saying, “I can play any part in the
world mentally, but I must always struggle to get [myself] out of the way, with my opinions, my
views, and my background” (Gordon, 1996, p. 4).

If we understand the actor as the simulator and the character as the target, we get a fairly
clear picture regarding what Gordon means by transformation, making adjustments for relative
differences, the egocentric shift, and running our own systems offline. Everything about the
simulator’s mental and emotional life must be suspended: facts, experiences, values, norms,
inputs, typical emotional reactions, and so on must all be quarantined for successful
uncontaminated simulation.

5.2.5. Gordon’s Ascent Routines

Gordon claims that it is entirely possible for people to ascribe mental states to themselves
without the concomitant possession of the relevant mental state concepts. One can, for example,
ascribe the mental state of belief to oneself without possessing the mental state concept of believing. Self-ascription of mental states in the absence correlative mental state concepts is made possible by a procedure that Gordon calls ascent routines. Ascent routines are a crucial component to Gordon’s account of the ST because the procedure allows people – even and especially children – to ascribe mental states to themselves (1) without prior possession of mental state concepts and (2) without the need for introspection, neither of which is required for successful simulation according to Gordon’s radical simulationist account.

By using an ascent routine, a person can determine whether or not she believes that p is just by asking herself the question of whether or not the fact of p. To exemplify this he says,

If someone were to ask me, (Q1) ‘Do you believe that Mickey Mouse has a tail?’ I would ask myself, (Q2) ‘Does Mickey Mouse have a tail?’ (with constraints on how I obtain the answer to Q2). If the answer to Q2 is Yes, then the presumptive answer to Q1 (the best I can do without taking into consideration the possible conflict between verbal and non-verbal behavior) is Yes (or, ‘Yes, I do believe Mickey has a tail’). The answer to Q1 is No if either the answer to Q2 is No or no answer is available within the constraints (Gordon, 1996 p. 15).

Notice that this procedure answers one question (Q1) by answering a different question (Q2), which is pitched at a lower semantic level, where Q1 is a question pertaining to one’s mental state (one’s belief about Mickey’s tail) and Q2 is just an object-level question directly about a fact in the world: in this case, Mickey’s tail. Gordon says that ascent routines are particularly interesting because the procedure allows one to answer a question about herself and her mental states by answering a question that has nothing to do with either herself or her mental states. Rather, it is just a question about an object “in the world” (such as Mickey Mouse’s tail).

Ascent routines allow kids to provide the ‘right’ answers to some mental state questions, which they developmentally lack the sophistication to understand. Gordon carefully emphasizes
that while ascent routines enable children to give right answers about their beliefs, ascent routines do not equip children to make genuine (comprehending) self-ascriptions of belief.

Gordon gives another example of using an ascent routine to ascribe a different mental state: that of ‘pain.’ He says that a child can say, “My foot hurts” when she injures her foot. She will understand that the foot that is hurting is ‘part of her’ without really grasping that the ‘hurting’ (the mental state of ‘pain’) is equally a part of her. He says, “the child understands the hurting to be a property of the foot just the way that its size, shape, and colour are properties of the foot” (Gordon, 1996, p. 17). There is nothing mental about the claim the child makes when she says her foot hurts. She is simply reporting the way the foot feels. Using an ascent routine, she can speak of her pain without having knowledge of pain as a mental state and without yet grasping the mental concept of pain.

When ascent routines work, they work, in part, because they require only very simplistic object-level thinking and do not require an understanding of mental states or mental state concepts. Whether the object-level thinking is about the existence of Mickey Mouse’s tail or the existence of a pain in a foot, the ascent routine is supposed to work in the same way. By employing the relevant ascent routine, a person can rather easily answer the lower semantic question regarding what is the matter of fact without inference to mental states, possession of mental state concepts, or introspection. Gordon suggests that ascent routines open up a logical space that allows us to conceive of mental locations such as: The tail is on Mickey. The pain is in the foot. With ascent routines these are simply facts that belong to some individual (Mickey Mouse) or belong to a part of some individual (the child’s foot). So a child can be trained to use mental state language (e.g. statements of belief) and ascent routines, which may enable a child to bootstrap her way into acquiring mental concepts without introspection and may represent a step
in the direction of acquiring a genuine (i.e. more sophisticated) understanding of the notion of a specific mental state, such as belief.

Elsewhere, Gordon describes how ascent routines give children a remarkable head start on learning self-ascription of mental states (Gordon, 1995a, p. 59). Here, Gordon attempts to elucidate how ascent routines are applicable to the mental state of desire or ‘want’. A young child who sees a banana may point at it and utter, “banana.” She may see Goldfish crackers and utter, “fish.” She may even point these objects when making her utterances. As adults, we train our children to say, “want banana” or “want fish.” Then we train our children to speak in more complete sentences and to say, “I want banana” or “I want fish.” We may then train our children to say “please” at the end of such sentences.

Gordon thinks that this is merely linguistic training, which does nothing to give the children mastery of the mental concepts of wanting or desiring. Further, he says, “it won’t even teach them that “I” refers to the speaker. So it isn’t sufficient for training children to make genuine ascriptions to themselves” (Gordon, 1995a, p. 59). At the time of the early training, the child has no concept of wanting or that it is she that is the “I” that is doing the wanting. By saying that ‘ascent routines enable a child to bootstrap their way into the acquisition of mental state concepts,’ I suppose that Gordon is trying to express that the way that we teach our children to talk – to string so-called proper sentences together – is not only useful for teaching a child language skills, but also is useful for aiding the child in the eventual acquisition and genuine ascriptions of mental state concepts. Eventually, as the child develops the ability to simulate, she can begin to see that some assertions of mental states (such as a belief) might contradict her own actual (unpretended) beliefs. It is then that she can move from an uncomprehending use (or
expression) of first-person mental state ascription words to a more sophisticated and genuine grasp (or possession) of actual mental state concepts.

The notion of ascent routines, however, is not entirely unproblematic. There are a few matters that I think Gordon needs to clarify to give us a more complete picture of ascent routines. Specifically, questions exist concerning how exactly ascent routines work in different scenarios with respect to varying mental states (See Goldman, 2006; and Nichols & Stich, 2002). Moreover, I question how they can work without possessing some kind of previously acquired mental state concept.

5.2.5.1. Ascent Routines Appear Not to Work for Self-Ascription of All Mental States.

First, as Goldman and others have noted (See Goldman, 2006; and Nichols & Stich, 2002), while Gordon briefly tells us how ascent routines may be useful for some self-ascriptions of some mental states, his accounting of how ascent routines are applicable to a wider array of mental states is woefully incomplete. Gordon initially gave a somewhat detailed description for how ascent routines work for the mental state of belief (as with Mickey Mouse’s tail) and less detailed descriptions of how they could work for the mental states of pain (as in the foot) or desire (as for the banana).

Goldman (2006) objects that Gordon has provided no reason to suppose that there is a distinctive first level utterance for each propositional attitude type that will be uttered only when the corresponding self-ascription of the propositional attitude is true. Goldman shows this by way of a counterexample illustrating that there is no apparent ascent routine for self-ascription of the mental state hope. In his example, he says, “Suppose the question is whether you hope team T won its game yesterday. Can you use the ascent routine to answer this question? Obviously,
answering the first-level question “Did team T win its game yesterday?” will not help you determine whether you hope team T won. What question, then, should you ask?” (Goldman, 2006, p. 240). As Goldman rightly observes, Gordon provides no reason for us to think that there is a distinctive question for each different type of attitude or mental state. Knowing whether or not team T won yesterday will not help answer the higher-level question regarding what one hopes. So it seems that the ascent routine cannot work for self-ascription of the mental state of hope.

In concurrence with Goldman, I return to an example given earlier.94 There, the example was meant to illustrate that we know our own mental states and that we express them with utterances such as, “I feel hungry. I now intend to drive to the store. I hope to find something there that I will want to eat.” These utterances express the mental states: feeling (hunger pains), intending, hoping, and wanting (e.g. desiring). The same trio of sentences will prove useful here. Let me address each of these in kind in an effort to show that ascent routines do not appear to be generally useful for all mental states.

I suppose that feeling hungry could be construed as a kind of pain state (i.e. hunger pains). Accordingly, we should expect that an ascent routine for hunger pain might work in the same way as in the injured foot example given by Gordon. There, he said that the child grasps the ‘hurting’ as a part of her without grasping that the mental state of ‘pain’ is also equally a part of her. He said, “the child understands the hurting to be a property of the foot just the way that its size, shape, and colour are properties of the foot” (Gordon, 1996, p. 17). The child, he says, reports only how the foot feels, but she has no (and employs no) prior concept of pain as a mental state.

94 Please see Chapter 2: The Theory of Mind, Section 2.2, this dissertation.
How does the ascent routine work in the case of feeling hungry – that is, experiencing a hunger pain? There is no property of the stomach to which the child can refer. The child does not say, “it is pink, soft, kidney bean shaped, and since it is rumbling or cramping, it must hungry.” Indeed the stomach, as an internal organ, is such that without some outside knowledge of anatomy or physiology, one could not even describe the properties of the stomach in the way that one could describe the properties of the more salient foot. Thus, it seems that there are at least some pains for which there could be no corresponding ascent routine – at least, under the description Gordon has provided thus far.

Similar problems exist for the mental state of intention. Intending is another mental state central to the ToM discourse. Indeed, Gordon (1986) discusses the expression of the mental state intend as just the utterance of “I shall…” However, it is difficult to imagine what ascent routine can exist for intending. The utterance, “I shall now drive to the store” expresses that one intends to drive to the store, where intend is strictly a mental state. In this case, there is no state of affairs (no “fact of the world”) to which one can refer. This is rather different than the case of belief. In the case of belief, if one is asked, “Do you believe Mickey has a tail?” one simply responds to the lower semantic question, “Does Mickey have a tail?” However, the mental state intend seems to not be like this at all. There is no fact of the matter “in the world” to refer to, nor any question that can be pitched at a lower semantic level. “I intend to…” is self-ascription of a mental state regarding one’s mental state in the specious present combine with a, heretofore unseen, prediction of one’s potential behavior.

Likewise (as Goldman notes, 2006), ascent routines appear problematic for the mental state of hope. When one says, “I hope there is something [at the store] that I will want to eat”, one makes an utterance about their self-ascription of the mental state hope. To phrase this slightly
differently, I might say, “I am not sure what I am in the mood to eat, but I will go to the store, and I hope to find something that looks appetizing to me at that time.” If we try to map this scenario onto Gordon’s (1996) ascent routine for belief, the ascent routine simply does not work. If we try to apply it, it would look something like this: If someone were to ask me, “Do you hope there is something appetizing at the store for you to eat?” I would, after examining the fact of the matter, instead answer a question pitched at a lower semantic level: “Is there something at the store that will be appetizing for me?” Of course, the answer to this question cannot be ascertained in the same way as “Does Mickey have a tail?” Knowing whether or not, in fact, there is something at the store that will strike the fancy of my appetite, will not help answer the higher-level question regarding what I hope. It seems that, in this example, an ascent routine would not work for the mental state hope or, at least, not in the same way that they work for the mental state belief.

Similarly, there is also the question of how the ascent routine could work for the mental state of desire. “I hope there is something at the store that I will want to eat.” Gordon briefly touches on this with his discussion of the child who is taught to say, “I want a banana” who is, nevertheless, ignorant regarding the mental state want (i.e. desire) when making such trained utterances. In Gordon’s example, the child sees a banana and utters, “want banana” almost regardless of whether or not there is any genuine wanting occurring. Though not explicitly stated, I think the implication is that the child’s utterance, “want banana” is akin to just saying something like, “there is a banana.” Again, Gordon says these utterances go a long way toward aiding in the child’s actual understanding of actual wants – that is, the acquisition of the mental state desire.
It seems to me, however, that ascent routines that are applicable to *desire* are problematic in the same or similar ways as for other mental states such as *hope*. In particular, there seem to be instances for which a person can self-ascribe the mental state of *desire* but where there is no corresponding question about something in the world. Thus, no ascent routine is obviously available in certain cases.

I wonder if there is a relevant ascent routine for instances in which one wants something that simply does not exist except, perhaps, in the imagination. For example, the child might say, “I want a three-eyed purple schnoogledorfer.” In this case, unlike the banana, there is no schnoogledorfer to which the child can point. Also, unlike the banana, there is not an object in the world that can qualify as the object of *desire*. It seems as though Gordon needs to say a bit more about how ascent routines work for mental states such as these in order for us to gain a better understanding of how ascent routines work to negate the necessities of previously acquired mental state concepts and introspection for self-ascription of mental states.

Nichols and Stich note problems with ascent routines that are similar in nature. According to them, there is no way to transform questions concerning retrospective reports into the kinds of fact-questions that Gordon’s theory requires. They say, “this includes responses to questions like “how did you figure that out?”” (Nichols and Stich, 2002). Nichols and Stich also note, “[problems] for questions about current desires, intentions and imaginings, questions like: “what do you want to do?”; “what are you going to do?” and “what are you imagining?”” (Ibid).

So, there are difficulties with ascent routines when attempting to apply them to mental states other than just *belief*. Where Nichols and Stich discussed the problematic nature of ascent routine questions relating to retrospective reports and Goldman discussed that ascent routines are problematic when considering self-ascribing the mental state of *hope*, I have shown that ascent
routines appear potentially problematic for a wide range of other basic mental states that are central to the ToM discourse such as, among others, pain, intend, hope, and desire. Taking these objections together, it seems as though ascent routines are just inadequate as a general theory of how we self-ascribe mental states.

5.2.5.2. Gordon Attempts to Answer Goldman’s “Belief Only” Objection.

In a more recent article, Gordon responds Goldman’s “belief only” objection, which claims that ascent routines are not generalizable to self-ascriptions of mental states other than belief such as, for example, hope (See Gordon, 2007). Gordon thinks that he can answer the challenge by eliminating an unstated, but widely assumed, constraint placed on ascent routines. This unstated constraint is an assumption that an ascent routine will involve a sentence about the world that is a different sentence than the sentence that is uttered about one’s mental states, and that the two sentences cannot be the same sentence. Gordon says, “Constraint C assumes the lower level utterance about “the world” and the higher level utterance about the speaker are not utterances of the same sentence” (Gordon, 2007, p. 156). When the implicit assumption of Constraint C is removed, we will supposedly find an algorithm that will presumably yield an ascent routine for hope and potentially any other propositional attitude.95

---

95 I think that Gordon’s claim that there will be an ascent routine for any propositional attitude is problematic for two reasons. First, Gordon seems to have altered his claim about how ascent routines are applicable. Gordon’s earlier work on ascent routines is explicitly aimed at mental states in general, and not strictly propositional attitudes (See Gordon 1995, 1995a, 1996). This was more in line with ToM literature and work on mental states and mental state concepts (See Baron-Cohen 1989, 1992, 1995). After criticisms by Nichols and Stich (2002) and Goldman (2006), Gordon adjusted his position so that ascent routines now appear to be for propositional attitudes (Gordon, 2007). In response, one might say that mental states just are propositional attitudes. This, however, is far from established and leads me to my second point. Many argue against the characterization of mental states as propositional attitudes (See Ben-Yami, 1997 and Merricks, 2009). Moreover, the work on mindreading and mentalizing explicitly discusses mental states that are not propositional attitudes (See Baron-Cohen 1989, 1991, 1995, 2001 2003; Frith, 1991, 2003; and Frith & Happé 1994, 1999). So, it seems that there exists a vast array of
Additionally, Gordon says that if we remove our implicit assumption of Constraint C, we must also change our implicit conception of an ascent routine. He says, “We cannot think of it as a simple formula that carries the speaker from a sentence that appears to be about “the world” to a sentence that appears to be about the speaker’s mind” (Gordon, 2007, p. 156). Indeed, Gordon thinks he is taking a step forward toward a bolder conception of ascent routines. He says,

No formula is needed to transform a sentence that expresses attitude \( \varphi(p) \) into a sentence that correctly describes one’s attitude, specifically, a sentence of the form, \( I \varphi \) \( that \) \( p \). In general, our early verbal expressions of attitudes are shaped, by explicit training and other learning mechanisms, into sentences that accurately describe our attitudes. Even if we are not actually (with the right communicative intentions) ascribing these attitudes to ourselves, even if we lack the conceptual resources to do so, we are steered towards uttering true sentences of the form, \( I \varphi \) \( that \) \( p \) (Ibid).

