MORAL AGENCY AND RESPONSIBILITY:

LESSONS FROM AUTISM SPECTRUM DISORDER

AN ABSTRACT

SUBMITTED ON THE FOURTH DAY OF APRIL 2016

TO THE DEPARTMENT OF PHILOSOPHY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF LIBERAL ARTS

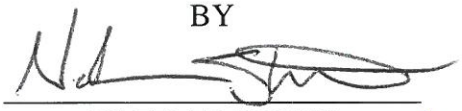
OF TULANE UNIVERSITY

FOR THE DEGREE

OF

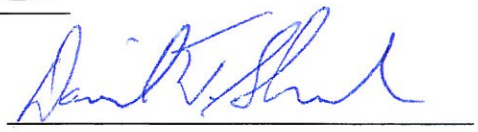
DOCTOR OF PHILOSOPHY

BY

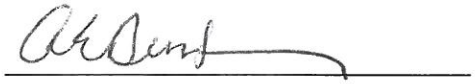


NATHAN STOUT

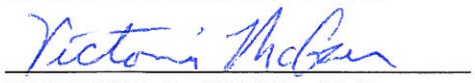
APPROVED:



DAVID SHOEMAKER, Ph.D.
Director



ALISON DENHAM, D. Phil.



VICTORIA MCGEER, Ph.D.

MICHAEL MCKENNA, Ph.D.

The following dissertation is a study of the moral psychology of autism and its implications for theories of moral responsibility. Many philosophers have recently attempted to examine and analyze the conditions of responsible agency by attending to the empirical data regarding "marginal" agents in an effort to see which agential capacities or incapacities lead these individuals to be included or excluded from the class of responsible agents. The dissertation takes up this methodology by focusing specifically on individuals with autism spectrum disorders (ASD). Some previous philosophical work has attempted to draw conclusions from autism in this way, but, in my opinion, these authors have done so by appealing to an inadequate view of the nature of the disorder. The following work aims to correct this by offering philosophers a comprehensive picture of ASD and then identifying the precise ways in which such a picture can inform philosophical work on moral agency and responsibility.

Part I of the project is a detailed treatment of the empirical data on the cognitive and affective characteristics of ASD. It begins by surveying the various cognitive theories of autism that are on offer in the literature: the Theory of Mind hypothesis, the Executive Dysfunction hypothesis, and the Weak Central Coherence hypothesis. Each of these theories purports to give a unifying theory of autism, but all of them struggle to explain various features of the disorder. Because none of these dominant theories gives an adequate picture of the nature of ASD, I propose (in chapter 2) a novel theory which unifies these three approaches while at the same time offering an explanation for a number of features of ASD which these other views cannot offer. According to the view I defend, the cognitive features of ASD can be explained by a deficit in counterfactual thinking ability. More specifically, I argue that we can make sense of the various

diagnostic features of ASD by understanding them as originating as a result of a deficit in the ability to represent counterfactual states of affairs.

After offering this novel theory of the cognitive features of autism, I turn to a discussion of the affective features of the disorder. The dominant view among philosophers is that ASD is an empathy disorder, and, as a result, individuals with ASD are often characterized as lacking in emotional responsiveness more generally. I resist this characterization (in chapter 3), and I argue, instead, that individuals with ASD are actually quite emotionally capable. The available empirical data suggests that individuals with ASD (those without co-morbid disorders, at least) demonstrate typical emotional profiles in most circumstances but struggle with understanding, expressing, and experiencing complex social emotions. This phenomenon can be explained, I argue, by the fact that such emotions demand a robust capacity for counterfactual representation which individuals with ASD lack.

In Part II of the dissertation, I attempt to bring this picture of the psychology of ASD to bear on a number of influential theories of moral responsibility. The philosophical literature on moral responsibility can be divided, roughly and with exceptions, into three types of theories. First, reasons-responsive theories are those that posit that the possession of certain rational capacities is necessary and sufficient for morally responsible agency. The second type of approach, exemplified by "real self" theories, claims that an agent is morally responsible for those actions which issue from her real, or deep, self. The crucial feature of these views is their claim that agents are only responsible for actions which mesh, in the appropriate way, with other features of their agency. Finally, quality of will theories of responsibility aim to explain moral

responsibility in terms of the quality of the agent's will at the time of her action. So, on these views, if an agent is capable of expressing good or ill will through her actions, then she is an appropriate candidate for moral praise or blame for those actions.

I address each of these three approaches to moral responsibility (in chapters 4, 5, and 6, respectively), and I argue that individuals with ASD present considerable challenges for each type of view. Because none of these theories seem able to give an adequate account of the moral responsibility of individuals with autism I argue that a new approach to the question of what responsible agency consists in is required. In the final chapter of the dissertation, I discuss a number of features that such an approach should have, and I propose several possibilities for what an adequate view might look like.

MORAL AGENCY AND RESPONSIBILITY:

LESSONS FROM AUTISM SPECTRUM DISORDER

A DISSERTATION

SUBMITTED ON THE FOURTH DAY OF APRIL 2016

TO THE DEPARTMENT OF PHILOSOPHY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF LIBERAL ARTS

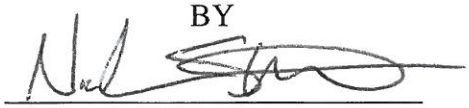
OF TULANE UNIVERSITY

FOR THE DEGREE

OF

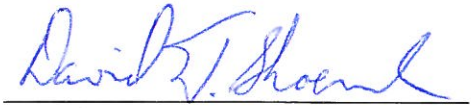
DOCTOR OF PHILOSOPHY

BY



NATHAN STOUT

APPROVED:



DAVID SHOEMAKER, Ph.D.
Director

ALISON DENHAM, D. Phil.

VICTORIA MCGEER, Ph.D.

MICHAEL MCKENNA, Ph.D.

Table of Contents

ACKNOWLEDGEMENTS

I owe a debt of gratitude to many people for the completion of this project. The largest such debt is to my wife, Jennifer. Without her unwavering support and enthusiastic encouragement this work may never have been undertaken at all and certainly would not have been seen through to its end. Countless times during the writing of this dissertation she helped me to see a way forward when it seemed that there was no end in sight. I also owe a tremendous thanks to my parents, Phillip and Carol Stout. Together they instilled in me the curiosity that led me to philosophy and the analytical mindset needed to pursue it. They have always been exceedingly supportive of my academic career, and I'm deeply grateful to them for it.

There are many people to whom I owe an intellectual debt of gratitude as well. The most significant of these is to the director of this project, David Shoemaker. His intellectual influence can be seen throughout this work, and it could not have been finished without him. His penetrating comments and criticisms along the way have made this project far better than I could have hoped to make it on my own, and his support as an advisor has been unmatched. As philosophical role models go, he is as good as they get. I am also extraordinarily grateful to the other members of my dissertation committee, Alison Denham, Michael McKenna, and Victoria McGeer. Their collective philosophical talent is genuinely amazing, and their feedback throughout this work has been profoundly challenging and immensely helpful. I also owe my thanks to two others who played an

important role in my academic career: Kevin Lowery, who helped to cultivate my interest in philosophy as an undergraduate, and Fritz Allhoff, who was a great source of both encouragement and opportunity during my time as an M.A. student and beyond.

I am also extremely grateful to my fellow graduate students for their helpful feedback along the way. I am especially thankful for the many IPA-fueled conversations that I had with Frankie Worrell in the early stages of my research which helped to get this project off the ground. He has been an admirable philosophical interlocutor and a very good friend throughout my time at Tulane. I am also grateful to Nathan Biebel, Nick Sars, Chris Boom, Tom Mulligan, Jesse Hill, and Dan Tigard, for conversations and comments along the way.

Finally, I am deeply grateful for the institutional support that I received during the writing of this dissertation. Specifically, I owe my thanks to the Murphy Institute's Center for Ethics and Public Affairs for generously funding my research during the 2013-2014 academic year, to Tulane's School of Liberal Arts for two summer research grants, and to the Philosophy Department for funding both research and travel expenses over the past several years.

Chapter 1: Why Autism?

The Project, Its Motivation, and Methodology

My concern in the following work is, at its most basic, simply to examine the nature of moral responsibility. In so doing, I will be concerned with identifying and understanding what I take to be necessary features of responsible agency, and I will be especially interested in examining the psychological capacities that underlie them. A staggering amount of work has been done by philosophers who study this particular sub-discipline, and I will by no means be able to give a full picture of the literature in what follows. Nonetheless, I hope to be able to bring to light certain considerations that may pose problems for several conceptions of moral responsibility that are presently on offer. Doing so will, I believe, help to point us toward a more adequate or complete theory of responsible agency.

In order to accomplish this, I plan to center my discussion on some important empirical literature regarding Autism Spectrum Disorder (ASD). ASD is a pervasive developmental disorder that is, according to the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5)*, characterized by "persistent deficits in social communication and social interaction," as well as "restricted, repetitive patterns of behavior, interests, or

activities." (DSM-5, 50) But why should philosophers care about autism? What, if anything, might we find philosophically interesting or helpful about ASD that would make such a project worthwhile?

One insightful answer to such questions can be found in Jeanette Kennett's article, "Autism, Empathy and Moral Agency." In it, she raises the following puzzle: psychopaths are often characterized in the philosophical literature as deeply amoral due to (on some accounts) their lack of empathy, but all of the evidence suggests that individuals with autism, who often have a very clear capacity for moral concern, lack empathy to the same extent (at least) as psychopaths. So, given that both groups lack empathy, what might be the explanation for the plausible view that individuals with ASD[1] are moral agents while psychopaths are not? In other words, Kennett suggests that examining the nature of the difference between individuals with ASD and psychopaths can be instructive in coming to an understanding of what is required for moral agency. This is, I think, a profoundly important problem, and to offer an adequate solution would be to do a great service to those interested in studying the nature of responsible agency. The remainder of this chapter, then, will be devoted to examining this problem and using it to motivate my larger project. In section 1, I explore Kennett's characterization of and solution to the problem in greater detail. In section 2, I give a brief account of an important objection to Kennett's solution from Victoria McGeer. Finally, in section 3, I

[1] There is an ongoing debate among members of the autistic community and their advocates regarding the use of the phrase "individuals with ASD," as opposed to "autistic individuals," to which philosophers ought to be sensitive. Some find the former phrase objectionable because it seems to imply that autism is a type of disease that ought to be cured or eradicated. Others find the latter phrase objectionable because it seems to suggest that one's identity consists in being autistic rather than in other features of the person him or herself. Most often, I will use the phrase "individuals with ASD" for the simple reason that it is the phrase more commonly used in the psychology and neuroscience literature. However, I by no means intend for this phrase imply any sort of judgment regarding the value of autistic persons or of what it is like to be autistic more generally.

will offer some preliminary arguments as to why I think that neither Kennett nor McGeer offer an adequate account of the relevant differences between psychopaths and individuals with autism and that this is largely because neither of them is working with a complete picture of autism's distinctive psychological features. If I am right, then the challenge of explaining the implications of autism for theories of responsible agency remains, and this fact, I think, is sufficient to get my project off the ground.

§1. Kennett on Psychopathy and Autism

Of psychopaths, Kennett writes, "Their behavior is characterized by impulsivity and irresponsibility; they are habitual liars, are indifferent to the rights of others, and display lack of remorse for wrongdoing. When they do something wrong, they do not really see what the fuss is about, and may engage in rationalizations and blame-shifting." (Kennett 2002, 341) She then goes on to consider evidence which suggests that a lack of empathy is at the root of this sort of behavior.

Precisely what is meant by the term 'empathy,' however, may not be entirely clear, and Kennett considers several different senses which may be relevant in discussions of psychopathy. She claims that the psychological evidence suggests that psychopaths "lack the capacity to enter sympathetically into the concerns and feelings of others," (341) but that this is interpreted in a variety of ways. For example, she cites some evidence which suggests that the correct interpretation is that psychopaths are "unable to think in terms of the interests of others or to regard those interests as reason-giving." (Elliot 1992, 210) There is also some suggestion in the literature, she notes, that what psychopaths lack is the "affective capacity involved in recognizing and being moved by another's distress." (Adshead 1996) Whatever the correct view may be, Kennett notes

that what is common throughout the literature is the view that psychopaths lack empathy, in *some* sense, and that this, at bottom, amounts to the lack of "an adequate pathway to other people's minds." (341)

The relevant pathway, according to Kennett, consists in the process of simulation that occurs between typically functioning individuals. This process involves entering imaginatively into the mind of another, and 'catching' her emotional states (a phenomenon that Kennett refers to as 'emotional seepage or contagion') If empathy is an essential part of gaining the kind of understanding necessary for moral agency, then it must involve the "imaginative process of simulation and its resulting emotional seepage and reciprocal awareness," (345) she thinks, because only then would it both provide us with an understanding of another person's mental states and allow us to experience the other's emotions in the relevant way. There is further evidence that psychopaths lack this emotional aspect of empathy. For example, R. J. R. Blair published a study in 1997 showing that psychopaths showed decreased psychophysiological responsiveness to distress cues in others. Additionally, there is ample anecdotal evidence demonstrating that psychopaths simply do not respond emotionally or even recognize emotions in others.[2] If it turns out that psychopaths lack this capacity (and Kennett thinks they do), then this would be evidence in favor of the view that emotional contagion through simulation is a necessary condition for moral agency (a broadly Humean view). Enter autism.

In response to this notion, Kennett says the following:

[2] In one such anecdote, a psychopath was shown a picture of a fearful face and asked to identify the emotion that the person was feeling. The psychopath responded that he didn't know what the emotion was but that it was the "face people pulled just before he killed them." See Ronson 2011.

> An incapacity in this area cannot be the complete explanation of the psychopath's moral failings, since another group of people, autistic people, who even more conspicuously lack empathy as I have so far described it, do in some cases seem capable of compensating for this deficit and becoming conscientious, though often clumsy moral agents.

In short, if it is true that empathy is a necessary condition for moral agency, then anyone who lacks it should fail to be a moral agent. However, individuals with autism also lack the capacity for empathy, and since these individuals are, in some cases, clearly moral agents, it cannot be the case that empathy is a necessary condition for moral agency. This argument, of course, will only hold water if individuals with autism actually do lack empathy in the relevant sense, and Kennett presents some (albeit anecdotal) evidence that they indeed do.

I will not survey her treatment of autistic moral psychology here as I will be striving to give a comprehensive account of my own later in the work. For now, it will suffice to say that Kennett picks out a few features of the psychology of ASD which she takes to be relevant to her argument. The first of these is the apparent difficulty with simulation that is observed in individuals with ASD. While many individuals with ASD may come to develop a theory of mind (henceforth, ToM), Kennett maintains that this is distinct from the simulation capacity that is necessary for empathy.[3] In addition to experiencing difficulties in simulation, Kennett cites some evidence that individuals with autism often report feelings of alienation or aloneness with respect to other people and that they have pronounced difficulties in social understanding and awareness. In these

[3] I will have much more to say about this in subsequent chapters. For now, it is worth noting that the psychological literature makes a distinction between cognitive empathy and emotional empathy. ToM seems to be a necessary component of cognitive empathy, but it is less clear what role it plays in emotional empathy. The distinction between these is a convoluted one, and Kennett glosses over it rather quickly.

respects, she claims, individuals with autism share much in common with psychopaths. She writes,

> Both sets of literature [those on autism and psychopathy] speak of a tendency to treat other people as tools or instruments, a lack of strong emotional connectedness to others and impaired capacity for friendship, and they link these impairments to failures of empathy. Indeed, in some of these respects those with autism are significantly worse off than psychopaths, who are usually perfectly competent at casual social interactions, and often possess a facile easy charm. Psychopaths are also usually well able to read the intentions and predict the behavior of others. (Kennett 2002, 349)

So, if individuals with autism share so much in common with psychopaths, what could account for the observable difference in their moral behavior? Kennett's suggestion for solving this puzzle comes by way of an appeal to the fact that most individuals with autism have a strong inclination to behave in a way that is very rule-driven and ordered. "The development and application of rules of conduct," she says, "whether by teachers and parents or later by themselves, play a prominent role in assisting them to negotiate the otherwise confusing social landscape." (350) This feature of ASD's psychology, a respect for rules, provides individuals with ASD, Kennett claims, with an alternative route to moral autonomy. Individuals with ASD are capable of developing moral rules by which to govern their behavior by reasoning explicitly about the constraints of reason in light of the "hard-won realization that other people have needs and feelings different from [one's] own." (352) Thus, she says, what follows from comparing these various features of autism and psychopathy is actually that moral agency requires not a specific brand of emotional capacity but rather the ability to act in accordance with the reasons that one has, and these reasons include those given to us by the interests of others. Psychopaths lack the capacity to understand the constraints that reason places on their actions, and this fact, rather than the fact that they lack empathy, explains their

amoralism. High-functioning individuals with ASD, on the other hand, retain this capacity, and are able to be autonomous moral agents as a result.

§2. McGeer's Objection

Kennett's view is initially plausible, but it has been the subject of an important objection from Victoria McGeer. McGeer agrees that Kennett's challenge poses an important problem for any Humean view of moral agency, but she rejects the notion that the challenge must lead to a Kantian view of agency. Instead, she seeks to show that there is a better way in which to interpret certain features of autistic psychology and that this interpretation will vindicate a view in which affective capacities play a necessary role in moral agency.

In doing this, McGeer picks up on the last feature of autistic psychology discussed above, namely the seemingly deep respect for moral rules. Kennett's suggestion is that this respect is the result of individuals with ASD exercising their capacity for reason in such a way as to arrive at an understanding of the moral law in a manner that is altogether different than typically developing individuals. McGeer suggests that the rule-oriented behavior of individuals with ASD has a decidedly different origin. Given the difficulty that these individuals have with social cognition, a difficulty which arises, she thinks, largely because of their less-developed capacity for ToM, processing the social world can be a difficult, and sometimes painful, experience for individuals with ASD. As a result, McGeer suggests the following interpretation of the rule-driven behavior of those with ASD:

> Rules and routines help keep things the same, making the world emotionally and cognitively more approachable. To this end, autistic individuals are highly motivated to follow rules and are very concerned that others do so too. My … suggestion, then, is that a good part of the behavior we identify as manifesting

> moral sensibility among individuals with autism may stem from a need to abide by whatever rules they have been taught without sharing our understanding of the ends those rules are meant to serve. In other words, for many such individuals, it may well be an open question as to how deeply their moral judgments and behavior are genuinely guided by moral concern. (McGeer 2008, 240)

So, far from being the case that individuals with ASD come to have a reverence for the moral law as a result of a Kantian exercise of rational capacity, it may simply be the case, McGeer claims, that these individuals exhibit rigid rule-driven behavior as a way of coping with the social difficulties that the disorder brings.

Ultimately, this leads McGeer to a pluralist approach to moral agency. She sees the rational capacity to which Kennett is appealing as a necessary condition for agency, but she maintains that affective capacities play a critical role in agency as well and that individuals with ASD demonstrate such capacities. The respect for reason that these individuals display, she suggests, derives from the practical realization that reason is necessary for achieving the ends and goals which one finds affectively salient. To this end, she writes,

> [T]he agential capacity for responding to reason is rooted in the capacity for valuing certain ends, and valuing certain ends is fundamentally rooted in the depth and quality of one's affective life. Since the affective lives of autistic individuals are substantially different from the lives of normally developed individuals, we should expect to see differences in the sort of ends that are valued and in the priorities assigned to these ends. (247)

In a more speculative fashion, McGeer proposes that we understand moral agency in individuals with ASD as being of a distinct kind, a kind which is characterized by what she calls a "concern for cosmic order." This sort of concern, she thinks, while being dramatically different from the kind that neurotypical agents enjoy, is sufficient to explain the difference between psychopaths and individuals with ASD with which the puzzle began.

I will return to McGeer's speculative proposal later in this work. For now, my goal in bringing it up is simply to highlight the problem that is posed for Kennett's solution to the puzzle and to show, at least, that it is a problematic solution at best. Indeed, in the end, Kennett largely agrees with the objection that McGeer raises.[4] For my part, I think that McGeer's explanation of the sort of moral concern shown by individuals with ASD is correct to a certain extent. Autistic persons show a great concern for order and repetition, and the experimental evidence, as we will see, bears this out. However, I also think that both of these theorists miss the mark in important ways, and in the following section I will gesture toward some reasons for this.

§3. The Challenge Moving Forward

As I see it, Kennett is absolutely correct about two things. First, she is certainly correct that there is an empirical difference between the moral capacities of individuals with ASD and those of psychopaths. Second, I think she is correct that understanding this difference may be instructive in helping us to understand the nature of moral agency. However, I believe that her account of the nature of the difference between the moral sensibilities associated with these disorders is largely incorrect. Moreover, I believe that attending to the differences between the two can teach us much more about responsible agency than Kennett has so far said. I will briefly outline some of the problems with Kennett's account in this section in the hope of setting up the rest of my project.

One important problem for Kennett (and, to the extent that she accepts Kennett's view, for McGeer as well) is that she relies very heavily on anecdotal evidence in presenting her view on ASD. While much can be learned from such evidence, it is

[4] See Kennett 2008.

problematic to put too much weight on it. There is often a great deal of comorbidity with other disorders (I will cover one of these, alexithymia, in detail in chapter 3) in individuals with ASD, and given its nature as a spectrum disorder, it is very difficult to glean much helpful information from first-personal accounts of individuals with ASD. This is, of course, not to say that these accounts are unhelpful. Indeed, they can offer a great deal of "inside out" insight into the psychology of the disorder, and they can help us to see how certain traits are manifested. However, such accounts must be understood, I think, in light of the available empirical research on ASD in which control groups help to rule out the misunderstandings that may exist in purely anecdotal accounts.

Another problem that may prove important for both Kennett and McGeer arises from the fact that they both identify a lack of empathy as the cause of the psychopath's amoralism. This may well be correct, but in doing so they each focus almost exclusively on the emotional aspects of agency and their role for both psychopaths and individuals with ASD. This is, of course, understandable given that Kennett's goal is to show that certain affective capacities are not necessary for moral agency, but it may well be the case that certain cognitive failures are what actually drive the conclusion that the psychopath fails to be responsible. Indeed, several theorists take such a view. Antony Duff, for example, argues that psychopathy is primarily a disorder of understanding and focuses on what he takes to be a cognitive defect in the psychopath. He writes, "A psychopath is seriously defective in practical understanding and rationality: he is cut off by his deficiency from a central dimension of human life, just as the intellectually defective are cut off from the dimension of intellectual understanding and thought." (Duff 1977, 199) There is also reason to believe that psychopaths are impaired in their ability to

make moral judgments as evidenced by their inability to make the moral/conventional distinction,[5] and while there is some dispute about precisely what follows from this distinction (or if such a distinction exists)[6] it is at least not obvious that the problem does not stem from a primarily cognitive deficit in psychopaths. Admittedly, Kennett notes that this may be one way to understand the moral incapacities of the psychopath, but she nevertheless focuses primarily on the emotional empathic deficits. However, if the aim of Kennett's paper is to explain the difference in moral agency between psychopaths and individuals with ASD, and it turns out that the psychopath is not a moral agent by virtue of his or her cognitive incapacities, then it will be important to compare those incapacities with the same in individuals with ASD, and this is something that Kennett does not do.

One final problem with the solutions offered by both Kennett and McGeer, I think, is that they each seem to be working with incomplete pictures of the emotional capacities possessed by individuals with ASD. I believe, and will aim to show in chapter 3, that the emotional lives of these individuals are far more robust than either Kennett or McGeer seem to allow. For example, the important capacity according to Kennett, is the capacity to imaginatively simulate the emotions of others and the "emotional seepage" that accompanies such simulation. Now, it is widely held that individuals with ASD have great difficulty with simulation, but there is some evidence to suggest that they are nevertheless prone to the emotional contagion described by Kennett.[7] McGeer, in fact,

[5] See Blair 1995.

[6] See Shoemaker 2011

[7] See Blair 1999

notes this point[8] but does not pursue it in her response to Kennett. I will return to this fact in great detail later on, but I bring it up here simply to show that the affective lives of individuals with ASD are, I think, far more robust than either Kennett or McGeer let on. If this is the case, then before we can draw any firm conclusions about responsible agency from ASD we must make every effort to understand the complexities involved. This project aims to do just that.

Given all of the above, it seems to me that what is missing from this discussion is a detailed, comprehensive, and systematic treatment of the empirical literature on ASD. Therefore, Part I of this work will aim to offer just that. In chapter 2, I survey the several cognitive theories of ASD on offer in the empirical literature. It is claimed by many that individuals with ASD have pervasive deficits in their ability to impute mental states to themselves and others (i.e. impaired ToM) and that this deficit explains the behavioral features of ASD. Others have claimed that the primary deficit in ASD is a deficit in executive function ability. Still others have suggested that the core deficit in ASD is not a deficit at all, but a type of processing bias in which individuals with ASD process information in a piecemeal fashion, focusing on details at the expense of forming a coherent picture of how they fit together (this view is known as the weak central coherence thesis). Each of these hypotheses aims to give a unifying picture of ASD, but each of them falls short in various respects. My aim in chapter 2, then, is to examine the empirical basis for these views, and to give an account of the common ground that they share. That common ground, I argue, is a deficit in counterfactual thinking. More specifically, I claim that it is a deficit in the ability to represent counterfactual states of

[8] See McGeer 2008, fn. 7.

affairs, and I survey evidence that such a deficit is present in ASD and examine the explanatory reach of the hypothesis that this deficit plays a central role in the disorder.

In chapter 3, I examine the empirical data on the affective capacities of individuals with ASD, and I challenge the picture of the emotional lives of individuals with ASD offered by Kennett and McGeer. The available evidence suggests, I argue, that individuals with ASD actually have robust emotional capacities. When co-morbid affective disorders are controlled for, individuals with ASD display the ability to recognize the emotions of others, to understand their own emotions, and to respond empathically to others. In fact, the affective abnormalities in ASD are, I claim, quite limited, and this is a striking departure from the commonly accepted view.

Part II of this work, then, will be aimed at bringing the empirical psychology discussed in Part I to bear on some of the prevailing theoretical views of responsible agency. As I noted above, I believe that Kennett was correct in saying that attending to the facts about ASD can be instructive in understanding agency, but once we have a more fully developed view of autistic agency I think that we will have much more to learn than she expected.

One especially instructive way of formulating the issue of what it means to be a responsible agent is as follows: "To be a responsible agent is to be worthy of X for Y in virtue of Z."[9] Typically, the X variable is thought to denote praise or blame, or some conception thereof. The Y variable may denote an action, attitude, omission, or something similar for which an agent may be praised or blamed. My goal in this project, however, is to examine the proposed theories for which the Z variable is supposed to be a

[9] This is David Shoemaker's formulation. See, Shoemaker 2015, 17.

placeholder. That is, I want to look at what autism research can tell us about the features of agents in virtue of which they may be appropriately praised or blamed for their actions. As I see it, the proposed theories can be separated, roughly and with exceptions, into three broad categories. *Reasons-responsive theories* claim that agents are responsible for their actions insofar as they possess the rational capacity to recognize reasons for acting and to guide their actions in accordance with those reasons. *Real Self theories* claim that agents are responsible for acts which flow from the agent's "real self," or for actions that come from an appropriate mesh between certain key aspects of her agency. Finally, *quality of will theories* hold that the matter of an agent's responsibility for a particular action is decided by determining whether or not the action expressed an objectionable quality of will on the agent's part. I treat each of these three approaches in turn in Part II. Chapter 4 discusses the implications of ASD for reasons-responsive views, Chapter 5 does the same for mesh theories, and Chapter 6 takes on the quality of will approach.

This manner of proceeding, however, raises one immediate worry: there are very few "pure" theories of responsible agency on offer in the literature. That is, many theories tend to mix elements of two or more of the three approaches just mentioned. Because of this feature of the literature, I will focus on discussing particular theories that can plausibly be taken to be *paradigm* cases of the type of theory in question. So, even though the focus will be on particular accounts, I intend for the arguments presented to be generalizable to other views which take the same sort of approach as the paradigm view. Thus, I will be primarily interested in assessing the implications of ASD for theory *types* as doing so will prove to be more illuminating for our understanding of morally responsible agency more generally.

§ Conclusion

So, the project that I will be pursuing in what follows will be aimed at furthering our understanding of the nature of responsible agency. I intend to accomplish this by diving headlong into the empirical literature on autism in order to build a comprehensive moral psychological profile of the disorder. After presenting said profile, I will attempt to show that the facts of ASD's psychology can be used to probe many of the prevailing views of responsibility that are current in the philosophical literature. Ultimately, what I think will follow from this is, contra Kennett, a view in which a particular set of affective capacities must be seen as necessary conditions for responsible agency, and I will argue explicitly for this in the concluding chapter of this work. I have appealed to Kennett's argument regarding psychopathy and autism (and to McGeer's important objection) in order to provide a motivating framework for what follows. I take Kennett to have shown, importantly, that there is an observable behavioral difference between individuals with ASD and psychopaths which leads us to different conclusions regarding that status as responsible agents despite some apparent similarities. However, I believe that she is mistaken about the nature of the difference, and, therefore, it is an open question as to the extent to which the peculiar moral psychology of individuals with ASD can teach us anything about responsible agency. I aim to close that question, and I think that doing so will help to arrive at a more complete philosophical account or responsible agency.