Although Gordon says that no formula is required and that we ought not assume that an ascent routine moves us from sentences about “the world” to sentences about the speaker, he acknowledges that ascent routine sometimes do appear to adhere to just such a formula and work in just that way. Recall his original example regarding Mickey Mouse’s tail. Here, the input sentence, “Mickey Mouse has a tail,” is clearly a sentence about the world; and the output sentence, “I believe that Mickey Mouse has a tail,” is clearly a sentence about the mental state of the speaker. In addition to this, Gordon cites some other propositional attitudes for which the corresponding ascent routines fall within Constraint C. He provides the following examples:

“\( \text{It’s surprising that they won.} \)’” > “I am surprised that they won.”
“\( \text{It’s upsetting that they won.} \)’” > “I am upset that they won.”
“\( \text{It’s regrettable that they won.} \)’” > “I regret that they won.” (Ibid. p. 157)

mental states that do not adhere to the standard form of propositional attitudes (i.e. “\( I \, V \) \( that \) \( p \)” where \( V \) is a verb used to ascribe a mental state and \( p \) is a proposition).

\(^96\) I concede that I, too, implicitly assumed both Constraint C and held the implicit conception of ascent routines that Gordon addresses in this 2007 article. While the article recasts ascent routines and may overcome the Goldman “belief only” objection, it raises new concerns, which I will address below.
In these cases, the input sentences could be interpreted, syntactically, to be nothing more than mere comments on the fact stated within the sentence, “They won.” whereas the output sentences are clearly sentences regarding the mental state of the speaker. Being that the input and output sentences are not the same sentence, these input/output combinations do adhere to Constraint C.

Contrarily, Gordon indicates, there are some cases in which the way that we express some “fact about the world” uses precisely the same sentence that we would use to self-ascribe a particular mental state. In these cases, we must discard the implicitly assumed Constraint C. For example, he says,

The form of speech most English-speaking people use to give verbal expression to a hope uses the same sentence as is used to ascribe the corresponding hope to oneself: for example, “I hope they won” – [or exclamatorily, “I hope they won!” or colloquially, “I sure hope they won!”]. Certainly the expressive use of “I hope the team won!” appears to be an utterance about the world (or a possible state of affairs). It may at the same time also be an assertion about oneself, a self-ascription of hope, though it need not be (Gordon, 2007, p. 158).

With this, Gordon is attempting to show that sometimes when one utters, “I hope they won” one is looking out into the world and the sentence is merely asserting something about the team and winning and the world. Other times, when one utters, “I hope they won!” one is looking inward and expressing something about oneself – that is, one is self-ascribing the mental state hope to oneself.

From these examples, it seems that, for the ascent routine to apply to self-ascriptions other than belief, we need a new model. The ascent routine model is not (as was the case with Mickey’s tail): “p.” > “I believe that p.” Gordon says, “Rather it is the “I φ that p” > “I φ that p” schema, where the [input sentence] represents the non-ascriptive (i.e. the assertive) use of the form and the [output sentence] represents the … explicit expressive self-ascription of a belief that
p” (Gordon, 2007, p. 160). Using this schema, Gordon gives us many examples: “I hope that p.” > “I hope that p.” “I fear that p.” > “I fear that p.” “I wish that p.” > “I wish that p.” “I regret that p.” > “I regret that p.” “I intend that p.” > “I intend that p.” (Ibid.). Gordon thinks that these examples indicate no difference in the input and output sentences beyond the intent of the speaker. He says, “When I actually intend to be attributing to myself the hope that the team won, my brain calls into play the same processing that gives rise to the non-ascriptive “I hope they won!”” (Ibid.). Finally, as noted, these utterances – as non-ascriptive speech acts – are conceptually undemanding. Gordon says we are trained to speak in this manner well before we acquire mental state concepts (Gordon, 1995a, p. 59). They can be uttered regardless of whether or not the speaker knows what “hope/fear/wish/regret that p” even means as a predicate and despite their ignorance, according to Gordon, the speaker’s resulting utterance usually “gets it right.”

In my view, Gordon’s response to the “belief only objection” is not entirely unproblematic. Next, I will briefly address what I see as some potential problems with Gordon’s revised account of ascent routines (2007) and, before I conclude the section, I will circle back to my claim that self ascriptions of mental states require possession of mental state concepts.

5.2.5.3. Some Problems with the Recast Ascent Routine: “I φ that p.” > “I φ that p.”

In letting go of Constraint C, Gordon now allows that ascent routines can be comprised of input and output sentences that are of the same sentence in the form of: “I φ that p.” > “I φ that p.” One problem that I see with the recast ascent routine is that, in cases where the input and output sentences are the same sentence, it appears to require some form of introspective mental state attribution (i.e. some self-ascription of a mental state derived from “looking inward,” – that
is, introspecting) for the speaker to discern whether she is expressing her attitude or merely asserting or proclaiming it.\footnote{Recall that Gordon’s account of the ST explicitly argues against the need for introspection. Hence, he equivocates in his claim that the simulator must “look inward” which, I assert, of course requires introspection.}

At first glance, the reader might argue that it is nonsensical to say that one must ascribe a mental state to oneself in order to know whether or not one is ascribing a mental state to oneself. But Gordon is quite clear when he says that these examples indicate no difference in the input and output sentences beyond the intent of the speaker. Recall, he says, “When I actually intend to be attributing to myself the hope that the team won, my brain calls into play the same processing that gives rise to the non-ascriptive “I hope they won!” (Gordon, 2007, p. 160). Within the ToM discourse, and even in Gordon’s view, intend counts as a mental state. Thus, Gordon sets up a sort of regress (though perhaps not an infinite one) whereby in order for a speaker to discern whether her sentence is merely asserting or proclaiming (“I \( \varphi \) that \( p \).”) on one hand, or it is expressing an attitude (“I \( \varphi \) that \( p \).”) on the other hand, she must first determine what she intends: the non-ascriptive \( p \) or the self-attribution of \( p \). To determine what she intends, she must look within (i.e. introspect). However, Gordon argues vehemently against both the need for introspection and the need for prior acquisition and possession of mental state concepts.

The case for my observation is bolstered by the work of Donald Davidson. Davidson suggested that the meaning ascribed to any individual sentence could only be determined by attributing meanings to many, perhaps all, of the individual's assertions, as well as their mental states and attitudes (Davidson, 2001). From this, I gather that the meaning of either sentence (“I \( \varphi \) that \( p \).” or “I \( \varphi \) that \( p \).”) is only determined by knowing the mental states and attitudes that inform the assertions or expressions of the speaker. Further, the fact that Gordon allows the input
and output sentences to be *the same sentence* leaves the intent of the sentence to be open to interpretation – that is, is the speaker expressing an attitude or merely reporting on it? The sentence can be interpreted as the assertion of *p* or an expression of an attitude, *p*. Given the Principle of Communicative Relevance, Wilson & Sperber argue that there should never be more than one interpretation of a given sentence because “A speaker who wants her utterance to be as easy as possible to understand should formulate it … so that the first interpretation to satisfy the hearer's expectation of relevance is the one she intended to convey” (Wilson & Sperber, 2004, p. 614). In a scenario whereby an input sentence is meant as a mere assertion and the output sentence (which is the same sentence) is meant as an expression of a propositional attitude, a hearer (and perhaps even the speaker) must determine the relevance of the sentence, which puts the hearer to an unjustified processing effort to determine the meaning of the sentence.

A second problem I find with the ascent routine (“I φ that *p*.” > “I φ that *p*.”), as Gordon has recast it, is that input/output combinations such as these are nugatory or hollow in at least two ways. First, the input sentences, at least in the examples provided by Gordon, (e.g. “It’s surprising that they won.” “It’s upsetting that they won.” It’s regrettable that they won.”) are perhaps factually defective because, despite syntactic appearances, are neither definitively true nor definitively false. More precisely, these sentences can be either true or false. In response to each of these proclamations, one could ask, “To *whom* is it surprising/upsetting/regrettable?” Each of these sentences will have a different truth-value depending on the referent of the sentence. The opposing team or fans of the opposing team may find it surprising, upsetting, or regrettable that the team won, in which case the sentence would be true. However the winning team and the fans of the winning team would likely not find it surprising, upsetting, or regrettable that the team won, in which case the same sentence would be false.
There is a second, perhaps more important, way in which the input/output sentences are nugatory and hollow. Gordon says, “In general, our early verbal expressions of attitudes are shaped, by explicit training and other learning mechanisms, into sentences that accurately describe our attitudes. Even if we are not actually (with the right communicative intentions) ascribing these attitudes to ourselves, even if we lack the conceptual resources to do so, we are steered towards uttering true sentences of the form, \( I \varphi \text{ that } p \)” (Gordon, 2007, p. 156). Children, according to Gordon, are trained to utter such things without any concept of their own mental states – it’s as if they engage in “mental state talk” without grasping what the mental state is. Under these circumstances, we don’t say, “Oh that child is so precocious! She knows and comprehends the concepts of belief and desire!” We don’t say that because we know the child’s utterances are empty – they are devoid of the depth that comes along with these utterances when they are actually self-ascribing specific and certain mental states. Why should it be any different for an adult when making mental-state-empty assertions? When an adult utters a sentence that sounds like mental state talk, but is not mental state talk, the utterance ought not be confused with the self-ascription of a mental state (i.e. it’s just empty talk).

Third, I think that Gordon takes himself out of the game with this. Let us recall that the primary discussion at hand is a discussion about what it is that underpins the ToM. It is about what underpins mentalizing. It is a discussion about how we know our own mental states and predict our behavior based on them and about how we know others’ mental states and predict their behavior based on them. Ascent routines as Gordon has now couched them, are specifically no longer about mental states. They are utterances about things in the world, or they are proclamations, or they are assertions, but they are specifically not about mental states, do not require mental state conceptual resources, and are not about mindreading. Ascent routines might
say something about how we ascribe propositional attitudes and mental states to ourselves by just uttering sentences but they are not furthering the discussion on how it is that we mentalize or mindread.

Last, but by no means least, there is a fourth and final important objection to be made regarding Gordon’s ascent routines as he has recast them in response to Goldman’s “belief only” objection. Specifically, I do not think that Gordon has succeeded in answering Goldman’s objection because Gordon has committed a sort of *ignoratio elenchi* – that is, he has missed the point of Goldman’s argument and refutes, or tries to refute, a claim other than that which was originally at issue. Let me explain.

Ascent Routines (as originally put forth by Gordon) were supposed to answer the question of how we can answer questions about our own mental states without the need for possession of mental state concepts and introspection. Goldman’s challenge, as I understand it, questions how ascent routines can help us to arrive at knowing and self-ascribing our mental states (other than belief) without prior possession of the relevant mental state concepts and without introspection. As I have elucidated, Gordon’s article (2007), in response, goes on to contrast merely making assertions about things in the world versus actually expressing (and so self-ascribing) one’s mental states. Gordon’s response informs us as to how we might classify sentences that appear to be about mental states as either assertions in the non-ascriptive sense or expressions in the self-ascriptive sense. However, the problem is not, “How do we classify…” The problem is, rather, “How do we know and self-ascribe our own mental states without possession of mental state concepts and introspection?” It seems to me that Gordon has simply missed the point of Goldman’s “belief only” objection and has simply not answered the question that Goldman’s challenge posed.
Gordon’s form of “radical simulation” insists upon three things: (1) that there is no need for prior acquisition of mental state concepts, (2) that simulation does not require any amount of introspection, and (3) that simulation is a complete transformation (as opposed to a transference) from the self to the target such that no analogical inference “from me to you” is needed. It is unclear to me whether Gordon thinks that ascent routines help to do away with the need for (3) analogical inference in the context of a simulation procedure. However, it is clear to me that Gordon thinks that ascent routines do away with (1) the need for prior acquisition of mental state concepts and (2) introspection for the purpose of self-ascribing mental states. At least in this respect, it seems that ascent routines (as currently elucidated by Gordon) fail to do the job, but perhaps Gordon could clarify these matters.

5.2.6. Concluding Remarks Concerning Gordon

In this section, I have described the ST as formulated by Robert Gordon. The most important aspects of his account are these: First, a simulating system need not recognize its own mental states, be in possession of the relevant mental state concepts, nor have any experiential or introspective basis for applying them. Second, one must quarantine the relevant facts about themselves when transforming themselves into the simulated other (i.e. the target). Third, ascent routines may be used for making ascriptions of mental states in lieu of possessing and employing mental state concepts. For Gordon, simulation does not require either introspection or possession and application of mental state concepts. These are the reasons that Gordon is often considered a “radical simulationist.” These are also the reasons that I do not agree with Gordon’s view.

In my view, introspection is absolutely critical to simulation for two reasons. First, without introspection I cannot ascribe any mental state to myself, and first-person self-attribution of
mental states just does not work. For example, if one was to ask me, “What do you want for lunch?” “Are you worried about something?” Or, “Do you hope that Brad will be at the party?” it is clear to me that I would need to first “look inside myself” (to use Gordon’s mental location language) to determine the mental states that are mine, such as: want, worry, and hope. My view is that I must be able to allude to my various potential mental states as points of reference in order to determine my current mental state(s) simply because the mental states under examination are mine. Again, using Gordon’s mental location language, my mental states are in me or come from within myself. To answer questions concerning my mental states – or even just to self-ascribe them in the absence of any questions – I must first survey some inner fact or facts about myself in the specious present, and distinguish those inner facts about myself from other possible scenarios of potential inner facts about myself. Secondly, and importantly, Gordon claims that, in order to perform accurate, uncontaminated simulation (i.e. simulation that is not corrupted by one’s own idiosyncratic mental states), (a) one must hold back the relevant features of oneself to instead focus on those of the target and (b) one must adjust for relevant differences between oneself and the target (Gordon, 1986). It is clear to me that one cannot do (a) or (b) without introspection. Without introspection, how can one know which relevant features of oneself to hold back or which differences to adjust for? The idea that I can know anything about my inner self without the aid of introspection and a survey of some inner fact or facts about myself makes little sense to me. Accordingly, in the remainder of this chapter, I will discuss arguments showing the importance of introspection to the simulation routine.

---

98 Recall that Gordon might say that ascent routines can account for such mental states and vacate the need for introspection. Ascent routines were described and refuted in the foregoing sections of this chapter (5.2.5 – 5.2.5.3).
In the next section of this chapter, I will be discussing the ST as reformulated by Alvin Goldman, who holds a different view of the ST. While Gordon rejects introspection entirely, as well as any analogical inference ‘from me to you’, Goldman’s version of the ST actually requires introspection, both the possession and application of mental state concepts, and the analogical inference ‘from me to you’. Ultimately, as I will show, it is Goldman’s version of the ST that is most supportable when we consider the special case of ASD.

5.3. Alvin Goldman’s Simulation Theory of Theory of Mind: Simulation with Introspection

Soon after Gordon published his work on the ST, other philosophers followed his lead writing in support of the ST. Jane Heal was an early proponent of the ST (1986), as was Paul Harris (1989). Alvin Goldman (also 1989) was also an early proponent and defender of the ST. As simulationists, these philosophers share the five key ideas found in Gordon’s version of the ST (stated above). They share the core simulationist idea that mindreaders do not theorize about the mental states of targets to explain and predict behavior. Rather, they simulate a target by using their own cognitive mechanisms to create mental states of their own and then impute them to the target – that is, they use their own mind as a model for the mind of the target so as to explain and predict what will transpire in the target. Of the simulation theorists mentioned, I find Goldman to offer the most compelling account of the ST and I think that it is especially supported by the case of ASD.

Goldman (2006) provided the most thoroughly developed and empirically supported case for the ST, defending it as the most plausible explanation of that which underpins the mindreading ability. Goldman describes this work as an interdisciplinary effort to reformulate, refine, and
update the case for simulation-based mindreading. He does this in light of both empirical findings and philosophical analysis (Goldman, 2009a).

He (Ibid.) includes discussions of a variety of other theories of ToM and presents arguments showing them to be implausible explanations of ToM. Goldman writes not just in favor of the ST in general, but of his particular version of the ST, which differs from Gordon’s original account. Goldman defends the ST in part by describing mindreading as process-driven rather than theory-driven, and creates a new distinction of the simulation activity: that of high-level simulation versus low-level simulation. Goldman then includes a discussion of mirror neurons, why he thinks that they are relevant to the TT/ST debate, and why he thinks that mirror neurons work to support ST over TT. Finally, Goldman especially distinguishes himself from Gordon by including the need for introspection in the mindreading ability, thereby establishing himself as a less-radical simulationist than is Gordon.