PART I:

The Psychology of Autism Spectrum Disorder

Chapter 2: Toward a Unifying Cognitive Theory of ASD

Having set out the basic aim of, and motivation for, this project, I will now turn to an account of the various psychological features of ASD, as understanding autistic psychology will be the first step toward setting the stage for the views on agency and responsibility which I will consider later in this work. In order to do this, I have separated the discussion into two parts. The first, which will be the subject of the present chapter, will focus on the cognitive features of the psychology of ASD. Here I hope to highlight some of the significant cognitive (in)capacities that are generally observed in studies involving autistic populations. Doing so will help to shed light on some views of agency and responsibility which focus on the necessity of various cognitive capabilities for these concepts. The second part of the discussion, to take place in the following chapter, will then focus on several emotional aspects of autistic psychology. The goal there will be to discover what the available evidence has to tell us about the emotional lives of individuals with ASD in an effort to see whether or not the distinctive emotional capacities associated with the disorder can enlighten our understanding of the role of the emotions in responsible agency. However, before moving on to the details of the empirical literature, it may be helpful to first say a bit about the nature and history of autism more generally.

Autism was first recognized as a distinct disorder by Leo Kanner in his 1943 study, "Autistic Disturbances of Affective Contact." In it, Kanner reported observations on eleven separate cases of children who exhibited a distinct set of behavioral symptoms. "The outstanding, 'pathognomic,' fundamental disorder," Kanner wrote, "is the children's *inability to relate themselves*, in the ordinary way to people and situations from the beginning of life … There is from the start an *extreme autistic aloneness* that, whenever possible, disregards, ignores, shuts out anything that comes to the child from the outside." (242, emphasis in original) In addition to this aloneness, Kanner observed a variety of other common traits among the children studied. They all, for instance, exhibited unusual communicative traits such as speaking of themselves in the second person, repeating particular words or phrases, or parroting the speech of others (also called echolalia). He also observed in them a deeply felt need for order such that they would go to great lengths to keep things the way they were and feel great anxiety at changes in their environment or routine.

Many of these features first noted by Kanner have now been codified as diagnostic criteria for the disorder. In general, autism is now defined as "A pervasive developmental disorder characterized by a pattern of deficits that include impaired (delayed and deviant) communication skills; failure to develop social relationships; and restricted, repetitive, and stereotypical behaviors." (Accardo and Whitman 2002, 39) These features are generally evaluated by using two diagnostic criteria. Drawing directly from the DSM-5, these criteria are as follows:

A. Persistent deficits in social communication and social interaction across multiple contexts, as manifested by the following, currently or by history…
    1. Deficits in social-emotional reciprocity, ranging, for example, from abnormal social approach and failure of normal back-and-forth

conversation; to reduced sharing of interests, emotions, or affect; to failure to initiate or respond to social interactions.

2. Deficits in nonverbal communicative behaviors used for social interaction, ranging, for example, from poorly integrated verbal and nonverbal communication; to abnormalities in eye contact and body language or deficits in understanding and use of gestures; to a total lack of facial expressions and nonverbal communication.
3. Deficits in developing, maintaining, and understanding relationships, ranging, for example, from difficulties adjusting behavior to suit various social contexts; to difficulties in sharing imaginative play or in making friends; to absence of interest in peers.

B. Restricted, repetitive patterns of behavior, interests, or activities, as manifested by at least two of the following, currently or by history…
    1. Stereotyped or repetitive motor movements, use of objects or speech…
    2. Insistence on sameness, inflexible adherence to routines, or ritualized patterns of verbal or nonverbal behavior …
    3. Highly restricted, fixated interests that are abnormal in intensity or focus…
    4. Hyper- or hyporeactivity to sensory input or unusual interest in sensory aspects of the environment. (DSM-5, 50)

Importantly, the criteria above reflect a change in our understanding of autism since Kanner's original study. Most notably, we now recognize autism as a *spectrum* disorder in which those who have the disorder may exhibit only a select few of the symptoms commonly associated with it. Additionally, the recently released DSM-5, quoted above, no longer recognizes Asperger's Syndrome as distinct from autism and instead includes it under the umbrella of ASD. All of this is to say that the common view of autism has changed dramatically over the years and now admits a wider range of individuals. This phenomenon has also led to increased diagnosis and public awareness.[1] Viewing autism as a spectrum disorder has led psychologists to designate certain individuals with autism as "high-functioning." That is, those who exhibit fewer of the deficits typically associated

[1] The prevalence of ASD is now estimated to be at nearly 1% of the general population. See DSM-5, 55. For an excellent summary of the history of ASD and some of the prevailing cultural views associated with it see Wolff 2004.

with ASD (or exhibit them less severely) are often said to be "on the high-functioning end" of the autism spectrum.[2]

With these general, diagnostic facts in mind, we can now proceed to take a closer look at some of the relevant cognitive features of the disorder. The discussion will proceed as follows: In §1, I outline the three most influential cognitive theories of ASD. In §2, I examine some further evidence from the empirical literature and present a new hypothesis which, I think, helps to unify these theories. In §3, I attempt to show that my speculative hypothesis carries a great deal of explanatory power which may be helpful in understanding several other symptoms of ASD and that it could have serious implications in the moral domain. Finally, in §4, I address a pair of questions which may persist regarding my hypothesis.

§1. Cognitive Theories of Autism

With the increase in research devoted to ASD, several theories of autistic cognition have come about which attempt to explain the behavioral symptoms of the disorder. Of these theories, three have proven especially influential. They are the Theory of Mind hypothesis, the Executive Dysfunction hypothesis, and the Weak Central Coherence hypothesis. In this section, I will treat each of these briefly, highlighting their respective strengths and weaknesses. [3]

§1.1. Theory of Mind

I propose to begin the discussion of ASD's cognitive psychological features by paying close attention to what is, perhaps, the most widely accepted cognitive theory of

[2] For the purposes of this work, I will be mostly concerned with such individuals, since my aim is to theorize about morally responsible agency.

[3] For an illuminating treatment of the several cognitive theories of ASD see, Rajendran and Mitchell 2007.

the disorder: the Theory of Mind hypothesis. The evidence from ASD research shows, rather clearly, that those with ASD have a pronounced deficit in their ability to impute mental states to other people. Numerous studies have shown that individuals with ASD lack the ability to take up the perspective of others, to know what another person is thinking. To put it another way, individuals with ASD seem to lack a theory of mind (ToM).[4]

The most famous study to have shown this was published by Simon Baron-Cohen, Alan Leslie, and Uta Frith in 1985. The procedure for the study was simple. The authors write,

> "There were two doll protagonists, Sally and Anne. … Sally first placed a marble into her basket. Then she left the scene, and the marble was transferred by Anne and hidden in her box. Then, when Sally returned, the experimenter asked the critical Belief Question: 'Where will Sally look for her marble?'"(41)

The idea behind the belief question was to determine whether or not the child subjects could make a prediction about the doll's actions based on the doll's false belief, a belief different than that held by the participants, about the location of the marble.[5] If the child answered correctly, then it was concluded that the child was not deficient in ToM. The results of the experiment were significant. The authors report that, "23 out of 27 normal children and 12 out of 14 Down's Syndrome children *passed* the Belief Question on both trials … By contrast, 16 of the 20 autistic children *failed* the Belief Question." (42)

Nonetheless, simple false belief tasks like the Sally-Anne scenario only yielded such clear results when the participants were young children. Because of this, researchers

[4] Philosophers have been interested in ToM for some time and for a variety of reasons. See, for example, Adams 2013. My aim here is to examine how it relates to moral cognition only.

[5] Children were also asked control questions in order to ensure that they understood the scenario correctly (e.g. "Where was the marble originally?" "Where is the marble now?")

began developing more sophisticated ToM tests. For example, in 1999, Baron-Cohen et al. developed a "faux pas" test in which children were asked to identify whether or not a character in a variety of stories committed a social faux pas where doing so required the participants to make inferences based on knowledge of one of the characters' mental state.[6] Whereas the simple false belief task showed a ToM deficit for participants between the ages of 4 and 6, the Faux Pas test proved to yield such results for older participants between the ages of 9 and 11 as well. Nevertheless, since both of these studies showed deficits only in young subjects, it seemed as though deficits in ToM were not the result of a pervasive incapacity for mental state reasoning but were rather the result of delayed development of such a capacity.

More recent studies suggest, however, that ToM deficits persist into adulthood for high-functioning autistic individuals, albeit in subtle ways. Such findings suggest that the explanation for better performance on ToM tasks later in life is the result of the development of compensatory heuristics rather than the development of a ToM. For example, Moran et al. found that high-functioning individuals with ASD demonstrate impairments in ToM when making moral judgments relative to neurotypical individuals when asked to evaluate the permissibility of various actions. They write,

> "[Neurotypical] participants exculpated protagonists for accidental harms caused on the basis of innocent intentions, whereas ASD individuals were less willing to make such exculpatory moral judgments. In judging accidental harms, ASD participants, relative to NT participants, appeared to show an underreliance on the information about innocent intentions and, as a result, an overreliance on negative outcomes." (Moran et al. 2011, 2690)

[6] The following is an example of one of the stories presented to the participants in the study: "Kim helped her Mum make an apple pie for her uncle when he came to visit. She carried it out of the kitchen. 'I made it just for you,' said Kim. 'Mmm,' replied Uncle Tom, 'That looks lovely. I love pies, except for apple, of course!'

What kind of pie had Kim made? Did Uncle Tom know that the pie was an apple pie?"(416).

They take this as evidence, then, that a deficit in ToM persists in individuals with ASD into adulthood.[7]

Additionally, research using functional magnetic resonance imaging (fMRI) data seems to support this conclusion. Studies have shown that a circumscribed region of the brain is involved in social cognition and that this region is also responsible for our capacity for ToM.[8] More specifically, it seems that a particular area of this region, the temporo-parietal junction (TPJ), may actually be selectively responsible for this capacity.[9] If this is the case, then it should follow that individuals with ASD would show decreased activity in the TPJ when presented with tasks that require the use of ToM. A recent study by Koster-Hale et al. provides evidence that this is indeed the case. In this study, researchers measured brain activity in subjects who were making moral judgments about both accidental and intentional harms. In the neurotypical subjects, differences in brain activity were detected between judgments regarding these various types of harms. TPJ activity in these subjects was higher for accidental harms than for intentional harms. However, in the subjects with ASD, there was no difference whatsoever in TPJ activity with respect to judgments regarding accidental and intentional harms. In sum, the authors write,

> [Neurotypical] adults showed a higher response to accidental than intentional harms in the RTPJ, suggesting increased activity in the face of mitigating mental state information. In contrast, ASD adults showed equal activation to both types of stories, suggesting that accidental harms did not elicit more consideration of mental states than intentional harms. (Koster-Hale et al. 2013, 5651)

---

[7] I will have much more to say about both compensatory systems and moral judgment below.

[8] See, for example, Gallagher and Frith (2003).

[9] See Saxe and Kanwisher (2003).

It seems, then, that the neurological data regarding ASD further confirms the view that individuals with ASD show significant impairments in theory of mind. This is the consensus view among those who study the disorder, and the evidence presented so far (and, indeed, further evidence could be offered[10]) is compelling. It seems clearly to be the case that autistic persons have difficulties in understanding the minds of others. Nonetheless, to say that individuals with ASD experience deficits in ToM is not to say that these deficits are the cause of the behavioral and social symptoms of the disorder. It remains to be seen whether this impairment is fully responsible for the difficulty that individuals with ASD have in social situations or if ToM difficulties are merely the result of some other cognitive deficit. One candidate for such a deficit comes in the form of the Executive Dysfunction hypothesis to which I turn next.

§1.2. Executive Dysfunction

While it is clear that individuals with ASD suffer deficits in ToM, it is unclear how, if this is taken to be the primary cognitive deficit, it could account for the non-social/communicative symptoms of the disorder. That is, a deficit in ToM would not, some say, be able to explain by itself the repetitive and restricted behavioral patterns that are often seen in individuals with ASD. Because of this, many have focused their attention on another important cognitive feature of ASD, namely, that some of its behavioral traits seem to be similar to those displayed by individuals who exhibit problems in executive function (EF, also referred to as executive control). While there is some disagreement about how precisely to define EF, it is generally thought of as a set of cognitive processes "that direct behavior regulation and orchestration of attaining a future

[10] See Frith, Morton, and Leslie 1991 for a review.

goal." (Kenworthy et al. 2008, 321) And it may include such processes as "working memory, inhibition, cognitive flexibility, monitoring, planning, and generativity." (321) Many of these processes have been shown to be impaired in individuals who have suffered frontal lobe damage, and the behavioral similarities between such individuals and those with ASD led researches to begin exploring the connection between EF and autism. As Rajendran and Mitchell note, "In contrast to the theory of mind hypothesis of autism the Executive Function account was not born from neurotypical research; rather, its conception came from researchers who noted that some symptoms of autism were similar to those associated with specific brain injury." (Rajendran and Mitchell 2007, 231) Studying these processes is quite difficult for a number of reasons, but it seems to be the case that individuals with ASD show some signs of EF deficits. In this section, I will highlight some of the key empirical findings in favor of this view.

To begin, it must be said that the picture which emerges from the evidence regarding EF in autism is a muddy one. Nevertheless, the theory has proven to be influential, and the evidence in its favor is forceful. In one of the first studies of executive function in ASD, for example, Ozonoff, Pennington, and Rogers (1991) found that individuals with ASD performed significantly worse on EF tests designed to measure planning and set shifting (or, mental flexibility). In the planning task, the Tower of Hanoi, subjects were given a device like that shown in FIGURE 1 and asked to replicate the tower on a different peg by moving one disc at a time and never placing a larger disc on top of a smaller disc. The task is designed to test planning abilities since successfully completing the tower requires subjects to foresee which future configurations will result from a given move. The authors found that subjects with ASD performed markedly less

well than controls on the task and inferred that ASD is, at least partially, characterized by deficits in planning.

FIGURE 1: TOWER OF HANOI
(The devices used for those with ASD are typically simpler than the one shown, consisting of three discs for the "easy" tasks and four discs for the "difficult" tasks.)

In a second EF test, the Wisconsin Card Sorting Task (WCST), Ozonoff, Pennington, and Rogers showed that individuals with ASD were also impaired in set-shifting relative to controls. They describe the task as follows:

> Four cards, varying along the dimensions of color, shape and number, were placed in front of the subject. Subjects were given two decks of cards that varied along these same dimensions and asked to match the cards in the deck with one of the four "key" cards. The experimenter told the subject if he had placed a card correctly or incorrectly, but did not reveal the sorting strategy to the subject. Once the subject had categorized 10 consecutive cards correctly, the sorting principle was changed without the subject's knowledge. The previous strategy then received negative feedback and the subject was expected to switch to the new categorization principle. (Ozonoff, Pennington, and Rogers 1991, 1089)

The purpose of the task, then, is to test the ability of the subject to switch mental sets. The test showed that the ASD group made significantly more perseverative responses than the control groups and, thus, were said to have "specific difficulty shifting cognitive set." (1092) So, from this study, it seemed to be the case that EF deficits were indeed present in those with ASD. Moreover, they argued that these were, in fact, the primary

deficits in the disorder rather than ToM deficits since their findings suggested that EF deficits were both more widespread in the ASD group than ToM deficits and more specific to the ASD group.

In the time since the early EF studies in ASD like the one referenced here, ample attention has been devoted to the Executive Dysfunction hypothesis, and the results have been mixed. In an extraordinarily helpful and thorough review, Elisabeth Hill showed just how scattered the evidence in favor of the theory is by considering the data from a large sample of research.[11] In addition to surveying the available research on planning and mental flexibility, she provided a comprehensive review of studies regarding additional EF processes such as inhibition, generativity, and self-monitoring. The picture that emerges from her review is a convoluted one, and I will offer a brief outline of it here.

With respect to planning, Hill reviewed nine separate studies which consisted of sixteen separate planning tasks. Of the sixteen tests, ten of them suggested impairments in the autistic group. Such mixed results could be the result of several factors. First, planning tasks like the tower task discussed above require more executive functions than simply planning. They require, for example, the ability to inhibit prepotent responses in favor of responses that would achieve the ultimate aim of the task as well as a capacity for working memory. So, it could be the case that the tasks are not measuring only one factor, and this seems to be a recurring problem for other tasks as well (the WCST, for instance, certainly requires inhibition of prepotent responses). Additionally, there is some evidence that performance on planning tasks may be more closely correlated with IQ than

[11] See Hill 2004

with autism, and this is something that has not been adequately accounted for in the literature.

With respect to mental flexibility, Hill suggests that a similarly complex picture presents itself. At first glance, the data suggests that there is a pervasive deficit in mental flexibility in autistic populations, but this is complicated by a number of features as well. While it does seem to be the case that individuals "experience an autism-specific 'stuck-in-set' perseveration," (Hill 2004, 202) the literature presents some problems. For instance, according to Hill, it may be the case that the available evidence does not sufficiently account for correlations between verbal IQ and mental flexibility. Moreover, given that there is a high incidence of learning disability associated with ASD, studies must make sure that they account for any additive effects that such disabilities might cause.

In addition to tests of planning and set-shifting, the literature on EF in autism also includes research on several other executive functions. For example, there is some suggestion that individuals with ASD show significant deficits in inhibition. While the results of inhibition tasks are subject to the same sorts of difficulties discussed with respect to planning and mental flexibility, there is some consensus that those with ASD show significant difficulties in the inhibition of prepotent responses. Some evidence of this can be seen from studies testing autistic groups' performance on the "Windows Task." As described by Hill,

> In this task a participant can win a desired object (chocolate) by pointing to one of two boxes, one of which can be seen to contain the chocolate. However, in order to win the chocolate the participant must point to the empty box, that is the one without the chocolate. Children with autism had significant difficulty inhibiting their prepotent desire to point to the chocolate, the move that meant that the experimenter, rather than the child retained the chocolate. (Hill 2004, 202)

In addition to inhibition, there is some evidence that individuals with ASD show deficits in generativity as well. Generativity is the executive function which involves the ability to generate new and novel ideas and behaviors, and such a deficit, were it present, would go a long way toward explaining the propensity for repetitive and perseverative behaviors.

All of this is to say that several of the executive functions which allow neurotypical individuals to navigate their everyday lives show some signs of impairment in those with ASD. These functions are notoriously difficult to test for a number of reasons: single tasks often involve the use of more than one executive function, executive function can be affected by factors other than autism, such as IQ, and executive functions typically have a quite long developmental trajectory such that subjects of different ages may show strikingly different EF profiles. Nonetheless, it is difficult to deny that individuals with ASD do seem to perform differently than controls on EF tasks and that such performance might be useful in explaining the non-social behavioral features of the disorder. Moreover, there is certainly a similarity between the behavioral patterns of individuals with ASD and those with frontal lobe damage which would suggest that the frontal lobes (and especially the pre-frontal cortex) might play a role in ASD. For these reasons, the Executive Dysfunction hypothesis is likely to remain an influential theory of autistic cognition.[12] It remains an open question, however, what the relationship is between the purported EF deficits and the now well-established ToM hypothesis outlined above.

[12] See Kenworthy et al. 2008 for a review of studies subsequent to the Hill review cited above for further information.

§1.3. Weak Central Coherence

A third cognitive theory of ASD which has proven to be influential is the Weak Central Coherence (WCC) hypothesis. First proposed by Uta Frith,[13] the WCC hypothesis claims that ASD is not characterized by any particular deficit but by a cognitive style that is different from that of typically developed individuals in the way that information is integrated. According to WCC, the cognitive style of ASD is characterized by a tendency to process information in a piecemeal fashion rather than integrating it into a coherent, general whole. In other words, upon hearing a narrative of some kind or perceiving visual information, individuals with ASD pick out the specific details and focus primarily on these rather than processing these inputs globally, unifying them, and searching for some general meaning.

There is evidence that individuals with ASD do indeed process information in this fashion. For example, individuals with ASD perform well on "embedded figure" tasks in which participants are asked to identify or locate a hidden figure or shape that is embedded or hidden in a picture with a more general overall meaning (e.g. a triangle embedded in a picture of a clock). (Frith and Happé 1994) Moreover, this style of information processing could potentially explain a number of the behavioral features of ASD. The desire for sameness, for example could plausibly be linked to a detail oriented processing style. Likewise for such features of ASD as having restricted patterns of interest or the fact that individuals with ASD are often exceptionally talented in subjects that involve systematizing detailed information (often referred to as "savant" skills). While little research exists on WCC in comparison to the EF and ToM hypotheses, it has

[13] See Frith 1989

nevertheless been well-received by researchers and has become an established basis of cognitive research into ASD.

Given that each of these theories offers explanations for different, though sometimes overlapping, sets of autistic traits, it has been proposed that ASD ought not to be viewed as characterized by a single, unifying deficit but instead as fractionable.[14] According to the fractionation hypothesis, each of the widely accepted cognitive theories discussed so far offers sufficient explanation for only a portion of the behavioral symptoms of the disorder. The reason for this is that the disorder itself is best understood as a loosely related set of independent components that often, but not always, co-occur. In short, the search for a unifying cognitive deficit in ASD has failed, on this view, because no such deficit exists. Rather, the various symptoms that make up an autism diagnosis are underwritten by multiple cognitive features which may, at times, present themselves independently of, or to different degrees than, one another. The fractionation hypothesis is, no doubt, still under dispute,[15] but it represents one possible way of reconciling the empirical data on the three most prominent cognitive theories of ASD. Another means of reconciling this data, of course, is by positing some other cognitive deficit which may be primary in ASD and play causal role for each of the major theories under discussion. I will pursue such an explanation in the following section.

§2. Common Cognitive Ground: Counterfactual Thinking

How might we think of the relationship between the WCC, EF, and ToM hypotheses? As Ozonoff, Pennington, and Rogers put it,

[14] See, for example, Happé and Ronald 2008 or Brundson and Happé 2014.

[15] See, for example, Hobson 2014.

> Several possibilities exist, among which are: (1) one deficit is primary and causes the other, which is secondary; (2) one deficit is primary, but does not cause the other, which is a correlated deficit caused by brain damage to a neuroanatomically proximal system; (3) a third deficit is primary and causes both the executive function and theory of mind impairments; and (4) both executive function and theory of mind deficits are independent primary deficits of autism. (Ozonoff, Pennington, and Rogers 1991, 1099)

Here I would like to propose a solution along the lines of (3). In order to do so, I will consider the following question: is there an underlying cognitive ability that might plausibly be construed as playing a causal role in the empirical evidence in favor of the ToM, EF, and WCC hypotheses? I think the answer is yes, and I believe that the ability in question is the ability to represent counterfactual states of affairs.

Representing counterfactual states of affairs is, of course, closely related to the ability to engage in counterfactual conditional reasoning, that is, reasoning by way of conditionals like the following: *If it were the case that X, then Y*. Counterfactual conditionals, thus, are conditionals containing false antecedents which are normally expressed in the subjunctive mood. My contention here is that it is plausible to hold that the culprit behind the ToM, EF, and CC deficits in ASD is an inability to represent the counterfactual states of affairs which may serve as antecedents in counterfactual conditionals. The remainder of this section will aim to show that this type of counterfactual thinking deficit could play a unifying causal role for all three of the cognitive theories under discussion. If I am correct, then the fractionation hypothesis can be rejected in favor of a simple, elegant unifying theory.

If it is to be shown that counterfactual representation plays a unifying cognitive role in ASD, it must first be shown that individuals with ASD show signs of deficits in counterfactual thinking. Again, counterfactual thinking involves suspending one's

presently observed reality and thinking about states of affairs that do not match that reality. (Harris, German, and Mills 1996) Humans, of course, use this sort of thinking quite often in everyday life for a range of reasons including reasoning about possible causes of events, possible outcomes of actions, or even for fantasizing about some imagined reality. Interestingly, the evidence regarding counterfactual thinking in ASD is somewhat mixed. For example, Leevers and Harris showed that those with ASD do not demonstrate deficits in deductive, syllogistic reasoning which contains counterfactual premises.[16] Additionally, Peterson and Bowler (2000) found that autistic subjects were as adept as control groups at performing subtractive counterfactual reasoning by observing subjects' ability to correctly answer questions of the form, "If X had not happened, then what state of affairs, Y, would have occurred?" 50% of participants with ASD failed these questions as opposed to 37% of typically developed participants, a difference that did not rise to the level of significance. Interestingly, of those participants with ASD who passed the counterfactual question only 44% were also able to pass a standard false belief test (compared to 76% in the typically developed control group). Peterson and Bowler explained these results by hypothesizing that the false belief task requires the spontaneous generation of the counterfactual state of affairs while the counterfactual state of affairs is stated explicitly in the counterfactual reasoning tasks. So, based on these two studies, it seems as though children with ASD are able to engage in counterfactual reasoning of some sort. However, in a more recent study, Grant, Riggs, and Boucher (2004) replicated the Peterson and Bowler results but added to that study by including non-standard false belief tasks. In these tasks, the false belief of the character in the story

[16] See Leevers and Harris 2000. Similar results were reported in Scott, Baron-Cohen, and Leslie 1999.

was made explicit[17] so as to test the Peterson and Bowler hypothesis that the explicit features of the counterfactual reasoning tasks were the reason for the improved performance of individuals with ASD as compared to the standard false belief tasks. Altering the experiment in this way proved successful, and the authors summarized their results as follows:

> [I]t appears that autistic children's difficulties in passing standard false-belief tasks derive from the cognitive requirements of the tasks and not from a flawed conceptual understanding of belief. Moreover, the cognitive difficulties that [children with autism] experience on standard false-belief tasks are associated both with the particular cognitive demands of counterfactual conditional reasoning and with drawing inferences/generating propositions where critical information is not made explicit. (Grant, Riggs, and Boucher 2004, 184)

What all of this suggests is that individuals with ASD are able to reason about counterfactual states of affairs when such reasoning is either deductive as in the Leevers and Harris study or involves states of affairs in which the counterfactual components are made explicit. When the counterfactual elements are not made explicit, however, individuals with ASD show clear deficits in counterfactual thinking.[18] Peterson and Bowler propose that this can be explained by deficits in generativity, but an alternate explanation may be that individuals with ASD are deficient in the ability to represent counterfactual states when not provided with their component parts, an explanation supported by the findings of Grant et al. Thus, given that counterfactual thinking involves conditional reasoning of the form, "If it were the case that X, then Y," where X is some state of affairs that is inconsistent with present reality, my contention is that individuals

[17] The following is an example of one such story: "This is Mary. Mary wants to find her kitten. Mary's kitten is really in the bedroom. Mary thinks her kitten is in the kitchen. Where will Mary look for her kitten?" (Grant, Riggs, and Boucher 2004, 179)

[18] For additional evidence, see Begeer et al. 2009.

with ASD may be seen as deficient in their ability to represent the state of affairs corresponding to X in the above conditional in cases where X is not made explicit. In real life interactions, of course, such a state is rarely made explicit, and therefore a deficit of this kind could provide an explanation for the behavioral symptoms of ASD.

§2.1. Counterfactual thinking and ToM

The preceding suggests that a deficit in counterfactual thinking is indeed part of the autistic profile. What remains to be shown, however, is how this deficit might be seen as underlying the three dominant cognitive theories of the disorder. In order to see how this might be the case, I will begin by examining the relationship between counterfactual thinking and ToM.

The available data suggests that there may be a close connection between counterfactual thinking and both ToM and EF. It is evident, upon reflection, that representing the mental state of another person or of oneself involves representing a state of affairs that does not match one's present, observed reality insofar as it requires one to represent the content of that mental state. This can be seen by simply considering the simple false belief task like that in the Sally/Anne scenario above. Does imputing a false belief to the doll in the story require the use of counterfactual reasoning? There is, I think, a straightforward sense in which it might, since a subject might reason as follows: *The marble is in the box, therefore Sally will look in the box. However, if it were the case that Sally believed that the marble is in the basket, then she would look in the basket.* The latter conditional is, of course, a counterfactual one, and if a subject were unable to make use of such a conditional, then she would be expected to fail the simple false belief task. I think that the same is true of the more advanced ToM tasks as well. In the Faux Pas test,

for example, a subject might reason thusly: *If person A had believed X, then he would not have done/said Y*. So, it seems to me that there is good reason to think that some type of counterfactual reasoning ability is, indeed, required for ToM tasks. Additionally, there is some compelling empirical evidence that this is true. In a 1998 study, Riggs et al. tested typically developing children using both false belief tasks and counterfactual reasoning tasks. The study made use of a standard false belief task similar to the Sally/Anne story above in which an object is moved without a character's knowledge. The other task used was the "Post Office" story which went as follows:

> Sally and Peter were in the house but Peter wasn't feeling very well, so he went to bed. Sally then went to the shops to get some medicine (the experimenter moved the doll out of sight behind his back). While Sally was at the shops, the phone rang and the man from the Post Office asked Peter to come and help put out a fire. Peter got out of bed and went to the Post Office. (Riggs et al. 1998, 76)

After hearing this story, children were asked both a false belief question (Where does Sally think Peter is?) and a counterfactual physical state question (If there had been no fire, where would Peter be?). The results showed a high degree of correlation between performance on the physical state counterfactual task and performance on the false belief task. The experimenters then ran additional experiments to factor out the effects of verbal mental age, narrative competence, and difficulty with general conditional reasoning. Even after these additional experiments, Riggs et al. concluded that deficits in counterfactual reasoning could be seen as underlying poor performance on false belief tasks.[19] Therefore, in order to have robust ToM abilities one must have a developed ability to represent counterfactual states of affairs. Moreover, this view is well supported by neurological data which has shown that counterfactual thinking implicates the same

[19] Similar results are reported in Guajardo, Parker, and Turley-Ames 2009.

brain structures – or, at least, structures that are proximal to – those involved in both ToM and EF.[20]

In examining the nature of the relationship between counterfactual thinking and ToM, many have proposed that certain executive functions may play an important role as well. As Riggs et al. and others have shown, performance on counterfactual thinking tasks predicts performance on ToM tasks in many cases, and given that certain features of executive function have been shown by Ozonoff, Pennington, and Rogers to predict ToM task performance,[21] it has been proposed that EF plays a mediating role between counterfactual thinking and ToM.[22] Specifically, it has been shown that inhibition, working memory, and mental flexibility all serve to mediate this relationship in some way. For example, Drayton, Turley-Ames, and Guajardo reported the following results:

> [A]nalyses confirmed that counterfactual thinking was a significant predictor of both [false belief] performance and working memory. Children who had higher counterfactual reasoning scores performed better on FB tasks and on the working memory measures. Moreover, working memory was a significant positive predictor of FB … This equation suggests that working memory explains, in part, why changes in counterfactual reasoning skills effect changes in FB performance. (Drayton, Turley-Ames, and Guajardo 2011, 541)

Similar results were reported in the study for inhibitory control as well, however, the mediating relationship in that case was not as pronounced. The presence of this mediating relationship is important because it may help to illuminate the way in which a deficit in counterfactual thinking can be seen to underlie the EF theory of ASD as well.

---

[20] See, Gallagher and Frith 2003; Saxe and Kanwisher 2003; Beldarrain, et al. 2005; De Brigard, et al. 2013; Van Hoeck et al. 2014.

[21] See also, Carlson and Moses 2001.

[22] See German and Nichols 2003; Guajardo, Parker, and Turley-Ames 2009; Drayton, Turley-Ames, and Guajardo 2011.

§2.2. Counterfactual thinking and EF

The above considerations constitute fairly strong evidence that performance on ToM tasks is underwritten by the ability to think counterfactually, but this is less clear in the case of EF. Nevertheless, there is reason to think that EF relies, to some extent, on counterfactual thinking as well. In the first place, at least one process that has traditionally been associated with EF clearly requires counterfactual thinking, namely planning. For example, In the Tower of Hanoi task successfully completing the tower would seemingly require an ability to represent future configurations of discs in order to decide on a given move to make. In other words, planning a successful path to finishing the task requires reasoning as follows: *If I were to make move A, then configuration X would result. If the tower were configured in manner X, then I could make move B. If I were to make move B, then configuration Y would result,* etc. This is quite clearly a chain of counterfactual inferences, and, therefore, deficits in counterfactual thinking would explain poorer performance on the task. There may be reason to suspect that other traditional EF tasks (such as the WCST and the Windows task) rely implicitly on counterfactual reasoning as well, though this is less clear than the case of planning. However, there is evidence that suggests that counterfactual thinking is much more fundamental to EF than simply being implicated in the traditional EF tasks.

While the precise nature of EF is disputed,[23] the view that counterfactual thinking has a foundational role in EF becomes more plausible if one adopts a problem-solving model of EF.[24] According to the problem-solving framework, EF consists of a

[23] See Jurado and Roselli 2007 for a review.

[24] See Zelazo et al. 1997.

hierarchically arranged set of processes which all function together in order to solve problems and instantiate goal-directed behavior. At the top of the hierarchy is the Problem Representation phase in which the individual represents the problem to be solved. This phase is followed by Planning, the phase in which a plan is devised for solving the represented problem. From there, the model proceeds to the Execution and Evaluation phases. The former is directed toward implementing the plan, and the latter is directed toward detecting errors and making corrections. On this understanding of EF, it is clear that counterfactual thinking plays a significant role. This is true for at least two reasons. First, a representation of a problem may, of course, require counterfactual thinking, and this is especially true in the case of any dilemma (e.g. If I perform act A, then X will follow, but if I perform act B, the Y will follow). Second, and more importantly, as I have just now suggested, the planning phase necessarily requires counterfactual thinking insofar as planning requires representing counterfactual, future states of affairs. What a model like this suggests is that one plausible way of thinking about EF is to view it as a system in which processes such as inhibition, working memory, and mental flexibility serve to mediate the relationship between our counterfactual representations and our actions in much the same way that these have been shown to mediate the relationship between counterfactual thinking and false belief understanding. On this proposal, a problem is represented, which may or may not involve counterfactuals, and a plan is established by way of counterfactual thinking. These are then executed and evaluated by carrying out processes such as inhibitory control, working memory, cognitive flexibility, etc. If this is the case, then a deficit in counterfactual thinking could account for apparent EF deficits insofar as the widely

examined EF functions would be left with nothing to execute or evaluate or, alternatively, would operate only on a rigid, factual set of information. An individual who is unable to represent counterfactually based problems or represent the counterfactual states necessary to make plans would, therefore, predictably behave in restricted, repetitive ways.

§ 2.3. Counterfactual thinking and WCC

Finally, it remains to be shown whether or not counterfactual thinking may play a causal role in WCC. To date, there have not been any studies conducted examining the relationship between central coherence and counterfactual thinking, and, because of this, any attempt to articulate a relationship between the two will be necessarily speculative. Nevertheless, there seems to be some basis for believing that deficits in counterfactual thinking play a causal role in WCC. Recall that WCC theory is premised on the claim that unlike typically developed individuals who process information globally, people with ASD process information in a localized manner. As Frith and Happé write, "[A]n individual does not normally take a situation detail by detail … In all ordinary instances he has an overmastering tendency simply to get a general impression of the whole; and, on the basis of this, he constructs the probable detail." (Frith and Happé 1994, 121 quoting Bartlett 1932, 206) WCC theory argues that individuals with ASD have an informational processing bias toward details rather than general pictures and are, therefore, not inclined to search for general meaning in the manner described by Bartlett.

This generalized way of processing information, however, rests on the individual's ability to interpolate local information from one's global view of the information being processed (to, in effect, fill in the details), and interpolation of this sort involves a straightforwardly counterfactual ability. It requires the individual to represent

details that have not been observed, and which, therefore, do not match the subject's observed reality, and to place them within the generalized whole. An individual who is deficient in counterfactual thinking, then, would be expected to prefer localized information processing since the interpolation of information in global processing would be difficult. Given the lack of data on this matter, future research into the relationship between counterfactual thinking and information processing styles, and especially whether counterfactual deficits predict localized processing biases, could prove to be extremely valuable.

In light of the evidence reviewed so far, a strong case can be made for the view that counterfactual thinking may play a causal role in each of the cognitive deficits that have traditionally been posited as characterizing ASD. If this is true, then the recent attempts to fractionate ASD may be premature since it may be the case that a theory which gets us closer to a unified picture is available. The view that counterfactual thinking plays a unifying cognitive role in ASD (call this the CT hypothesis), moreover, becomes even more plausible when one considers the further explanatory power that it has for giving an account of a number of other widely recognized behavioral traits of ASD. The following section will aim to make this explanatory power evident.

§3. The Explanatory Reach of the Counterfactual Thinking Hypothesis

One important desideratum of any theoretical hypothesis is that it be able to explain as many of the relevant phenomena as possible. Recall, for example, that one of the primary reasons for rethinking the ToM hypothesis was that it could not explain the non-social features of ASD. Since the CT hypothesis purports to play a causal role for both ToM and EF, it can explain all of the features of the disorder that these theories

explain. However, it seems to me that it can offer an elegant explanation for several other features of ASD as well. It will have something to say about various affective features of the disorder (an issue I will return to in chapter 3), and it will may have far-reaching implications for moral reasoning (to be considered at length in chapter 4) as well. Additionally, though, I think that it can make some headway in explaining the well-documented deficits in pretense that autistic children experience, and I believe it can offer a compelling explanation for the deficits in episodic thinking which have inspired a growing literature in recent years.

Regarding pretense, several studies have identified a deficit in autistic children when it comes to pretend play. Indeed, this was observed by Kanner in his original 1943 study, and it resulted in a deficit in imagination being named as one of the primary diagnostic criteria for the disorder in a subsequent review by Wing and Gould (1979).[25] Since that time, substantial evidence has emerged which verifies that children with ASD do indeed demonstrate impairments in pretend play relative to properly matched control groups.[26] For example, one study tested pretend play by observing sessions of free play in children with ASD and compared them with observations of groups of typically developed children and children with Down's syndrome. "On average, the group of nine individuals with autism involved in the study were engaged in pretend play on less than 1 percent of [the] 60 sampling points, compared with 5 percent and 4 percent of occasions for Down's syndrome and typically developing individuals respectively." (Jarrold 2003, 381) The difference in group performance, thus, turned out to be significant. So, while

[25] As was noted earlier, the DSM-5 no longer includes imagination impairments as one of the primary diagnostic criteria.

[26] For a helpful review of some of this literature, see Jarrold 2003.

there are several questions regarding pretense in ASD that still need to be answered, the available evidence suggests a difficulty exists, and this is especially true with respect to spontaneous pretense.

The underlying cause of this deficit in pretense is not well understood, but it seems to me that it could be, at least in part, explained by the CT hypothesis. If a child is deficient in her capacity to represent counterfactual states of affairs, then we would expect her to be deficient in the ability to be sufficiently imaginative to engage well in pretend play. That is, if pretending requires the ability to, say, substitute one object for another (e.g. to pretend that the floor is lava), then this is straightforwardly a matter of counterfactual thinking, and we should expect to see the deficits that have so far been observed.

One possible problem with this, however, is that it does seem that children with ASD are able to engage in pretend play when they are explicitly prompted to do so (e.g. A child is given a doll and a set of props and asked to show the experimenter how the doll might wear a hat. The child then must choose a prop to stand in for the hat[27]). There are a number of difficulties for tests involving prompts,[28] but even if we grant that these tests show legitimate engagement in pretense this would be in line with the evidence on counterfactual reasoning. This is because it has been shown that individuals with ASD are more successful at counterfactual reasoning when the counterfactual state of affairs is made explicit, and insofar as prompted pretend play makes the grounds for pretense

[27] See Jarrold 2003 for a discussion of such experiments.

[28] For example, it is unclear whether the children are, in fact, pretending in these cases or if they are simply making the best of the props that they have been given without actually understanding the object as standing in for some other object.

explicit we should not be surprised to see improved performance on the part of those with ASD. Thus, it seems to me that the CT hypothesis can account for the empirical data on pretense in a simple and plausible way.

In addition to offering an explanation of the pretense deficit, the CT hypothesis can, I believe, tell a plausible story about another deficit in ASD that has garnered some recent attention, namely, the deficit in episodic thinking. The first serious work on episodic thinking in ASD was aimed at testing episodic memory, a type of memory involving personally experienced events (or episodes). As Ryan, et al. describe it, "Episodic … recollection involves thinking about a past event – it is personal, emotional, populated with players and specific places, imbued with detail, and it often has relevance to our sense of self and the meaning of our lives." (Ryan, Hoscheidt, and Nadel 2008, 5) As such, episodic memory is an integral part of our everyday lives, and the available evidence suggests that individuals with ASD experience impairments in it. For example, in a 2007 study, Bruck et al. administered a questionnaire to a group of high-functioning children with ASD which asked questions about events in the children's recent and distant pasts, verifying the truth of the answers with the children's parents. During testing the children also participated in a magic show, and the experimenters, after a sufficient delay, asked questions about this event to gauge the children's recollection. The results were striking, and the authors summarized them as follows:

> We found that children with ASD showed deficits in memory for personally experienced events. Relative to typically developing, age-matched peers, ASD children showed these deficits for events in their far past as well as their recent past. The deficits were found in terms of the number of details provided for various events as well as the accuracy of replies to open-ended, specific, and yes/no questions. For events from their distant past, not only did children with ASD show deficits in the number of details provided, but also more of these

> children compared to normally developing children failed to recall the events at all. (Bruck et al. 2007, 89)

Other studies have generated similar results. For example, Bowler, Gardiner, and Gaigg (2007) found both that adults with Asperger's syndrome do, indeed, show deficits in episodic memory and that this memory was qualitatively the same as that experienced by typically developed individuals. Additionally, results from Crane and Goddard (2008) confirm that adults with ASD demonstrate an episodic memory deficit while also showing that these same individuals possess intact semantic memory. Thus, the evidence in favor of the existence of an episodic memory deficit in ASD is quite compelling.

In addition to episodic memory, there seems to be a deficit in episodic future thinking in ASD as well. In the first study testing episodic future thinking, Lind and Bowler tested a group of high-functioning individuals with ASD by presenting subjects with both a memory and a future thinking prompt (e.g. Try to remember something that happened to you. Try to imagine something that will happen to you.). "In each condition," they write, "participants were asked to try to remember/imagine events from (a) today, (b) yesterday/tomorrow, (c) a week ago/in a week, (d) a month ago/in a month, (e) a year ago/in a year, (f) 5 years ago/in 5 years, and (g) 10 years ago/in 10 years." (Lind and Bowler 2010, 899) Their findings confirmed the autistic deficit in episodic memory, but they also showed that the same sort of deficit exists in episodic future thinking, and these results were replicated in a later study by Terrett et al. (2013) Interstingly, this later study hypothesized that the underlying cause of deficits in episodic future thinking would be a deficit in EF, specifically in set-shifting, since, the authors suspected, it may be the case that episodic future thinking requires subjects to shift between memories and expected future events in order to represent future episodes.

However, the study showed that no correlation existed between set-shifting capacity and the capacity for episodic future thinking. This is important for my purposes as it shows that at least one EF explanation for episodic future thinking deficits is off the table.

So, can the CT hypothesis offer a plausible explanation for these deficits in episodic cognition? I think that it can, and the explanation it offers is a simple and straightforward one. What are past and future episodes, after all, if not counterfactual states of affairs? They are, in an obvious sense, states of affairs that are not reflective of a present reality.[29] So, to the extent that an individual is deficient in the ability to represent counterfactual states of affairs we should expect that person to also be deficient in representing past and future episodes that run counter to a present reality. Moreover, there is some evidence that episodic cognition implicates some of the same core brain structures that have been cropping up throughout the course of this discussion,[30] and there is a growing body of experimental evidence which suggests that counterfactual thinking plays a causal role in the episodic cognition system.[31] All of this lends further plausibility to the notion that deficient counterfactual thinking may provide the greatest explanatory reach as the primary cognitive deficit in ASD.

§5. Conclusion

My aim in this chapter has been to present and defend a speculative hypothesis which could help to unify the available evidence regarding the various cognitive features

[29] There is one sense in which past episodes are not counterfactual since such events *actually did* take place. Nevertheless, since my claim is that individuals with ASD have trouble *representing* states that do not reflect *present* reality, past episodes will count as being counterfactual states for my purposes here.

[30] Lind and Bowler, for example, note that episodic memory and episodic future thinking occur "within medial prefrontal, medial temporal, and parietal regions, as well as the lateral prefrontal cortex, lateral temporal cortex, and occipital lobe." (Lind and Bowler 2010, 896)

[31] See Van Hoeck et al. 2010; Van Hoeck et al. 2013; Schacter et al. 2015.

of ASD. I began by surveying the three dominant cognitive theories of the disorder: the Theory of Mind hypothesis, the Executive Dysfunction hypothesis, and the Weak Central Coherence hypothesis. I then suggested that the underlying feature of all of these hypotheses seemed to be an inability in those with ASD to represent counterfactual states of affairs, a suggestion that I have termed, the Counterfactual Thinking Hypothesis. Others have argued for the view that counterfactual reasoning is behind deficits in ToM, but, to my knowledge, none have suggested it as a candidate for a unifying feature of the disorder. Such a suggestion is, I think, very plausible, and I offered evidence that counterfactual reasoning may be behind other features of autism as well – namely, deficits in pretense and episodic cognition.

My hypothesis is, by nature, speculative, and further research is needed to either prove or disprove it. For example, there have been no studies conducted which examine the relationship between central coherence and counterfactual thinking. Such studies could potentially shed light on the nature of the counterfactual thinking deficit in ASD and its relation to information processing styles. On a more basic level, greater understanding of counterfactual thinking in ASD would help us to understand the extent of the deficit and just how far-reaching its implications may be. One way of studying this might be to systematically examine the use of subjunctive language in autistic populations. Clinical impressions of language use in ASD suggest that the use of the subjunctive mood is severely limited,[32] but, shockingly, no experimental data exists to corroborate this. To the extent that language use reflects patterns of thought, the sub-

[32] I am grateful to clinicians Rebekah Moilanen and Penny Roberts for their insight on this matter.

standard use of the subjunctive mood might betray deficits in counterfactual thinking in ASD. This is a matter that warrants further investigation.

Finally, investigation into counterfactual thinking in other clinical populations could potentially be helpful in understanding its relationship with ASD. Studying this type of thinking in those with mild or severe learning disabilities might help us to determine whether or not counterfactual thinking deficits are specific to ASD. However, such studies must be clear about the precise nature of any observed counterfactual thinking deficit, since such a deficit may present in two differing ways. First, it may be the case that an individual is able to represent counterfactual states of affairs but deficient in making any kind of inference from these states. Second, it may be the case that an individual is capable of making sound inferences but incapable of representing the requisite counterfactual antecedents. I have been arguing here that the former deficit is operative in ASD, and any future studies on other clinical populations will need to identify the relevant deficit precisely.

If the CT hypothesis is correct, then it could have considerable implications for future ASD research and for the development of interventions. What has been said so far, I think, puts the hypothesis on firm ground. However, it is important to note, that my strong thesis that counterfactual thinking is *the* unifying deficit in autism need not be accepted for the theoretical conclusions regarding moral responsibility to hold. Instead, all that I need in order to defend the conclusions in Part II is the weaker thesis which holds that counterfactual thinking deficits are indeed present in ASD and that they are more central to the disorder than has so far been recognized. This is, clearly, a much weaker thesis, but if it is correct (and I think the preceding discussion shows this to be the

case) it would still be an important finding that could have far-reaching implications for autism research. The explanatory power of this hypothesis is difficult to deny, and the following chapters will, I hope, shed even more light on that power, as many of the affective features of ASD can, I think, be better understood by reference to the CT hypothesis (in either its strong or weak form), and once these features are understood, several important conclusions regarding the moral psychology of the disorder will be soon to follow.

Chapter 3:

Understanding the Emotional Profile of ASD

Having gained some understanding of a number of the cognitive features of ASD and my proposed unifying hypothesis regarding those features, I would now like to examine the basic affective features of the disorder. While there are concrete, identifiable theoretical splits in the literature on cognition in ASD (i.e. those between Theory of Mind (ToM), Executive Function (EF), and Weak Central Coherence (WCC) accounts), no such split seems to appear with respect to the affective features. Instead, widespread disagreement persists on the nature and extent of particular emotional deficits in the disorder. This makes discussing them rather more difficult, and the sheer complexity of the emotional profile that is characteristic of ASD serves to make the difficulty even greater. Nevertheless, there do seem to be several features of the emotional lives of individuals with ASD that lend themselves to some degree of generalization. It is these features with which I will be concerned in this chapter.

My aim will be to address the emotional profile of ASD by examining it under the rubric of self- and other-oriented emotional features. When investigating the emotional lives of individuals, there are several features that, I think, are of primary interest, and if we want to understand an individual's emotional profile, there is a discrete set of

questions that could be asked which would tell us most of what we want to know. Among these are the following: Which emotions does the individual feel? How does the individual understand these emotions? Does the individual recognize emotional displays in other people? How does the individual respond to such displays in others, if at all? The former two questions serve to pick out the relevant self-oriented features of the emotional profile, and the latter two pick out the other-oriented features.

In §1, I will attempt to answer the first two questions with respect to individuals with ASD. In that section I will first examine the empirical evidence which indicates that there are a number of emotions which individuals with ASD clearly experience. However, there is also considerable evidence which suggests that those with ASD do not experience many of the complex, social emotions, and I will do my best to make sense of this evidence in §1 as well. I will then turn to a discussion of the way in which individuals with ASD process emotions. Although much of the evidence suggests that these individuals are impaired in their ability to understand their emotions, there is, nevertheless, substantial disagreement over the extent to which this is the case, and I will attempt to offer an account of the source of this disagreement.

In §2 I will turn my attention to the other-oriented features of the autistic emotional profile. To begin, I will take on the problem of emotion recognition in ASD. Many researchers have sought to test whether those with ASD are deficient in their ability to recognize emotion in other people, and many have drawn conclusions on opposing sides of this issue. In light of this disagreement, I offer a discussion of some new research that can, I think, help to resolve this dispute. From here, I move on to answering the last of the four questions posed above: how, if at all, do individuals with

ASD respond to the emotions of others? Answering this question will involve examining the capacity for empathy in ASD, and this issue will occupy a large portion of the chapter.

Finally, in §3, I discuss how the evidence regarding the emotional features of ASD might be understood in light of the Counterfactual Thinking Hypothesis presented in the previous chapter. I argue that there is good reason to suspect that some of the deficiencies experienced by those with ASD in these various features of affective psychology can be explained by appealing to deficits in counterfactual thinking. If this is true, then it will be further evidence in favor of the CT hypothesis.

§1. ASD and the Self-Oriented Features of Emotion

In this section, I will be interested in answering two related questions: Which emotions do those with ASD experience, and how do they understand these emotions? I turn to the first now.

§1.1. Emotional Experience and Emotion Expression

Determining which emotions those with ASD have access to is, at bottom, a matter of understanding the emotional phenomenology of the disorder, and there are two basic ways of studying this. The first is to simply ask individuals with ASD to report their emotional state in response to a variety of stimuli, and the second is to study the way in which individuals with ASD *express* emotions. Studies that take the first strategy are less common because they require participants who are conversant enough to accurately report their emotions. As a result, there is far more evidence available relating to emotional expression in ASD than there is evidence of first hand reports. The difficulty with focusing on expression as a means of delineating the range of emotions to which

individuals with ASD are prone is that this strategy rests on the assumption that emotional expressions are accurately reflective of the emotional state of the individual who displays them. I will accept this assumption for the time being, but I will return to the problem briefly in §2. For now, I propose to begin the discussion of emotional phenomenology by examining the issue of emotion expression.

One of the early diagnostic features of ASD was that those with the disorder seemed to be "affectively flat."[1] That is, they seemed not to express emotion in a way that typically developed individuals might. For example, Attwood, Frith, and Hermelin (1988) conducted a study in which the use of gestures was compared in children with ASD and children with Down's syndrome and which showed that children with ASD understood and used basic instrumental gestures (e.g. pointing to an object). However, not a single autistic child in the study used any sort of expressive gesture (i.e. a gesture intended to convey some feeling state such as putting one's arm around another person who is sad or injured) while those with Down's syndrome were able to do so frequently. In another study, Dawson et al. observed interactions between children with ASD and their mothers and found that although these children did display emotion, they were less likely to do so in response to expressions from the mothers. As the authors put it, "The autistic child often fails to combine emotion and eye contact in a single act that conveys communicative intent. The autistic child expresses emotion as frequently but does not readily communicate emotion to others." (Dawson et al. 1990, 344) One problem with both of these studies, however, is that they only include ASD participants that also have learning disabilities (i.e. IQ's lower than 70). There is some evidence available that seems

[1] This is the phrase used in the DSM-III-R as discussed by Capps et al. 1993.

to indicate that high-functioning children with ASD are much more emotionally expressive than these studies suggest.

In one compelling study, Capps et al. (1993) gathered parental reports of affective expressiveness in high-functioning children with ASD. They then showed videos of emotion-eliciting scenarios to those same children and recorded spontaneous facial affect expressions. What they found was that the children with ASD showed at least as much spontaneous facial expression at appropriate times during the videos, and the parents reported, on the whole, that the children were not, in fact, affectively flat. The authors write,

> These findings corroborate a growing body of clinical and empirical evidence that counters the view that children with autism are affectively flat … [O]bservation of the older autistic children, and parent reports on the emotion behaviors of both young and older children with autism, suggest that they have strong emotional reactions … [O]lder nonretarded autistic children displayed *more* facial affect while watching the empathy videos than did comparison children … Parents' perceptions were consistent with these observations in that autistic children were not viewed as affectively flat. Rather, their parents perceived them to be highly expressive… (Capps et al. 1993, 482)

So, it seems to be the case that deficits in expression are tied more closely to individuals who are on the lower-functioning end of the autism spectrum. Other studies have been undertaken which have helped to corroborate this fact. Yirmiya et al. (1992), for example, tested the capacity for empathy in high-functioning individuals with ASD and, in so doing, asked subjects to verbally report emotional experiences that they had had, and many of the autistic participants were able to do so. Several other studies have subsequently taken place and have produced similar results.[2] So, it would seem that high-functioning individuals with ASD do not experience the same deficits in emotion

[2] See Begeer, et al. 2008 for a discussion of these.

expression that those with learning disabilities demonstrate. However, the fact that children with high-functioning forms of ASD perform this well on expressive tasks in the laboratory may not be reason enough to expect that this might generalize to more natural circumstances. As Begeer, et al. put it, "Children with [high-functioning ASD] show remarkable adequate emotional expressive skills. However, these data were generally obtained under laboratory conditions, and may present an overly positive perspective…" (Begeer et al. 2008, 346) Nevertheless, despite the severity of many of the affective deficits in ASD, emotional expression seems to be left mostly intact by comparison.

So, the view that individuals with ASD simply do not express emotions is, at least with respect to those high on the spectrum, false. This, however, does not tell us anything about the range of emotions being expressed. Paying attention to the types of expression taking place, however, can tell us a great deal about the emotions to which those with ASD are prone, and I will do my best to outline the relevant findings in this regard now.

One thing that is well-established in the literature is that individuals with ASD seem clearly to experience the basic emotions of happiness, sadness, anger, and fear.[3] For example, in the survey of parental reports mentioned above, Capps et al. found that parents consistently reported observing all of these simple emotions in their children, and other studies have arrived at similar findings. Rieffe, et al. (2007) tested high-functioning children with ASD in their ability to experience these basic emotions by gathering self-reported data from participants regarding their emotional experience. The results of their study showed that individuals with ASD clearly had access to simple emotions, but they did show some difficulty in distinguishing these from one another. So, most researchers

[3] See Nuske, Vivanti, and Dissanayake 2013 for a helpful review on this matter.

are in agreement that experience of simple emotions is unaffected in autism. The more pressing question for the purposes of this chapter is whether or not there are any emotions which those with ASD are not prone to experience.

One especially insightful treatment of this question can be found in a recent monograph by Hobson, et al. (2006) In it, the researchers conduct a series of studies aimed at understanding the ability of those with ASD to identify with others and, thus, develop a sense of self-awareness, and they do so by testing their capacity to experience and understand social emotions – primarily, guilt, pride, and shame/embarrassment.[4] The results of these studies are important, and so I will devote some extended treatment to them here. In the first study,[5] Hobson, et al. focused on parental reports of emotion expression in children with ASD. Like the Capps, et al. study cited above, they found no difference in reports of basic emotions (again, these are fear, anger, happiness, and sadness) between ASD groups and control groups. However, the results regarding more complex emotions were far more mixed. In addition to the basic emotions, the authors asked the parents to report on their children's expression of jealousy, pride, shyness, embarrassment, shame, pity, concern, and guilt, and they found some pronounced differences between groups as a result. First, they found that the children with ASD were seen as expressing both jealousy and pride, but the expression of pride was somewhat abnormal in the sense that the children with ASD often acted proud of their accomplishments but seldom saw fit to share these accomplishments with others.

[4] Note that the authors often refer to embarrassment, shame, and coyness as having the same essential basis, namely, that something about the subject is being observed by an audience, and, therefore, to experience any of these three requires sensitivity to the perceptions of the audience.

[5] See Hobson 2006, chapter 3.

Additionally, the results of the parent interviews showed no significant group difference with respect to shyness. However, significant group differences appeared in the reports regarding both pity and concern, but the source of the significance was not an absence of these emotions in the ASD group but atypical expression, or differences in the degree, of the emotion instead. So, the children with ASD did show signs of experiencing both pity and concern for others, but these occurred in a manner different than that in the control group. Where the authors did note a relative absence of emotion was with respect to embarrassment, shame, and guilt. For these three emotions, none of the parents of the ASD group reported that they were clearly present in the children.[6] The authors take this to be evidence that "in virtue of their limited propensity to identify with others, children with autism relatively lack organization of feeling into self/other structures of experience." (Hobson 2006, 71)

The second of their studies, however, produced some seemingly conflicting results.[7] Given that they claimed that what is distinctive about the autistic emotional profile is that it lacks an intersubjective organization, they sought in this study to test whether or not high-functioning adolescents with ASD could recognize instances of pride, guilt, and shame/embarrassment in others and whether or not they could report instances of experiencing these themselves. In the first part of the study, participants were shown videos of actors pretending to experience one of these three emotions, and they

[6] It ought to be noted that this is a point of disagreement between Hobson, et al. and the study from Capps, et al. In the Capps study, parents reported observing both shame and guilt in children with ASD. There are many reasons why such disagreement might arise, not the least of which is the fact that parental reports may be subject to some editorializing on the part of parents who may unknowingly project their own emotions onto their children. I won't address this methodological worry here, but I will say that I find the Hobson study much more compelling, especially in conjunction with the other studies in the monograph, and so I take their results to be more accurate in my discussion.