In the remainder of this section, I will discuss Goldman’s version of the ST. First, I will discuss Goldman’s description of mindreading as a process-driven activity rather than a theory-driven activity, and his distinction between high-level and low-level simulation. Next, I will briefly discuss mirror neurons. Finally, at the conclusion of this section, I will discuss Goldman’s inclusion of (and insistence upon the essentialness of) introspection, showing how this distinguishes his account from Gordon’s account of the ST. Here, I will focus the discussion on the importance of introspection to Goldman’s account of mindreading and to my argument, which claims that a simulation account of mindreading that includes introspection is likely the best account of ToM.

---

Although I think that Goldman takes mirror neurons to be an important component of his version of the ST, I will not discuss mirror neurons at length, except to say what they are and why I am not including them in my discussion more thoroughly.
5.3.1. Mindreading is Process-Driven not Theory-Driven

Like Gordon, Goldman contemplates the controversial question regarding how the mindreading task is executed, seeking to show that it is executed by simulation of the target’s mental states rather than theorizing about them. Goldman in part defends the ST on grounds that there is a difference between theory-driven mentalizing, which must be used for systems different than oneself, and process-driven mentalizing, which can be applied to systems resembling oneself (Goldman, 1992). I elucidate this difference in what follows.

According to Goldman, theory-driven mindreading is guided by, and requires, a theory about the systems of the target. In contrast, process-driven mindreading is accomplished via simulation that is guided primarily by the cognizer’s natural mental processes, which may or may not include various inputs to the simulation such as beliefs. Presumably, if and when one human behavior control system is being used to model others, then general information (e.g. theories) about the behavior control systems of the target is unnecessary. The simulation thus is said to be process-driven, rather than theory-driven because no underlying theory (nor reference to any theory) is required to understand the mental states of the target or predict the behavior of the target. He says, “if the process or mechanism driving the simulation is similar enough to the process or mechanism driving the target, and if the initial states are also sufficiently similar, the simulation might produce an isomorphic final state to that of the target without the help of theorizing” (Goldman, 2012, p. 11).

So, if the processes, systems, and mechanisms of the target are enough “like mine” (or, to restate, if the target is enough “like me”), I am able to use a process-driven simulation to accomplish the simulation and mental state attribution of my target without the need or help of theorizing. This seems right to me. It seems patently obvious that I simulate other human beings...
with a fair amount of ease and accuracy and, in so doing, accurately explain and predict their behavior. However, if I want to understand, explain, and predict what a humpback whale, for example, is going to do next, I will need to refer to and deploy some kind of theory concerning cetology and humpback whale behavior to make such a prediction. To understand and predict the behavior of a humpback whale, I take myself to need a theory of its typical behavior. However, to understand and predict the behavior (and mental states) of a fellow human being, I do not take myself to need any such theory, since my own systems that are, ostensibly, similar enough to those of my target are available to me to use in the simulative activity.

The basic idea of simulation as process-driven mindreading is that we take the resources already present in our own brain, which it uses to guide our own behavior (often in accordance to our mental states), and we modify these resources to work as representations of other people. Thus, we do not store general information or theories about what makes people believe certain things, behave in certain ways, or function characteristically. Simulation is thus said to be “process-driven,” rather than “theory-driven,” because we take our own characteristic functioning “offline” and apply it to the target. If no reference to an underlying theory is needed, and mentalizing and explaining and predicting the behavior of a target can be accomplished via simulation, then it can plausibly be suggested that the mindreading task is potentially process-driven rather than theory-driven.

Drawing out the distinction between process-driven versus theory-driven simulation, Goldman offers an articulation of a possible way that the mindreading task is executed without theorizing.

---

100 Dennett gives a similar example. He says, “If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what “comes to me” in my make believe state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges” (1987, p. 100). Here, Dennett is illustrating the need for some kind of theory to make sophisticated predictions about systems that are substantially different from oneself.
5.3.2. **High-Level versus Low-Level Mindreading**

Another way in which Goldman’s account of the ST is distinguishable from Gordon’s account is a distinction that Goldman makes between high-level and low-level simulation. Goldman (2006) offers a two-pronged theory of mindreading that differentiates between “high-level” simulation based mindreading and “low-level” simulation based mindreading. For Goldman, both levels of mindreading count as simulation – that is, they each exemplify simulational characteristics, broadly construed, but their underlying systems are different. I will discuss these each in turn.

High-level mindreading and its underlying framework are more complex than low-level mindreading in a few ways. First, high-level mindreading is typically a highly conscious and voluntary activity that involves consciously and knowingly putting oneself in the position of the target via deliberate enactment of the imagination (See Goldman 2006, 2008). Secondly, high-level mindreading generally involves simulating the propositional attitudes of the target (such as the target’s “wants,” “beliefs,” “hopes,” or “intentions”) towards a given thing (a –that clause, which completes a sentence) (See Shanton and Goldman, 2010). Thirdly, high-level mindreading may entail deployment of information held in episodic memory as well as current and relevant factors and information about both the target and the situation of the target. Goldman says that high-level mindreading is the kind of simulation we engage in when we want to engage in pretense or perspective taking themes (Goldman, 2008). In short, high-level mindreading simulations are a deliberate undertaking on the part of the simulator, requiring conscious construction of appropriate pretend states to ascertain and attribute propositional attitudes and mental states to a specific target in a given situation.
High-level mindreading processes are to be contrasted with low-level mindreading processes. While, again, Goldman says each of these *counts as simulation*, they are very different in how they are deployed, when (in which situations) they are deployed, and in their underlying mechanisms. In fact, the two are so different that Goldman says “There is no parallel here between high-level processes and [low-level] processes” (Goldman, 2008). Where high-level simulational processes are available to consciousness and, generally, are an intentional activity of which the simulator is keenly aware, low-level simulational processes tend to be automatic and unconscious. Shanton and Goldman say that low-level simulation is a fairly simple – even a “primitive” – way of attributing mental states to others (2010). In particular, hey say, “[low-level simulational processes] do not have associate propositional contents, which [would] inevitably introduce greater complexity” (Shanton and Goldman, 2010). The low-level simulational process remains beneath the surface of consciousness and does not involve any computation or inference on the part of the simulator in order to assign mental states to a target. By contrast, the kind of complexity introduced by having associate propositional content would require a high-level process, of which the simulator would be cognizant while consciously constructing pretend states in order to attribute propositional attitudes to her target.

Goldman proposes that the underlying framework of low-level simulational processes are hard-wired and primitive neural mechanisms. These neural mechanisms are normally produced and associated with one’s own behavior or emotions and are apparently also automatically and unconsciously activated in response to observing the behavior or emotions of a given target (Goldman, 2006).

Deployment of a low-level simulational process typically involves structurally simple mental states. For example, a low-level simulational process is deployed when the simulator observes
certain salient facial expressions of a target, ones that will ordinarily belie her underlying emotional state. Goldman notes that there is some evidence that the six basic emotions (fear, anger, disgust, sadness, happiness, and surprise) have fairly uniform or characteristic facial expressions, and that most people can discriminate these emotions based on these distinctive expressions. Assigning a mental state (or, in this case, an emotional state) to a target strictly on the basis of her facial expression is one example of a low-level simulative process. Also, there is some evidence from cognitive neuroscientists that when an observer witnesses the facial features of a given emotion, she will simultaneously experience the emotion of the target—at least, at the unconscious neural level (See Goldman, 2006, 2008, 2012; and Shanton and Goldman, 2010).

The “neural activity” referred to (which is involved in this simultaneous experience in the simulator when observing a target) purportedly comes from something in the brain called “mirror neurons.” Although Goldman does not claim that these mirror neurons are responsible for all simulation, he does propose that mirror neurons are a core example of low-level mindreading (2008). Mirror neurons produce mental mimicry. He says, “Low-level mindreading, then, can be viewed as an elaboration of a primitive tendency to engage in automatic mental mimicry… a deep-seated property of the social brain, and low-level mindreading builds on its foundations” (Goldman, 2012). Before turning to another way in which Goldman distinguishes himself from Gordon (that which involves the notion of introspection), I will address the subject of mirror neurons in a brief subsection.

5.3.3. Mirror Neurons: Important for Goldman, but Controversial

In this subsection I will first briefly describe mirror neurons and their discovery. I will then discuss why mirror neurons seem to be an important discovery to any theory of ToM generally,
and why Goldman (and others) understand mirror neurons to be an important discovery relating specifically to the ST. Finally, I will describe some skepticism concerning mirror neurons and say why it is that I am not including a more in depth discussion about them in the ST chapter or in this dissertation as a whole.

Giacomo Rizzolatti, a professor of neuroscience at the University of Parma in Italy, accidentally discovered mirror neurons in 1996. At the time of the discovery, Dr. Rizzolatti was working with macaque monkeys. He first noted that when the monkey reached for a peanut, a specific neuron in the monkey’s brain fired. He then noticed that when he reached for the peanut, the same neuron in the monkey’s brain fired just by the monkey watching him reach for the peanut. This suggested to Rizzolatti that these specific neurons in the brain would fire whether in correlation with an action of the monkey, or in correlation with the monkey’s mere observation of the same action by someone else. Further studies by the Parma group confirmed that these ‘mirror neurons’ were found not only in monkeys, but also in human beings (Rizzolatti & Craighero, 2004).

Since their discovery, mirror neurons have been a matter of great excitement in the neuroscience and cognitive science communities. Mirror neurons have been viewed as a potential explanation for the fact that we are apparently able to understand and interact with other human beings (Gallese, et.al., 2004; Keysers & Gazzola, 2009; Rizzolatti & Craighero, 2004). It has been thought that mirror neurons could offer a neurological explanation of our observations that we explain and predict other people’s behaviors and mental states, and that we seem to feel other people’s predicaments, emotions, and movements (Spaulding, 2012; Glaser, “Mirror Neurons”). Mirror neurons, it has been claimed, could account for the rich and complex level of empathy human beings experience when watching sport, movies, dance, video games, or
even just the emotions of others (Glaser, “Monkey See Monkey Do”). We wince when the quarterback gets tackled; we cry or laugh during movies; and the muscles of our legs flex when we watch dancers dance or boxers fight. Dr. Marco Iacoboni claims that we do all of these things because the mirror neurons in our brains help us to connect to other people by helping us to understand not only to their actions, but also their feelings and mental states (Iacoboni, 2009; Iacoboni, “Mirror Neurons”). On some accounts, mirroring activity in the brain via mirror neurons appears to be a promising neurological explanation for the human capacities of empathy and of ToM in general.101

Mirror neurons are thought by some to be responsible for understanding and imitating actions, goals, and other mental states (See Calvo-Merino et.al., 2004; Gallese, 2007; Goldman, 2009a; Iacoboni, 2009; Rizzolatti & Sinigaglia, 2010). As such, they also appear promising as a sort of confirmation of ST because simulation theorists argue that we understand others via mental simulation. They contend that we adopt pretend mental states that we think the target has and then use these pretend mental states to explain and predict the target’s behavior – that is we mentalize by simulating or mirroring their mental states. Gallese and Goldman claim that mirror neuron activity does not proceed by theoretical inference but that these neurons create a state in the brain of the observer that matches the neural state in the brain of the target (1998). Hence, mirror neuron activity resembles a simulation heuristic and is often considered strong evidence favoring the ST over the TT (Gallese & Goldman, 1998; Goldman, 2006, 2009; Gordon, 2005; Hurley, 2005; Iacoboni, 2009).

101 See Shannon Spaulding for her summary of mirror neuron research as it relates to ToM research. In short, she states, “An appealing tentative suggestion is that we understand what another person is doing, feeling, and experiencing because when we observe the other person parts of our brains are activating as if we were doing what that other person is doing. Our brain activity mirrors the other person’s brain activity such that it is as if we are acting, feeling, or experiencing how the target is acting, feeling or experiencing” (Spaulding, 2012, pp. 518-519).
Additionally, as detailed earlier in this section, Goldman distinguishes between high-level and low-level simulation, where low-level mindreading is a relatively primitive, simple, and automatic form of mindreading, which is largely below the level of consciousness versus high-level simulation, which typically entails voluntary and conscious putting (imagining) oneself in place of the target. Goldman thinks that low-level simulational mindreading rests substantially on the mirroring system. He also thinks that, given his definition of mental simulation, mirroring represents a clear form of (and qualifies as an instance of) interpersonal mental simulation (Goldman, *Ibid*, pp., 132, 143-144). Specifically, according to Gallese and Goldman (1998, p. 498), “mirror neurons represent a primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mindreading.” Thus, since the discovery and explication of mirror neurons, simulationists such as Goldman have viewed them as potential evidence for low-level simulation and the ST.

Despite the excitement over the discovery of mirror neurons, their potential to explain mindreading at the neuron level, and their potential to substantiate the ST, there are many reasons for being skeptical about just how important they are and just what role, if any, they play in mindreading. Goldman admits that, “no detailed positive account of the mirroring process is presently available” (Goldman, 2006, p. 127). He also seems to express some skepticism about the role, if any, that mirror neurons play in high-level simulation based mindreading. He is very careful to state that mirror neurons “potentially” play a role in (perhaps only) low-level simulation. Further, Goldman says, “…the definition of mirroring (as I construe it) does not imply mindreading, but leaves open the possibility that some mindreading is based on mirroring” (Goldman, 2006, p. 137).
Other researchers are also reining in their excitement about what mirror neurons might or might not explain. Rizzolatti and the Parma Group seem to be taking a step back from their claims about what mirror neurons can account for – that is, that they may account for understanding action goals, but may not account for understanding non-movement oriented goals such as desires, intentions, sensory goals or other forms normally associated with mentalizing (Rizzolatti & Sinigaglia, 2010). Some researchers have argued that it is dubious as to whether the mirror system allows us to understand the emotions, intentions, perspectives and even the actions of others (Hickok, 2009; Hickok et. al., 2011).

There seems to be much continued controversy regarding the claims that mirror neurons supports mindreading. Certainly Goldman argues that there is ample evidence that mirror neurons do support mindreading in some important way. Others, such as cognitive neuroscientist Rebecca Saxe (2009) and philosopher Pierre Jacob (2008) argue that there is no evidence whatsoever that mirror neurons cause mindreading. As a philosopher writing a dissertation in support of the ST over the TT, I would happily concur with Goldman and write more favorably about the role that mirror neurons play in mindreading and in substantiating the ST story. Unfortunately, in the end, I concur with Hickock who says, “the idea that mirror neurons cause mindreading is an interesting hypothesis, but not one that has been thoroughly vetted” (Hickok and Hauser, 2010). It seems to me that much more research must be done in this area before anything about mirror neurons and the role they may play in supporting mindreading can be stated conclusively.

Claims regarding how mirror neurons work (and to what end), as well as suppositions regarding whether mirror neurons support mindreading or serve as evidence favoring simulationist accounts of ToM, have come under fire and sparked controversy over both the
claims and the supporting research. Given the amount of controversy over such issues, to make a case for or against the claim that mirror neurons support the ST would require a separate dissertation. As such, I neither defend nor refute them at this time and so leave them aside for now.

5.3.4. Introspection is Imperative for Simulation and Mindreading and is a Critical Component of Goldman’s Version of the Simulation Theory

Thus far, in this section, I have discussed Goldman’s account of the ST and some ways in which it varies from Gordon’s account of the ST. I will now introduce the most crucial component of Goldman’s account of the ST that distinguishes it from Gordon’s account: the notion of introspection.

As I have noted, Gordon’s version of the ST presents and argues for a ‘radical’ form of simulationism according to which our concepts of the mental are acquired through a process of simulation without subjects needing to have introspective access to their own mental states (Gordon, 1996a). In a move that Carruthers calls “ambitious,” Gordon claims that a simulator can actually represent the beliefs and desires of a target without introspective access to ones own mental states and without engaging in any sort of analogical inference from oneself to the other (Carruthers 1996, p. 33; also see Gordon 1986, 1992b, 1995a, & 1995b).

As I explained earlier in this chapter, Gordon’s version of the ST suggests not just transference of the simulator’s mental states to those of the target, but an actual transformation from the simulator to the target. Again, Gordon completely rejects the idea that introspection needs to play any role in simulation based mindreading, and so argues against the view that we make inferences from what we ourselves would do in the imagined circumstances to what others
will do in the same or similar circumstances (Gordon, 1992b, 1995a). Gordon claims that
simulation involves only the kind of imaginative identification that can operate without
introspective awareness or self-attribution of mental states. In short, Gordon expunges
introspection from his version of the ST and dispenses with the requirement for prior or
concomitant possession of mental concepts in mindreading by rejecting a final step of simulation
in which a (simulated) mental state is classified under a mental concept. (Goldman, 2006, p. 259)

In contrast to Gordon’s account, Goldman’s version of the ST requires introspection. Where
Gordon rejects introspection entirely, Goldman claims that it is essential that (1) the simulator
recognizes her own mental states, (2) possesses the relative mental state concept, and (3) has
either some experiential or some introspective basis for applying it. Further, whereas Gordon
rejects any analogical inference “from me to you,” Goldman’s account of the ST presupposes
first-person awareness of one’s own mental states for both first-person and third-person mental
state attribution, and so actually embraces the inference from simulator to target as being a kind
of argument from analogy.

Goldman readily acknowledges that many psychologists and cognitive scientists are, in
general, very leery of the notion that human beings possess some kind of privileged access to
one’s own mental states. He says, “psychology has, historically, challenged the legitimacy of
introspection as a method of psychological science. This challenge was so potent during the
behaviorist era, the very word introspection became – and largely remains - taboo” (Goldman,
2006, p. 229). Nevertheless, Goldman insists that introspection is a common (even standard or
‘normal’) way in which normal human beings are able to recognize, discern, or detect broad or
universal categories of mental states and identify them with one’s current instantiation of a
mental states (e.g. belief, desire, fear, pain, anger). In his early work, Goldman develops the idea
that these instances of particular mental states have intrinsic, almost qualia-like, qualities that are introspectible. He says, “introspection is the standard first-person method for mental state recognition and could play a pivotal role in the account of mental concepts” (Goldman, 2006, p. 259).

Goldman stresses the importance of – and even assigns priority to – introspection, especially for first-person mindreading. It is the method typically developing humans use to recognize one’s own mental states. Such recognition in first-person mental state attribution comes prior to and forms the practical basis for third-person mental state attribution. As such, it follows that introspection must also play an important role in third-person mental state attribution. This is because if introspection is required for first-person mindreading and first-person mindreading is prior to and forms the practical basis of third-person mindreading, then third-person mindreading must derive from the same basis as first-person mindreading. So, just as introspection plays a crucial role in first-person mindreading, third-person mindreading must also include introspection. On Goldman’s account, a mindreader must first introspectively access their offline products of simulation before projecting them onto the target or inferring that these products emulate those of the target.

Central to Goldman’s thesis is that self-attribution of mental concepts always (at least partially) employs introspection via either introspection-derived or introspection-associated mental representations (Goldman, 2006, p. 260). Goldman regards introspection as a necessary condition for mindreading. He takes introspection to be a necessary component, playing this predominant role in the self-attribution of mental states, which, he says, accommodates the role of self-attribution in simulative routines when simulating a target (Goldman, 2006, p. 274). For

---

102 This is just as in the case where A is required for B, and B is required for C, then A must be required for C.
example, Goldman maintains that before a mindreader can attribute a pretend mental state to a given target, *she must first introspect* the state generated by her cognitive system and determine its type and content (Shanton and Goldman, 2010 emphasis mine). So, on Goldman’s account, one first recognizes one’s own mental states under actual or imagined conditions and then infers, on the basis of an assumed similarity or analogous inference “from me to you,” that the target is in a similar mental state as that which one is recognizing in oneself under the actual or imagined conditions.

As is clear, introspection is a crucial part of Goldman’s version of the ST. For Goldman, introspection plays a vital role in one’s self-attribution of mental states and first-person mindreading. Moreover, because these precede and establish the practical basis for third-person mental state attribution and third-person mindreading, introspection also plays a vital role in simulational processes and routines. Lastly, owed to the necessity of introspection for third-person mindreading, Goldman says that simulation *does* depend on judgments of similarity to the self and analogous inferences “from me to you.” So, Goldman says, “It is unwise to insist [as Gordon does] on a version of the ST that categorically denies this [kind of introspection and analogical inference]” (Goldman, 2006, p. 187).

As shown, there are a few key differences between Gordon and Goldman’s respective versions of the ST. In particular, one difference is that Goldman takes pains to explicitly clarify mentalizing as a process-driven, not a theory-driven activity on the part of the mindreader. A second difference is that Goldman distinguishes between high-level simulational processes and low-level simulational processes for simulation-based mindreading and offers an account of mirror neurons in relation to, and as a potential explanation of, low-level simulational processes. A final difference is that Goldman assigns primary importance to the roles played by both
introspection and analogy from the simulator to the target when mindreading via simulational processes and routines. As I will show, Goldman’s inclusion, indeed requirement, of introspection for both first-person and third-person mindreading is not only crucial for Goldman’s account of the ST, but it is also central to my argument in this dissertation, which lends support to the ST.

5.4. The Theory Theorist’s Criticism of the Simulation Theory: Collapse

A few words must be said regarding perhaps the most common criticism of the ST. Many theory theorists argue that in one way or another, at some point, simulation routines require the use of (or reference to) a theory. Some proponents of the TT seem to construe ‘theory’ in a very narrow sense, asserting that the ‘theories’ in question are organized and based on generalizations and principles that have the same status as the laws of science and, in particular, psychological laws. Other theory theorists seem to construe ‘theory’ in a much more broad sense, one that includes any kind of information about or reference to scientific psychology or folk psychology. Although there are different iterations of the collapse objection by different theory theorists of different ilk, all of them end with the same basic claim, viz., that the ST merely collapses into the TT. Here, I will briefly discuss a couple of versions of the ‘collapse’ argument.

First, some theory theorists, such as Davies and Stone (1995a, 1995b) and Perner (1991), claim that the ST collapses into the TT because some reference to or reliance upon theoreticity is required before the inception of any simulation routine. Before a person can even begin a simulation routine, they must first theorize – that is, a theory is required to set up the simulational routine in the first place and even before it begins.
Davies and Stone acknowledge the possibility of simulation-based mindreading, but hold that such simulation routines are first *initiated* by theorizing. This is because, on both Gordon’s and Goldman’s accounts, when a simulator is initiating a simulation routine to understand a target’s mental states and explain or predict the target’s behavior, the simulator must first determine which facts are relevant to the simulation and which facts are not. Moreover, the simulator must determine which facts of the simulation are idiosyncratic, pertaining only to her, and must quarantine them (i.e. refrain from attributing them to her target), thereby preventing them from contaminating the simulation.

Davies and Stone maintain that such determinations require theories. They say, “taking these variables into consideration requires reflecting on a number of *theoretical considerations*” [emphasis mine] (Davies and Stone, 1995a, p. 19). Elsewhere, to elucidate this point, they refer to Perner who discusses a cognitive developmental task known as the *Three Mountain Task*, which uses a tabletop model of a mountain scene to test for the ability to understand a visual array from another’s perspective.\(^{103}\) Davies and Stone say, “to imagine what a certain mountain scene looks like from your point of view, I may need to do some theorizing: “Since the big mountain is over there, he will see it on the right side; and of the small mountain farther away from him he will see only half behind the large one” (Perner, 1991, p. 269)” (Davies and Stone, 1995b, p. 23).

Their claim is that to understand another’s visual perspective, one must rely on some form of

---

\(^{103}\) The Three Mountain task was devised by Piaget (1956) to test whether a child’s thinking was egocentric. To complete the task, the child sits at a table before a tabletop model of three mountains. The mountains are of different heights and each has a different feature: one with snow on top, one with a hut on top and another with a red cross on top. The child is allowed to walk round the model, to look at it, and then sits down at one side of the table. A doll is then placed at various positions of the table. The child is then shown photographs of the mountains taken from different positions, and asked to indicate which photograph shows the doll’s visual perspective. Piaget assumed that if the child correctly picked out the card showing the doll’s view, she was not egocentric. If, however, the child picked the card showing her own visual perspective, Piaget assumed she was egocentric.
theorizing and that this cannot be accomplished by simulation – that putting oneself in the shoes of another is insufficient to predict the visual perspective of a target and a theory is required to adequately complete the task. I suppose that what they have in mind here is that, to understand the mountain scene from a different perspective, an attributor will (probably) need to consult some theory of (perhaps) geometry and triangulation. If they were correct in this assumption, they would be correct in maintaining that imagining the mountain scene from the point of view of the target requires some reflection on theoretical considerations. The question is: Does understanding how the mountain scene appears from the perspective of another require a theory? More importantly: Is their objection that simulation cannot afford an attributor the ability to predict a visual array from an alternate perspective (and, thus, that the ST fails) compelling? I do not think the objection is compelling for two reasons. First, even if some theorizing is needed to understand another’s visual perspective, this would not undermine the ST because certainly a version of the ST could allow for some theorizing in visual perspective taking, but deny that we rely on theorizing when we explain and predict the behavior of another. Secondly, I think that there is an alternate explanation of how one might imagine a mountain scene from another’s point of view – one that does not include any theoretical considerations. Certainly, the ST can explain how one can attribute a certain perspective of the mountain scene to a target without ever consulting a theory. Given this, it follows that we can simulate humans without relying on a theory.

Let me explain this. First, I wish to show that understanding a target’s perceptual states – what they can and cannot see – is relevant to the discussion as one aspect of mentalistic understanding, and, more importantly, can be accomplished and executed by some means other than theorizing, namely by simulating. Nichols et al. hold that any cognitive component can be
taken offline. They say, “any cognitive component can be viewed as a possible engine for simulation and (in principle) be taken “off line” – that is, detached from its usual function – to perform some other function or produce another capacity”(Nichols et al., 1996, p. 41). If they are right, then it follows that we can take our visual component offline and use it (qua mental imagery) to understand the visual array from the perspective of the target. This is bolstered by Currie, who argues specifically that mental imagery derives from offline simulation. He says, “an episode of mental imagery runs the visual system offline, disconnected from standard inputs and outputs … just as the simulation of belief and desire can mirror the causation of affective states by real belief and desire [in a target], so having a mental image of something [e.g. a mountain scene from another’s perspective] can produce the affective states that [actually seeing it] tends to cause” (See Currie, 1995 & 1997). So it seems that on Currie’s account understanding the mountain scene from the perspective of another involves a certain amount of visual imagery as mental imagery and that engaging in such mental imagery to understand a target’s visual perspective can be accomplished on the simulation account.

Currie conceded that the foregoing could be a problem for the ST because people with ASD seem to be able to tell whether another person, with a different visual perspective than her own, can see something. But, according to the ST, we would expect people with ASD to do poorly on a visual perspective test owed to their inability to simulate accurately. Currie indicates, “one possible response suggested to me by Simon Blackburn: knowing where another creature is looking has such significance for survival that it is plausible that there is a dedicated mechanism hard-wired to solve line-of-sight- problems” (Currie, 1996, pg. 250). In fact, Currie and Blackburn are right because Baron-Cohen’s work points to precisely this: a dedicated mechanism
or pair of mechanisms that he claims are both hard-wired and used to solve line-of-sight problems.

As mentioned, Baron-Cohen discusses four discrete subsystems that serve as separate components of the human mindreading system (Baron-Cohen, 1995, pp. 31–58). These subsystems are processing subsystems specialized for tracking properties of the world. Importantly, these subsystems are partial components that inform the ToM. Even more importantly, on Baron-Cohen’s account of them, the subsystems do not seem to rely on any form of theorizing whatsoever. Instead, the subsystems rely on simulating.

As I said, I take it that Davies and Stone are asserting that a simulator will need to refer (probably) to some theory of geometry and triangulation in order to imagine a certain mountain scene from the viewpoint of another. Were this the case, it would undeniably follow that some form of theorizing was being employed to imagine the mountain scene from the perspective of the target. What I am proposing here is that the mechanisms used to accomplish this are not theoretical in any sense. Rather, the mechanisms that can accomplish this task are two of Baron-Cohen’s discrete subsystems. Specifically, I suggest that the subsystems of Eye Direction Detector (EDD) and Shared Attention Mechanism (SAM) aid in ascertaining the view of the mountain scene from the target’s perspective and, further, that these mechanisms do not require any theory, tacit or otherwise. Let me explain how these subsystems could work together in lieu of theorizing to allow one to imagine the mountain scene from another’s perspective.

Baron-Cohen describes EDD as a specialized part of the human visual system that detects the presence of eyes and computes whether eyes are directed toward it or toward something else (Ibid, p. 39). EDD is relevant to this discussion because EDD can explain how it is that a

---

104 For more details see this dissertation Chapter 4, Section 3.1.
simulator can detect what her target is looking at without reference to any sort of theory. EDD seems to function automatically in a reflex-like fashion. The evidence that EDD is not theory-driven comes from a vast amount of empirical data showing that even 2-month-old infants detect the eyes and the eye directions of their mothers, especially when nursing (Ibid) and surely, it would seem ludicrous to infer that 2-month-old infants are employing a rational theory to detect their mother’s eyes and the direction of them. Baron-Cohen goes on to say, “For the infant, then, I assume that the eyes must seem to pop out of the face it is observing. This is probably the case for adults too” (Ibid). In adults, EDD is more sophisticated because adults can compute whether the eyes of the target are directed at something else. According to Baron-Cohen, adults do this via simulation. Specifically, they “infer from their own case that if another organism’s eyes are directed at something then that organism sees that thing” (Ibid). The adult, in employing EDD takes her own experiences and “generalizes them to [the target] by analogy with the self” (Baron-Cohen, 1995, p. 43). This just is a case of simulation.

The second subsystem that I suggest is involved when an agent considers the mountain scene from the perspective of her target is SAM. SAM, in conjunction with EDD, helps the simulator to build triadic representations (Baron-Cohen, 1995, p. 44). A triadic representation specifies the relations between the Self, a target, and a third object. In this case, the triadic representation would specify that the simulator and the target are both attending to the same object: the mountain scene. Baron-Cohen suggests that SAM operates first by gathering information from EDD and then by the simulator “comparing the [target’s] perceptual state with the Self’s current perceptual state” (Ibid p. 46). SAM takes the dyadic representations of the Self’s current perceptual state and maps them onto (or fuses them with) the analogously assumed dyadic representations of the target’s perceptual state, creating the triadic representation. Nowhere does
Baron-Cohen say (or even imply) that any measure of theorizing (tacit or otherwise) is required for the use of EDD, SAM, or EDD and SAM together. In fact, Baron-Cohen again suggests that SAM works by comparing the perceived perceptual state of the target to one’s own perceptual state and using oneself as a model for what the target must see. Further, I suggest that theorizing is not even needed for the use of SAM and EDD. This is a function that not only infants can engage in, but also one that animals can and do engage in. When a lioness on the Serengeti sees a gazelle, she stalks her prey in cooperative efforts with another lioness. Each lioness must detect the eye gaze of the other and determine that they are attending to the same gazelle. Surely, the lionesses are not engaging in some high-level rational theorizing to ascertain this prior to coordinating their hunt to pounce upon their prey.

In short, while Davies and Stone claim that some theorizing is required to imagine what the tabletop mountain scene model looks like from a target’s point of view, I maintain that no such theorizing is necessary. The simulator’s recognition of the target’s eye gaze and resultant visual perception can be executed by EDD in conjunction with SAM, and together these comprise a simulative process, one that exploits the same neural mechanisms used when the simulator herself undergoes the same perceptions.

Given Baron-Cohen’s EDD and SAM, we have a possible explanation of how an attributor might accomplish understanding the mountain scene from a target’s perspective without ever consulting or referring to a theory: We employ highly automatic and reflexive cognitive mechanisms to do so. ToM capabilities and theoretical activity are reserved, as a rational activity, to human beings and, perhaps, the higher primates. Infants and even animals – neither of which are developed or skilled enough to engage in theorizing about the mental states of others – can perform the tasks of understanding what a target it looking at, what the target is attending to,
and what it must look like from the perspective of the target. *No theory – tacit or otherwise – is required* to perform this task. So, upon the whole, theorizing is not required to satisfactorily assure an attributor of her understanding of the mountain scene from the perspective of her target. Moreover, as I mentioned previously, even some form of theorizing is required for such visual perspective taking tasks, this in no way undermines my case for simulation, which claims merely that no theorizing is required for mental state attributions.