[7] See Hobson 2006, chapter 4.

were asked to describe what the actor was feeling. The authors found that there was no significant difference between the ASD group and the control group with respect to any of the emotions depicted. In the second part of the study, then, the authors sought to examine how, or if, the participants related their own personal experience of pride and guilt to others. So, they asked each participant if s/he could tell about a time when s/he felt one of the emotions in question. Once again, no significant group differences were found. That is, children with ASD were as able, on the whole, to report times of experiencing either pride or guilt. However, the authors did note that many of the responses from the children with ASD were atypical. Moreover, in the case of guilt, a majority of the participants in the ASD group could not provide a personal example of the emotion. The authors summarize the outcome, saying,

> The results serve to highlight how children and adolescents with autism can recognize, experience and reflect upon *certain* qualities of pride, and in a minority of cases also guilt. At the same time, the study failed to provide decisive evidence whether they can recognize, experience, and reflect upon the more person-centered and interpersonally engaged qualities of the feelings. (Hobson 2006, 89)

As a result, the authors performed a third study in which they aimed to examine these person-centered qualities directly.

In this third study,[8] the authors directly tested the children's emotional responses to situations that were designed to elicit pride, guilt, or coyness in relation to another person.[9] In the scenarios designed to elicit pride, the group differences were insignificant. Participants with ASD seemed to show pride as often as their typically developed peers.

[8] See Hobson 2006, chapter 5.

[9] For example, in order to test pride, children were asked to draw a house. The drawing was complemented by the evaluator who then asked the children to show the drawing to another evaluator who was sitting across the room. The second evaluator offered effusive praise for the drawing, and the reaction of the child was scored.

However, the way in which pride was manifested was somewhat abnormal in the sense that the participants with ASD seemed to get much less enjoyment out of sharing their achievements with others than did participants from the control group. Group differences also were not present in tests of coyness/embarrassment, and even in the typically developed participants the authors reported difficulties in clearly identifying elements of this emotion. However, the results with respect to scenarios designed to elicit guilt did produce significant results. Participants with ASD were far less likely in these scenarios to manifest signs of guilt toward other individuals. In fact, only one out of the twelve participants in the ASD group showed clear signs of guilt while such manifestations were present in ten of the twelve neurotypical participants. In light of these results, the authors write,

> The evidence suggests that although children with autism may be observed to show self-conscious reactions to certain events … there is tentative indication that their expressions of pride are less strongly elicited in a person-directed context, and more substantial evidence that they show guilt to a significantly lesser degree and with a different quality of emotional relatedness in an interpersonal setting. (Hobson 2006, 106)

So what is the overall lesson to be drawn from the Hobson, et al. studies? The authors suggest a variety of conclusions. First, what seems to be shown is that individuals with ASD rarely show person-focused emotions, but they do, nonetheless, seem to be able to appreciate *something* about the social emotions (after all, they had no problem describing these emotions in others and were often able to relate their own experiences of them). The studies also suggest that those with ASD are not completely unselfconscious in the way that people often think they are. That is, they are clearly responsive to situations in which they are the object of another person's attention, and they can appreciate certain facts about such situations. However, what is striking is the way in

which those with ASD seem to manifest different qualities of emotions such as pride and guilt and to manifest them in a way that removes focus from the interpersonal features of those emotions.

To summarize this section, then, it is clear that individuals with ASD are quite emotionally expressive, and this is especially true the higher up the autism spectrum an individual is. Moreover, individuals with ASD seem to clearly experience all of the same simple emotions (e.g. anger, sadness, happiness, etc.) that typically developed individuals do. However, it is less clear whether those with ASD experience more complex (and particularly social) emotions in a way that is similar to their neurotypical peers. They clearly have access to some facets of these emotions, but their experience of them seems to be atypical in an important way. So, from the available evidence, we have a partial picture of the self-oriented emotional profile of those with ASD. This picture gives insight into the emotional experience and expression of these individuals, but in order to fill out our discussion of the self-oriented emotional profile of the disorder, we need to have some grasp on the way in which individuals with ASD *understand* these emotions.

§1.2. Understanding Emotions in ASD

Some indication of the way in which those with ASD understand emotions has already been given in the course of taking stock of their ability to experience emotion. That is, it seems clear that they can understand a number of simple emotions, in themselves and others, but they seem to have difficulties with emotional understanding as the emotions become more complex. However, much more than this must be said in order to grasp the nature of emotion processing deficits in ASD.

To begin, there is evidence that the emotional understanding of individuals with ASD is more nuanced than has been said so far. For example, in the Rieffe et al. study mentioned above, the authors performed an experiment designed to test whether or not individuals with ASD could describe their own emotions and whether they could distinguish multiple emotional perspectives in scenarios that elicit multiple emotions (say, sadness and anger). The results showed that those with ASD could indeed describe feeling the basic emotions, but they were less able than neurotypical individuals to differentiate emotions with a negative valence from one another. In other words, there was evidence that the simple emotions of anger, sadness, and fear were all experienced as a general "bad" feeling. Moreover, those with ASD were less likely to identify multiple emotional perspectives than typically developed individuals. However, individuals with ASD performed at similar levels to controls in scenarios that were designed to elicit multiple emotions of an opposite valence, say, happiness and sadness. The researchers also found differences in the ability of those in the ASD group to talk about their emotional experience. To this end, they write, "[I]t was found that children with autism … were less able to generate emotionally charged situations from their own experience, provided fewer emotionally charged social situations, … and acknowledged fewer different emotional perspectives in the multiple emotion scenarios." (Rieffe et al. 2007, 462)

In another study testing the way in which individuals with ASD talk about their emotional experience, Losh and Capps (2006) sought to study both the content and form of accounts of emotional experience given by children with ASD as compared to the same in neurotypical children. What they found was striking. In accord with the Hobson

et al. and Rieffe et al. studies outlined so far, they found that children with ASD were less able to describe complex emotions. However, they also found that in offering an account of emotional experience children with ASD were less likely to give an appropriate context for their emotions and offered fewer personal details of events in which the emotions occurred. Additionally, the results of the study showed that the emotional accounts given by children with ASD significantly less often came in the form of narrative. Consider, for example, the following account of a 12 year old boy with ASD,

> *Evaluator:* Tell me about a time you felt guilty.
> *Child:* Well, probably like, like when I do something wrong and then, like, later I might get in trouble for it. Then I kinda feel guilty. (Losh and Capps 2006, 814)

This account demonstrates an understanding of when guilt might be appropriate, but it does not give the impression that the understanding comes from a personal experience, and no reference to any particular episode of guilt is made. Accounts such as this differed considerably from accounts given by neurotypical children which often took the form of a story and contained personal details of the children's experience. This led the authors to conclude that "autistic children's script-like emotional accounts, lacking reference even to the causes of their emotions, leave in question the children's depth of understanding of all types of emotional experience." (816)

In an extremely important study, Hill, Berthoz, and Frith (2004) propose that the lack of understanding emotions that occurs in ASD can be understood as evincing what is known as the alexithymia construct. Alexithymia is a set of characteristics that reflect deficits in the cognitive processing of emotions and emotion regulation, and it is commonly seen as being comprised by the following features: "1) difficulty identifying and describing subjective feelings; 2) difficulty distinguishing between feelings and the

bodily sensations of emotional arousal; 3) constricted imaginal capacities, as evidenced by a paucity of fantasies; and 4) an externally oriented cognitive style." (Taylor 2000) These characteristic features of the alexithymia construct can have far-reaching impacts on the emotional competence of individuals who possess them. For instance, they may

> [R]eflect deficits both in the cognitive-experiential component of emotion response systems and at the level of interpersonal regulation of emotion. Unable to identify accurately their own subjective feelings, not only are individuals with high degrees of alexithymia limited in their ability to reflect on and regulate their emotions, but they also verbally communicate emotional distress to other people very poorly, thereby failing to enlist others for aid or comfort. In turn, the lack of emotion-sharing may contribute to the difficulty in identifying emotions. (Taylor 2000)

Given that several of these characteristics can be readily observed in individuals with ASD, Hill, Berthoz, and Frith administered the standard test for alexithymia, the revised Twenty-Item Toronto Alexithymia Scale (TAS-20),[10] to a large sample of high-functioning adults with ASD, relatives of these individuals, and adults without ASD. The results were overwhelming. Of the participants in the ASD group, only 14.8% fell within the non-impaired category as compared to 82.9% of those in the neurotypical group, and nearly half of the participants with ASD were considered severely impaired. So, the alexithymia construct appears to be a pervasive part of the emotional profile of those with ASD, and, recognizing this fact can help to make sense of the deficits in emotion understanding experienced by these individuals. It remains an open question, however, just how far reaching the consequences of alexithymia may be for those with autism, and I will return to this question after sketching the remaining emotional features of ASD.

[10] The TAS-20 is a self-report questionnaire which consists of three sections designed to test for difficulty identifying feelings, difficulty describing feelings, and externally-oriented thinking. Questions are answered on a 5-point Likert scale, and some of the questions are negatively keyed.

§2. ASD and the Other-Oriented Features of Emotion

Having offered an answer to the first and second questions posed in the introduction to this chapter with respect to individuals with ASD (i.e. Which emotions do they feel and how do they understand them?) I now turn to the remaining two questions. I will begin by surveying the literature on emotion recognition in ASD before moving on to a discussion of the capacity, or lack of capacity, that those with ASD have for empathy. Hopefully, after doing so, we will have an adequate answer to both the question of whether or not individuals with ASD can recognize emotional displays in others and whether (and how) they are able to respond to such displays. Once these have been answered, I will aim to explore the implications of the empirical data on emotion for the counterfactual hypothesis posed in the previous chapter.

§2.1. ASD and Emotion Recognition

When it comes to views on emotion recognition deficits in autism, the data can best be described by one word: equivocal. A large number of studies testing emotion recognition have been performed, and a significant portion of these have shown that individuals with ASD have difficulty recognizing emotions from the facial expressions of others. However, a large percentage of the studies undertaken have also shown no difference in this ability between ASD groups and control groups. In a recent meta-analysis of the available data, Uljarevic and Hamilton (2013) examined 48 studies on visual emotion recognition in ASD, and, after controlling for publication bias, found that the data leans only very slightly in favor of the conclusion that a recognition deficit occurs in autism.

Of those studies that find emotion recognition deficits in ASD, most of them lean on data from facial recognition tasks. However, many studies have implemented tasks designed to test other features of emotion recognition as well, such as recognizing emotions in the voice, body language, or even music.[11] And several experiments have been undertaken which have sought to test methods for improving emotion recognition in individuals with ASD using various computer programs with some success.[12] Despite the concentrated effort on the part of researchers, however, evidence in favor of the prevalence of emotion recognition deficits in ASD remains spotty at best. What is clear, though, is that very many (even, perhaps, a slight majority of) individuals with ASD have difficulty recognizing emotional displays in other people and that a great many of them also experience no such deficit. Several explanations have been offered for this discrepancy in the research. Some suggest that it is the result of deficits in cognition, or motivation, or integration which may be present in individuals with ASD. (Begeer 2008) Alternatively, Kennedy and Adolphs (2012) have found that there is a great deal of test-retest unreliability among individuals with ASD and that this, in conjunction with the general heterogeneity of autism spectrum disorders, might be a contributing factor to the equivocal nature of the emotion recognition data. However, no consensus has been reached regarding the source of this equivocation.

To my mind, the most compelling explanation of the discrepancy in studies of emotion recognition is one that comes by way of acknowledging the prevalence of alexithymia in the autistic population. There is some evidence available that suggests that

[11] See Begeer, et al. 2008 and Nuske, Vivanti, and Dissanayake 2013 for discussions of these.

[12] See, for example, Golan, et al. 2010 or Hopkins, et al. 2011.

highly alexithymic individuals demonstrate deficits in emotion recognition. For example, in a recent study, Montebarocci et al. (2011) found that individuals with high TAS-20 scores performed more poorly on emotion recognition tasks than those with normal scores. They write,

> Results of the present study … indicate that high alexithymic participants showed reduced ability to identify posed facial expression of emotion. As expected, recognition accuracy varied as a function of emotion category … surprise and happiness obtained the highest accuracy scores and anger and fear obtained the lowest scores. (Montebarocci et al. 2011, 249)

That negative emotions were less readily recognized by individuals with high rates of alexithymia ought to pique the reader's interest in light of the discussion of emotional experience and expression above. Recall that those with ASD demonstrated difficulties in distinguishing between their own negative emotions in the studies that I noted previously. This suggests, I think, that alexithymia has a much stronger link to the emotional features of ASD than most have previously thought, and, indeed, there have been findings reported recently that lend credence to this view.

In a recent study, Cook et al. (2013) used a facial recognition task similar to those used in previous experiments, but they matched the test groups for levels of alexithymia (32 total participants, 16 with ASD, 16 without ASD, and 5 alexithymic participants in each group). The results of the study were substantial. The researchers found that deficits in emotion recognition correlated highly with levels of alexithymia and did not correlate with ASD. They write,

> Together, the results of these analyses strongly argue that alexithymia, and not autism, is associated with impaired expression recognition. Autism did not correlate with attribution precision and failed to account for significant variance in the regression analyses. In contrast, alexithymia correlated with expression-attribution precision and remained a highly significant predictor after the

> influence of demographic variables and autism had been accounted for. (Cook et al. 2013, 728)

This conclusion is consistent with previous results from Bird, Press, and Richardson (2011) who found that high levels of alexithymia in ASD populations play a significant role in the atypical facial gaze that is often observed in those with ASD. So, it would seem that emotion recognition deficits may not be specific to ASD but to alexithymia instead. This hypothesis is a natural one, given that alexithymia involves both a difficulty in identifying emotions and a difficulty in distinguishing these from their physiological sensations, it is predictable, to some degree, that it would lead to a difficulty in distinguishing the emotions of others as well. Given the role that we have already seen that alexithymia plays in emotional understanding, its role in emotion recognition deficits in ASD makes one wonder just how much of the emotional profile of autism is actually an emotional profile of alexithymia.

One suggestion as to how we might best understand the relationship between ASD and alexithymia is the "alexithymia hypothesis" proposed by Bird and Cook. "The alexithymia hypothesis," they write, "suggests that, where observed, the 'emotional symptoms of autism' are in fact due to the greater proportion of individuals with severe alexithymia in the autistic population." (Bird and Cook 2013, 2) In other words, their claim is that the emotional symptoms of ASD are actually emotional symptoms of alexithymia. This is, of course, a bold hypothesis, but it strikes me as compelling for reasons that will become clear. However, before considering it further, I will turn to a discussion of the final question posed at the outset of this chapter, namely, how, if at all, do those with ASD respond to the emotions of others?

§2.2. ASD and Empathy

As I noted briefly in chapter 1, precisely what is meant by the term 'empathy' is notoriously difficult to pin down. So, before discussing the role of empathy in ASD, I will take some time to address some of the various understandings of the term.

Recall that Kennett thinks that the essential features of empathy are the "imaginative process of simulation and its resulting emotional seepage and reciprocal awareness." (Kennett 2002, 345) While this is surely an important part of empathy, it is far too loose a definition to proceed with here. So, I will begin by examining what the psychology literature has to say about the nature of empathy and then by bolstering this definition with accounts on offer from the philosophical community.

One thing that is clear from the various ways in which empathy is discussed in the psychology literature is that it requires both cognitive and affective abilities. For example, for psychologists, one way of thinking of empathy is to see it as consisting of three features: "a cognitive ability to discriminate among affective states of others; a second, more advanced cognitive ability to assume the perspective and role of another person; and an emotional response." (Yirmiya et al. 1991, citing Feshbach 1982) Other conceptual models of empathy, however, split the affective component into separate parts as well, producing a model with four separate sub-components. Hirvelä and Helkama describe the various parts (empathic concern, perspective taking, fantasy, and personal distress) of such a conception, saying,

> [Empathic concern] (or sympathy[13]) refers to emotional compassion, while [perspective taking] is the tendency to try to intellectually look at everyday issues from another person's viewpoint. [Fantasy] taps identification with fictional

[13] It should be noted that many theorists refrain from including sympathy as a component of empathy, preferring to treat the two separately.

> characters in, for example, novels and movies. The fourth component, [personal distress], is self-oriented and refers to the proclivity to experience personal discomfort in response to negative emotions of other people. (Hirvelä and Helkama 2011, 560. Citing Davis 1994)

So, on this conception, empathy consists of the cognitive ability to take another's perspective and to "imaginatively transpose oneself into fictional situations." (561) This is, of course, quite similar to Kennett's cognitive criteria of imaginative simulation and reciprocal awareness, cited above. It also consists in an ability to feel concern for and discomfort as a result of another's emotional state. In the remainder of this chapter, I will be focused primarily on emotional empathy rather than cognitive empathy. This is because a discussion of cognitive empathy essentially boils down to a discussion of ToM, and I have already discussed ToM at length in chapter two.[14]

So far, then, with respect to emotional empathy, the authors I have so far cited have identified three strictly emotional components. Kennett thinks that the essential emotional feature is emotional seepage (or contagion, as it is often referred to), and Hirvelä and Helkama identify empathic concern and personal distress as central (the blanket requirement of "an emotional response" that Yirmiya et al. use is, of course, too broad to be of any conceptual help). Do these three features give a comprehensive view of emotional empathy? Not quite, and more needs to be said about each before moving on to their role in ASD.

First, a distinction needs to be made between emotional contagion and what Stueber refers to as "affective empathy." The difference between the two arises from the

[14] This way of putting it is probably too crude as cognitive empathy almost certainly involves more than simply the ability to mind read. However, most of those interested in cognitive empathy (see, for example, Goldman 2006) are interested in it for the sake of its import for what it means for us to have access to other minds, and this is, at bottom, a question of ToM. For a thorough discussion of this type of empathy see Stueber 2014.

fact that emotional contagion requires no awareness of the fact that one's emotion is being caused by the emotions of another. Feeling sad in a room full of mourners could be the result of emotional contagion, but one need not realize the source of one's sadness in order for this to be the case. Affective empathy, on the other hand, requires something more. As Stueber puts it,

> In contrast to mere emotional contagion, [affective] empathy presupposes the ability to differentiate between oneself and the other. It requires that one is minimally aware of the fact that one is having an emotional experience due to the perception of the other's emotion, or more generally due to attending to his situation … In order for my happiness or unhappiness to be genuinely empathic it has to be happiness or unhappiness about what makes the other person happy. (Stueber 2014)

Thus, affective empathy, as a component of emotional empathy, is narrower than simple emotional contagion and requires more from the empathizer as well. Affective empathy, then, is the vicarious sharing of affect that results from the awareness of another's emotional state or situation, and it requires "the involvement of psychological processes that make a person have feelings that are more congruent with another's situation than with his own situation." (Stueber 2014, citing Hoffman 2000)

A further clarification needs to be made between affective empathy and empathic concern as well. Whereas affective empathy involves experiencing emotions that are more congruent with those of the other, empathic concern need not. In order to experience empathic concern, one needs only to feel increased concern for the other's situation regardless of whether this concern is congruent with the other's actual emotional state. This brings us finally to the criterion of personal distress which is similar to empathic concern but for the fact that, while empathic concern is other-oriented, personal

distress is self-oriented. That is, personal distress is simply the experience of an emotion in response *to* another person's situation but not *for* her situation.

So far, then, I have drawn on various sources in order to suggest that our concept of empathy can be chopped up into six distinct segments. Two of these (perspective taking and fantasy) are cognitive, and four (emotional contagion, affective empathy, empathic concern, and personal distress) are emotional. In his recent book, however, David Shoemaker (2015) proposes a different segmentation according to the varying levels of engagement that one might have with another. First, he suggests that we can divide empathy into two primary categories: detached and identifying. Detached empathy, on his account, corresponds to what I have here been referring to as perspective taking, and it is purely cognitive. Identifying empathy, on the other hand, is more complex. It consists, he says, of both a cognitive and an emotional component. Its cognitive component, "evaluational empathy," requires that the empathizer not only take the perspective of the other person but also adopt that person's evaluative stance, which amounts to her general view of the worth of various things, such that the empathizer sees the other person's ends as worth pursuing *from the other's perspective*. The emotional component of identifying empathy is similar in that it requires adoption by the empathizer of a particular stance of the other.

As I understand him, what Shoemaker has in mind is something like what I have so far been referring to as affective empathy. However, while I have taken a rather relaxed view of this component of emotional empathy (i.e. one which requires only congruent emotions and the ability to distinguish between self and other), Shoemaker's conception is much more taxing in that it requires not a loose alignment of the

empathizer's emotional state with that of the other but an adoption, to at least some extent, of the other's emotional cares.[15] This strikes me as too stringent a requirement for emotional empathy, and it seems to me that there is good reason to set the empathy bar rather lower. However, I will reserve arguing more fully for this position for Part II of this project.[16] What I have said so far should, I think, be sufficient to clarify the conceptual terrain of the empathy literature. So, I will now turn to the data regarding the ability of individuals with ASD to engage in these various aspects of empathy.

One thing that should be clear in light of the evidence presented in the previous chapter is that individuals with ASD are deficient in cognitive empathy. Insofar as perspective taking is a function of ToM, it should not be surprising to learn that those with ASD show impairments in their ability to take another's perspective. Moreover, given the evidence regarding pretense already presented, it should also not be surprising to learn that those with ASD have difficulty with the imaginative transposition of themselves into fictional scenarios, a feature of the fantasy component of empathy. Finally, to the extent that Shoemaker's notion of evaluational empathy requires the ability to represent mental states in other minds in order to take up their evaluative stance, we should expect individuals with ASD to show deficits in this component of empathy as well. In short, individuals with ASD have predictable deficiencies in cognitive empathy.[17]

---

[15] Note that this is not an unusual disagreement among different conceptions of empathy. Some theorists hold that it is necessary to hold the exact same emotion as the person one is empathizing with while others simply hold that the empathizer must experience an appropriate emotional response (i.e. pity in response to another's sadness). See Baron-Cohen and Wheelwright 2004.

[16] How precisely to understand affective empathy will play a large role in the later discussion of responsible agency. I bring it up here simply as a way of noting the wide-ranging way in which philosophers and psychologists have conceptualized empathy, and what I have said so far is sufficient in that respect.

[17] For confirmation of this, see Baron-Cohen and Wheelwright 2004 and Rogers et al. 2007.

What, however, of their capacity for emotional empathy? In which, if any, of the features of emotional empathy do those with ASD have the ability to engage?

As it happens, the evidence suggests that many features of emotional empathy are retained in autism. To begin, there is good reason to think that individuals with ASD retain some capacity for emotional contagion. For example, in the Hobson, et al. study cited above parents of children with ASD reported, in several instances, that the children seemed to be affected by the moods of others. Of a child with ASD, one parent reported, "He's very keen at picking up other people's moods I think. If someone's upset, he gets very upset … I mean if you're rushing around trying to get ready, he'd get equally hyper." Of another child participating in the study, a different parent reported, "I know that when I am distressed, because I am distressed, I get impatient, and he reacts in the same way." (Hobson 2006, 61) So, this provides some (albeit anecdotal) evidence to support the notion that individuals with ASD are prone to emotional contagion insofar as emotional contagion involves "catching" another's emotions and is independent of any awareness of the source of the emotion. However, there is further (and more compelling) evidence for this view to be found in the neuroscience literature. In a recent study, Hadjikhani, et al. used fMRI data gathered from participants who were viewing videos of facial expressions of individuals experiencing pain. What they found was that there was no difference in brain activation between individuals with ASD and neurotypical control groups. "In both groups," they write, "we observed activation in the pain-matrix in areas consistently associated with empathy-for-pain tasks." (Hadjikhani et al. 2014, 6) These results led the researchers to conclude that, "[In] ASD, basic automatic processes involved in shared representations of pain are preserved. Our results suggest that rather

than a global deficit in empathy and sharing, individuals with ASD show capacity for emotional empathy." (8) This is far too general a conclusion given the difficulties involved in characterizing emotional empathy that I have discussed so far, but the results of the study do suggest a retained capacity for the more limited empathic response involved in emotional *contagion*.

Nevertheless, there is some countervailing evidence with respect to emotional contagion. For example, Scambler et al. conducted a study in which the emotional responses of children with ASD to a number of affective displays by experimenters were recorded. "The current findings demonstrate," they suggest, "that the children with autism responded to the emotional presses with emotional contagion approximately half as often as the other two groups." (Scambler et al. 2007, 561) Results such as these can be explained, however, by appealing to the alexithymia-based difficulty in emotion recognition discussed above. The responses used to measure emotional contagion in this study may have been such that the participants found identifying them to be too cognitively demanding.[18] Since the study did not control for alexithymia, it may have been the case that emotion recognition deficits on the more cognitively demanding tasks contributed to the decreased performance of the ASD group insofar as "catching" another person's emotions presupposes an ability to recognize what the relevant emotion is. I will return to this issue below.

In addition to emotional contagion, there is some clear evidence that individuals with ASD have intact personal distress experiences in response to the emotional states of

[18] The researchers used non-standard tasks including one in which the experimenter expressed either positive or negative affect upon opening a box (saying either, "oohhh" or "aahhh" upon seeing an unidentified object inside) as well as another which involved expressing either disgust or delight in response to eating "yummy" or "yucky" foods.

others as well. One example of such evidence can be found in a study conducted by James Blair (1999) in which he sought to test the psychophysiological responsiveness of children with ASD to emotional distress cues in others. Children were shown images of other people showing distress cues (like a crying face), and the researchers measured the skin conductivity, a physiological measure of emotional arousal, of the participants upon seeing the images. The results showed significant increases in physiological arousal upon seeing the distress cues as compared with levels of arousal in response to neutral stimuli. Moreover, Blair reports that,

> [T]here was some indication that at least some of the children with autism were finding the pictures of others in distress as aversive. Thus, two of the children with autism tested placed their hands in front of their eyes when a distress cue picture was presented to them and refused to look at it. When asked, they specifically stated that they did not like these pictures. (Blair 1999, 483)

Additional evidence of this can be found in a study from Rogers, et al. in which researchers tested a group of individuals with Asperger's syndrome using a self-report measure of the various features of empathy called the Interpersonal Reactivity Index (IRI) which requires subjects to answer questions about themselves and the way in which they are affected by various situations.[19] They found that the participants with AS actually showed a greater degree of personal distress than those in the control groups. "This indicates," they write, "a greater tendency to have self-oriented feelings of anxiety and discomfort in response to tense interpersonal settings." (Rogers et al. 2007, 713) These results were subsequently duplicated by Dziobek, et al. (2008) in a similar study, and so, it seems clear that the personal distress feature of affective empathy is retained in ASD.

[19] For example, the personal distress element of the questionnaire asks subjects to indicate the extent to which statements such as, "Being in a tense emotional situation scares me," apply to them.

Importantly, the IRI is also a measure of empathic concern and both the Rogers, et al. and the Dziobek, et al. studies reported no significant differences between the ASD and control groups with respect to empathic concern. Moreover, Dziobek, et al. administered an additional empathy measure, the Multi-faceted Empathy Test, and found similar results. The lack of difference in emotional features of empathy led the researchers to remark that, "The disparity between intact emotional empathy found in the current study and prevailing beliefs about a lack of empathy and compassion in individuals with autistic conditions is remarkable." (Dziobek et al. 2008, 471)

This, of course, leads one to wonder why these prevailing beliefs about the empathic responsiveness of individuals with ASD exist. The answer, I think, can be found by attending to the complications of alexithymia for the study of autistic conditions. I have already cited some compelling evidence suggesting that alexithymia is responsible for any emotion recognition deficits seen in the ASD population, but this may not be the only feature of the emotional profile of ASD to which alexithymia contributes. In fact, the Alexithymia Hypothesis of Bird and Cook, mentioned above, suggests that empathy deficits, where observed, are symptomatic of alexithymia, and not ASD, in much the same way. If this is true, then one explanation for the difference between the common perception that individuals with ASD are not empathic and the experimental data which suggests that they are quite so might be explained by the high occurrence of alexithymia in the autistic population.

One reason to suspect that this hypothesis regarding empathy may be true arises from the fact that alexithymia has been seen to cause empathy deficits independently of an ASD diagnosis. For example, it has been repeatedly shown that high rates of

alexithymia are found among those who have been diagnosed with anorexia nervosa (AN),[20] and, more recently, there has been research conducted which found that individuals who suffer from this condition show marked deficits in emotional empathy.[21] Additionally, there has been a sizeable amount of data collected on emotional empathy from another demographic which has long been known to have a high incidence of alexithymia, namely, those who have suffered traumatic brain injury (TBI). In a recent study, Wood and Williams (2007) tested a group of 121 individuals who had undergone TBI and found that 57.9% of their participants scored in the alexithymic range on the TAS-20 compared to 15.4% for the control group (and an estimated 7-10% of the general population). In a subsequent study, Wood and Williams (2008) tested the capacity for emotional empathy in a group of TBI sufferers, and they found that over 60% of participants in the TBI group showed low emotional empathy abilities. In an attempt to link these studies, Williams and Wood (2010) performed a third experiment on a group of participants with TBI. Administering both the TAS-20 and a self-report empathy test, they found that 71.8% percent of participants who scored in the alexithymic range also reported low emotional empathy scores. All of this is to say that there is significant evidence from outside of the autism literature which links empathy deficits to the presence of alexithymia. This being the case, it is not much of a leap to suppose that alexithymia, given its high incidence in the autistic population, may be responsible for empathy deficits in ASD as well.

---

[20] See, for example, Beadle, et al. 2013.

[21] See Morris, et al. 2014. The empathy test used in this study was the Socio-Emotional Questionnaire (SEQ) which has been used to reliably gauge the subject's ability to "feel the same emotion when it is felt by others" (49). Thus, the relevant type of emotional empathy being measured in this study seems to be something like affective empathy as I have defined it above.