I suppose that Davies and Stone could object by saying that an attributor must rely on some form of theorizing even just to pick out a target and to be sure that she is simulating the correct target. Here again, the simulation theorist could grant that this is the case and it still would not undermine the ST claim that theorizing is not required for mental state attribution. Even so, I dare say that no theory is required. My internal focal processes and decision-making cognitive processes have already identified and selected my target. For example, let us assume that I observe four people in line at the grocery store, three of them are waiting quietly and patiently in line, but one of them is shifting her weight, while huffing and puffing and repeatedly looking at her watch. *No theory is needed* to select my target because her motions have drawn my attention to her. I do not need to employ a theory to pick her out of the line as the one I wish to attribute mental states to any more than I would need to play eeny-meeny-miney-moe to select her as my target, and I suspect that such selection is highly automatic in nature. My observations of her behavior (or perhaps my idle curiosity) have drawn my attention to her and I have zeroed in on her as a target without theorizing. And what happens in terms of my mentalizing about her and my resultant behavioral predictions can all come as a result of the process-driven simulation that is first instantiated by my attention mechanism, none of which requires a theory for deployment.
No theory is required for (what appears to be) highly automatic and non-reflective target selection or for the mentalizing which follows target selection.

A second version of the collapse argument claims that theorizing is a necessary component of simulating. So, what the simulation theorist calls “simulating” involves some theory. This objection comes largely from Dennett (1987), who wondered, for example, how one could engage in any sort of simulational routine without in essence theorizing. In his oft-quoted explication, Dennett says,

If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what ‘comes to me’ in my make believe state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges. Why should my making believe I have your beliefs be any different? In both cases, knowledge of the imitated object is needed to drive the make-believe “simulation,” and the knowledge must be organized into something rather like a theory (Dennett, 1987, p. 100).

So, I take it that Dennett is saying that when one attempts to imagine oneself as something else (such as a bridge), the accuracy of said imaginings will be dependent upon how savvy one is regarding the inner workings of and scientific laws relevant to that thing which is under examination. Moreover, especially in his example of the suspension bridge, Dennett is supposing that one would need a sophisticated understanding of both physics and engineering (each of which indubitably require a great deal of theoretical knowledge) in order to understand and accurately predict the behavior of the suspension bridge under the certain circumstance of wind.

I will grant the case of the suspension bridge to Dennett: to understand and predict how a certain suspension bridge will behave under certain conditions, such as various wind speeds, I would need to refer to and employ the theories of physics and engineering. It seems right to me that human beings will tend to employ theories when we are attempting to understand and predict the behavior of things, beings, or systems that are dramatically different from human
beings. I mentioned something like this earlier in this chapter when I discussed attempting to understand, explain, and predict the behavior of whales. I said that for me to do so, I would likely need to refer to and deploy some kind of theory concerning cetology and whale behavior. Similarly, I would grant that laws of physics, the laws of gravity, and the scientific laws of astronomy will allow me to understand the behavior of particles in the Kuiper belt and, subsequently, to predict their behavior (i.e. trajectories) in the event that they should collide with other, larger, particles within the belt. In my view, these things are quite without dispute: I am going to employ some theory to understand the behaviors of systems that are very different from my own such as in the cases of suspension bridges, whales, and asteroids.

However, because the systems of other human beings are presumably enough like my own system (and in highly relevant ways), I do not require a psychological theory to understand the mental states and predict the behavior of other human beings. If this claim were right, it would break the analogy that Dennet is making.

Suppose for example, that I take as my target a person from a different culture, with different features, of a different age and sex, who is a non-English speaker. This target, being of my species, is still going to be “like me” (unlike a bridge, whale, or asteroid) and we will share enough relevant features for me to accurately read at least some of his or her mental states and predict some of his or her behavior. A great deal of research has shown that the basic emotions anger, fear, disgust, amusement, sadness and surprise are easily and universally recognizable, suggesting that these emotions and their vocalizations are similar across all human cultures and shared by both sexes. So, as Goldman indicates, I will easily engage in low-level (i.e. automatic or reflexive) simulations regarding the unfamiliar person and will do so with a fair amount of accuracy. Further, I think that even some high-level (i.e. highly conscious and voluntary)
simulations regarding the complete stranger will also be successful because of the universally recognizable nature of the basic emotions.

For example, if I focus my attention on the complete stranger in a restaurant taking her first bite of food from her plate, I will be able to identify her basic underlying emotions because they will be betrayed by her facial expressions and vocalizations. Witnessing her expressions and vocalizations (and also referring to myself in a conscious, high-level simulation), I will be able to detect whether she likes the dish, or not. I will be able to detect pleasure or disgust and, engaging in such high-level simulation, I will be able to predict her behavior with relative ease: If she likes the dish, she will take another bite. But if she appears disgusted, she will not take another bite and may even push the plate away. Similarly, to invoke an early example of Gordon’s, I need no psychological theory to predict how strangers, such as Mr. Tees and Mr. Crane, will feel about and react to missing their respective flights after a prolonged taxi ride.

This is in stark difference from performing a simulation regarding Dennett’s suspension bridge. When attempting to simulate systems, entities, or beings that are radically or fundamentally different than myself, I will – of course – refer to or depend upon a theory to understand it and predict its behavior. Notably, Dennett’s simulation example does not even pertain to a human being, but to an inanimate object – one that is indubitably subject to the laws of science. Understanding these systems and predicting the behavior of these systems will require theory-driven simulations.

However, and in strong contrast, to understand the mental states and predict the behavior of a fellow human being, I need no such theory, since my own systems that are, ostensibly, similar enough to those of my target are available to me to use in the simulative activity. Goldman says, “If a simulating system resembles the target, [the simulation] might succeed by engaging some
of the processes or operations that it shares with the target. It won’t need a theory to do this, neither a theory of the target nor a theory of itself” (Goldman, 2006, p. 32). My own systems will suffice as a model of my target because my target is “enough like me” that the model (in this case: myself) will suffice to understand the being under observation. In fact, I can really only simulate targets that are “like me” with any accuracy. Noting the same, Gallese says, “A subject’s mental states or processes simulate another’s mental states or processes just in case it resembles the second state or process in some significant respect” (Gallese & Sinigaglia, 2011, pg. 513) (See also Goldman: 2006, 2008 and 2009). Note that my own systems will never suffice as a model for a suspension bridge or a humpback whale because I am nothing like a suspension bridge or a humpback whale. Those systems are fundamentally different from my own and resemblance is a key factor in successful simulation.

Goldman says that process-driven simulation of a target can be successful in producing a final state that is identical or isomorphic to that of the target, provided that the attributor and target are driven by the same or relevantly similar cognitive processes and share identical or relevantly similar initial states (Goldman, 2006, p. 32 paraphrased). I can do this with ease and with fair acuity because the cognitive processes of another typically developed human being are (I assume) extremely similar to my own cognitive processes. Note how very different this is from attempting to understand (and predict the behaviors of) entities with radically different cognitive processes, such as whales, or entities with no cognitive processes whatsoever, such as

---

105 For now, I leave aside questions regarding whether every human is “enough like me.” Obviously, there will be cases in which the human target is not “enough like me” for a simulation routine to be successful (e.g. cases of severe mental retardation or mental illness).

106 Notice that targets bear empirically discoverable properties, the observance of which can cause the attributor to either treat them as minded and therefore as targets subject to the simulation heuristic, or to treat them as not-minded and therefore not as targets subject to the simulation heuristic. Knowledge or even assumption of ‘resemblance-to-me’ is not a prerequisite to engaging in the simulation routine; it is just that the simulations routine can be successful if ‘resemblance-to-me’ conditions exist.
suspension bridges and asteroids. Goldman’s distinction between theory-driven simulation and process-driven simulation helps to undermine Dennett’s ‘suspension bridge’ counterexample because it illustrates a specific difference in the way that we understand different systems. We must theorize to understand how Dennett’s bridge will behave during gusts of high winds, but we can use our own internal cognitive processes to understand the mental states and subsequent behaviors other human beings. We can grant to Dennett that theories are required to understand suspension bridges and the like, yet retain the claim that no theory is needed to underscore our ToM ability. Process-driven simulations are free of theorizing.

A third and final version of the collapse objection comes from Churchland. Notably, Churchland is by no means a friend of the TT. However, in his argument against the TT he offers a discussion of simulating and theorizing, in which he says,

A simulation itself, even a successful one, provides no explanation [of human behavior]. What explanatory understanding requires is an appreciation of the general patterns that comprehend the individual events in both cases [of first-person and third-person understanding]. And that brings us back to the idea of a moderately general theory (Churchland, 2006, p. 320).

So, according to Churchland, perhaps a simulation can begin and proceed without theorizing (contrary to Davies and Stone and Perner), and perhaps a simulating is not just the same thing as theorizing (contrary to Dennett), but – in any case – simulations do not provide explanations, and deriving explanations relies upon an understanding of general patterns which, in turn, requires a theory. Thus, Churchland claims that the ST collapses into the TT because, in the end, the ST cannot provide an explanation for how we explain and predict behavior. When it comes to
explaining human behavior, the simulator must rely on theorizing to form explanations regarding the behavior of their target because simulations do not provide explanations. ¹⁰⁷

I suggest that Churchland perhaps errs in a couple of ways. The elucidation of these errors will serve to undermine his claim.

It seems as though Churchland assumes that the kinds of simulations run by mental state attributors are relevantly similar to the class of simulations that are characteristically used in the physical and engineering sciences because the simulations used in the latter do not provide direct explanations, but only indirect explanations that are mediated by some form of theory. ¹⁰⁸ If the simulation routine used by mental state attributors were the same kind of simulations used in the physical and engineering sciences, then Churchland’s claim would be correct because the latter are dependent upon theories for explanatory power. However, it would be wrong to assume that the two kinds of simulations are relevantly similar because the kind of simulations employed for mindreading in particular (and in the social sciences in general) appear not to be as dependent upon theories for explanatory power in the way that simulations used in the physical and engineering sciences are so dependent. In the case of mindreading simulations, the mental state attributor (qua simulator) has a prominent explanatory role. This is because the underlying cognitive processes of the simulator are relevant to those of the target and as such are instrumental in contributing explanations of the anticipated behavior of the target. Hence, the attributor need not refer to the patterns of her target to form her explanations of the target’s

¹⁰⁷ To clarify: Davies & Stone and Perner claim that the ST collapses because theorizing is required before simulation can begin. Dennett claims that the ST collapses because simulation does not seem to work without being a kind of theorizing. And Churchland claims that the ST collapses because, simulationists are trying to provide an explanation for how we explain and predict behavior of targets and explanations require theories.

¹⁰⁸ This comparison is used by Richard David-Rus (2014) who distinguishes the general explanatory virtues of simulations and simulational routines used in social sciences from the kinds of simulations used in the physical and engineering sciences.
mental states and behaviors, but she can refer to herself and her own cognitive processes as a model instead.

The simulator uses herself as well as her own cognitive processes as a manipulable model of the target. Notably, the simulation run by the attributor, in the capacity of a model, can be used in multiple ways and in various contexts. For example, the simulation routine could be used to determine causes and causal factors of the behavior of the model, and so effectively answering ‘why x?’ about the causes and causal factors of target’s behavior, and could do so without theorizing. Furthermore, because Goldman allows for the analogical inference ‘from me to you’ and for introspection, he, especially, has an additional response to Churchland’s objection available to him. Goldman could say, for example, that what the simulator learns about herself via introspection (whilst using herself as a model of the target) can be immediately put to use in interpreting and explaining the mental states and behavior of the target. Moreover I think that the simulator could accomplish the foregoing without reviewing (or even establishing) patterns of her own behavior because often one can explain and predict one’s own behavior in a novel situation – one in which no pattern has been established – by so-called “knee jerk reactions” (those that are automatic or immediate reactions made without examining the causes or facts) to a given situation. So, introspective reflection upon the simulator’s first-person experiences and mental states will serve as a platform for understanding and explaining the mental states and behaviors of the target.

Additionally, Churchland errs by providing a false premise in his argument: that simulations do not provide explanations. It has long been thought that theories are the bearers of explanations, generating understanding of the object, system, process, or being under observation and inquiry. However, the long (and perhaps fatigued) debate over explanation in philosophy of
science has revealed that *theories are not* the only bearers of explanation and that, in fact, *simulation models can also be considered bearers of explanation*. Winsberg, for example, claims that, “simulations are used for heuristic purposes, for predictions, and for generating an understanding of targets and their behavior” (Winsberg, 2015). I take it that generating an understanding of the target includes understanding the behavior of the target for prediction purposes (i.e. providing an explanation of the target’s behavior). Saying that simulations in general can be used for “*generating understanding*” of a target’s behavior is just to say that simulations generate explanations and, hence, that simulations can be considered to have explanatory power. Contrary to what Churchland claims, simulations *can and do* provide explanations. So, if (as Winsberg claims) simulations are capable of predicting a target’s behavior and generating an understanding it, and if such understanding is rightfully considered an explanation, then, contrary to Churchland’s claim, simulations can and do provide explanations.

I am not alone in my assertion that simulations yield explanations. Weirich (2011) also discusses the explanatory power of models and simulations. Where one could hold that simulations can only yield functional explanations or potential explanations, Weirich holds that simulations provide objective explanations and not just tentative ones or mere explanatory attempts. Simulations rely on their underlying models. Simulations draw on the explanatory power of the underlying model by identifying the factors at work in the model and then providing explanations by identifying the factors at work in the target.\(^\text{109}\) Again, the simulator uses herself and her own cognitive processes as a manipulable model of the target, and the

\(^{109}\) Weirich’s work considers the nature of simulations in general, varieties of simulation, and uses of simulations for representation, prediction, and explanation of a variety of kinds of targets. While he does not specifically discuss the kinds of simulations employed in a simulative routine (as suggested by proponents of the ST), I nevertheless take his findings to be analogous to the case of simulations used for ToM purposes.
introspective attributor can reflect upon her own mental states to explain and predict the mental states and behaviors of her target.

The foregoing has shown that the ST offers an explanation of ToM as simulation-based and can do so without reliance upon theoreticity or collapsing into the TT. I need to be careful here by stating that, while I do think that the ST most adequately explains that which underpins our apparent ability to mindread, I am not making the claim that simulation routines are not prone to erroneous outcomes. I do not know of any simulationist who claims that simulations run perfectly or that they consistently produce accurate understandings of a target’s mental states or accurate predictions of a target’s behavior. Indeed, the simulation routine prescribed by the ST can produce inaccurate results, which I will discuss in the next section.

5.5. Inaccurate or Failed Mindreading on the Simulation Theory Account

At this point in the chapter, there is an important empirical issue to which I must attend. One might ask, under the ST interpretation of ToM, to what kinds of errors is mentalizing prone? That is, when employing simulation for mindreading purposes, what is apt to go wrong and why might that be the case? By all ToM accounts, whether by theory theorists or by simulationists, there is widespread agreement that we both attribute mental states to others and also experience having mental states, which we ascribe to ourselves. However, the role that first-person experience of mental states plays in the mindreading of others is neither perfectly clear nor agreed upon. Questions remain concerning the priority and role of first-person experience and self-attribution of one’s own mental states. Specifically, contrary to Gordon’s claims, must we first be in possession of relevant mental state concepts, introspectively aware of our own mental states, and be able to attribute it to ourselves before we are able to attribute the same (or a
similar) mental state to another person? Baron-Cohen, for example, asks, “Are we first aware of our own mental states and then extend this experience to others by analogy, merely “simulating” being them when we want to take into account what they might be thinking or feeling? (Baron-Cohen, 1995, p. 130).

My position is that this is the case. As stated previously in this chapter, Goldman’s account includes and indeed depends upon the notion of introspection. I concur with Goldman’s account. It seems to me that, during the course of cognitive development, one is first introspectively aware of one’s own first-person experience and is first able to attribute mental states to oneself, and that this comes prior to the ability to attribute mental states to others. I think that precisely because we are first aware of our own first-person experiences and our own mental states via introspection, our mentalizing may be prone to a specific kind of error, which I discuss below.

Separately, Epley and Caruso, in particular, have done an abundance of work in this area (2009). They claim, “there are three critical barriers to having a ToM or empathic accuracy: (1) activating the ability to mentalize, (2) adjusting an egocentric default, and (3) accessing accurate information about others” (Epley and Caruso, 2009, p. 296). These are each necessary for ToM and so failure to do any one of these three will indubitably result in inaccurate mindreading. It is my contention that it is failing to adjust for an egocentric default that lies at the heart of many errors of the errors that people with ASD experience in mentalizing. This failure can be seen as a failure to quarantine (or screen off) one’s own idiosyncratic mental states.

---

110 In psychology, ‘empathic’ accuracy refers to how accurately one person can infer the thoughts and feelings of another person and is seen as an important aspect of “everyday mind reading.” Research on empathic accuracy has explored its relationship with mentalizing. It is generally held that in order to accurately infer another’s psychological state, one must be able to both share that state (affect sharing), and understand cognitively how to label that state (mentalizing).

111 Typically, when evaluating the mental states of others, one uses one’s own mental states as the starting point of the judgment by default. For accurate mentalizing, this ‘egocentric default’ must be subsequently corrected or adjusted to reflect the mental states of the target (rather than one’s own idiosyncratic mental states).
allowing one’s own mental states to leak into the simulation, thereby contaminating it and erroneously attributing a particular mental state to the target or a failing to accurately predict the behavior of the target. Because Goldman’s version of the ST requires introspection for simulation, his account can offer this explanation of failed simulation: It comes as a result of failing to quarantine when engaging in high-level mindreading tasks.\textsuperscript{112} Let me explain.

In the case of high-level mindreading, a mindreader attempts to construct pretend or unreal mental states that correspond to those of the target. It is preferable to characterize pretense not as a distinct propositional attitude but, rather, as a kind of cognitive operation or process the outputs of which are pretend states that are generated in imagination.\textsuperscript{113} Pretense enters the simulation routine when an attributor generates pretend mental states of a target so as to predict a target’s decision to do a certain action. Because the “mental states” generated are those of the target, they are merely pretend states of the simulator. The simulator pretends to have the same initial mental state of the target and makes a decision based on those pretend states. Having made a decision in pretend mode, she predicts that her target will decide to take a certain action. When mindreading accurately, a simulator generates a pretend mental state in herself – that is, she generates the mental state she \textit{would} have if she were her target, but (again) it is merely a pretend state because it is not the actual mental state of the simulator. This pretend mental state is then attributed to the target as the target’s actual mental state, but the mental state is merely pretended

\footnotesize
\textsuperscript{112} I take it that more complex mindreading tasks, by their nature, require high-level simulation routines and so I refer to them as ‘high-level mindreading tasks’ here. Please see this dissertation, section 5.3.2, for a discussion of high-level versus low-level simulation.

\textsuperscript{113} Goldman says, “If pretense is a distinct attitude, how can we intelligently talk about pretend-belief, pretend-desire, pretend-hope, and so forth as simulationists are wont to do? If pretense is a separate attitude, each hyphenated phrase would designate a compound attitude, and it is unclear what such a compound would be” (Goldman, 2006, p. 47).

211
by the simulator since it is not her actual mental state. By contrast, it seems that an attributor’s inability to engage in pretense (enough to put herself in her target’s shoes) puts her simulations at risk of being contaminated by her own actual mental states – mental states which must be quarantined in order to attribute mental states to others and predict their ensuing behavior. The ability to engage in pretense is critical for accuracy of the simulation routine because without this ability one is prone to make certain kinds of errors.

Interestingly, young children especially, but also occasionally even adults, might sometimes exhibit difficulty distinguishing between pretense and reality, yielding these specific errors. Pretending requires the ability to “act as if” something is the case, whether it is the case or not. Furthermore, pretending requires the ability to “tell the difference” between what is real and what is merely being pretended while it is being pretended (Leslie, 1987, p. 413). In the case of simulative role taking, according to Gordon, when pretense and reality get blurred then we should expect a quite distinctive and weird sort of error: a blurring of the distinction between oneself and the other” (Gordon, 1995b, p. 731).

The ability to engage in pretense (by itself) is not enough to prevent this kind of error. In order for a simulation to yield accurate mental state attributions, the simulator must also be able to quarantine (or ‘bracket’) all or most of her own actual mental states (unless they are also shared by the target), preventing them from entering the simulation. If the simulator is unable to quarantine her mental states during simulation, she may blur her real mental states with the pretend mental states she is trying to assume, again producing the “quite distinctive and weird sort of error” that Gordon mentions – that is, she will blur the distinction between herself and her

---

114 She is, in pretense, momentarily transforming herself into being her target so as to take on the mental states of her target. It is a pretend state because no actual output of behavior results from said transformation.
target, much in the same way that an actor can blur the distinction between a role she is playing and her private life.

However, introspection aids in preventing this distinctive and weird sort of blurring error. Introspection can facilitate the quarantining of one’s own idiosyncratic mental states and prevent the blur between one’s real mental states and the mental states generated in pretense that are as attributable to a target. If appropriate introspection is required to produce accurate mindreading during simulative routines, then errors in introspection may be the culprit responsible for inaccurate mindreading with respect to failed or misfired simulation. There are three ways that introspection might play a role in failed simulation, which result in inaccurate mindreading. First, one must be able to introspect in the first place and, in so doing, recognize one’s own mental states and identify them as one’s own. Inability to know one’s own real mental states would surely impede one’s ability to differentiate them from the real or pretended mental states of one’s target. So, outright failure to introspect may be responsible for a failure to engage in the kind of pretense that is necessary to enter into the simulation routine in the first place.

Second, during simulative role taking, one must not only introspect in order to recognize and identify one’s own mental states, but also one must be able to select them as the particular set of mental states that must be quarantined. Erroneous selection of mental states during introspection may be responsible for failing to quarantine and effectively screening off one’s own idiosyncratic mental states so as to accurately take on the (real or pretended) mental states of the target. One must first introspect to know what one’s own mental states are in order to then accurately select them as the particular set of mental states that must be quarantined, knowing that they belong to her but not the target.
Failing to differentiate one’s mental states from those of the target will result in the simulator’s first-person experiences and first-person mental state attributions leaking into and contaminating the simulation. One must first ‘look inward’ via introspection, recognize a particular mental state or set of mental states as belonging to oneself, distinguish them from those which belong to the target, quarantine one’s own mental states, and generate pretend mental states to attribute to her target. Failure to do so results in the exact kind of blurring between oneself and one’s target that Gordon discusses. However, via introspection the simulator can recognize and identify her own mental states, select the appropriate idiosyncratic mental states to quarantine or bracket from the simulation routine, and then engage in pretense in order to attribute pretend mental states to her target.

To be clear: the aforementioned introspection failures are not the only kind that result in specifically egocentric errors – errors in which an attributor ascribes mental states to others in a way that reflects her own mental states (e.g. failing to recognize which mental states belong to her and distinguishing them from those of the target). There is a third kind of introspection failure that is an especial culprit in these cases. What is primarily responsible for egocentric errors is what I shall call “hyper-introspection.”

Hyper-introspection occurs when one is constantly looking inward, observing and examining one’s own mental and emotional states and mental processes. It is an over-concentration on one’s own view of reality and one’s own mental states, resulting in an inability to “get out of one’s self” sufficiently enough to engage in pretense – the ability to engage in pretense is effectively disabled. Adequately screening off this egocentric hyper-self-awareness becomes very difficult or, in some cases, impossible. If one is a hyper-introspectionist, then (a) this inability to distance one’s self from one’s own view of reality may preclude engaging in the pretense required to take
on the perspective of another, and (b) one might be too focused on one’s own mental states to even be able to quarantine them.

Something similar to the hyper-introspectionism that I am suggesting is found in people who suffer from narcissistic personality disorder (hereafter, ‘NPD’), and with similar results. NPD is characterized by an enduring inner-experience, impaired empathy, and a pattern of self-centered egotistical behavior, making it difficult or impossible to recognize or identify with the feelings of others (DSM IV, 2013). Often when a person suffers from the extreme self-centeredness and self-importance that is generally associated with NPD, that person may be so overly self-focused (or so overly focused on her own life and her own problems) that she is unable to attend to the problems of others or to understand the plights of others. My contention is that hyper-introspection (while attempting third-person mental state attribution) yields a similar “self-centered” result – that is, a hyper-introspectionist may not be able to see beyond her own mental states and may not be able to screen them off so as to prevent them from contaminating the simulation routine.

Hyper-introspectionism constitutes one of Epley and Caruso’s “critical barriers” to accurate simulation because hyper-introspectionism would make it difficult, if not impossible, for the simulator to “adjust for the egocentric default.” Our default position, according to Epley and Caruso, is introspective awareness of our own egocentric and idiosyncratic mental states and processes. In order to engage in a simulation that results in accurate mindreading, the simulator must adjust her focus from her default focus on her own mental states and processes to a focus on the imagined mental states and processes of her target. A failure to adjust the egocentric default and a failure to re-center the focus of mentalizing onto the target would result in misfired or an altogether failed simulation.
The ST easily accounts for erroneous mindreading that reflects the mental states of the attributor as ‘misfired simulations.’ According to Goldman, “When genuine states [of the simulator] mistakenly seep into the initial set of pretend states, mistaken attributions (to the target) are all too likely. Thus, ST readily predicts a pattern of error that is empirically observed. On the other hand, it is hard for the TT (without ad hoc assumptions) to make the same prediction. This is why the empirical findings support ST” (Goldman, 2008, p. 785). The TT cannot so easily account for such erroneous or egocentric mindreading. I imagine that the theory theorist would have to supply some theory-based account explaining the projection of an attributor’s own mental states onto her target. To my knowledge, nowhere does any theory theorist offer such an account of how it is or why it is that an attributor’s own mental states may seep into theory-based mindreading. When an attributor ascribes mental states to others in a way that reflects her own current mental states, her mental states will have a causal impact on their attributions that they should not have - an egocentric bias (i.e. a bias that results from failing to inhibit one’s own genuine mental states) can easily occur. This is explainable on the ST account, but not on the TT account, since no TT account includes an inference “from me to you” in the way that the ST includes such inferences.

The idea that quarantine or bracketing is required for successful simulation is supported elsewhere in the literature as well. For example, Flanagan aptly notes that the way we understand each other is not necessarily by having each other’s experiences but by conceiving of each other’s experiences in the imagination. He says, “What we speak of as taking on the point of view of another involves imaginatively taking on what we think things are like for the other, and this typically requires bracketing out to some extent what things are like for ourselves” (Flanagan, 1992, p. 106). Additionally, paraphrasing Batson, if one focuses too much on one’s
own feelings (or mental states), one might erroneously project them onto the target. For example, Batson says, “a simulator may lose sight of whether or not the target is in distress over something and, instead, focus on what the character of her own distress would be in that situation” (Batson, 2009, p. 274).

This focus on the self may evoke egocentric simulation as a result of the failure to quarantine. A version of the ST that includes introspection can accommodate an account of hyper-introspectionism,\(^{115}\) and predicts that if one fails to quarantine one’s own mental states when constructing pretend input states, then egocentric biases in mental state attribution can result (Goldman & Shanton, 2010).\(^{116}\) Goldman discusses this at length in his more recent work. He says,

For simulation to be successful and accurate, the exercise must not only include pretend or surrogate states that correspond to those of the target but also exclude the mindreader’s own genuine states that don’t correspond to the ones of the target. This implies the possibility of two kinds of error or failure; failure to include states possessed by the target and failure to exclude states lacked by the target. The second type of error will occur if a mindreader allows a genuine state of his own, which he “knows” that the target lacks, to creep into the simulation and contaminate it. This is called quarantine failure. There is strong evidence that quarantine failure is a serious problem for mental-state attributors. This supports ST because quarantine failure is a likely affliction if mindreading is executed by simulation but should pose no comparable threat if mindreading is executed by theorizing (Goldman, 2012).

The ability, then, to produce accurate inferences via simulation requires the skill to collect a set of potentially ambiguous cues about the target’s mental states, pair them with and parse them from one’s own mental states, and project them in imagination onto the target (Myers and

\(^{115}\) Obviously, any version of the ST that does not include an account of introspection would not accept hyper-introspectionism as a viable explanation for the simulator’s failure to quarantine her own idiosyncratic mental states. My contention is that a version of the ST, such as Goldman’s, which includes introspection, can account for cases where either too much introspection or not enough introspection negatively impacts the accuracy of the simulation output.

\(^{116}\) Note that Goldman and Shanton do not use the term ‘hyper-introspectionism’. This term is one that I employ.
Hodges, 2009, p. 281). To simulate accurately, it seems that we need to keep two emotional response systems up and running simultaneously – that is, our own and that of our target. A simulator must decouple the emotions of others from her own; that is, she must quarantine (e.g. screen off, filter, or bracket) her own emotions and emotional systems in order to understand those of others in such a way as to accurately explain and predict the behavior of the target in a way that is non-idiiosyncratic.

My position is that to prevent egocentric errors in mindreading, one must be able to do all of these things: parse a target’s mental states from one’s own; keep two emotional response systems up and running simultaneously; decouple the mental states of others from one’s own mental states; and quarantine one’s own mental states and emotions so that they do not contaminate a simulation routine. To do all of these things and avoid egocentric errors in mindreading, one must be introspecting. Further, to know one’s own mental states, to differentiate between the two systems (one’s own versus the target’s), and to recognize one’s own mental states (as one’s own) so as to be able to bracket them during the simulation, one must be introspecting. Moreover, if one is either hypo-introspectionist (i.e. one who introspects less than normal) or hyper-introspectionist (i.e. one who introspects in an excessive or exaggerated way), the simulation will misfire and the mindreading will be inaccurate. In the case of hypo-introspection, the simulator is either incapable of introspection or introspects in an impaired way. As a result, the hypo-introspectionist will be unable to successfully simulate because, as Goldman insists, introspection is a necessary component for both first-person and third-person mindreading tasks. In the case of specifically hyper-introspection, the kind of mindreading errors that will most likely occur will be egocentric in nature. This is because the hyper-introspectionist person is
constantly ‘looking inward’ and her mental state attributions will be tainted by an egocentric bias.

The egocentric mindreading errors of a hyper-introspectionist person, which reflect the past or current mental states of the simulator instead of the target, are fairly prevalent. This is a point that supports the simulationist account of high-level mindreading, according to Goldman. He says, “their own mental states somehow have a causal impact on their attributions that they shouldn’t have, yielding an inappropriate tilt or bias. …An example of the bias is the so-called “curse of knowledge,” in which an attributor informed of a target’s ignorance of a certain matter nonetheless seems to assume that the target has the same knowledge that she, the attributor has” (Goldman, 2008, p. 785).

It seems fair to ask why these quarantine failures are so prevalent. Here, Epley and Caruso are particularly clear, as they offer several possible causes that could explain such prevalence. First, they say that one’s own initial egocentric assessment of the target - and egocentric evaluation of the mental states at hand - must serve as the starting point of a judgment that is subsequently corrected or adjusted. “When asked how to predict how other people feel, people’s own current feelings heavily influence their evaluations… Such egocentric projections suggest that people make predictions about how others will feel by first imagining what they themselves would feel in the other’s situation” (Epley and Caruso, 2009, p. 300-301). I claim that if the simulator is hyper-introspectionist, she will “get stuck” in this first stage of first imagining how she will feel because – again - the hyper-introspectionist person is constantly ‘looking inward’. This “getting stuck” at the stages of simulation (that of first reflecting on her own feelings and mental states, or of first imagining what she would feel in the target’s situation) could result in a completely
egocentric assessment of her target’s mental states. Her hyper-introspectionist egocentrism would explain her failure to quarantine her own mental states during the simulation routine.

Second, a major barrier to accurate simulation is that it can be incredibly difficult for a simulator to first ignore her own perspective to try to experience, simulate, or infer the perceptions of another person. Epley and Caruso say that one’s egocentric experiences are “an easy default that can prevent us from considering the perceptions of a target in a different psychological state” (Epley and Caruso, Ibid, p. 297). The easy default judgment is that of one’s own perspective. “Overcoming any default in judgment requires deliberate reasoning that takes time, attention, and motivation to expend mental effort. Anything that hinders people’s ability to engage in such deliberate reasoning should therefore increase reliance on an egocentric default among perspective takers” (Epley and Caruso, Ibid.). I believe that hyper-introspectionism is sufficient to hinder a simulator’s ability to engage in the kind of reasoning Epley and Caruso describe because hyper-introspectionism would diminish the attributor’s ability to take the time and (especially) the attention to make the appropriate adjustments from her egocentric defaults. Hyper-introspectionism may even obliterate the motivation to overcome egocentric default judgments. Further, hyper-introspectionism is the exact kind of hindrance that would increase the simulator’s reliance on her egocentric default. So, it seems that hyper-introspectionism would most certainly exacerbate this difficulty often rendering the task of accurate mindreading nearly impossible.

Third, Epley and Caruso claim that the egocentric default must be adjusted for accurate simulation. “Insufficient adjustment can explain why one might tend to overestimate the extent to which others share their attitudes, beliefs, and knowledge” (Epley and Caruso, 2009, p. 302). Assessments based on one’s own ideology, emotion, direct experience, or preexisting attitudes
and knowledge, or other mental states must be either dismissed entirely or simply adjusted to reflect those of the target rather than of the self – even when the target may have a substantially different psychological point of view. “Overcoming egocentrism and one’s own current state is therefore the litmus test for attempted perspective taking” (Epley and Caruso, 2009, p. 297). Without said adjustment, the simulation can fall to the default perspective of “self” and the ensuing simulation is partially, if not completely, egocentric, reflecting the mental states of the simulator rather than those of the target. If the simulator is particularly hyper-introspectionist, she may often remain in the default mode of attention to “self” and may only be capable of egocentric and contaminated simulation.