Further evidence in favor of the view that the empathy deficits seen in ASD are the result of alexithymia can be found in fMRI data collected in a recent study from Bird, et al. (2010) In this study, the researchers used an empathy-for-pain model similar to that used in the Hadjikhani, et al. study described above. What they found was that irregularities in the empathic brain activation patterns were predicted by TAS-20 scores for participants both with and without ASD. Moreover, ASD seemed to play no role at all in irregular activation patterns in response to witnessing others in pain. These are extremely important results, and, in conjunction with the other alexithymia-related evidence outlined so far, give good reason to believe that deficits in emotional contagion, personal distress, and empathic concern are best explained not as symptoms of autism but instead as symptoms of alexithymia which is co-morbid in a high percentage of individuals with ASD.

This leaves open, of course, the possibility that affective empathy (in the sense specified above) is impaired in a way that is specific to ASD. Recall that affective empathy requires both that the empathizer comes to have an emotional state congruent with that of the other and that she be aware of the other as the source of her emotional state. While alexithymia may account for any deficit in the emotional contagion required to bring about congruence, it cannot account, it would seem for any deficit which may persist in the ability to recognize another as the source of one's emotions. So, is there reason to suspect that individuals with ASD might be deficient in this ability? An answer to this question might be found by attending again to the series of studies from Hobson, et al. discussed above. As I noted in that discussion, Hobson, et al. found that children with ASD were both sensitive to various social emotions and self-conscious when

confronted with situations in which they were the focus of another's attention. This suggests to me that there is indeed some awareness of others as the source of one's emotions in individuals with ASD, at least insofar as they are able to recognize themselves as subject to the attitudes and evaluations of another person. However, it may be recalled, also, that in the Hobson, et al. studies the children with ASD displayed social emotions in a way that was far less other-oriented than did the neurotypical children which suggests that while individuals with ASD may be aware of others as the source of their emotions they may find it more cognitively demanding to attend to these social features of emotions. Indeed, one participant in the Hill, et al. study cited above suggests that this is the case saying, "I get so mad when people say, 'got no feelings, can't relate to me.' I have feelings – told very deep … Trouble is wires crossed so show all this in perhaps odd bizarre fashion or in misplaced way." (Hill, Berthoz, and Frith 2004, 233) Therefore, it seems to me that the capacity for empathy in individuals with ASD (who are not also alexithymic) can be best summarized as follows: individuals with ASD are deficient in cognitive empathy in all its forms (i.e. fantasy, perspective-taking, and evaluative stance-taking), yet their capacity for emotional empathy is mostly intact, consisting of preserved abilities for emotional concern, personal distress, emotional contagion, and (though perhaps limited by cognitive difficulties) affective empathy. This is, obviously, a far different picture of empathy in ASD than that presented by Kennett, but I think it is one which is more justified by the currently available empirical data on the disorder.

Clearly, the emotional profile of ASD is extraordinarily complex, and I have offered a great deal of information about it so far in this chapter. So, before moving on to

discuss how the evidence that I have presented here relates to the conclusions reached in the previous chapter, I would like to pause, briefly, in order to take stock of what has so far been said.

I set out to answer four questions regarding the emotional profile of individuals with ASD: Which emotions do they feel, how do they understand these emotions, can they recognize emotions in others, and how, if at all, do they respond to others' emotions? In order to discuss the first, I turned to an investigation into emotion expression in ASD and presented evidence showing that high-functioning individuals with ASD are, in fact, quite expressive. In addition to expressiveness, I highlighted various studies which consulted both parental reports of emotional experience in ASD as well as self-reports, and descriptions, of such experiences. From these, I concluded that individuals with ASD are prone to experience all of the same simple emotions (anger, fear, sadness, happiness, etc.) as neurotypical individuals but that they have atypical experiences of more complex, social emotions (such as guilt and pride). I then turned to the issue of emotional understanding in ASD in an effort to answer the second of my four questions. I offered evidence suggesting that individuals with ASD have difficulty in understanding complex emotions, differentiating multiple emotional perspectives, and distinguishing negatively charged emotions from one another. I also cited some evidence which shows that those with ASD are less able to provide adequate descriptions of, and context for, their emotional experiences. One way of explaining this, I suggested, was by attending to the fact that an inordinately large percentage of the ASD population also shows signs of co-morbid alexithymia, and I outlined the ways in which this affects emotional understanding.

Having provided answers for the first two, self-oriented, questions about the emotional lives of those with ASD, I turned to the latter two, other-oriented, questions. With respect to the ability of individuals with ASD to recognize emotions in others, I outlined just how equivocal the available empirical data on the matter is. I then introduced the Alexithymia Hypothesis, proposed by Bird and Cook, as a way of explaining the uncertainty in the literature and as a way of showing that deficits in emotion recognition are attributable to alexithymia and not to ASD. Finally, in answering the fourth question I turned to an extended discussion of empathy. I identified several conceptual components of empathy, and I presented evidence which, I think, shows that individuals with ASD have deficits in cognitive empathy (i.e. fantasy, perspective taking, and evaluational stance-taking) but that their capacity for emotional empathy (including personal distress, empathic concern, emotional contagion, and, to a limited extent, affective empathy) remains basically intact. Importantly, I also presented evidence which seems to justify the view that, where observed, empathy deficits in ASD are a result of alexithymia and not autism, thus reinforcing the Alexithymia Hypothesis of Bird and Cook. So, with all of this in mind, I will now turn to a discussion of how this emotional profile of ASD is best understood in light of the Counterfactual Thinking (CT) hypothesis that I defended in chapter 2.

§3. Counterfactual Elements in the Emotional Profile of ASD

To begin, I have already noted one way in which CT might play an explanatory role in what has been discussed here, namely as a way of explaining the deficits in cognitive empathy in ASD. That is, insofar as counterfactual reasoning underlies both ToM and pretense, as I argued in the previous chapter, it can readily offer an explanation

for the deficits in cognitive empathy relating to perspective taking and fantasy. Moreover, insofar as evaluational empathy requires ToM, the CT hypothesis can offer an explanation for that deficit as well. Are there any features of the *affective* profile of ASD, however, for which CT might offer a similar explanation? I think that there are.

First, it seems to me that appealing to deficits in counterfactual representation might help to explain the relatively poor performance of those with ASD on tests of self-reported emotional experience and emotion description. Consider, for example, the study from Rieffe, et al. reported above. The results of this study, it will be recalled, suggested that, "… children with autism … were less able to generate emotionally charged situations from their own experience, provided fewer emotionally charged social situations, … and acknowledged fewer different emotional perspectives in the multiple emotion scenarios." (Rieffe et al. 2007 462) One explanation for why this might have been the case may be found not in the distinctiveness of the emotional capacities of the participants but in the particular demands of the tasks being used. First, in testing the children's ability to generate emotionally charged situations from their own experience, the testers asked questions such as, "Can you tell me about the last time you felt … [emotion]?" (458) However, providing a robust, detail-filled answer to this question requires well-functioning episodic memory, and we have already seen that this is often impaired in individuals with ASD. Additionally, in testing the participants' ability to differentiate multiple emotional perspectives, Rieffe, et al. asked the following prompt,

> Now I'm going to tell you some stories about things that might not have happened to you, but I want you to listen carefully and *try to imagine what it would be like if they really happened to you*. Ok? After each story, I'll ask you how you would feel *if it really happened to you*. (459, emphasis added)

The italicized portion of this prompt, however, is straightforwardly counterfactual, and anyone who experiences deficiencies in representing counterfactual states of affairs would be expected to perform more poorly in responding to it than those who do not. A similar worry arises with respect to the study by Losh and Capps cited above. Recall that in that study participants were asked to describe instances in which they experienced various emotions and that participants with ASD were less able to provide such instances and, when they did, their descriptions lacked narrative content. This task, once again, is one which requires fully functioning episodic memory and so, it may be comparatively more difficult for those with ASD than for typically developed individuals. So, in studies such as these, it may not be the case that participants are demonstrating deficits in understanding emotions. Instead, it may just be the case that the tasks used by the testers are such that they are more cognitively demanding for participants with ASD.

Second, there may be reason to believe that CT might play an explanatory role with respect to what has been said so far about the actual phenomenology of the emotions in ASD. More specifically, it might help to explain why individuals with ASD seem not to experience, or to experience atypically, a number of the complex, social emotions described by Hobson, et al. in the studies cited above. This is because, it would seem, some such emotions either contain a counterfactual element or require cognitive capacities which are underwritten by a capacity for counterfactual representation. Take embarrassment as an example. In order to be embarrassed one must be able to represent someone else as holding a certain opinion of oneself, and this, of course, requires ToM abilities. More specifically, it requires the ability to take the evaluative stance of another and to judge that, via that stance, another person is evaluating one negatively. Shame, a

somewhat moralized version of embarrassment, functions much the same way. In order to feel ashamed, one must realize that one's actions reflect poorly on one's character in the eyes of others. So, it would be expected that an individual who was deficient in this ToM capacity would manifest embarrassment and shame in unordinary ways, if at all, and this was shown to be the case for those with ASD in the Hobson et al. studies.

Some social emotions, however, require counterfactual reasoning abilities in a way that is wholly independent of ToM. To see how this might be, consider the case of guilt. Coming up with an adequate definition of the concept of guilt is notoriously difficult, and I do not wish to enter into the debate over how, precisely, to do so here.[22] Nonetheless, I think that there are some necessary features of guilt that can be (somewhat) uncontroversially specified. As I see it, guilt requires first that one be able to judge that one has either violated a norm, be it moral or conventional, or caused harm to another. Second, one must feel led, presumably by regret, to apologize for one's transgression, and, finally, to make amends for the wrong.[23] Given these features, a deficit in the ability to represent counterfactual states of affairs might explain the atypical nature of guilt reported in the participants with ASD in the Hobson, et al. studies. While individuals with ASD are clearly capable of judging that they have violated norms or caused harm, the other features of guilt may not be so readily available to them. In particular, the notion of regret seems to hinge in a clear sense on the capacity for counterfactual reasoning. As Bernard Williams writes,

---

[22] For a helpful summary of the various conceptions of guilt in the psychology literature, see Tilghman-Osborne, Cole and Felton 2010.

[23] This does not, of course, entail that the guilty party *actually does* make amends, only that she feels driven to do so. There are surely other features of guilt that may be necessary, but what has been said so far will suffice for my purposes here.

> The constitutive thought of regret in general is something like, 'how much better if it had been otherwise,' and the feeling can in principle apply to anything of which one can form some conception of how it might have been otherwise, together with consciousness of how things would then have been better. (Williams 1976, 27)

This conception of regret is not without its detractors, however. For example, Daniel Jacobson argues that Williams is incorrect and that regret involves, fundamentally, "the syndrome of painful feelings of self-reproach … accompanied by the wish to undo the error and the intention to act differently next time." (Jacobson 2013, 104) For either of these understandings, however, the emotion of regret seems to involve some counterfactual element. The ability to conceive of how something might have been otherwise, as required by Williams, necessitates the ability to reason counterfactually, and if this ability is impaired, then regret (and, consequently, guilt) would surely be impaired as well. Likewise, in Jacobson's formulation of both the wish to undo one's error and the intention to act differently in the future involve representing counterfactual states in a particular way. Regret, therefore, seems clearly to rely on counterfactual thinking capacities. Indeed, if we think of individuals with ASD as having the ability to judge themselves to have acted wrongly while lacking the ability to be moved by considerations of how it would have been better if they had not so acted, then this might account for the peculiarities reported by parents in the Hobson, et al. studies. For example, one parent in these studies reported, regarding a child with ASD, "No, you couldn't call it guilty, he just sort of knows that he's done wrong. I could say that he just acknowledges that it was wrong." (Hobson et al. 2006, 68) This report would certainly be in line with what I have said here with respect to guilt. The child in question judges that he has acted wrongly, but he seems to lack the drive to apologize or to make amends.

Moreover, according to the parental report, the child does not seem to feel anything that is phenomenologically similar to guilt. The CT hypothesis offers a ready-made explanation for this limited manifestation of guilt by claiming that this lack stems from an inability to conceive that things would have been better had the child not acted as he did.

While there may certainly be some who disagree with this conception of guilt as requiring regret, it nevertheless seems to be the case that, if the CT hypothesis is correct, individuals with ASD would show deficits in expressions of regret. Very little research has been done on counterfactual-based emotions in ASD, but in one such study (apparently the only of its kind) by Begeer, et al. (2014), it was found that children with ASD performed less well than neurotypical children on tasks designed to elicit counterfactual-based emotions (specifically relief, disappointment, contentment, and regret) for all emotions tested. Interestingly, however, the results showed that, while the participants with ASD performed below controls on all emotions, it was the positively valenced emotions that proved most difficult. Importantly, the authors claim that virtually all complex emotions are

> … emotions that normally depend on a psychological appraisal of another person. Thus, second-order emotions, which include pride, jealousy, and embarrassment, cannot be understood merely in terms of an individual's current first-order mental attitudes (i.e. thoughts, beliefs, and preferences) or situational determinants. Instead, second-order emotions require a contrasting psychological appraisal, attribution, or perspective. (Begeer et al. 2014, 302)

If this is true, then it would mean that all complex emotions are based, in one way or another, on the ability to reason counterfactually in the way in which I have described, and this fact lends a great deal more credibility to the CT hypothesis given the apparent deficits in complex emotional experience in ASD.

So, I think there are good reasons to suspect that the relative inability to represent counterfactual states of affairs in ASD may play a causal role in a number of the emotional irregularities that I have described so far in this chapter. However, I have also placed a great deal of weight on, and confidence in, the Alexithymia Hypothesis and its explanatory role in the emotional features of ASD. Because of this, some extended discussion is needed regarding the relationship between my CT hypothesis and the Alexithymia Hypothesis. I turn to that discussion now.

Bird and Cook are careful to emphasize the distinctness of alexithymia and ASD, saying,

> Despite their frequent co-occurrence, alexithymia and autism are independent constructs. Alexithymia is neither necessary nor sufficient for an autism diagnosis, nor is it universal among autistic individuals. Conversely, many individuals show severe degrees of alexithymia without demonstrating autistic symptoms. (Bird and Cook 2013, 1)

Moreover, in a recent review of the literature on emotions in ASD, Nuske, Vivanti, Dissanayake (2013) determined that the emotional impairments often described are neither universal in, nor specific to ASD. This finding would, of course, pair well with the view that alexithymia underlies the emotional impairments that are often observed. However, what has yet to be explained is the simple fact that the rate of alexithymia in ASD is extraordinarily high, and explaining this rate of occurrence would be extremely valuable in coming to a better understanding of autism.

Recall from *§1.2* of this chapter that the alexithymia construct is characterized by the following four elements: “1) difficulty identifying and describing subjective feelings; 2) difficulty distinguishing between feelings and the bodily sensations of emotional arousal; 3) constricted imaginal capacities, as evidenced by a paucity of fantasies; and 4)

an externally oriented cognitive style." (Taylor 2000) With these in mind, what, if any, connections can be drawn between these elements and the capacity for representing counterfactual states of affairs? First, there is a clear connection between counterfactual ability and imaginal capacities. That is, if one is lacking in one's ability to represent states of affairs that part from reality, then one will clearly be lacking in both imagination and fantasy. Second, there may be reason to think that those with ASD have a largely externally oriented cognitive style as well. Recall from the previous chapter some of the particular behavioral features of ASD such as, "stereotyped or repetitive motor movements, use of objects or speech," or "Hyper- or hyporeactivity to sensory input or unusual interest in sensory aspects of the environment." (DSM-5, 50) Such behaviors are clearly externally oriented, and it is not unreasonable to think that they flow from a cognitive style which is similarly oriented. Indeed, many clinicians describe their clients with ASD as displaying a very concrete, non-subjective way of thinking, a trait which would be expected from an individual who was not prone to posit counterfactual states of affairs.[24] This is because deficient counterfactual thinking is likely to lead to a decreased propensity to think in terms of subjective states which may or may not be reflective of the actual state of the world. So, it would seem that at least two of the features of alexithymia might also be construed as features of a counterfactual reasoning deficit. It is less clear, however, whether any such relationship might be found with respect to the remaining two features.

One way of conceiving of the remaining two features of alexithymia (i.e. difficulty identifying and describing emotions and difficulty distinguishing emotions

[24] I am grateful to clinicians, Penny Roberts and Rebekah Moilanen, for alerting me to this fact.

from their bodily sensations), first postulated by Lane and Schwartz (1987), is to see them as arising from an inability to make mental representations of emotions. They posit a developmental theory of emotional awareness which consists of the following five levels: "…physical sensations, action tendencies, single emotions, blends of emotion, and blends of blends of emotional experience (the capacity to appreciate complexity in the experience of self and others)." (Lane et al. 1997, 837) Representations of emotions are made by differentiating and integrating emotional experience, and it is in this respect that they suppose those with alexithymia to be deficient. To this end, they write, "[T]he alexithymic individual avoids reflecting on and generating symbolic representations of experience." (Lane and Schwartz 1987) If this is truly the source of these features of alexithymia, then it is not difficult to see how impaired counterfactual reasoning might play a role here. If one is impaired in representing counterfactual states, then this may well involve impaired symbolization abilities such as those required to formulate mental representations of emotional experience.

Admittedly, the connection here is tenuous, and I am not trying to step out on a limb and proclaim that counterfactual deficits *cause* alexithymia. What I am suggesting is that there is some overlapping conceptual territory between counterfactual disabilities and alexithymic traits, and insofar as individuals with ASD present deficits in counterfactual thinking this might help to explain the high occurrence of alexithymia in ASD. To my knowledge, however, no one has tested any such connection, but to do so would, I think, be beneficial in understanding the role of counterfactual reasoning in the emotions. One way of testing the relationship might be by attempting to identify any correlations between the counterfactual cognitive abilities (e.g. planning, episodic memory, ToM,

etc.) and alexithymia. If it turned out to be the case that groups who were deficient in these abilities performed at typical levels on tests of alexithymia, this might be suggestive of some other cause. If there was, however, a causal relationship between counterfactual reasoning and alexithymia, we might have some explanation for the increased prevalence of alexithymia in the ASD population. That is, if we conceive of the autism spectrum as being characterized by degrees of counterfactual thinking impairment, then we might be able to identify a threshold of impairment above which individuals on the spectrum are able to overcome alexithymic traits, in effect showing an inverse relationship between TAS-20 scores and counterfactual ability.

One final note on the relationship between the emotional features of ASD and the CT hypothesis must be addressed before closing this chapter, and it is with respect to empathy. As I have said, it seems to me that the Alexithymia Hypothesis offers the best explanation for empathy deficits in ASD. However, even apart from alexithymia, there is a straightforward way in which the CT hypothesis might still be able to explain the apparent empathy deficit. Crudely put, to engage in affective empathy is simply to ask oneself, "How would I feel if I were in her shoes?" and to subsequently feel whatever emotion is appropriate. Given this way of putting the matter, CT deficits may be able to explain the limited ability for affective empathy experienced by individuals with ASD. It may simply be the case that they find it difficult to answer the relevant "How would I feel if…" question since this question requires counterfactual representation of oneself as experiencing another's circumstances. In all, then, there seems to be ample reason to suspect that many of the emotional features of ASD are underwritten by the same

counterfactual impairments that I proposed were behind the cognitive features of the disorder in chapter 2.

§4. Conclusion

The nature of autism's effects on an individual's affective capacities is, as can be seen from the preceding, quite complex, and, therefore, does not lend well to a short, distilled summary statement. Nevertheless, I will close by saying, in general, that individuals with ASD are much more emotionally capable than they are widely believed to be. They are emotionally expressive, they experience and understand a wide range of emotions, they recognize emotions in others, and they are capable, to a large degree, of emotional empathy. The primary deficit that those with ASD experience, then, pertains to their capacity for experiencing some of the more complex emotions, and this limits them somewhat in their ability to engage in affective empathy. Moreover, impairments in counterfactual reasoning, I have argued, give rise to impairments in cognitive empathy which likely contributes to the view that individuals with ASD are not empathic. This picture is complicated by the fact that a very large proportion of the ASD population demonstrates co-morbid alexithymia, and I have offered evidence in favor of the view that where deficits in emotional understanding, emotion recognition, and emotional empathy are observed they are symptoms of alexithymia rather than ASD. Finally, I have suggested that the emotional features which are not alexithymia-based arise as a result of impaired counterfactual thinking, and I have also speculated that there may be a significant overlap between this impairment and the alexithymia construct.

So, this chapter concludes my discussion of both the cognitive and affective psychology of autism. I have tried to offer a way of understanding the relevant work on

ASD such that a unifying thread can be traced through it. I think that unifying thread is a deficit in counterfactual thinking, and I have speculated that further research into counterfactual thinking in ASD would do a great deal to improve our understanding of the disorder. I have tried to give a comprehensive picture of an enormous psychology literature in a relatively small amount of space, and this surely means that some things have been glossed over rather hastily. Despite that fact, what I have said so far provides enough of an understanding of the nature of ASD to allow me to move on to a discussion of the distinctly *moral* psychology that characterizes it. In the following chapter, therefore, I will offer an account of moral judgment and moral motivation in ASD which will then allow for a discussion of the implications of autism for a number of views of morally responsible agency.

PART II:

Autism Spectrum Disorder and Theories of Moral Responsibility

Prologue: Methodological Notes

Before moving on to examine particular philosophical accounts of responsibility, it is necessary to pause and make certain methodological features of the ensuing discussion explicit. Up to this point, I have been claiming that my goal in Part II will be to "examine the implications" of ASD for theories of moral responsibility, or to use the psychology of ASD to "test theories of responsibility." Phrases such as these, however, are somewhat unclear. So, what, precisely, might it mean to examine the implications of ASD for theories of moral responsibility? What I have in mind is something like the following: each type of theory of responsibility will claim that certain features of agents are necessary or sufficient for moral responsibility, and, therefore, each will make a determination of some kind regarding the status of individuals with ASD. My goal, then, will be to examine the extent to which each type of theory "gets things right" in the case of autism.

Proceeding in this way, of course, presents a rather obvious worry. In order to assess the extent to which the theories get things right we need to have a somewhat settled notion of what the right answer is, and it is not at all clear that we have such an answer in the case of individuals with ASD. So, there is a deep methodological objection lurking behind the arguments in the following chapters insofar as the theorists to be discussed might simply shrug and claim that the arguments that I advance against their views give us as much reason to think that individuals with ASD are *not* responsible

agents as they do to think that the theories discussed are mistaken. As it is often quipped, one philosopher's *modus ponens* is another's *modus tollens*. So, perhaps all that can be shown from the evidence presented in Part I are some interesting reasons to think that those with ASD are not among the class of responsible agents. If this is correct, then to ask whether or not a theory gets things right with respect to ASD is to simply put the cart before the horse. Rather, the proponent of a given theory might claim that if the view gets things correct in the case of more ordinary agents, then it will get things right about ASD, whatever determination it may make. The theory, they may say, is just weightier than our intuitive judgments in the case of autism.

This is a difficult charge to dismiss, and, I think, it points to a deep methodological divide in the literature on moral responsibility. As such, I won't be able to offer a decisive argument here, but it seems to me that there is good reason to think that the burden of proof is on those who suggest that the order of explanation is from theory to cases rather than the other direction. This is because, despite the various incapacities and deficits described in Part I, high-functioning individuals with ASD look and act very much like neurotypical agents across a wide variety of circumstances, and where differences arise, they do not, intuitively, seem to be such that they warrant seeing these individuals as anything other than responsible agents. Moreover, quite a lot is at stake in the question of morally responsible agency. To be denied one's status as a morally responsible agent is to be denied one's membership in a moral community, and denying an individual (or group of individuals, as the case may be) this status requires a very high burden of proof. I do not think that this burden has been met.

To see why, we can look to the methodology that the theorists to be discussed in the coming chapters use. Each of them, as we will see, identifies some capacity of an agent or some feature of the agent's psychology as either necessary or sufficient (or both) for that agent's being morally responsible. The features in question might be the relationship between first- and second-order desires, or the quality of the agent's will, or the agent's ability to recognize and act on reasons of a certain kind, or something else entirely, as we will see. Then, having identified some crucial capacity or psychological feature, each of the theorists uses as evidence in favor of his or her candidate for the crucial element the fact that it can give plausible explanations for a number of intuitive cases. In doing so, then, each view identifies the central conditions for responsible agency.

Now, the methodological objection described above says that, given their failure to satisfy these conditions in the right sort of way, we ought to hold on to our theory and simply conclude that individuals with ASD are not responsible rather than rejecting the theory. However, it seems to me that there is more argumentative work to be done before this response is licensed. Specifically, a reason needs to be given for why we should think that imaginary cases like those that will be discussed (of evil neuroscientists, willing addicts and estranged desires) or other actual cases (of children or the insane, say) should be more authoritative in informing our conception of moral responsibility than are real world cases of autistic persons. Given the prevalence of autistic agents in the world and the strong intuition that high-functioning autistic people are capable of responsible agency across a great variety of circumstances, it seems to me that the inability of a theory to give us a plausible account of them should count at least as heavily *against* a

view as its ability to handle the sorts of cases discussed here count in its favor. Moreover, the argumentative burden is on those who deny this to tell us why. So, the theorists that will be discussed in the following chapters may well be entitled to advance the methodological objection, but before acting on that entitlement, they need to meet a burden of proof that has yet to be met.

Now, there may be a response in the offing for those who make the methodological objection, as they may simply claim that the cases they use, rather than being more authoritative, simply give a much clearer and stronger intuition than the case of ASD which is much more difficult and complex.[1] However, I think this claim should be resisted, as the intuitions generated by very high-functioning autistic persons seem to me to be every bit as strong as those generated by the cases which ground the several theories to be discussed in Part II. One need spend only a small amount of time interacting with autistic people, I think, in order to share this intuition, and the intuition becomes even stronger when one considers that many high-functioning autistic people are not diagnosed until later in life due to the fact that they look and act so much like everyone else. In a recent short film,[2] a group of autistic women who received late diagnoses (most after the age of 21 and some well into their 30's) were interviewed, and their responses were striking.

On why they think they went undiagnosed:

> I was quite functional. I didn't cause many problems. I kept to myself quite a lot. I got through school fine. I got good grades. I had a couple of friends. It was never really a problem.
>
> I fit in really well. I learned how to analyze situations.

[1] I'm grateful to both David Shoemaker and Michael McKenna for raising this worry.
[2] See Belcher 2015

> Female autism is just so well hidden.
>
> I did a really good job of fitting in. I spent a lot of my time really intently watching what the other girls were doing and making sure that I acted like them and dressed like them and did everything that they did even if it didn't make sense to me.
>
> I learned very quickly what was socially acceptable, and I learned how to be smart about my stimming[3] … Really, through watching other people I learned how to *be* in the world. I learned how to make eye contact, and I learned how to speak. And really, I learned to conform, which is kind of sad. I was always afraid of being caught out, but I guess I covered that up by being the loudest and the funniest and the most outspoken – at university, at school – and I think people got the idea that I was really confident, but that in itself was my way of hiding this incredible fear and terror that I would be found out for who I was and what I really felt. An untrained eye can see a girl with autism and think that she blends in fine, but I can pick it straightaway … My personal experience has been that many women and girls like me, with autism, are deeply compassionate caring and kind individuals and we can appear to be completely normal until something upsets us, and then you can tell because I think we feel so much deeper than most people and think about things on levels that other people wouldn't even think about going to.
>
> I memorized these rules and adapted them to my life in order to get by. So, uh, I would learn them through friends, watching other friends talk to one another. I learned it through TV. I learned it through movies I learned it through the books that I read. I was constantly trying to pick up on new ways to help myself interact better socially with my peers. A couple of my neighborhood friends who I would pick out certain personality traits about them that were seen in a positive light or that I saw have positive consequences, and I would just literally adopt that exact trait. I remember my one friend it was how she was able to kind of be funny. Um, learning humor was hard for me, but I also saw how she did it and I would copy her exact sentences and her exact gestures, and I would copy everything that she did. Then I would try it out, and if it went well, then it's in my rule book for life.

On what it felt like to receive their diagnosis:

> What it felt like to be diagnosed was a combination of relief and hopelessness, and a bit of, a lot of grief, actually.
>
> I first felt really relieved. Um, it gave me a lot of answers to things that I had questions about for a while … I moved past feeling the relief and on to kind of the

[3] "Stimming" is a bit of vernacular used to refer to the ways in which autistic people try to control their sensory stimulation. In more severe cases, this could include rocking back and forth or flapping one's hands or even hitting oneself. It could also include more discrete actions like touching items made of materials whose tactile sensations are soothing.

> anger, and depression, and confusion, and guilt, and there were so many feelings going on, to now more of embracing it and seeing that this is who I've been my whole life and it doesn't make me any different. It's just now I see things differently.
>
> I forgave myself, and I let go of a lot of things from the past that I couldn't have reconciled without my diagnosis.

These responses do not read like the responses of profoundly marginal agents. They read like the responses of individuals who care deeply about others and about themselves and who are trying to find a way to navigate the world around them but who manage to do so successfully in ways that are, in many cases, fundamentally different than neurotypical agents. In short, watching and listening to autistic persons gives the strong intuition that they are indeed responsible agents.

Nevertheless, they do clearly have barriers to their success in navigating the social world and in understanding other agents. The evidence presented in Part I speaks to just how pervasive these difficulties can be, and first personal accounts support that evidence well. For example, one woman from the film, when asked how it feels to be a woman with autism responded with the following:

> What does it feel like to be a woman with autism? Hmm, what does it feel like to be a woman *without* autism? Sometimes I feel like I come from a different culture, and when it's really bad, I feel like I'm a different species. Social situations can sometimes feel to me like a math equation, and I'm looking at it on a blackboard, and it's full of equations, and I don't even know how to count. I don't even have that basic knowledge, and I look at it and everyone else seems to have this knowledge of what to do like maybe they have a calculator, and I'm there trying to slowly work through the most basic things.