This is not to say that we can really ever get completely out of our own way. I gather that our high-level mindreading will always have some default egocentric bias as its starting point. Probably, we quite often resort to imagining how we would react in the situation to guess how the target will react. Batson calls this “egocentric anchoring” and claims that it must be “followed by an adjustment” for successful simulation (Batson, 2009, p. 276).

The simulation story offers an explanation of how accurate mindreading is possible (Goldman, 2008). It shows how mindreading can occur without the attributor possessing or deploying any theory of the target’s mental operations. Moreover, the ST is prepared to adequately deal with, or explain, erroneous mental state attribution. The resources for this come especially from Goldman’s account of the ST because it includes the notion of introspection and introduces the problem of quarantine failure. It is Goldman’s insistence on introspection as a necessary condition to mindreading that opens the door to the possibility of hyper-introspectionism and the failure to quarantine as an explanation of egocentric and idiosyncratic simulation and its resultant contaminated (failed) mental state attribution.
In the next section, I note that there are some particular pathologies – such as ASD – in which quarantine failure is highly prevalent. In particular, I note that people with ASD exhibit hyper-introspectionism and so may consistently fail to make the required adjustment from the egocentric bias. The result of this is that people with ASD have extreme difficulty with mindreading tasks. This result comports with the portion of the diagnostic criteria for ASD describing the difficulty in understanding social cues regularly experienced by people who suffer from ASD. In the next section, I will further explore the idea that people with ASD struggle with egocentrism and mindreading that is tainted by an egocentric bias.

5.5.1. Egocentrism as Some Evidence for Autistic Hyper-Introspectionism

As stated previously, the case of ASD may have important implications for the broader debate regarding what mechanism, process, or ability underwrites the normal ToM, and, in fact, I will use ASD as a main point of adjudication (i.e. a deciding factor) for the debate between the TT and the ST. Currie is particularly illustrative on this point. He says, “Suppose, for example, that people with autism are strikingly deficient on a certain task T. And suppose we find strong empirical support for the hypothesis that this deficiency is caused by difficulties with simulation. Then we have evidence for the hypothesis that the performance of normal subjects on T is to be explained in part by simulative competence” (Currie, 1996, p. 244 emphasis mine). This is to say that if people with ASD cannot do task T (well or reliably), and if we establish that this is because they either cannot simulate or they cannot do so competently or skillfully, then we are entitled to claim that typically developed persons without ASD who perform well on task T do

117 Please see Chapter 3, Appendix K, section A and Appendix L for the specificities relating to the social deficits experienced by people with ASD from the DSM-V (2013).
so, at least in part, via simulation.\textsuperscript{118} If, however, proficiency on task T were explained by theoretical competence, and people with ASD are strikingly deficient at task T, we should expect that the difficulties people with ASD experience with task T is caused by difficulties with theorizing. But this is not the case. We know that people with ASD are strikingly deficient on ToM tasks. Since we know from Simon Baron-Cohen’s work on EMB (See Chapter 4, Section 3.4 and Section 4) that people with ASD tend to excel at systemizing and theorizing, it seems to follow that theorizing cannot underpin the ToM ability. In terms of the TT/ST debate, this essentially negates the possibility that the TT is the best account of ToM and holds the door open for the ST as the better account of ToM.

Additionally, we can review the host of core deficits that are central to ASD and we can ask the following question: How many of these deficits are as a result of a failure of ToM in the sense that ToM is underwritten by pretense and simulation? If it turns out that there is such a correlation between the core deficits of ASD and the absence of simulative competency, and if there is also good evidence that core deficits central to ASD can be linked to such incompetent simulation, then the ST gets a sort of confirmation. In what follows, I discuss the egocentrism of ASD, correlate that to hyper-introspectionism, and ultimately show that the ST is a favorable account of ToM.

In earlier works, both Baron-Cohen (1989) and Carruthers (1990b) claim that individuals with ASD seem to lack awareness of their own mental states. But this just seems not to be the case. Consider, for example, the autistic tendency toward radical self-absorption coupled with an apparent inability to discern when a listening audience is not equally absorbed in a subject matter

\textsuperscript{118} Accurate simulation, I claim, requires proficiently quarantining one’s own mental states and employing empathic accuracy.
upon which she is fixated. As I will explain below, a person with ASD tends to assume that the contents of her mind are identical to the contents of the mind of another, and that the other’s mental states are identical to her current mental states. Frith and Happé each call this “autistic egocentrism” and is consistent with my claim that persons with ASD are hyper-introspective. Autistic egocentrism or hyper-introspectionism may explain why, on the ST account, people with ASD apparently fail to quarantine their own idiosyncratic mental states in order to perform the simulation task necessary to accurately attribute mental states to their target. A brief discussion of scholarly work indicating such egocentrism (which I claim leads to hyper-introspectionism) is necessary to corroborate my claim.

Uta Frith has done an abundant amount of work regarding the extreme egocentrism generally found in people with ASD, and she poses some interesting and relevant hypotheses. She develops her work from the work of Asperger (1952), whom she quotes as saying, “…these children are egocentric in the extreme. They follow only their own wishes, interests, and spontaneous impulses, without considering restrictions or prescriptions imposed from the outside… they have a genuine defect in their understanding of the other person” (Asperger quoted in Frith, 1991, p. 81). Additionally, it is well established that part of the diagnostic criteria for ASD is difficulty with sharing content particularly in the context of pretense. Frith surmises that people with ASD make no distinction between what is in their own mind and what is in anybody else’s mind. So, the question of sharing content does not even arise (Frith, 2003, p. 102). It may be that people with ASD are not aware of the selves of others and that their self-awareness is a different kind of self-awareness than that which we see in individuals without ASD. It may be that people with ASD simply do not know that their inner experiences are different from the inner experiences of non-autistic people. Note that it is not the case that
researches believe that people with ASD don’t think that others have a mental life. Rather, researchers believe that people with ASD just assume that others know what they know, like what they like, are interested in what they are interested in, and so on. When a person with ASD reflects on the mental states of others, she does so with an extreme egocentric bias. I believe that it is this extreme sort of egocentrism, correlated with hyper-introspectionism, that “locks up” people with ASD inside of themselves, making it virtually impossible for them to “get out of their own way” enough to simulate others.

An important question is whether there is any evidence for a different kind of self-awareness in people with ASD. Asperger certainly thought so. The results of his experiments led to his claim that “these children are totally egocentric because they do not distinguish their own from other peoples mental states and do not recognize that they may differ” (Asperger quoted in Frith, 1991, fn. p. 81, emphasis added). Taking from this, Frith hypothesizes that autistic self-awareness is all self and, while persons with ASD is aware that others have mental lives, they do not reflect on the “self” that is in other “selves” or recognize relevant differences between their mental states and the mental states of others. Frith calls this special kind of self-awareness “self knowledge.” She says that people with ASD who acquire a ToM first and foremost attain knowledge of their own mind and that they possess detailed introspective knowledge of their own minds, but not of other minds. If this is true, she says, “we could explain the autobiographies of autistic people that go into remarkable detail about their own inner states, far more so than most autobiographers, and why they hardly even speculate about how they may have affected people who play an important role in their lives” (Frith, 2003, p. 210).

Developmentally, in individuals without ASD, awareness of one’s own mind tends to arise concomitantly with awareness of other minds. However, in people with ASD, Frith indicates that
the awareness halts at self-awareness and detailed introspective knowledge and that the awareness of other minds does not develop in the same way as in typically developing people (Ibid).

Similarly, Dewey indicates that even though people with ASD (of normal intelligence) may know that other people’s minds may have some factual knowledge that they lack, they nevertheless assume that others will know what is in their minds without being told and that, conversely, they sometimes feel accountable for knowledge which nobody would expect them to have (Dewey, 1991, p. 200). It seems that the kind of egocentrism I have been describing, combined with hyper-introspectionism, would easily lead to the kind of results reported by Dewey. Dewey does not claim that people with ASD lack awareness of other minds altogether, but that they assume or expect that the contents and mental states of others are the same as theirs.

Hill and Frith note something akin to this. They report on current studies indicating that most people with ASD “are impaired in the intuitive understanding that people have mental states different from their own” (2006, p. 283). According to Happé, this leads to a further impairment in mental state attribution in which people with ASD attribute mental states and process social events in a highly idiosyncratic and egocentric ways, in contrast to the spontaneous and intuitive interpretations seen in typically developing persons (Happé 1999 and Kaland et.al., 2002).

Taking the empirical research under consideration, the simulationist (i.e. the advocate of the ST) could argue that whereas, ordinarily, a typically developing person will use her own mental states to simulate the mental states of others, this process is truncated for someone with ASD. As a result of the egocentric and hyper-introspectionist nature of her mental state attributions, the simulations of a person with ASD are flawed – reflecting primarily her own mental states and not those of her target. Ultimately simulations performed by someone with ASD misfire because she
simulates under the assumption that the mental states of the target are identical to her own and she interprets the mental states of her target in highly idiosyncratic and egocentric ways.

Of course, simulations are used for multifarious purposes apart from mindreading and in many domains outside of the psychological domain. A question remains concerning how well people with ASD perform on various kinds of simulational tasks outside of the psychological domain - that is, since they struggle with simulation specifically for mindreading, do they also struggle with simulation more generally (i.e. simulation tasks that fall outside of the psychological domain)?

I remain agnostic as to whether or not people with ASD will do poorly on various tasks outside of the psychological domain that require some form of non-psychological simulation. I suspect that there are some of these tasks – models that rely heavily on theorizing and systemizing, in mathematics, science, and engineering – on which they will do quite well; and that there are others – models that require pretense and role-playing - on which they will do less well. There is, for example, no reason that I know of why a person with ASD could not run a computer simulation or succeed with an airplane simulator. My claim herein is not that people with ASD have difficulty with simulating in general. My claim, which I have shown, is that people with ASD exhibit extreme egocentrism, difficulty with holding two mental representations simultaneously while screening off (quarantining) one of them, a deficit in their ability to engage in pretense, and a deficit in their ability to empathize. All of these are well-documented impairments and deficits of ASD. It just so happens simulation for mindreading requires (1) pretense, (2) empathy, (3) the ability to take a non-egocentric stance/perspective, and (4) the ability to quarantine one's own idiosyncratic mental states, and people with ASD have difficulties with 1-4. People with ASD (generally) do not have the basic nuts and bolts required
to run mental simulations and produce accurate mental state attributions – but, again, this is less attributable to “an inability to simulate” than it is attributable to the well-documented deficits and impairments experienced by people with ASD – in short, an absence of the qualities required for mentalizing on the ST account.

In the preceding section of this chapter, I described the need to adjust for one’s egocentric bias (as discussed by Epley and Caruso) and to the need to quarantine one’s own idiosyncratic mental states in order to produce accurate simulations of others. It seems that if people with ASD are egocentric in the extreme (as Asperger showed), strictly and egocentrically self-aware (as Frith showed), and hyper-introspectionist (as I have described), then simulation routines for mindreading would be exceptionally difficult for people with ASD. They would presumably be at a constant risk of importing, and thus projecting, too much of their own particular mental states (that is, too much of the organization of their own minds) onto the target. Mindreading would be difficult, if not impossible, to perform. Quarantine failure, which I have previously attributed to hyper-introspectionism, can likewise be attributed to extreme egocentric bias. This, I hold, is where mindreading goes amiss when someone with ASD attempts to (consciously or unconsciously) employ their ToM: A person with ASD, given their egocentrism and hyper-introspectionism, will be unable to quarantine their own mental states and so is highly likely to self-contaminate their own predictions of others. The failure to quarantine will of course result in contaminated predictions (i.e. contaminate simulations). Lastly, I have remained largely agnostic as to the question concerning how well people with ASD can be expected to perform on simulation tasks outside of the psychological domain. This is because my claim is not that people with ASD have difficulty with simulation in general, but because it is well established that people with ASD have specific deficits that would naturally impede their ability to simulate
particularly for mindreading tasks (e.g. egocentrism and difficulty with empathy, pretense, and holding two mental representations simultaneously while quarantining one of them).

Taking the foregoing into consideration, I affirm Currie’s view: We know that people with autism are strikingly deficient at ToM tasks and we have sufficient empirical evidence to suggest that this deficiency is caused by difficulties with simulation. So we have evidence for the hypothesis that the proficient performance of typically developed subjects on ToM tasks is to be explained – at least in part – by simulative competence. Hence we have a sort of confirmation of the ST.

Thus far in this section I have been discussing how extreme egocentrism is correlated with the failure to quarantine, leading to flawed simulations and failed mindreading. Now, I will turn my attention to the question of empathy and empathic accuracy as needed for accurate simulations. Since people with ASD are particularly weak in these areas, and both empathy and empathic accuracy are required for accurate simulational routines, it follows that they would simulate poorly in the absence of empathy and empathic accuracy.

First, recall that Baron-Cohen’s work on the EMB theory of autism included the findings that not only are people with ASD especially proficient theorizers, but also that people with ASD are especially deficient empathizers. His findings concur with the diagnostic criteria for ASD, which includes frequent difficulty with pretense and empathy.\(^\text{119}\) It so happens that empathy and empathic accuracy are also both necessary for simulating routine that leads to accurate mindreading (see Epley and Caruso, 2009; Frith, 1991, 2003; and Frith & Happé, 1994, 1999).

Many other researchers have also found that people with ASD struggle with empathic accuracy. For example, Bauminger and Kasari (1999) found that participants in a clinical group

\(^{119}\) See Appendices to Chapter 3, this dissertation.
with ASD “revealed idiosyncratic ways of interpreting…situations where mental states were involved.” Others have found that high-functioning people with ASD – even those who could pass second-order standard ToM tests – tend to respond “with significantly more incorrect, irrelevant, and idiosyncratic ways on mental state inferences, as compared with normal developing controls” (Kaland, et.al., 2002, p. 524). Myers and Hodges, note that effective empathic accuracy is not about the empathizer feeling some measure of concern for another person. Rather, empathic accuracy relates to “the perceiver’s motivation and ability to create coherent mental representations of the target person” (Myers and Hodges, 2009, p. 285). Arguably, people with ASD lack the ability to create such representations.\textsuperscript{120} Tantum asserted, “lack of empathy may result in inappropriate emotional reactions” (Tantam, 1991, p. 176). As noted, very often, people with ASD have highly inappropriate emotional reactions while attempting to navigate social situations. I suspect that these inappropriate reactions could be as a result of a paucity of empathy, as Tatum notes. However, I think that an additional contributing factor to these perceived inappropriate emotional reactions might be a dearth of the kind of empathic accuracy that the simulationist claims is required for accurate mindreading.

Frith also indicates that people with ASD lack the ability for empathy (Frith, 2003, pp. 111-112). Frith is careful to clarify that by “empathy” she does not mean the kind of instinctive sympathy one might feel when observing another in pain but, rather, what she calls \textit{intentional empathy}. Intentional empathy is to be contrasted with instinctive empathy (most often called “sympathy”). Intentional empathy does not require the ability to sympathize or feel pity, but, instead, relies on the ability to attribute mental states to others. She says, “Intentional empathy

\textsuperscript{120} I do not claim that people with ASD lack the motivation or desire to create coherent mental representations of targets. Quite to the contrary, and based on anecdotal observations both in clinical and non-clinical settings, it seems that many people with high-functioning ASD have an extremely strong motivation and desire to create these accurate mental representations of other and struggle with frustrations over their inability to do so.
does require the ability to mentalize and is thus dependent on the instinctive orientation to other people’s mental states. This is the way that Baron-Cohen uses the term “empathizing” when characterizing the social impairments of autism. This type of empathy is linked to understanding the reasons for another person’s sadness or fear, and responding with the appropriate response” (Frith, 2003, p. 112). Instinctive sympathy (such as feeling another’s pain) is not required for proficient intentional empathy, but the ability to attribute mental states to others is. Frith says that if one is proficient with ToM (i.e. if one can attribute mental states to another), one will be able to understand and give the appropriate responses, even if unable to sympathize and feel distress over the other person’s misfortune or pain. However, she says, “if you cannot readily and consistently attribute mental states to others, then you will be lost as to how to respond even if you do feel sympathy for the person. In this case, people may wrongly think that you are callous” (Ibid). So instinctive sympathy is not required for intentional empathy, but it appears that – on Frith’s view – intentional empathy is required for instinctive sympathy. One can be highly instinctively empathetic (i.e. sympathetic) to the pain of others, yet still experience a poverty of intentional empathy. On Frith’s view, it is difficulty with (or absence of) intentional empathy that gives rise to the social deficits and social awkwardness so inherent in people with ASD.