There is an important distinction to be made between morally responsible agency and moral responsibility for actions. There is one sense in which these two are not distinct at all, namely in the sense that only morally responsible agents can be morally responsible for their actions. However, the two may come apart insofar as it may be

possible for morally responsible agents to perform actions for which they are not morally responsible. For example, an agent who lies to a colleague is not typically considered morally responsible for doing so if, say, she did so under post-hypnotic suggestion. Similarly, an agent who is otherwise a morally responsible agent may not be morally responsible for some action performed under extreme duress, and this suggests that the conditions under which one is morally responsible for some action are not the same as the conditions under which one is a morally responsible agent. This distinction is central to the problem at hand, because the above responses generate, in my opinion, two very strong intuitions regarding individuals with ASD. First, they generate a clear intuition that high-functioning autistic persons are morally responsible agents, and, second, they generate a clear intuition that there are a range of actions for which autistic persons are not morally responsible. Together, these intuitions will be problematic for the theories of responsibility to be discussed in the coming chapters.

In order to keep these intuitions clear, it will be helpful to have a paradigm case of a high-functioning autistic agent to refer back to throughout the course of the theoretical discussions ahead. Having such a case in hand will help to make the arguments in the following chapters clearer and stronger, and it will help to give a concrete picture of the sort of agent that is problematic for the dominant approaches to moral responsibility. Given the high degree of variability among autistic persons along the autism spectrum, it will be essential to maintain a focus on those on the high-functioning end, and a paradigm case will be helpful in that regard.

> **The Paradigm Autistic Agent:** Adam is a high-functioning autistic man. He has an IQ in the average to above-average range and lives on his own. He has pervasive impairments in counterfactual representational abilities, and as a result he has a difficult time interpreting and navigating social interactions. It is difficult

> for him to make sense of the desires, beliefs, and intentions of other people and, sometimes, of his own. Additionally, when Adam is in social settings he sometimes finds sensory and emotional stimuli overwhelming and has often seemed withdrawn or has acted in ways that people find odd. However, he has lots of experience in social situations and over time he has developed compensatory strategies that help him to overcome these difficulties. He can generally engage in social behavior that is pleasant and appropriate, and in many cases others do not even recognize that he is autistic. Additionally, Adam can reliably judge right from wrong, and he has a robust emotional life. He feels all of the same emotions as his peers and responds empathically to the emotions of others when he is able to interpret their emotional state correctly, but his more complex emotions can be difficult for him to understand (though he is able to do so with great effort) and are sometimes manifested atypically. Finally, Adam's counterfactual deficits also make certain executive abilities difficult. He sometimes struggles to make long term plans, and he can become fixated on things that interest him. However, his compensatory strategies have made it so that these qualities do not lead him to engage in compulsive behaviors.

The case of Adam, it seems to me, is a plausible description of a real-life autistic person, and he produces the clear intuitions expressed above. Absent some preexisting theoretical commitment, it is difficult to see why anyone should not count him among the class of responsible agents. However, given the above description, it seems clearly to be the case that there is some subset of actions or attitudes for which we would not hold him responsible. These intuitions are, I think, strong enough to support the burden of proof argument I have advanced here, and this case is rich enough to provide a point of reference for the autistic agents to be discussed in the following chapters. Thus, we are now in a position to see what implications the empirical data presented in Part I have for theoretical approaches to responsibility.

Chapter 4: ASD and Reasons-Responsiveness

In Part I, I argued for a unique understanding of the cognitive and affective psychology of Autism Spectrum Disorder. I attempted to show, in chapter 2, that the three dominant theories of autism can be unified to a large extent by appealing to a cognitive deficit in counterfactual thinking that individuals with ASD possess. In chapter 3, I argued that individuals with ASD are actually far more emotionally capable than they are often portrayed as being and that most deficits in emotional competence can be explained by the fact that individuals with ASD often also present co-morbid alexithymia. However, I argued further that deficits in counterfactual thinking may account for some atypical emotional features in ASD, particularly atypical social-emotional profiles. Thus, counterfactual thinking deficits help to unify both cognitive and affective evidence related to ASD.

Having presented and defended this picture of the psychology of ASD, I will now turn to the second part of my project which is to examine its implications for theories of moral responsibility. Such theories, as I suggested in chapter 1, can be split roughly into three categories: Reasons-Responsive theories, Real Self theories (also called Deep Self or Mesh theories), and Quality of Will theories, and each of these three approaches, recall, grounds moral responsibility on different features of agents. I will discuss each of

these at length, focusing, in this chapter, on the reasons-responsive approach, in Chapter 5 on the real self approach, and in chapter 6 on the quality of will approach.

The present chapter will proceed as follows. I will begin, in §1, by giving a detailed treatment of John Martin Fischer and Mark Ravizza's reasons-responsive theory of moral responsibility. Their seminal work, *Responsibility and Control*, has been hugely influential and is by far the most detailed and comprehensive reasons-responsiveness theory on offer.[1] As such, it will be the primary focus of the arguments in this chapter. More specifically, I will argue in §2 that Fischer and Ravizza's view is open to two basic challenges from the empirical data on ASD. First, I will argue that their account of moderate reasons-responsiveness is not able to give a satisfactory characterization of the receptivity and reactivity to reasons that individuals with ASD possess. Second, I argue that individuals with ASD, in some cases, do not meet the subjective requirement for responsibility that Fischer and Ravizza describe yet are morally responsible nonetheless. Each of these arguments aims to undermine an important feature of Fischer and Ravizza's view, and, if I am correct, then serious doubt will be cast on the view as a whole.

§1. Fischer and Ravizza's Reasons-Responsive View

The goal of the reasons-responsive theorist, most generally, is to give an account of what it means to be properly responsive to reasons and how such responsiveness grounds moral responsibility. Typically, this is done by giving an account of the sort of control that it is necessary for an agent to possess in order to be the apt target of praise

[1] There are, of course, other important reasons-responsive views on offer in the literature. For example, Susan Wolf presents such a view in her book, *Freedom within Reason* (1990), as does Ishtiyaque Haji in his *Moral Appraisability* (1998). However, Fischer and Ravizza's view is seen as the gold standard in the literature, and that is why it is the primary focus here.

and blame for her actions. So, for example, imagine an agent who, upon seeing that the sky is becoming dark and hearing tornado sirens in the distance, decides to stay inside her home. This agent perceives certain reasons for staying indoors (e.g. it will be dangerous to go outside) and acts on those reasons. Contrast this agent with one who is extremely agoraphobic and who stays indoors at all times due to his crippling fear of leaving his house. This latter agent, upon hearing the sirens and seeing the foreboding skies, also stays indoors, but he is not sensitive to reasons in the same way that our first agent is. This is because the agoraphobic agent would have remained in his home in cases where he lacked reason to do so as well as in cases where he had reason *not* to do so. The fact that his actions would have remained the same regardless of the reasons that there were for acting or not acting as he did seems to suggest that the agent is not properly responsive to reasons (at least insofar as they bear on staying indoors), and this seems to stem from his inability to control his actions in the relevant way.

The view that Fischer and Ravizza proffer aims to give an account of the control required for moral responsibility and can be clumsily described as an actual sequence, mechanism-based, externalist, moderate reasons-responsive theory of moral responsibility. Each of these qualifiers represents an important distinction in the view (which they term "semicompatibilism" in order to avoid the clumsiness of the phrase above), and, perhaps unsurprisingly, each of these distinctions has spawned internal debates among various proponents of reasons-responsive views. This section will be devoted to elucidating these distinctions in order to set the stage for a discussion of how ASD presents potential problems for reasons-responsive views.

Fischer and Ravizza begin by making explicit the type of control of which their theory is intended to give an account. They do this by appealing to an intuitive distinction between regulative control and guidance control. If we imagine an agent who is driving a car, we might imagine the agent as exercising two distinct types of control over the automobile. One type would consist in the agent's forming intentions to turn the car in one direction or the other as well as the decision to direct the car according to those intentions, this Fischer and Ravizza call "regulatory control." The other type of control simply consists in the ability to guide the car in such a way that would carry out the decisions made by the agent. Thus, Fischer and Ravizza write, "Guidance control of an action involves an agent's freely performing that action … Regulative control involves a *dual* power: for example, the power freely to do some act A, and the power freely to do something else instead." (Fischer and Ravizza 1998, 31) On this definition, then, regulative control requires the ability to do otherwise.

However, a particular type of counterexample, originally posed by Harry Frankfurt (1969), represents a compelling challenge to the notion that moral responsibility requires this ability. In the original Frankfurt case, we are asked to imagine that a man, Jones, wants to shoot another man, Smith, and that a third man, Black, wants Jones to act on this desire. Black, not wishing to take any risk that Jones will fail to act on his desire to shoot Smith, takes steps to ensure that Jones will do what Black wants him to do by implanting a device in Jones' brain that will cause Jones to decide to shoot Smith if there should be any sign that Jones is wavering in his decision to do so. In such a case, Jones lacks the ability to do otherwise than shoot Smith because if at any time he shows signs of not doing so, then Black's device will be activated and will cause Jones to

decide to shoot Smith. Now, imagine that Jones, for reasons that are entirely his own, carries out his plan to shoot Smith and that Black's device is never activated. Surely Jones is morally responsible for shooting Smith in this case since the presence of Black's device was wholly irrelevant to his doing so. However, if this is true, then it appears that Jones is responsible despite lacking the ability to do otherwise, and, so, Frankfurt argues, the ability to do otherwise is not a necessary condition for moral responsibility. Fischer and Ravizza, agreeing with Frankfurt, argue that cases such as this provide strong evidence for the claim that the only type of control necessary to ground moral responsibility is guidance control, and, so, their view aims to give an account of just what this sort of control comes to.

That guidance control is all that is necessary for moral responsibility is not the only thing that Fischer and Ravizza take Frankfurt cases to show. Consider once again the example of the agoraphobic agent discussed above. The feature that this case shares with the Frankfurt cases is the fact that in neither case is the agent able to do anything other than what she in fact does. So, to that extent, neither agent is responsive to reasons. However, in the case of the agoraphobic, what drives our judgment that he is not reasons-responsive is the fact that the agent's rational deliberative capacity seems to be impaired, or malfunctioning, in the sense that things that should count as reasons for or against acting simply do not in her case. In the Frankfurt case, however, what makes it that the agent is not properly responsive to reasons is the presence of the counterfactual intervener, Black. The reason that we hold the agent in the Frankfurt case to be morally responsible, Fischer and Ravizza claim, is that the intervener affects the counterfactual sequence in which an agent might act otherwise, but *in the actual sequence* in which the

agent acts, her act flows directly from her own deliberation. Thus, the Frankfurt cases tell in favor of an "actual sequence" theory of moral responsibility which holds that the only facts relevant to the responsibility of an agent for her actions are those that occur in the actual sequence in which her action occurs.

In addition to providing support for an actual sequence view, Fischer and Ravizza claim that Frankfurt cases give evidence for a further distinction as well. As I have just noted, the fact that the agent in the Frankfurt cases cannot act other than she in fact does suggests that the *agent* in those cases is not properly reasons-responsive. In order to make sense of the agent's moral responsibility in these cases, Fischer and Ravizza suggest that we must switch from an agent-based view to a mechanism-based view. That is, while it may be true that the agent is not reasons-responsive in Frankfurt cases, the *mechanism* (where this just refers to the process by which the action comes about) of the agent's action is. So, for Fischer and Ravizza, "an agent exhibits guidance control of an action insofar as the mechanism that actually issues in the action is his own, reasons-responsive mechanism." (1998, 39) What is left, then, is to elucidate precisely what sort of reasons-responsiveness Fischer and Ravizza have in mind.

In their discussion of the sort of reasons-responsiveness that is necessary and sufficient for the control required for moral responsibility, Fischer and Ravizza begin by differentiating between a mechanism's being either strongly reasons-responsive or weakly reasons-responsive. A mechanism which issues in action in the actual sequence is strongly reasons-responsive, they say, if, in any alternate sequence, it "were to operate and there were sufficient reason to do otherwise, the agent would *recognize* the sufficient reason to do otherwise and thus *choose* to do otherwise and *do* otherwise." (1998, 41)

This entails that a mechanism can fail to be strongly reasons-responsive if the agent fails to recognize the reasons that there are for acting, fails to make a decision which aligns with those reasons, or fails to act on that decision. Fischer and Ravizza refer to the first sort of failure as a failure to be "receptive" to reasons and to the latter two sorts as failure to be "reactive" to reasons. So, strong reasons-responsiveness requires that a mechanism allow an agent to recognize reasons and to translate them into action whenever sufficient reasons exist. This sort of view is defended by Robert Nozick who proposes that in order to be responsible or to act freely an agent's actions must "track value" or "rightness." Nozick explains this notion, saying,

> Person S's (doing of) act A tracks rightness when
> (1) Act A is right
> (2) S intentionally does A …
> (3) If A weren't permissible, S wouldn't intentionally do A
> (4) If A were mandatory, S would intentionally do A. (Nozick 1981, 319)

So, conditions (3) and (4) on Nozick's account suggest that being appropriately responsive to reasons is to be defined, in part, by reference to how an agent would act in a particular counterfactual scenario. However, Fischer and Ravizza argue that strong reasons-responsiveness is too strong a requirement for moral responsibility, for we can think of cases in which an agent is morally responsible but is not strongly reasons-responsive.

Consider, for example, a standard case of weakness of the will. Suppose that Bill loves shoplifting. He gets a thrill every time he takes something from a shop without paying for it, but his desires are not overwhelming or irresistible. Suppose, further, that Bill understands that what he is doing is wrong and harmful to the shop owners from whom he is stealing. He understands that he has sufficient reason to refrain from stealing,

but upon walking into a store one afternoon he gives in to his desire to steal and does so despite knowing the he ought not. Clearly, we think that Bill is morally responsible for stealing in this case. If strong reasons-responsiveness were necessary for moral responsibility, however, this judgment would be problematic. This is because the mechanism that issues in Bill's action turns out not to be strongly reasons-responsive. Bill is able to recognize the reasons he has not to shoplift, but his weak willed behavior is evidence of a failure to translate those reasons into action. So, if strong reasons-responsiveness is necessary for moral responsibility, then Bill is not morally responsible, but this response is unacceptable in light of our strong intuitions that he is.

One way of resolving this problem is to construe reasons-responsiveness as requiring something much weaker. Fischer and Ravizza consider this solution and define weak reasons-responsiveness as follows: an agent is weakly reasons-responsive if, holding fixed the mechanism that operates in the actual sequence, there exists "some possible scenario (or possible world) in which there is a sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise." (Fischer and Ravizza 1998, 44) So, consider, once again, Bill the shoplifter. Bill's weakness of will showed that his mechanism of action was not strongly reasons-responsive, but is it weakly reasons-responsive? To answer this, we simply need to know whether there is any scenario in which Bill would have reason to refrain from shoplifting, recognize this reason, and then refrain. Suppose that it is true that if a security guard were standing next to him, Bill would recognize this as a reason not to steal and refrain from stealing. If this is the case, then Bill satisfies the conditions for weak reasons-responsiveness and is thus morally responsible for his action. While this construal of reasons-responsiveness solves

the problem associated with strong reasons-responsiveness, it is prone to the opposite sort of worry, namely, that it is too weak a notion to fully ground moral responsibility. As a way of getting at this point, Fischer and Ravizza introduce what they call the "problem of strange patterns."

Consider, once again, our agoraphobic agent. Suppose that she would stay in her home if there were a threatening storm outside, if it was a beautiful, sunny day, if she were out of groceries, if her daughter was graduating from high school that day, and in a great many other scenarios in which she has good reasons or no reasons for staying indoors, or in which she has good reason not to do so. Clearly, such an agent seems not to be reasons-responsive. However, suppose that, holding fixed the mechanism that operates in these scenarios, the agoraphobic agent *would* leave her home if it were exactly 72 degrees outside and George Clooney was eating Greek yogurt on a day that the Detroit Red Wings win the Stanley Cup. If, under these circumstances, the agoraphobic agent would see herself as having a reason to leave her home, choose to do so, and then act on this choice, then she meets the conditions for weak reasons-responsiveness and is responsible for not leaving her home, say, to be present at her daughter's graduation. This seems to get things wrong, though, for why should the fact that an agent would exhibit such strange, idiosyncratic patterns of behavior in some possible world be sufficient to ground her moral responsibility in the actual sequence? The presence of just any possible scenario in which the agent's mechanism produces a different result is too weak a requirement for reasons-responsiveness.

So, strong reasons-responsiveness is too strong, and weak reasons-responsiveness is too weak. In order to carve out a middle ground, Fischer and Ravizza propose that what

is needed is a kind of moderate reasons-responsiveness, and they develop this notion by appealing to an asymmetry between reasons-receptivity and reasons-reactivity. More specifically, what is required for moral responsibility is a mechanism that is weakly reactive to reasons but which is receptive to reasons in a stronger sense. They suggest that by understanding these as asymmetrical we can make sense of both cases of weakness of will and cases of strange patterns. So, in the case of Bill, our weak-willed shoplifter, we can see that he meets the requirements of moderate reasons-responsiveness since he seems to be strongly receptive to reasons and, at least weakly reactive to them (as evidenced by his ability to refrain from stealing in the presence of a security guard). In contrast, our agoraphobic agent seems to fail to be moderately reasons-responsive for, although she is weakly reactive to reasons, she does not seem to be receptive to reasons in the stronger sense that Fischer and Ravizza have in mind. Thus, moderate reasons-responsiveness seems to give a plausible explanation of our intuitive judgments about each of these agents.

However, more needs to be said about this way of understanding receptivity and reactivity on Fischer and Ravizza's account. Just what is this stronger sense of receptivity that they claim an agent must have? In an attempt to characterize this, they write,

> In judging a mechanism's receptivity, we are not only concerned to see that a person acting on that mechanism recognizes a sufficient reason in one instance; we also want to see that the person exhibits an appropriate *pattern* of reasons-recognition. In other words, we want to know if (when acting on the actual mechanism) he recognizes how reasons fit together, sees why one reason is stronger than another, and understands how the acceptance of one reason as sufficient implies that a stronger reason must also be sufficient. (1998, 70-71)

A mechanism that shows this sort of pattern, Fischer and Ravizza call "regularly receptive." In order to determine whether or not a mechanism possesses such a pattern,

they ask us to imagine a third party conducting an imaginary interview where the agent in question is asked various questions about actual and hypothetical scenarios. If, they say, the agent's answers to questions in the imaginary interview produce a pattern of receptivity that is understandable by the third party, then the mechanism is regularly reasons-receptive. In short, on their view, "Regular reasons-receptivity … is reasons-receptivity that gives rise to a minimally comprehensible pattern, judged from some perspective that takes into account subjective features of the agent (i.e., the agent's preferences, values, and beliefs)…" (73) Furthermore, they claim, this pattern must be one that is "minimally grounded in reality."

Reactivity, on the other hand, requires no such pattern. Rather, if it is the case that a mechanism is reactive to reasons in *some* possible scenario, then that mechanism is reactive to reasons in *any* scenario. This is because, to use Fischer and Ravizza's phrase, reactivity is all of a piece. So, the fact that Bill's mechanism of action *would* react to reasons not to shoplift in the presence of a security guard suggests that the mechanism *can* react to reasons in the relevant way, and this is all that is required for moral responsibility according to Fischer and Ravizza. Given this, they summarize the asymmetry between receptivity and reactivity as follows:

> In the case of receptivity to reasons, the agent (holding fixed the relevant mechanism) must exhibit an understandable pattern of reasons-recognition, in order to render it plausible that his mechanism has the "cognitive power" to recognize the actual incentive to do otherwise. In the case of reactivity to reasons, the agent (when acting from the relevant mechanism" must simply display *some* reactivity, in order to render it plausible that his mechanism has the "executive power" to react to the actual incentive to do otherwise. (75)

So, on this view, an agent has control over her actions in the sense required for moral responsibility, only if those actions flow from a mechanism that is moderately reasons-

responsive in the sense just described and that belongs to the agent in the relevant way. All that is left is to make clear just what the notion of a mechanism "belonging to an agent" or "being the agent's own" amounts to.

Fischer and Ravizza characterize what it means for a mechanism to be an agent's own by appealing to certain historical facts about the agent. This feature of the account makes it an externalist approach insofar as it entails that the control required for moral responsibility involves more than merely that the internal states of the agent be arranged or configured in the right sort of way. Rather, on their view, a mechanism belongs to an agent only if the agent has "taken responsibility" for it, and the process of taking responsibility is a straightforwardly historical process which, they argue, involves three components. First, in order to take responsibility for a mechanism, an agent must come to see herself as the source of her actions – actions which have an effect on the world around her. Second, the agent must see herself "as a fair target of the reactive attitudes as a result of how [s]he exercises this agency in certain contexts." (1998, 211) These first two conditions of taking responsibility hang an individual's moral responsibility in an important way on her ability to see herself in a certain way. That is, in order to be responsible, an agent must have a conception of the self as one which engages with other people in morally significant ways. One must see oneself as one whose actions impact other people in the world and toward whom others may legitimately direct blaming and praising responses. In short, one must see oneself as a moral agent. The third, and final, criterion for taking responsibility, then, is that one's view of oneself as an agent of this sort must be sufficiently grounded on evidence. One's view of oneself as a moral agent must not be the result of delusion or something sufficiently similar.

Importantly, however, on Fischer and Ravizza's view taking responsibility, in the sense under discussion here, is not a matter of taking responsibility *for one's agency*. "When an agent takes responsibility," they write, "he obviously is *not* accepting responsibility for all his actions *whatever their source*; rather, he is accepting responsibility for only those actions which flow from a certain source." (215) In other words, what an agent takes responsibility for is the particular mechanism from which his actions flow. So, the agent must see the mechanism as her own, and she must see herself as a fair target of praise or blame on the basis of actions that flow from that mechanism. Once an agent has done this, Fischer and Ravizza claim, "it is almost as if the agent has some sort of 'standing policy' with respect to that kind of mechanism. Thus, when the agent subsequently acts from mechanisms of that kind, that mechanism is *his own* insofar as he has already taken responsibility for acting from that kind of mechanism." (216)

So, Fischer and Ravizza's theory of moral responsibility is an actual sequence, mechanism-based, externalist, moderate reasons-responsive view with a strong subjectivist criterion of mechanism ownership. The intricacy and nuance of the theory make it, perhaps, the most plausible (and certainly the most influential) reasons-responsive view on offer in the moral responsibility literature. However, I believe that it is open to at least two important challenges from the empirical research on ASD.

§2. The Challenge from ASD

As I have argued in Part I of this project, individuals with ASD display important deficits in counterfactual thinking. These deficits are such that they have implications for both the cognitive ability to recognize certain kinds of reasons as well as for a number of executive functions. This feature of ASD has important implications for Fischer and

Ravizza's characterization of moderate reasons-responsiveness which, recall, relies on an important asymmetry between reasons-receptivity, a cognitive power to recognize reasons, and reasons-reactivity, an executive power to react in light of reasons. The empirical data regarding ASD suggests, I will argue, that the receptivity/reactivity profile of individuals with the disorder is quite different than that of neurotypical individuals, and it is not clear that the account of that profile given by Fischer and Ravizza can accommodate individuals with ASD.

Additionally, Fischer and Ravizza's view relies on the subjectivist criterion of taking responsibility for one's mechanisms of action, and this criterion requires a robust ability to see oneself as the source of one's actions. The counterfactual thinking deficits experienced by those with ASD as well as the related affective irregularities suggest, I will argue, that this criterion is too strong. That is, it suggests that requiring that individuals see themselves in the particular way that Fischer and Ravizza think they must may not be necessary for moral responsibility, and this complication is made even more troubling by the pervasiveness of alexithymia among autistic populations. I will argue that individuals with ASD may be morally responsible for their actions despite failing to take responsibility for the mechanism of those actions, and if this is true it will follow that Fischer and Ravizza's theory does not, succeed in specifying necessary and sufficient conditions for moral responsibility. I will take each of these challenges in turn in the remainder of this section.

§2.1. Receptivity and Reactivity in ASD

Fischer and Ravizza claim that a mechanism's being moderately reasons-responsive is both a necessary and sufficient condition for the sort of control required for

moral responsibility. The first of the two challenges from ASD that I will present here denies the sufficiency of this condition by claiming to show that there may be individuals with ASD who are moderately reasons-responsive yet fail to be responsible for a range of actions. The reason that ASD presents a unique challenge to this notion, I contend, lies in the cognitive/executive construal of responsiveness that Fischer and Ravizza propose and the unique deficits that individuals with ASD experience in both cognitive ability and executive control.

To see how this may be, consider the moral deliberation mechanism of a high-functioning autistic agent. The first question that must be asked is whether or not the mechanism is regularly receptive to reasons in the sense that Fischer and Ravizza's view requires. That is, does the mechanism give rise to an understandable pattern of receptivity that is minimally grounded in reality? There is no evidence to suggest that individuals with ASD fail to meet this condition, and there is plenty of evidence to suggest that individuals with ASD show no impairments in their ability to recognize moral reasons. For example, individuals with ASD tend to be as adept at making the moral/conventional distinction as their typically developing counterparts.[2] That is, they are able to distinguish between transgressions on the part of others which involve what we commonly view as moral wrongs and those which involve wrongs that simply defy convention and this suggests, at least, provisionally, that individuals with ASD seem to be sensitive to the same sorts of considerations as neurotypical agents with respect to what matters morally. Moreover, individuals with ASD can recognize the moral/conventional distinction as adeptly as typically developed individuals even in cases involving authoritative

[2] See Blair 1996; Leslie, Mallon, and DiCorcia 2006.

permission (e.g. "Moral harm, *x*, would be wrong even if authority figure, *A*, said it was permissible."). (Zalla et al. 2011)

That individuals with ASD can make the moral/conventional distinction is enlightening only insofar as it suggests that autistic individuals are no worse off in this respect than neurotypical agents. There are important problems with the moral/conventional task that call into question what exactly it shows.[3] It may well be that all the task tells us is that agents who complete it successfully are able to successfully identify moral and conventional norms, but this does not entail that they are actually making *moral judgments.* Nevertheless, the fact that individuals with ASD perform at control levels on this task does suggest that they are at least receptive to moral considerations to a sufficient degree to allow them to successfully distinguish norms of morality from norms of convention, and this is not a trivial finding.

To say that individuals with ASD are able to make this distinction, however, is not to say that the ability to make certain fine-grained moral judgments is unimpaired in ASD. Despite being able to recognize the moral/conventional distinction, individuals with ASD tend to have problems with a number of features of moral judgment. For example, studies have found that these individuals have trouble distinguishing intentional from unintentional harms.[4] Moreover, they tend to have similar problems distinguishing attempted harms from neutral actions. In other words, when asked to make a moral judgment involving an act in which an agent accidentally causes some harm, individuals with ASD will tend to make the same judgment and prescribe the same degree of

[3] For a discussion of these problems, see, Shoemaker 2011.

[4] See Moran, et al. 2011; Koster-Hale et al. 2013.

punishment as they would in cases where an agent intentionally inflicts the same harm. Likewise, when asked to make a moral judgment in cases in which an agent tries but fails to cause harm to another, individuals with ASD will tend to make the same judgment as they would in morally neutral cases. This is, perhaps, not surprising since, in cases of accidental harms or failed attempts to intentionally harm, neurotypical individuals tend to assign blame or punishment based on information about inferred mental states, but accessing this information requires a robust counterfactual representation ability which, as we have seen, is impaired in ASD.

Additionally, in most every study on moral judgment in individuals with high-functioning ASD the results indicate that these individuals struggle to articulate adequate justifications for their moral judgments. In many cases, the explanation for any given moral judgment relies on an appeal to some rule or another and this in a very general condemnatory manner (e.g. "It's bad to lie."). One way of explaining this might be to appeal to the fact that high-functioning individuals with ASD become high-functioning in virtue of their ability to develop compensatory systems that make up for some of the cognitive or emotional impairments associated with autism. In order to make sense of their surroundings and to function in the world around them, individuals with ASD must come up with a way to relate to the world in spite of the psychological factors associated with ASD. One way that they are able to do this is by identifying a set of rules by which they can orient their interactions with other people. As a result, individuals with ASD are often seen as rigid rule-followers,[5] and this would explain the self-reported justifications for moral judgments like those just mentioned. More generally, it is important to

[5] For more on this, see McGeer 2008.

recognize that individuals on the high-functioning end of the autism spectrum cope with their autism and with their surroundings by developing cognitive strategies which allow them to circumvent some of the difficulties that their disorder entails.

One might be concerned that these atypical features of moral judgment in ASD are such that they preclude autistic agents from meeting Fischer and Ravizza's criterion for regular receptivity. After all, it matters to us whether or not an agent had a particular set of intentions in acting or whether she displayed good or ill will, and insofar as failing to be receptive to these features may amount to failing to be receptive to a particular set of moral reasons, it could be that individuals with ASD fail to meet the regular receptivity requirement. However, this is not the case. The fact that individuals with ASD are able to develop compensatory heuristics for moral judgments and to distinguish moral from conventional norms most of the time is evidence that they are capable of recognizing an understandable pattern of reasons that are minimally grounded in reality, and this is all that is required for regular receptivity according to Fischer and Ravizza. What this brings out, it seems, is an ambiguity in what Fischer and Ravizza take to constitute a moral reason. Fischer and Ravizza are explicit that their aim is to give an account of what it means to be responsible for a given *action*, and the fact that individuals with ASD make moral judgments in an atypical way does not preclude them from being regularly receptive to reasons so long as the judgments that are made yield coextensive sets of actions. Fischer and Ravizza's claim is *not* that regular receptivity requires making neurotypical moral judgments. Rather, it is that individuals must be receptive to reasons in such a way that would lead others to judge that it would be appropriate to respond to the individual with one or more of the reactive attitudes on the basis of the actions that

flow from the reasons she recognizes. It may well be the case that what individuals with ASD are doing in these atypical cases is not moral judgment, but something just as good. Nonetheless, insofar as this form of judgment yields actions that are consistent with those that flow from neurotypical moral judgments, there is no basis on Fischer and Ravizza's view for denying that individuals with ASD fail to meet the regular receptivity requirement.

Despite all of this, given the cognitive impairments present in ASD, surely there are cases in which individuals with autism are regularly receptive and weakly reactive to reasons yet are not responsible. This is because all that is required for regular receptivity, according to Fischer and Ravizza, is that a third party giving an imaginary interview could identify a regular pattern of receptiveness to reasons that is minimally grounded in reality. It is uncontroversial that such a pattern exists in high-functioning autistic individuals. Indeed, the compensatory systems developed by individuals with ASD seem to serve precisely this function. However the evidence of impairments in counterfactual thinking described above suggests that there is a range of reasons to which individuals with ASD may not be receptive, namely those reasons that require a robust ability to represent counterfactual states. It follows from this, then, that it is possible that individuals with ASD may act from a mechanism that is regularly receptive to reasons yet is not receptive to the reasons *that there are* to perform, or not perform, a given action if those reasons require an ability to think counterfactually.[6]

[6] Patrick Todd and Neal Tognazzini pose a similar challenge in their 2008 paper and suggest that Fischer and Ravizza's view needs to be amended in order to accommodate this worry. More on this and on why the challenge presented here goes beyond Todd and Tognazzini's below.

What is interesting about this case, however, is that the cognitive deficit that undermines receptivity for those with ASD undermines reactivity as well. This is trivially true in many cases (clearly, one cannot react to a reason that one does not recognize), but it is true in a wider range of cases than these as well. The reason for this lies in the nature of moral judgments more generally. Over the last decade, research in moral psychology has revealed that moral judgment appears to take place according to a dual process model. This dual process model has typically construed our moral judgments as occurring as a result of either an emotional or cognitive process.[7] On this view, cognitive processes tend to result in consequentialist moral judgments whereas emotional processes tend to result in deontological moral judgments. A dual system view such as this is helpful in explaining certain peculiarities in our moral judgments. So, for example, in Judith Jarvis Thomson's famous Bystander and Footbridge variants of the trolley problem (Thomson 1985), individuals tend, overwhelmingly, to make utilitarian and deontological judgments, respectively. If a dual-process model of moral judgment is correct, this difference in judgments can be easily explained since, it is posited, the latter case activates an emotional system that the former does not.

More recently, however, it has been proposed that this cognitive vs. emotional dual process framework be abandoned in favor of a model-based vs. model-free framework.[8] On this view, moral decision making (and, indeed, decision making in general) can be best understood in computational terms. In some cases, we make moral decisions by representing a model, or decision tree, which we then follow in order to

[7] See Haidt 2001; Green 2008; Cushman, Young, and Greene 2010.

[8] See Crockett 2013; Cushman 2013; Cushman 2015; Dolan & Dayan 2013.

reach some end to which we assign value. Decisions are then made according to whether or not certain courses of action conform to our model or help to achieve the desired end. In other cases, however, no model is constructed at all. Rather, in these cases we make decisions by assigning value to particular actions. So, via positive and negative feedback we begin to assign either positive or negative value to various action types through habituation and those actions which are met with positive feedback take on the form of moral rules. Fiery Cushman summarizes the distinction as follows:

> Goal-directed actions require a working model of the world. You pick a desirable outcome, and then form a plan to bring it about. Thus, they correspond to the class of model-based reinforcement learning algorithms. In contrast, habits are reactive stimulus-response pairings that are strengthened when followed by reward. Executing a habit does not require planning toward a valued outcome, and thus corresponds to the alternative class of model-free algorithms. (Cushman 2015, 59)

According to this framework, then, consequentialist moral judgments proceed from the model-based system while deontological moral judgments proceed from the model-free system since achieving desirable consequences requires having a working model of the world while adhering to moral rules does not.

This way of understanding moral judgment is, clearly, well suited to making sense of these judgments (and judgment and deliberation in general) in ASD. I argued at length in chapter 2 that ASD may be characterized by an impairment in the ability to represent counterfactual states and that this impairment is central to the observed deficits in executive function in individuals with the disorder given that it is a necessary component of the ability to represent problems and make plans. If this is true, then we should expect individuals with ASD to make decisions predominantly by way of a model-free process since the model-based process requires an ability to construct a model of the world as well as a plan for achieving goals, and, indeed, this seems to be correct. The

compensatory systems developed by individuals with ASD seem to be sophisticated model-free processes that these individuals are able to develop over time in order to navigate the world around them. That is, they seem to consist in a highly complex system of paired inputs and outputs. Through habituation, high-functioning individuals with ASD are able to identify certain types of appropriate responses to various situations, and, so, to develop a habituated, model-free algorithm that will aid them in responding to new situations which share features of previously experienced situations. This way of understanding deliberation and judgment in ASD meshes well with the observed repetitive and restricted behaviors and interests in ASD insofar as the model-free process favors rule-based behavior.

The reliance on model-free processes has implications for reasons reactivity in ASD as well. Since counterfactual thinking is *impaired* in ASD rather than *lacking* altogether it is possible that there may be cases in which a model-based system is required for making a successful moral judgment, and the autistic agent recognizes this and is able to form a model. However, it may be that, due to the impairment in counterfactual thinking, the model is ill-formed and not such that it could render the right sort of decision. Alternatively, it may be that the autistic agent is able to construct a model that is reasonably well formed but is unable to formulate a plan for action according to that model and is, therefore, unable to react to the reasons that she recognizes. In either case, the autistic agent seems to offer a genuine counterexample to Fischer and Ravizza's view of the sort of reasons reactivity that is required for moral responsibility. Recall that their claim was simply that to be reactive to reasons, a mechanism must be able to react to reasons in at least one scenario since doing so would

demonstrate an executive ability that is all of a piece. However, what these considerations about moral judgment in ASD seem to show is that such judgments implicate two *kinds* of executive power and that it is possible that a mechanism may fail to issue in morally responsible action if it possesses one but not the other. If this is true, then Fischer and Ravizza's claim that moral responsibility requires only weak reactivity seems to be false.

Michael McKenna and Alfred Mele have each separately made a similar objection to this portion of Fischer and Ravizza's view. In each case, the author asks us to imagine an agent who is reactive to one reason yet seems to be operating from a mechanism that is intuitively not responsible.[9] Importantly, neither of them takes this to be a reason to reject Fischer and Ravizza's view. Instead, they suggest that the picture of reactivity merely needs to be revised. As Mele puts it, "An attractive strategy for avoiding the (apparent) problem that I have been developing is to beef up the reasons-reactivity condition in such a way that … agents with … severe psychological maladies of the pertinent kind do not count as reasons-reactive enough to be morally responsible for the relevant behavior." (Mele 2006, 290) In response, Fischer accepts this revision to his and Ravizza's account of reactivity, saying,

> This posits a more refined notion of moderate reasons-responsiveness, with what might be called "spheres of responsiveness;" the "outer spheres" would not necessarily indicate sufficient responsiveness for moral responsibility. Of course, it may not be straightforward to characterize precisely the "borders" of the spheres; that is, it might not be easy to say exactly what degree of strength of the relevant sort of urge renders the agent in question immune to moral responsibility. (Fischer 2006, 328)

Despite this alteration, he maintains that this revision is consistent with thinking of reactivity as "all of a piece" since one way of interpreting this notion of spheres of

[9] Mele uses the case of an agoraphobic agent while McKenna rests his argument on a hypothetical case of an agent that is reactive to one and only one reason. See, Mele 2006; McKenna 2005.

responsibility is to say that it simply means that "an agent who can react to any reason may have great difficulty in doing so in any particular context." (Fischer 2006, 328n7) So, Fischer claims, accepting the objection from McKenna and Mele preserves the basic features of the Fischer and Ravizza view.

However, a similar response to the objection that I have presented here is not open to Fischer and Ravizza for at least two reasons. First, the empirical evidence regarding the dual process system of moral judgment shows that individuals react to reasons in two ways which are distinguishable from one another. On the basis of this evidence, then, the objection from ASD explicitly denies the "all of a piece" claim made by Fischer and Ravizza. Second, Mele's objection to Fischer and Ravizza is presented in terms of irresistible desires as is Fischer's response to it. Because of this, Fischer's response offers a picture of agents as possessing concentric spheres of responsibility where the agent, in the inner spheres, is not confronted with desires or urges which block her ability to exercise guidance control but, in the outer spheres, she may be. Thus, reactivity is a matter of degree, and, he thinks, we need not worry about those outer spheres where the agent is overcome by her urges and where our intuitions are fuzzy anyway. However, the objection from ASD rests not on the presence of overwhelming urges or on *degrees* of reactivity. Instead, it rests on the presence of an impairment in a particular *kind* of reactivity. Because of this, it makes little sense to talk of spheres of responsiveness where reactivity is weakened as one moves toward the outer spheres. Rather, moral judgment requires both kinds of reactivity regardless of which sphere one finds oneself in. So, the objection from ASD shows, I contend, that reactivity is not all of a piece and that, as a result, an individual can fail to be morally responsible while

meeting Fischer and Ravizza's conditions for reactivity, *even if these conditions are strengthened* in the way that McKenna and Mele suggest.

One response to this objection that Fischer and Ravizza might offer is this: the fact that moral judgment proceeds according to two separate processes suggests that there are two separate mechanisms at work, and so, insofar as individuals with ASD react to reasons on a model-free process but do not do so on a model-based process this suggests that their model-free mechanism is weakly reactive while their model-based mechanism is not. However, this is to oversimplify the dual process model of moral judgment. This is because the two systems are not simply independent systems which issue in particular judgments on their own. Rather, as Cushman puts it,

> The emerging picture is not … as simple as "two independent systems: model-based and model-free." For one thing, each of the "systems" clearly comprises multiple dissociable components … In addition, decision making in the model-free system appears to involve independent and opponent "go" and "no-go" processes. More critical to the current discussion is the evidence that model-free and model-based systems interact closely. The successful operation of each system largely depends on interactions with the other. (Cushman 2013, 280)

So, given that Fischer and Ravizza define a mechanism as simply the process by which an action is produced, there is no empirical basis for treating model-based and model-free processes as distinct mechanisms since both processes will be involved in the production of a given judgment. My claim here is that individuals with ASD rely much more heavily on a model-free system than do neurotypical individuals, not that they operate entirely on a model-free system. If this is true, then it would be implausible for Fischer and Ravizza to claim that a different mechanism is at work. One way around this difficulty might be for Fischer and Ravizza to claim that mechanisms are to be understood much more specifically than this. That is, they might claim that the mechanism issuing in action is to

be identified with the actual psychological details or brain states of the agent. In their book, they are quite vague about how precisely to identify the mechanism in question and to hold it fixed, but there are good reasons for them to avoid identifying mechanisms with actual states of the brain.[10]

So, if Fischer and Ravizza's account of weak reasons reactivity is not satisfactory what type of reactivity would do the trick? It seems to me that it would need to be something much stronger than what they suggest and considerably stronger than the type suggested by McKenna and Mele. What is needed, if we are going to get things right in the case of ASD is something that might be referred to as "regular reactivity," to riff on Fischer and Ravizza's terminology. Regular reactivity would involve a regular pattern of reacting to reasons which issues from a well-balanced dual process system of deliberation, i.e. a system in which the model-free and model-based systems are functionally typical and act in concert with one another. This requirement would avoid the problem of judging individuals with ASD to be responsible for actions when they lacked the capacity to react to the relevant reasons. However, accepting this version of reasons reactivity will prove problematic for Fischer and Ravizza.

I have argued that ASD seems to pose a challenge to Fischer and Ravizza's account of moderate reasons-responsiveness in two ways. First, it is clear that individuals with ASD demonstrate regular patterns of receptivity to reasons yet are not responsible for a range of actions given that their cognitive deficits rule out their having access to a certain range of reasons. So, Fischer and Ravizza's notion of regular receptivity is too weak and must be made stronger if it is to render the correct verdict regarding individuals

[10] See McKenna 2005 for a discussion of these reasons.

with ASD. One way of doing this would be to claim, as Patrick Todd and Neal Tognazzini do, that Fischer and Ravizza should adopt a stronger account of regular receptivity which says that an agent must be receptive to the actual reasons that there are for acting.[11] Second, given the empirical evidence for a dual process system of moral judgment together with the executive control deficits displayed by individuals with ASD, it seems to be the case Fischer and Ravizza's account of reasons reactivity is also too weak insofar as it may also count those with ASD as responsible in cases where they seem clearly not to be. One possible way around this, I have just suggested, is to adopt a view which says that regular reactivity to reasons is necessary for moral responsibility, thereby making this condition of reasons-responsiveness considerably stronger as well.

However, it seems to me that Fischer and Ravizza cannot take on both of these suggested revisions to their account. To do so would be to say that in order to be responsible an agent must be able to recognize the actual reasons there are for acting and then must react to those reasons according to a consistent pattern and in a way that implicates a well-balanced system of judgments. In other words, this way of proceeding strengthens both the requirement of regular receptivity *and* the requirement of weak reactivity. To modify both of these, though, is to require something too strong because it results (to a large extent, at least) in a loss of the asymmetry between receptivity and reactivity, and this asymmetry is what enables their conception of moderate reasons-responsiveness. In short, taking on the revisions that I have outlined here pushes Fischer and Ravizza toward a view that looks much like the strong reasons-responsiveness that

[11] They define this notion as follows: "An actually operative kind of mechanism is receptive to the actual reason if and only if there is a world in which the same mechanism operates and the same sufficient reason to do otherwise is present and the agent recognizes the sufficient reason to do otherwise." (Todd and Tognazzini 2008, 692)

they explicitly reject. It seems to me, therefore, that individuals with ASD pose a challenge to Fischer and Ravizza's view, and it is not clear that their account has the resources to overcome it.

§2.2. ASD and the Subjective Requirement of Responsibility

I take it that the challenge just presented offers compelling evidence against Fischer and Ravizza's view. However, I will briefly discuss another challenge that the literature on ASD may present for their account. Recall that another requirement for responsibility was that a mechanism must belong to the agent in the relevant way and that they define this sense of "belonging to" in terms of taking responsibility for the mechanism. Recall, also, that taking responsibility occurs in three stages: the agent must see her actions as having significance for other people, she must see herself as an apt target of the reactive attitudes of others, and her seeing herself in these ways must be based on evidence. If an agent fails to meet any of these three conditions, then she has not taken responsibility for the mechanism of her actions and is, therefore, not responsible for them. My claim in this section is that individuals with ASD may be responsible for their actions in certain cases despite failing to meet the second, and possibly the first, of these two conditions.

Mele (2006) has posed a similar challenge for Fischer and Ravizza's view by focusing in on the second of these three subjective criteria. He asks us to imagine an agent, Phil, who believes that determinism is true and that it is incompatible with moral responsibility and, so, does not see himself as an apt target of the reactive attitudes of others. If the second of Fischer and Ravizza's subjectivist criteria is true, he claims, then Phil is not morally responsible since he has not taken responsibility for his mechanism of

action. Assuming, for the sake of argument, that some people are in fact morally responsible, then it seems to follow that Phil is not among them since he has, by hypothesis, false beliefs about determinism and moral responsibility. However, this seems to clearly be the wrong conclusion in this case.

Fischer has two responses to this case. First, he intimates that it is not clear that the case of Phil offers a genuine challenge to his, and Ravizza's, view because it is not clear that we have strong intuitions in that case. He reiterates that the goal of the view is to arrive at a reflective equilibrium between common sense cases and theoretical principles, and he implies that Phil's case is not clearly one with a common sense intuition. Second, he claims that even if Mele's problem case does genuinely tell against the second of his three subjectivist criteria, that particular criterion may be dropped without threat to the overall theory. (Fischer 2006)

By appealing to the empirical data on ASD, however, I believe that a case can be constructed in which neither of the responses offered by Fischer are correct. Recall from chapter 3 that autistic populations experience an unusually high rate of alexithymia, a disorder which causes an inability to recognize one's own emotions. Recall, also, that the empirical data seems to show that individuals with ASD show abnormalities in their experience of complex social emotions (e.g. guilt, shame, regret, relief, disappointment, etc.) Now, as Fischer puts it, the three subjectivist criteria are supposed to be signs of a kind of "self-engagement characteristic of a morally responsible agent." (Fischer 2006, 327) However, given the affective deficits in ASD it is plausible to think that there may be individuals who lack this self-engagement but are nevertheless responsible for their

actions. To see this, imagine a case of an autistic person that approximates the paradigm case but with slightly more severe impairments.

> Rachel, a high-functioning autistic agent, understands and is able to make a number of moral judgments. She does this primarily by way of a model-free process of moral judgment in which she assigns value to particular action types. Throughout her life, she has consistently received negative feedback in circumstances in which she has laughed at funerals. As a result, she has assigned a great deal of value to not laughing at funerals. Suppose, further, that Rachel is alexithymic and so does not fully recognize various emotional states when she experiences them and that she has an atypical understanding of the various social emotions. Now, imagine that Rachel is attending the funeral of a distant relative who has passed away when a particular phrase that the priest utters strikes her as funny. She laughs audibly and deeply offends the next of kin of the deceased.

I have the strong intuition in this case that Rachel is morally responsible for laughing for two reasons: (1) she knowingly violated a moral rule that she can be reasonably expected to follow, and (2) it is a rule to which she has identified as having moral value.

The reason that this case is problematic for Fischer and Ravizza, I think, is that it is unlikely, given Rachel' incapacities, that she meets the second subjectivist criterion of their view, and it is plausible that she fails to meet the first criterion as well. Here is why, in order to understand oneself as the apt target of the reactive attitudes of others, one must understand how those attitudes are directed and must have some phenomenological acquaintance with them as well.[12] However, given that Rachel is both limited in her ability to recognize these emotions due to her alexithymia and has atypical phenomenological experience of them, it seems as though she would lack the ability to see that she would be the appropriate object of these emotions in others (this conclusion is even more plausible if we suppose that she has the typical deficits in mental state reasoning that those with ASD generally possess). So, we have a case of an individual

[12] Michael McKenna argues for a similar view in his book, *Conversation and Responsibility.*

who fails to meet the second subjectivist criterion and is nonetheless morally responsible. Moreover, that Rachel fails to meet this criterion seems to me much more clear than the claim that Mele's hypothetical agent fails in the same regard. If this is true, then we have headed off Fischer's first response to Mele to some degree.

The stronger claim that I want to provisionally make here, however, is that Rachel's incapacities seem to be such that she fails to meet the first of the three subjectivist criteria in a particular sense as well. Recall that the first criterion simply says that the agent must see her actions as having significance for others. That is, she must see them as bringing about certain consequences for other agents. This is certainly true of Rachel in most cases it would seem. That she assigns value to actions suggests that she sees those actions as having effects in the world. However, she also seems to be limited in her capacity to see certain *kinds* of effects that her actions might have. If we suppose that she is cut off from certain emotions that she feels, either by her inability to recognize them or her atypical experiences of them or both, then it is plausible to suppose that she cannot see her actions as causing these same emotions to exist in other people, and if that is true, then in a limited sense, she is unable to see her actions as efficacious for others. If this is true, then Fischer's second response to Mele, will not apply in this case either, since doing away with both the first and second subjectivist criteria would be decisive evidence against his and Ravizza's view in general.

The objection presented in this section is surely more controversial than the objection I offered in §2.1, and Fischer and Ravizza may well be able to respond to it in a way that would save this portion of their view. Nevertheless, given that the subjectivist portion of their view aims to offer a requirement of engagement of the agent with the rest

of the moral community, there is something intuitively plausible about the claim that autism (the word literally means "self"-ism) represents a case of a lack of such engagement. Appealing to the affective features of the disorder seems to me to be a plausible way of developing this intuitive notion.

§3. Conclusion

To conclude, the empirical evidence regarding ASD shows that both components of Fischer and Ravizza's account of moderate reasons-responsiveness fail to provide necessary and sufficient conditions for the type of control required for moral responsibility. Moreover, the most plausible revisions to the view that would accommodate this evidence are such that they would transform the view into one that Fischer and Ravizza explicitly deny. In addition to posing a challenge to their account of moderate reasons-responsiveness, the evidence of affective irregularities and the presence of alexithymia in ASD also pose a challenge for Fischer and Ravizza's account of taking responsibility insofar as there may be individuals who lack the capacity to meet the subjective requirements that Fischer and Ravizza specify but are nevertheless responsible agents. Despite my being heavily critical of their view, there may well be a way to proceed which would allow us to salvage some of the best features of the theory. Specifically, it may be that what I have offered here is a deep methodological worry for Fischer and Ravizza.

As should be clear, much of the reason that the evidence from ASD has been problematic for Fischer and Ravizza's view is that they are aiming to give a classical analysis of moral responsibility in terms of necessary conditions. However, it seems as though no matter what revisions to these conditions are made, the view remains

problematic in important ways in light of this evidence. So, perhaps a way to move forward would be to abandon this methodology of strict conceptual analysis in favor of a methodology which aims to examine how the theory would handle paradigm cases of responsible agency and then to modify it as needed when one moves on to more atypical cases. What this would give us is an account of *the degree to which* people with ASD are responsible agents and it would allow us to preserve the attractive features of the theory.[13] I am sympathetic to this way of proceeding, for surely there is much to like about Fischer and Ravizza's view (an agent's ability to recognize and respond to reasons is clearly relevant to her moral responsibility), and a positive account may, indeed, require such an approach. So, I grant that what I have said here does not amount to a death knell for the reasons-responsive approach. However, it does give us reason to suspect that reasons-responsiveness may not be the best candidate for an account of autistic persons, and, so a different approach may need to be considered. In the following chapter, then, I will do just that by considering the extent to which real self theories can accommodate the evidence from ASD.

[13] I'm grateful to Michael McKenna for this suggestion.

Chapter 5: ASD and the Real Self

Having shown that certain features of ASD pose considerable problems for reasons-responsive views of moral responsibility generally, and Fischer and Ravizza's view in particular, I will turn now to a discussion of another prominent type of theory which claims that moral responsibility is a function of an agent's actions or attitudes having their source in certain features of the agent. Accounts of this type are referred to as "real self" theories,[1] and in the contemporary literature, the number of distinct views reflecting this methodology is roughly equal to the number of theorists defending it. As a result, it is difficult to state this general approach with any greater specificity than the following: a real self view is one which claims that an agent is responsible for her actions or attitudes in virtue of the fact that they are expressive of her real self. The various real self views are distinguished from one another, then, by the way in which they specify what an agent's real self consists in.

In this chapter, I will discuss two prominent conceptions of the real self, one offered by Harry Frankfurt and the other by Gary Watson. Frankfurt's view attempts to ground the real self in a particular hierarchical relationship between an agent's desires, and Watson attempts to do the same by appealing to the interaction between the agent's

[1] These theories are sometimes also referred to as "deep self" theories or "mesh" theories, and I will use these interchangeably throughout this chapter.

desires and her evaluative judgments. I will argue that neither of these approaches gives an adequate account of autistic individuals and that the primary reason for this is that individuals with ASD are impaired in their ability to understand the nature of desire and, as a result, to reflect on their own desires in the right sort of way. Having presented the challenge that ASD poses for both Frankfurt and Watson, I attempt to offer a positive account of the real self that might succeed in explaining when actions or attitudes may be appropriately attributed to autistic persons. Finally, I pause, in an appendix, to consider one additional theoretical approach to responsibility which could be plausibly construed as a real self view, namely, the sort of view defended by T.M. Scanlon and Angela Smith which holds that actions are attributable to an agent only if they are connected in the right way to the agent's judgment sensitive attitudes.

§1. Frankfurt's Hierarchical Account

The most influential and widely-discussed real self view was first proposed by Harry Frankfurt in his, "Freedom of the Will and the Concept of a Person." There, Frankfurt's central claim was that both freedom of the will and personhood are best understood by reference to the structure of an agent's desires. Such a view can make sense of the intuitive idea that part of what it means for an agent to act freely is simply for her to act in the way that she wants to act. This view, however, is clearly too simplistic, and, so, Frankfurt offers a number of distinctions in order to shed light on just how an agent's desires are central to her being morally responsible for her actions.

He begins by making a distinction between an agent's will and an agent's second-order volitions. The will, according to Frankfurt, is simply an effective first order desire (i.e. a desire that actually leads an agent to act in accordance with it). A second-order

volition, then, is a desire for a particular first order desire to be one's will. According to Frankfurt, whether or not an agent is morally responsible is to be determined by the hierarchical relationship between one's effective first order desires and one's second-order volitions. If an agent's actions issue from a will that she wanted to have, then her will is free and those actions are *hers* in an important sense.

To show this more clearly, Frankfurt offers examples of three hypothetical drug addicts. Imagine, first, a case of a typical, unwilling addict who desperately wants not to succumb to his addiction. However, given the strength and irresistibility of his first order desire for the drug, he always succumbs to the addiction despite wanting his effective first order desire, or his will, to be something other than it in fact is. In short, the unwilling addict has a second-order volition (to not want the drug) that his will (to take the drug) not be the will that he has. Thus, he lacks an appropriate mesh between his second-order volition and his will, and, on Frankfurt's account, he is not responsible when he takes the drug.

We can contrast the unwilling addict, Frankfurt suggests, with the wanton addict. A wanton, on Frankfurt's account is an individual who lacks second-order volitions altogether. So, the wanton addict, though he has the same irresistible first order desire as the unwilling addict, simply does not care which of his first order desires issues in his acting. That is, the wanton is utterly indifferent to which of his first order desires constitutes his will, and as a result, the wanton fails to even be a person at all. As Frankfurt puts it, "When a *person* acts, the desire by which he is moved is either the will he wants or a will he wants to be without. When a *wanton* acts, it is neither." (Frankfurt

1971, 19, emphasis in original) The wanton addict, then, is not responsible simply because he fails to meet the criteria for being a person.

Finally, Frankfurt asks us to imagine a third addict, the willing addict, who shares the irresistible desire to take the drug but who is completely delighted with this desire. That is, the first order desire to take the drug is the desire that the willing addict wants to have. If this desire were to suddenly wane, then he would take any steps possible to ensure its return. Intuitively, it seems, we would be correct to hold this addict responsible for taking the drug in a way that we ought not to hold the unwilling addict responsible. Frankfurt's explanation for this is that the willing addict's taking of the drug issues from a hierarchical relationship between his second-order volition and his effective first order desire (in which the former is what causes the latter) which the unwilling addict lacks, and the arrangement of these desires makes it the case that he is responsible for his actions.

Frankfurt's picture of what it means for an agent to be morally responsible is an intuitively appealing one. It captures something striking about actions for which agents can be held responsible, namely, that they reflect certain important facts about the agent herself, or about who the agent *really* is. However, his view is not without its problems. The most notable among these is the fact that there seems to be nothing particularly special about second-order volitions that would license our affording them such a central role in grounding such important concepts as personhood and responsibility, or, as Gary Watson puts it, self-determination. That is, if second-order volitions are nothing more than desires, then why should we think that they are so important, and what would stop us from requiring that they be desires that an agent wants to have in the same way that

Frankfurt requires that the will be a desire that the agent wants to have? There is nothing in Frankfurt's picture, so far, that allows us to avoid such an infinitely ascending hierarchy of desires. In short, why should we think that second-order volitions are authoritative in the way that Frankfurt's view seems to suggest? Sensitive to this issue, Frankfurt attempts to stem the worry about a regress by invoking the notion of identification. Thus, he writes,

> It is possible, however, to terminate such a series of acts without cutting it off arbitrarily. When a person identifies himself *decisively* with one of his first-order desires, this commitment "resounds" throughout the potentially endless array of higher orders. Consider a person who, without reservation or conflict, wants to be motivated by the desire to concentrate on his work. The fact that his second-order volition to be moved by this desire is a decisive one means that there is no room for questions concerning the pertinence of desires or volitions of higher orders. (Frankfurt 1971, 21)

So, Frankfurt thinks it possible to non-arbitrarily cut off any potential regress by identifying with the desire that constitutes one's will.

This portion of Frankfurt's original view has been the subject of forceful criticism by Watson who denies that Frankfurt's proposed source of authority for second-order volitions can do the work that Frankfurt wants it to do.[2] Watson's central claim here is just that Frankfurt's proposed solution on the basis of identification with a particular desire simply cannot avoid the charge of arbitrariness. As Watson puts it,

> We wanted to know what prevents wantonness with regard to one's higher-order volitions. What gives these volitions any special relation to "oneself"? It is unhelpful to answer that one makes a "decisive commitment," where this just means that an interminable ascent to higher orders is not going to be permitted. This *is* arbitrary. (Watson 1975, 349)

In responding to this criticism, Frankfurt clarifies his view by claiming that the identification needed in order to assign second-order volitions the authority they need

[2] See Watson 1975. I'll have more to say about Watson's alternative proposal below.

comes in the way of a wholehearted decision by the agent to constitute herself in accordance with the desire in question. (Frankfurt 1987) Central to Frankfurt's response is the claim that when one makes a *decisive* commitment to a certain desire he does so "in the belief that no further accurate inquiry would require him to change his mind." (169) Frankfurt continues,

> It is therefore pointless to pursue the inquiry any further … [A] person can without arbitrariness terminate a potentially endless sequence of evaluations when he finds that there is not disturbing conflict, either between results already obtained or between a result already obtained and one he might reasonably expect to obtain if the sequence were to continue. Terminating the sequence at that point – the point at which there is no conflict or doubt – is not arbitrary." (169)

So, when an agent makes a decisive commitment to a particular desire, she is deciding to endorse that desire and is thereby creating herself in a distinctive way (Frankfurt uses the language "making up one's mind" here to convey the notion that the agent is literally configuring her mind in a certain fashion). When such a decision is made wholeheartedly – that is, when an agent, in deciding, establishes certain patterns of response or certain preferences in her mental economy – identification becomes an act sufficiently authoritative as to give the agent's second-order volitions the quality that Frankfurt's account suggests and to overcome the charge of arbitrariness leveled by Watson.