Happé makes an excellent point regarding the autistic inability to empathize. She describes the autobiographical work by Dr. Temple Grandin. Dr. Grandin goes into great detail about jumping into the dip vat and swimming through it so that she could understand just what the animals were feeling when they were forced into it. Happé notes that this is of interest because Dr. Grandin was unable to imagine the process but had to physically put herself through the experience in order to feel the same feeling. Happé says, “When we empathize with another
person, we generally mean that we feel with them, despite the fact that we are not actually suffering with them” (Happé, 1991, p. 210).

Thus it seems well established, or at least widely accepted, that people with ASD are deficient with empathy and empathic accuracy in a multitude of ways. They are classically unable to engage in pretense and in pretend play. They have difficulty with intentional empathy. They may struggle also with instinctive sympathy. In the foregoing example, Dr. Grandin struggled with instinctive sympathy to the extent that she had to employ a compensatory mechanism, such as jumping into the cattle dip vat, to gain a roughly rudimentary kind of sympathy.

Importantly, simulation requires – indeed, it is literally dependent upon – the ability to project oneself into the shoes of the target. Recall that on Gordon’s view, to simulate one must transform oneself into the target.\textsuperscript{121} One must take on pretend states and run one’s own mental state routines in an ‘offline mode.’ It seems that if a person could not engage in pretense, empathy, or have strong empathic accuracy, she could not engage in the kinds of pretense, imagination, and empathy that is absolutely required for running the kind of simulational routine that will result in accurate mindreading. Similarly, if one is egocentric and hyper-introspectionist in the extreme, one will likely never be able to quarantine one’s own idiosyncratic mental states and “get out of one’s own way” enough to produce uncontaminated simulations resulting in accurate mindreading. Without the capacity for pretense and imagination, without the ability to quarantine the “self” recognize other “selves,” and without the ability for empathy, any attempted simulation is likely to misfire because all such attempts are likely to be contaminated by egocentrism and hyper-introspectionism.

\textsuperscript{121} See section 5.2.3, this dissertation.
In short, any simulation attempted by a person with ASD is likely to be an erroneous simulation. It is not the case that people with ASD do not simulate. I maintain that people with ASD do simulate – or at least they attempt to simulate. I think, it is more likely the case that people with ASD are unable to quarantine and empathize and so their outputted simulations are consistently contaminated by their own idiosyncratic mental states. Extreme egocentrism and hyper-introspectionism lead to unqualified egocentric descriptions of mental states and contaminated simulation. This just is the hallmark of the autistic’s outlook on the social world. This contaminated simulation may account for the vast majority of the social impairments experienced by people with ASD. A person with ASD is hyperaware of her own mental states and hyper-introspectionist – so much so that she assumes that the mental states of others are just the same as her own mental states because she is unable to imagine and project otherwise. She simulates as if the contents of the target’s mind and the target’s mental states are, most likely, the contents of her own mind and her own mental states. She simulates under the assumption that the target’s knowledge and mental states are the ones that she possesses. Hence we find the social difficulties so prevalent in ASD. The autistic simulates, but she simulates badly.

This, I think, is how the ST gets corroboration: not because people with ASD cannot simulate, but because they do simulate – they just do it badly. Their brand of egocentric, hyper-introspectionist simulation produces a bad simulation. The autistic person makes the copy of the target, but makes a bad copy. Importantly, the case of ASD does not only serve to confirm the ST, it also provides a strong reason to prefer the ST to the TT: Since we know that people with ASD can theorize, and can do so accurately with great acuity, it would seem that, if the ToM were underpinned by theorizing (as the TT claims), they should be able to accurately mentalize. Yet, people with ASD find it difficult (and in some cases impossible) to mentalize. So, in this
light, it cannot be theorizing that forms the basis of ToM activity. However, since people with ASD have difficulty with both simulation and mentalizing, it makes sense that simulating does underpin ToM activity. Hence, it makes sense to think that the ST is the more correct theory of ToM.

5.5.2. Does a Bad Copy Still “Count” as a Copy?

In some sense, simulators are models. By a “model” in the general sense I mean a system or a thing that is used to follow, imitate, or understand a similar system or thing. When I say that simulators are models, I mean that a simulator introspectively uses her own cognitive system as an example of the cognitive systems of her targets. Analogously, a flight simulator is a model of an airplane. The cockpit of a flight simulator is a model of the cockpit of an actual airplane. Using the flight simulator to learn how to land an aircraft, the neophyte pilot can learn how to land an actual aircraft. Similarly, the simulating agent uses herself as a model for her target when she transforms herself in imagination and imagines herself as her target (i.e. she “puts herself in the shoes” of her target). I have said that failing to quarantine one’s own idiosyncratic mental states, failing to employ empathy, and failing to have empathic accuracy would result in erroneous simulations leading to contaminated simulations and incorrect third-person mental state ascriptions. This stems, I believe, from the fact that the person with ASD is using an inferior model. The model used constitutes an inferior model because the model reflects only the simulator, who happens to suffer from ASD, and whose mental states will not map onto those of the target, who doesn’t. Imagine if a flight simulator – such as the kind we might find in an

---

122 Of note, there is an important difference between the pilot using a flight simulator as a simulated model of an airplane versus an attributor using herself as a simulated model of her target: The pilot and the flight simulator are two separate entities, but the attributor and her model are one and the same entity.
arcade – was used as an actual flight simulator! The model employed by the learner would be so inferior that she could never learn to fly an actual airplane! I propose that anytime anyone simulates egocentrically with corresponding quarantine failure, the result is likely to be a bad copy because the model that was used reflected only the attributes of herself and does not reflect the attributes of her target. Thus, the failure of acquisition is not owed to an inability to simulate (i.e. to create a faithful copy); the failure of acquisition is owed to use of a flawed (i.e. idiosyncratic) model.

Not all mindreading has to be accurate or successful for the ST to be correct. Goldman claims that the ST does not need to claim that all, or even most, mindreading is – in fact – successful. But it is reasonable to assume that mindreading must have a high enough success rate to make it evolutionarily viable, and this apparently has been the case (See Goldman, 2006 and 2008). Hence, the accuracy of mental state attributions and the resemblance of simulating states seem relatively important, in terms of evolutionary viability. Even so, Goldman does not impose a strict requirement upon accuracy. He says, “We allow for some looseness. X can be a simulation of Y even if it is not an exact duplicate of Y. Moreover, X can be a simulation of Y even if it only aims to duplicate, copy, or replicate Y. Simulation does not require successful duplication. The simulated event or process (Y) may also be a merely possible event, not an actual one (e.g., as when we simulate hypothetical events” (Shanton and Goldman, 2010, p. 528). Goldman is not the only simulationist who maintains that inaccurate simulations may still count as simulations. In fact, most advocates of the simulation theory generally agree that simulation need not run perfectly or be perfectly accurate and allow for such erroneous or “misfired” simulations (See Currie, 1996, 2000; Gordon, 1986, 1992(b), 2005; Harris, 1992; and Heal, 1994, 1996, 1998).
If ever there was going to be a case of an erroneous, failed, or misfired simulation, certainly one way in which it would occur in conjunction with the existence of some pathology, which prevents the simulator from performing the key elements required for accurate simulation.\textsuperscript{123} This is exactly the kind of case we see in simulations attempted by people with the pathology of ASD who do not possess the ability or skill-set to pretend, imaginatively project, adjust for the egocentric bias, quarantine their own mental states, and employ empathy and empathic accuracy. As I have stated, a person with ASD is hyper-introspectionist, so she will consistently produce simulations that are contaminated by her own egocentric and idiosyncratic mental states. As a result, she will consistently produce a bad copy when attempting to simulate and, notably, a bad copy “still counts” \textit{as} a copy in terms of simulation and on every simulationist account.

\textbf{5.6. Concluding Remarks}

In this chapter, I have outlined the ST in general and discussed two major proponents of it: Robert Gordon and Alvin Goldman. Also, I have given an exposition of their theories including Gordon’s version of the ST which (1) does not require prior mental state concept possession and (2) does not allow for an analogical inference ‘from me to you’, and (3) does not allow for introspection; and Goldman’s version of the ST which does allow for each of these three. Along the way, I have taken pains to elucidate some intricacies of each account, such as describing Gordon’s ascent routines and Goldman’s interest in mirror neurons. I have discussed the most frequent criticism against the ST, which is the threat that, in one way or another, the ST collapses

\textsuperscript{123} Other factors such as exhaustion or emotional distress could also lead to erroneous, failed, or misfired simulations. Here, I focus solely on simulations that are error-prone owed to an underlying pathology.
into the TT owed to some reliance upon theorizing, and I have shown that the ST has the machinery to refute such criticisms.

I have addressed an important conceptual issue, explaining how it is that simulations can sometimes produce failed or inaccurate mindreading. This occurs both frequently and predictably in cases when the simulator fails to quarantine her own idiosyncratic mental states while simulating, thereby allowing the simulation to be contaminated. My claim is that hyper-introspectionism combined with extreme egocentrism is responsible for the misfires in simulation that are so commonly observed in people with ASD. In particular, given their propensity for hyper-introspectionism and extreme egocentrism, it is highly unlikely that a person with ASD would ever be able to quarantine their own mental states or to spontaneously and intuitively produce a simulation that is uncontaminated by their own mental states. The requirement to quarantine one’s own idiosyncratic mental states so as not to contaminate one’s simulation, and Goldman’s insistence that introspection is a necessary component to mindreading, are essential to my arguments, which favor the ST in general and, especially, Goldman’s account of the ST in particular. The expositions and arguments of this chapter, in conjunction with the expositions and arguments of the foregoing chapters work together, ultimately, to show that the simulationist offers a more favorable and plausible explanation of the ToM than that of theory theorist.
CHAPTER 6: AN ADJUDICATION OF THE DEBATE BETWEEN THEORY THEORISTS AND SIMULATION THEORISTS WITH CONCLUDING REMARKS

Thus far, in this dissertation, I have given detailed expositions of the ToM capacity: the perceived human ability to understand the mental states of ourselves and of others and the perceived human ability to use that understanding to explain and predict the behavior of others and of ourselves (i.e. mentalizing or mindreading). I have explained that there are, essentially, two camps of theorists who attempt to explain what it is that underpins this ToM ability: the TT and the ST. Since people with ASD present as a group of people who exhibit a deficiency or impairment of ToM, understanding why this is the case can help us to understand what underpins the ToM ability. Thus, I have given a detailed exposition of ASD and its classic triad of impairments.

After the expositions of ToM and ASD, I discussed at great length the TT and the two proponents of the TT who give the most plausible account: Alison Gopnik and Simon Baron-Cohen. I also argued that their accounts couldn’t be correct, showing some considerations that strongly suggest that it is not theorizing that underpins the ToM capacity. In fact, I was able to use Baron-Cohen’s account as grist not only against Gopnik’s account but also against the TT as a whole. This led me to the discussion of the competing theory, the ST. I then discussed at great length the ST and the two proponents of the ST who give the most plausible account: Robert Gordon and Alvin Goldman. I elucidated some key differences in their accounts, working to show that Goldman’s account, with its critical inclusion (indeed requirement) for introspection, is the more preferable account. The expositions and arguments of the foregoing chapters of this dissertation work together and culminate in my final argument, which I present below.
6.1. A Brief Overview of the Argument

If we grant that the ToM exists as a human capacity to understand the mental states of oneself and others and to explain and predict one’s own behavior as well as the behavior of others (i.e. ‘mindreading’ or ‘mentalizing’), then we ought to explain what it is that underwrites the presumed ToM ability. Currently, there is a lively debate over whether our mindreading abilities are better understood in terms of theorizing (as claimed by the TT), or in terms of simulating (as claimed by the ST). This debate presents as though it must be either the TT or the ST that can explain the ToM ability. I join this debate with my assertion that the special case of ASD can aid in adjudicating the debate.

I have elucidated some considerations that strongly suggest that the TT cannot be correct because, if theorizing were to underwrite the ToM ability, then we should expect that people who cannot mentalize also would not be able to theorize (because it is claimed that this ability is supported by a domain-general learning mechanism that allows for theorizing in multiple contexts). It is widely agreed upon that people with ASD have difficulties with mindreading, empathy, and pretense. These difficulties result in the highly predictable social deficits experienced by people with ASD. Using Baron-Cohen’s EMB theory of autism, I have shown that people with ASD actually tend to be especially competent theorizers in multiple domains outside of the psychological domain and even within the psychological domain and also tend to be especially poor empathizers. Under the TT interpretation of ToM, we should expect this not to be the case (in fact we would expect to see the opposite as true). But empirical data shows this to be the case, which suggests that the TT is not the correct account of the ToM ability. Baron-Cohen’s EMB does much of the heavy lifting to negate the possibility of the TT as the best
explanation of ToM. Moreover, it explains why it is that people with ASD are likely to employ some theoretically or systematically based compensatory mechanism that enables them to mentalize in lieu of simulating. Rather than simulating to mentalize (as would a person without ASD), people with ASD must compensate by ‘hacking out’ a rudimentary and compensatory system to achieve what the non-autistic person achieves via simulating (as is shown by cases such as that of Dr. Temple Grandin and others). My arguments against the TT and Baron-Cohen’s EMB theory of ASD work together to show that the TT cannot be the correct theory with regard to its supposition that theorizing underpins the ToM capability. This swings the door to an alternative theory wide open. I argue that the theory we ought to be considering as preferable to the TT is the ST.

On the ST account, because empathy is required for simulative perspective-taking, we should expect that one who is a poor mindreader would also be an especially incompetent empathizer, which is precisely what is gained by Baron-Cohen’s EMB. The case for ST is further bolstered because each and every diagnostic criterion for ASD includes deficits in both empathy and pretense – both of which are absolutely required for running an offline simulative routine. So, consideration of the case of ASD provides reasons to favor the ST over the TT.

A question remained as to which version of the ST is the most favorable. Again, we can adjudicate this matter by looking at the special case of ASD. Both Gordon’s account of the ST and Goldman’s account of the ST mention the importance of quarantine – that is, that the simulator must screen off her own idiosyncratic mental states – for accurate simulation yielding both accurate mental state attributions and accurate behavior predictions. I suggest that quarantine, in particular knowing what to quarantine and when, absolutely requires introspection. Moreover, I suggest that people with ASD do not adequately quarantine, owing to extreme
egocentric hyper-introspectionism. Because quarantine is central to the argument and because quarantine requires introspection, we should rule out any account of ToM and the ST that does not allow for introspection. Gordon does not allow for introspection; so, we should rule out his account. Not only does Goldman’s account of the ST allow for introspection, it actually requires it as a necessary component of the mindreading task and simulation routine. Hence, Goldman’s version of the ST shines as the most plausible, most favorable account of the ToM.

In conclusion, when taking the special case of ASD into consideration while adjudicating the TT/ST debate, I find not only that the TT cannot be the right theory of ToM, but importantly - that the ST serves as both a viable alternative to the TT and a highly favorable account of ToM. Furthermore, when taking into consideration some specific deficits experienced by people with ASD and traits common to people with ASD I find that Gordon’s version of the ST cannot be the right version of the ST. By contrast, Goldman’s account of the ST, which requires introspection as a necessary condition for mindreading, works to explain a primary deficit of ASD: an inability to mindread via simulation because of an inability to engage in empathy and pretense combined quarantine failure owed to extreme egocentric hyper- introspectionism. Goldman’s account explains a primary deficits of ASD and, in turn, the case of ASD promotes Goldman’s account of the ST as the most plausible account of the ToM capacity.
BIBLIOGRAPHY

American Psychiatric Association (1952). *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC.


--------- (2005). The real reason children love fantasy: Kids aren’t escapists, they’re little


-------. (1995a). Simulation without introspection or inference from me to you. In M. Davies


“Monkey See Monkey Do.” Narr. Robert Krulwich. NOVASCienceNOW. Public Broadcasting


------------- (2009a) The neural evidence for mirror neurons is weaker than I think you think it is. Philosophical Studies. 144(3), 447-456.