In later work, Frankfurt characterizes wholeheartedness not as a function of identification but as a function of a kind of self-satisfaction. (Frankfurt 1992) On his later view, to be wholehearted about something is to lack a certain ambivalence about it. "Ambivalence," he says, "is constituted by conflicting volitional movements or tendencies, either conscious or unconscious, that meet two conditions. First, they are inherently and hence unavoidably opposed; that is, they do not just happen to conflict on account of contingent circumstances. Second, they are both wholly internal to a person's

will rather than alien to him; that is, he is not passive with respect to them." (Frankfurt 1992, 8) What makes one wholehearted, then, is that one is satisfied with one's desires such that ambivalence of this sort is not present in the individual. Frankfurt describes this sense of satisfaction as simply an "absence of restlessness or resistance." (12) In order to be satisfied with one's desires, then, one must simply have no drive to change the desires that one has. This is, clearly, a much weaker condition than the making of a decisive commitment involved identification, but the evidence from ASD will, I think, pose similar problems for both versions of the view.

The important questions for our purposes are these: (1) are the desires of individuals with ASD such that they can take on the hierarchical relationship necessary for Frankfurt's account? And (2) are individuals with ASD able to engage in the process of identification, or be in the state of satisfaction, that Frankfurt thinks is necessary for responsibility? I will address each of these in what follows.

In order to answer these questions we need some understanding of how desire is understood by individuals with ASD. We have already seen some evidence, in Part I, that desire understanding may be impaired in ASD. First, since individuals with ASD show signs of impaired ToM (or, better, that they fail ToM tests due to the tests' counterfactual requirements), it is plausible that they might show deficits in the ability to understand desires as mental states in much the same way as belief because understanding desires would also require representing counterfactual or non-occurrent mental states. Second, given that desires commonly involve counterfactual representations in a way that other mental states – like belief, perhaps – do not (e.g. my desire to go to a concert might involve my representing some counterfactual state of affairs by imagining myself being

there), it might be expected that individuals with ASD would lack some understanding of desire on this basis as well. Indeed, this impaired understanding of desire seems to be borne out in the empirical literature.

While research that specifically targets desire in ASD is comparatively sparse, there is nevertheless some evidence available which suggests that individuals with ASD are less able to understand desire than are neurotypical individuals. For example, in a study conducted by Phillips, Baron-Cohen, and Rutter (1995) children with ASD showed signs of severe impairment relative to neurotypical and mentally handicapped participants in both the ability to recognize when the desires of others were unsatisfied and the ability to understand the representational content of desires. In the first experiment of this study, participants were presented with vignettes about actors whose desires were satisfied and others whose desires went unsatisfied. In half of the stories the desires of the actor were made explicit and in the other half participants were required to infer the desire prior to judging satisfaction. The researchers found that individuals with ASD performed at near control levels on the explicit tasks but showed signs of significant impairment on the implicit tasks (54% of autistic participants passed the explicit task while only 13% passed the implicit task whereas control participants passed the tasks at rates of 77% and 67% respectively). What is suggested by this study is that individuals with ASD are capable of matching goals with their corresponding successful outcomes, but that they seem much less able to infer the desires of others when the goal or desire is not provided explicitly.[3]

[3] It is worth noting that this is in line with my claim in Chapter 2 that the counterfactual thinking deficit in ASD presents itself when the counterfactual elements of a particular task are implicit rather than explicit. Recall that in cases where participants were provided with the counterfactual elements of a task, they were able to perform near control levels. This could provide a useful explanation for the deficits described here as well. Insofar as understanding the desires of others requires representing those desires as belonging to another person, this could explain the relative inability to infer desires in this study. Thus, impaired desire understanding may constitute another data point in favor of the counterfactual thinking hypothesis.

In a second experiment, the authors tested the ability of participants to recognize when another person's desire has changed. The goal of this experiment was to test the ability of participants to recognize the representational nature of desire. As they put it,

> The same entity (object, event, situation) can appear desirable to one person, while at the same time being undesirable to another … to understand this fact requires a concept of desire that includes its personal, subjective nature. One needs to understand that desirability is not a property of the object, but of the person's mental representation of the object. (Phillips, Baron-Cohen, and Rutter 1995, 160-161)

In order to test this ability, participants were presented with vignettes in which characters are described as beginning to perform some action, changing their minds, and then performing a different action, and participants were asked to identify the desire of the character before he or she changed his/her mind. The results showed that typically developed children and children with mental handicap performed at near-ceiling levels while children with ASD performed substantially worse. The authors summarized the results as follows:

> In this experiment, it was not the satisfaction conditions of desire that needed to be understood … it required participants to realize that desires exist at the mental level, representing aspects of the world as desirable or not desirable. It would appear that children with autism are impaired in the ability to understand this aspect of desire, compared with children with mental handicap and normal children. Although there was some indication that the task is less difficult than understanding false belief, an autism-specific deficit was still apparent. (165)

So, based on the research of Phillips et al., it would appear that individuals with ASD do, indeed, show signs of impairment in the ability to understand the nature of desires as states that occur in the mind.

Similar results were obtained in a more recent study conducted by Broekhof et al. (2015) which tested the ability of participants to recognize the subjective nature of desire. In this study, participants were given stories containing information about a character's

preferences. In half of these stories, the character's preference was the same as the preference of the participant, and in the other half, the preferences of the two were different. Participants were then asked questions about how the character in the story would behave on the basis of his or her desires. On tasks in which the desire of the character in the story matched the desire of the participant, participants with ASD performed at the same level as typically developed participants. However, on tasks where the preferences of the two were dissimilar, participants with ASD performed significantly worse than typically developed controls. What these results suggest, the authors claim, is that individuals with ASD seem to be impaired in the ability to recognize that the desires of others differ from their own. If this is correct, then it would support the argument from Phillips et al. that autistic individuals seem not to grasp the representational nature of desire. To put this another way, for autistic persons it seems to be the case that the property of being desirable attaches to the object itself rather than to one's representation of the object. So, on this conception, it would not be the case that an object is desirable *to me*. Rather, it would the case that the object simply *is desirable*, and to have a desire is to be drawn to (or to have some typical phenomenological experience of wanting) the object in virtue of its having this property. To help avoid confusion, I will refer to this unique sense of desirability as "desirability*" in the following discussion.

Given this evidence, coupled with the evidence presented in Part I, it can be concluded that individuals with ASD show signs of impaired understanding of desire. It is important to note that the experimental evidence with respect to desire is much the same as the evidence on belief understanding presented in chapter 2. Recall that the evidence on belief understanding in ASD showed an atypical understanding not only of

the beliefs of others but of one's own beliefs as well. While the studies cited here do not take up the question of whether individuals with ASD understand their own desires in an ordinary way, it can be reasonably inferred that they do not. Given the similarity between these experimental methods and those cited in chapter 2 regarding belief understanding, it is more likely that individuals with ASD are impaired in their understanding of desires in general, including their own desires. With this evidence presented, we can now turn to the issue before us: how, if at all does this impairment in desire understanding help us to answer the two questions relating to Frankfurt's account posed above? I will address each of these in turn.

Obviously, individuals with ASD have desires. Just as obviously, those desires can be effective in issuing in action. So, individuals with ASD clearly have a will in the Frankfurtian sense. The important question for Frankfurt's view is whether or not individuals with ASD can have second-order volitions. Given the evidence just presented, it is not clear what the answer to this question should be. No research has been conducted, to my knowledge, on the presence of second-order volitions in ASD, so our best hope for answering this question is to do so speculatively on the basis of desire understanding in ASD more generally, and it seems to me that there may be good reason to doubt that such desires are present. The general conclusion to draw from the evidence regarding desire understanding is that individuals with ASD seem impaired in their ability to understand that desires are mental representations.

The reason that this is problematic for Frankfurt lies in the fact that his view requires a substantial degree of self-reflection on one's desires *as the desires of one's self.* So, the crucial feature of the wanton is that he simply does not care to reflect on his

first order desires and, so, does not have second-order volitions as a result. The problem that the evidence on desire understanding in ASD poses comes out of this issue of self-reflection as well. However, while the wanton fails to reflect on his first order desires simply because he does not care what they are or what they will be in the future, my suggestion is that individuals with ASD do not engage in this sort of self-reflection because they do not have a sufficient grasp of desire as something that is fundamentally subjective. To reflect on desire in ASD, then, is to reflect on the desirability* of some object, but given this conception of desire, this is not to reflect on one's *will*. It is the latter sort of reflection that Frankfurt's account requires.

Perhaps an example will be helpful here. Suppose that our paradigm agent, Adam, desires a cigarette. On the basis of this desire, he lights up a smoke and puffs away. So, his will is to smoke the cigarette. Now, the important question, on Frankfurt's view, is whether or not Adam wants this to be his will. How will Adam answer this question? If the studies cited above give a correct picture of Adam's understanding of desire, then it is hard to see how he could make sense of this. It is possible that he might bemoan the fact that desirability* is a property of cigarettes. That is, he might desire that cigarettes not be desirable*, but, crucially, this is not a desire about his will. Rather, it is a desire about cigarettes. It is on a par with desiring that cigarettes not be unhealthy or desiring that they not smell bad, or something similar. As such, it is not a second-order volition at all. The key problem for Frankfurt's account that the evidence on desire in ASD presents, then, is this: given that desires are taken to be the desirability* of their objects, any second-order volition in ASD that purports to play the role that Frankfurt's view needs it to play will fail to do so because such desires would turn out not to be about one's will at all.

If one does not find the evidence or the arguments presented so far convincing, there may be good reason to think that individuals with ASD do not have second-order desires based on their counterfactual impairments as well. This is because desiring that one's will be different than it currently is would require that one be able to counterfactually represent oneself as having some other will. It would require that one imagine oneself in the future and judge that one's future self would be better off with different effective desires. This, of course, would require a robust ability for counterfactual representation, and this is precisely what individuals with ASD seem to lack. So, Adam, in order to have genuine second-order volitions would have to reason as follows: "I wish it were the case that I didn't desire cigarettes because if I didn't I would save a lot of money by not having to buy them every week, and I wouldn't have to worry about the health problems that I would encounter in the future after years of smoking." Clearly, though, this requires a strong ability for counterfactual thinking as well as a broadly intact capacity for episodic future thinking, and both of these, as I have shown, are impaired in ASD.

So, if this speculative conclusion regarding the absence of second-order volitions in ASD is correct, then how would autistic individuals be classified on Frankfurt's account? It seems as though his view would entail that individuals who have impairments such as these in desire understanding are, in fact, wantons, since, on his view, a wanton is just an individual who lacks second-order volitions. However, this is a repugnant implication of the view since not only does it entail that many individuals with ASD are not responsible agents but that they are not even persons. What is more, this implication can easily be falsified by anyone who has spent any amount of time conversing with a

high-functioning autistic person. So, if Frankfurt is committed to this view of individuals with ASD, then this is reason to reject the view.[4]

However, there may be a response open to Frankfurt here. If it is correct that someone like Adam, in the example above, could have the desire that desirability not be a property of cigarettes, then perhaps a desire of this sort could play the functional role of second-order volitions and prevent wantonness in individuals with ASD. After all, Adam does appear to be relevantly different than Frankfurt's case of the willing addict. The fact that he wants cigarettes not to be desirable* sets him apart, in some way, from the person who is wholly delighted that they have this property. So, if desires about the desirability* of particular objects play a functionally equivalent role to second-order volitions in Frankfurt's view, then perhaps he can say that individuals with ASD can be morally responsible agents by virtue of their having *these* desires even in the absence of second-order volitions. This response might be sufficient to show that individuals with ASD have something like second-order volitions, but it will not succeed in saving Frankfurt's view from the broader challenge from ASD. If desires such as these are to play the role of second-order volitions, it is not enough for them to have the right sort of content; they must also occupy the authoritative position that Frankfurt assigns to second-order volitions.

On Frankfurt's view, second-order volitions attain their authoritative role when the agent identifies wholeheartedly with them or when she is satisfied with them. So, this

[4] It is important to note that the repugnance argument only goes so far. Frankfurt's conception of a person does not entail that non-persons are not moral agents or that they lack moral rights. It merely entails that their will is not free in important ways and that, therefore, they are not morally responsible. However, to deny this of someone is still to deny her of something important, and to do so requires a great deal of justification. As I will make clear at the end of this chapter, it is not clear that Frankfurt has met this justificatory burden.

brings us to the second question posed above: can individuals with ASD engage in this form of identification? It is difficult to see how they could. Recall that wholehearted identification involves a decisive commitment to constitute oneself according to the desire in question. It is simply implausible to suppose that one can constitute oneself according to a first order desire like those that I have just described. So, perhaps Adam could arrive at the conclusion that no further accurate deliberation would lead him to give up his desire that cigarettes not be desirable*, but there seems to be no plausible story to tell about how this sort of commitment to a desire like this could *constitute* Adam in any way since Adam (or some essential feature of Adam) is not the object of the desire. More importantly, Adam's desire for cigarettes and his desire that cigarettes not be desirable* do not seem to pose any sort of agential conflict in the way that the first order desire and the second-order volition of Frankfurt's unwilling addict seem to do. The unwilling addict wants his will to be something other than it is, and it is the fact that it is *his will* which poses puzzles about his responsibility. In Adam's case, however, no such conflict arises. His desire to smoke is no more incompatible with his desire that cigarettes not be desirable* than it would be with a desire that cigarettes be less expensive, say.[5]

The problem here, then, is that desires like Adam's are not such that they lend themselves to either identification or satisfaction in Frankfurt's sense. There is an additional problem for Frankfurt, however, that is worth pointing out. Even if we suppose that individuals with ASD do have second-order volitions, it is not clear that they would

[5] In defense of Frankfurt, perhaps one might claim that certain rueful attitudes could help to show that individuals with ASD do, in fact, have second-order volitions. So, perhaps individuals with ASD might demonstrate second-order volitions by experiencing regret, or by wishing that particular objects were not desirable. Notice, though, that neither of these seems likely to occur given the pervasive counterfactual deficits in individuals with ASD. So, again, even if the argument from desire understanding fails, it seems as though there is still strong evidence that my counterfactual thinking hypothesis tells against Frankfurt's view on independent grounds. I'm grateful to David Shoemaker for pointing this out.

be able to identify with these in the robust sense required by Frankfurt's view. If identifying with one's desires involves making up one's mind, in Frankfurt's near-literal sense, then this would seem to require that desire be understood as reflecting something about the agent's mind rather than as a response to the property of desirability* in some object, but this seems to be precisely the understanding that is missing (or impaired) in ASD. So, it looks as though Frankfurt's account will be faced with a serious challenge *even if* we assume that individuals with ASD have second-order volitions, because it is questionable whether such volitions would have the necessary authority in ASD.

Thus, the evidence on desire understanding in ASD casts doubt on the claim that desires could ground the real self of autistic individuals. So, if real self views are to meet the challenge of correctly characterizing autistic agents, we will need to look to some other agential feature to do the grounding work.

§2. Watson's Dual System Account

In his response to Frankfurt, Watson suggests that an agent's values and their connection to her desires may be a better candidate for identifying the real self. Given his objections, canvassed above, to the hierarchical nature of Frankfurt's view, Watson offers a competing, dual-system view in which an agent is seen as being free in the sense required for moral responsibility only in cases where what the agent values and what the agent is most motivated to do are aligned. So, rather than understanding responsibility as stemming from a properly ordered hierarchical relationship between desires, Watson conceives of it as stemming from the fact that an agent's motivational system and her evaluative system are acting in harmony with one another. As he puts it, "If there are sources of motivation independent of the agent's values, then it is possible that

sometimes he is motivated to do things he does not deem worth doing. This possibility is the basis for the principal problem of free action: a person may be obstructed by his own will." (Watson 1975, 345) Watson defines these two systems of an agent as follows:

> *The valuation system* of an agent is that set of considerations which, when combined with his factual beliefs (and probability estimates), yields judgments of the form: the thing for me to do in these circumstances, all things considered, is *a* … *The motivational system* of an agent is that set of considerations which move him to action. We identify his motivational system by identifying what motivates him. The possibility of unfree action consists in the fact that an agent's valuational system and motivational system may not completely coincide. Those systems harmonize to the extent that what determines the agent's all-things-considered judgments also determines his actions. (347)

At various points in his paper Watson invokes a variety of features that he takes to be essential to values. One helpful view of what values consist in can be found in John Doris' recent work, and Doris' view is, I think, largely in line with Watson's.[6] Doris writes, "[V]alues are associated with desires that exhibit some degree of strength, duration, ultimacy, and non-fungibility, while playing a determinative-justificatory role in planning." (Doris 2015, 28) So, for both Watson and Doris, an agent is responsible to the extent that her actions reflect her values, which are to be understood in terms of these central features. In cases where her actions originate from something other than her values, she is not responsible.

So, if Watson's dual-system view of the real self is to accurately account for the responsibility of individuals with ASD, we need to know whether the motivational and

[6] It seems to me that most of the central features of Doris' definition of value can be found in Watson's view. For example, Watson writes, "In part, to value something is, in the appropriate circumstances, to want it…," (346) and this seems to captures Doris' appeal to desire. Watson also seems to include a non-fungibility criteria when he writes, "…we judge that to cease to have such appetites is to lose something of worth," (344) as well as a planning criteria, saying, "… we all have more or less long-term aims and normative principles that we are willing to defend. It is such things as these that are to be identified with our values." (346) Additionally, the fact that Watson seems to lean on the notion of judgments is also consistent with Doris' criteria of values playing a justificatory role in planning. So, Doris' definition of value seems to me to be mostly in line with what Watson defends.

valuational systems of autistic individuals typically interact in the way that Watson has in mind. If they do not, then an appeal to the distinction between these two systems will not result in an adequate characterization of the real self in autism. Given the empirical evidence on desire understanding, it seems to me that this is indeed the case.

Consider the role of desire in the views of Watson and Doris. For each of them, desires play a central role in giving us access to those things which agents value. In other words, desires with a sufficient degree of strength, duration, ultimacy, etc. serve as markers of value. When we see such desires, we can infer that some value is nearby. The only way that desires could do this is if they reflect something important about the mental life of the agent. Indeed, Watson thinks that this must be the case and writes, "Apparently the difference [between values and desires] will have to do with the agent's attitude towards the various things he is disposed to try to get." (Watson 1975, 346) The importance of the agent's attitudes toward her desires is central to Watson's account as it is what leads to the possibility of an agent's experiencing an "estranged desire," a desire which does not reflect the agent's values. As instances of estranged desires Watson offers examples of a mother who suddenly has a desire to drown her bawling infant in the bathtub or a man who believes that his sexual urges are the work of the devil. In each of these cases it is just false that the agent values what is desired – clearly the mother does not value drowning her baby, for example. In order for an agent to be estranged from her desires (and, thus, in order for the problem of free agency to exist) it must be the case that the agent can reflect on her desires *through the filter of her values*.[7]

[7] I'm grateful to David Shoemaker for this way of putting it.

Here, then, as in Frankfurt's view, we find a necessary condition of self-reflection in the exercise of free agency. In order to be free, the agent must be able to reflect on her desires in light of her evaluative judgments, and this opens up the possibility of her being obstructed by her own will. This requirement of self-reflection, however, makes Watson's view vulnerable to the same sort of objection that I posed for Frankfurt above. If individuals with ASD lack cognitive access to desires as mind-dependent representations, then self-reflection of this kind is ruled out. In other words, given the desire understanding impairments outlined above, reflecting on desire for individuals with ASD would amount to reflecting on a certain property of some object or action rather than reflecting on oneself. If this is right, then it would make little difference whether the desire conformed to one's values. Instead, the object would be seen as something that is desirable regardless of one's attitude toward it, and, so, the problem of estrangement simply wouldn't arise. The same issue posed for Frankfurt, therefore, comes up for Watson as well. Reflection on desire in ASD is simply not, or so it would seem, *self*-reflection. To put the point another way, if one lacks an adequate understanding of one's mental states, then one will not be able to reflect on those states in the way that Watson's and Frankfurt's views require. Whereas the evidence on desire understanding in ASD led Frankfurt to the repugnant conclusion that autistic individuals are not persons, it seems to lead Watson to the absurd conclusion that autistic persons *never* act freely.

§3. Toward a Conception of the Autistic Deep Self

The arguments to this point cast serious doubt on the plausibility of both Frankfurt's and Watson's general accounts of the real self. However, the objection that I

have posed against Watson raises an important problem that requires a solution. The account of the interaction between desires and values in ASD that I have been defending entails that the desires of individuals with ASD are not regulated by their values since such regulation would require a robust self-reflective ability. If this is true, though, and these sorts of desires are not regulated or suppressed, then we should expect to see autistic persons acting on desires like those Watson cites. That is, we should expect, e.g. to see autistic mothers drowning their bawling infants when the desire strikes them. Clearly, though, this is not the case. Autistic persons, though impaired in social interaction, generally display no tendencies toward violent or antisocial behavior. So, clearly these sorts of desires, if they exist, are being successfully regulated or suppressed, and an explanation of how this is accomplished is needed.

One way of explaining the ability to regulate or suppress these sorts of desires is by appealing to the evidence on inhibitory control in individuals with ASD. Recall that the experimental data outlined in chapter 2 seemed to show that autistic participants had impairments in their ability to inhibit *prepotent* responses (i.e. responses that the individual is strongly predisposed to have) but that they performed at control levels on all other inhibitory tasks. (Hill 2004a; Hill 2004b) So, one immediate explanation for why individuals with ASD do not act on desires like those that Watson describes is that the responses to such desires (e.g. drowning one's infant) are not prepotent responses, and, therefore, they do not override inhibitory control. Of course, this only pushes the problem back a step as we still need some explanation for why these responses are not prepotent, or, better, an explanation for why other response *are*, and the most natural and plausible explanation, it seems to me, arises from the intact capacity for affective empathy in

autistic persons as well as the somewhat atypical nature of moral judgment in ASD. It is likely that these produce strong prepotent responses such that aberrant desires are able to be suppressed.

So, as a first pass, perhaps we can say that individuals with ASD are able to regulate desires according to the extent to which those desires conform to standing emotional dispositions, specifically, in the moral case, dispositions characterized by prosocial attitudes toward others. Given that autistic persons are prone to emotional contagion (and, thus, able to be moved by the emotions of others), experience personal distress in response to others' negative emotions, and experience emotional concern for others' well-being, it is likely that they, like most moral agents, have strong prosocial prepotent responses in the majority of circumstances[8] (even if their ability to felicitously navigate social situations is compromised). In short, autistic people are clearly able to engage in caring relationships, and they are just as clearly able to feel a strong emotional concern for others. Therefore, it is likely these emotional dispositions result in prepotent responses that would be sufficient to suppress rogue desires like those with which Watson is concerned.

A second explanation for why autistic persons are able to inhibit desires such as these may come out of the atypical nature of moral judgments that I have argued is present in ASD. More specifically, I argued in the previous chapter that individuals with ASD rely more heavily than do neurotypical people on model-free judgments due to the fact that these do not require the use of models that must be supported by counterfactual abilities. Importantly, recall, model-free judgments arise due to habituation over time as a

[8] Indeed, this seems to be what is borne out in the literature on moral judgment as we saw in chapter 4.

result of positive and negative feedback. This fact is important because it allows judgments to proceed independently of any counterfactual or self-reflective considerations regarding desire satisfaction. So, it could be the case that, due to their habituation in the agent, model-free judgments produce strong prepotent responses that would be sufficient to override anomalous desires.

Thus, there are two plausible candidate features of the psychology of ASD that allow us to avoid the problematic implication of my objection to Watson, and the crucial feature of each of these is that neither requires that the agent be able to reflect on her desires in order for the action to be hers. Instead, all that would be required is that the desire be in line with the agent's standing emotional dispositions or with her habituated patterns of judgment. Notice, however, that what this gives us is the basis for an alternate conception of the real self, one that can accommodate the empirical data on ASD in a way that the views considered here could not. So, applying this to Watson's case of the mother who desires to drown her bawling infant, we can see clearly that her desire does not align with her standing emotional dispositions (e.g. to care for her child, to be concerned with her child's well-being, etc.), and, likewise, it does not, presumably, match up with her pattern of model-free judgment. However, had the desire been such that it could not have been suppressed, even by these strong prepotent responses, then it seems plausible to say that she would not have acted freely had she acted on that desire.

This conception of the real self is similar to, though importantly different from, the account proposed by David Shoemaker in his recent work.[9] He locates the real self in an agent's cares and commitments such that an agent is responsible for an attitude if it is

[9] See Shoemaker 2015a and Shoemaker 2015b.

"causally dependent on, and its content is harmonious with, at least one of the agent's cares, commitments, or care-commitment clusters." (Shoemaker 2015b, 59) For Shoemaker, cares are "dispositions to respond emotionally in sync with the fortunes of the cared-for object," (51) and commitments are simply the values that are demonstrated by the sum total of the agent's evaluative judgments (that is, the agent's evaluative stance) and, as a result, comprise the agent's evaluative stance. Both cares and commitments, on his view, are indicators of what *matters to* an agent, and, so, they mark something important about the deep self of the agent.

Clearly, then, Shoemaker's appeal to cares corresponds closely to the emotional dispositions that I have been discussing here. However, there is an important difference between an agent's commitments, as he conceives them, and the pattern of judgments to which I have appealed. The agent's pattern of judgments is broader, it seems to me, than Shoemaker's notion of commitments. What I have in mind is simply the judgments that the agent makes over time under relevantly similar conditions. This pattern need not reveal any commitments about the agent's conception of the good and, so, does not constitute an evaluative stance that the agent takes, but it does reveal something important and genuine about the way in which the agent typically governs herself across a wide range of circumstances. If an action or attitude is in line with this pattern, then it can be reasonably attributed to some feature of her psychology that is genuinely hers. Another crucial difference between the agent's patterns of judgment and her commitments, as Shoemaker conceives them, is that the agent's pattern of model-free judgments are modulated in important ways by her emotions. Shoemaker draws on Watson's view in describing commitments, and, so, these turn out to be fundamentally

rational. Importantly, though, model-free judgments are simply judgments that have been habituated via positive and negative feedback, and a central part of such feedback is the emotional response that it occasions in the agent who receives it. So, one important upshot of the preliminary view I am offering here is that the emotions play a necessary role in an action or attitude's being attributed to the agent whereas, on Shoemaker's account this need not be the case.

Admittedly, the notion of the real self just presented is inchoate, and it remains to be seen whether it can be generalized beyond the case of autism such that it can constitute a stand-alone account of the real self. Nevertheless, if a plausible account of the autistic real self is to be offered, it seems to me that it must draw on the features that I have discussed here. The agent's emotional dispositions, along with her established patterns of model-free judgment seem to be the only agential features that could support strong enough prepotent responses to preclude the agent from acting on aberrant desires such as those described by Watson. If this is the case, then it seems as though, contrary to Kennett's view (discussed in chapter 1), the most plausible route to moral agency for those with ASD runs through their intact emotional capacities and not through a rationalist moral concern as she suggests. So, even if the account offered here fails as a general conception of the real self, it will succeed in making headway on the problem that originally motivated this project, and this is a non-trivial accomplishment.

§4. Conclusion

In this chapter, I have presented the two most influential approaches to characterizing an agent's "real self" which claim that we can ground that self in the hierarchical relationship between an agent's desires or in the interaction between her

desires and values. Each of these approaches, I have argued, seem to imply, implausibly, that individuals with ASD are not moral agents (or, in Frankfurt's case, even persons). Therefore, each of them fails to give an adequate account of the moral responsibility of individuals with ASD. Really, this failure should not be surprising. The "self" is a difficult notion to pin down even in neurotypical agents, but in autism it is especially convoluted. Given the cognitive deficits described in Chapter 2 and the emotional irregularities (coupled, in many cases, with co-morbid alexithymia) outlined in Chapter 3, it is unclear whether any unified autistic "self" even exists. I have tried to offer an account of what the real self in ASD, if there is one, might consist in. However, even if this conception fails as a general account of the real self, it seems that we ought not to take this to mean that autistic individuals are not morally responsible given the intuitive, anecdotal evidence that individuals with ASD are capable of moral agency. One thing that the considerations offer here clearly show, I think, is that a real self view comes much closer to giving an adequate account of the responsibility of autistic persons than do the reasons-responsive views discussed in the previous chapter. However, it may be the case that some other theoretical approach to responsibility captures the evidence from ASD better, and in order to determine whether this is the case I will turn to a third type of theory, quality of will theories, in the next chapter in order to see whether they offer a better chance of getting things right with respect to autism.

§5. Appendix: Scanlon and Smith on Judgment Sensitivity

Before moving on to the next chapter, it is worth pausing to consider one final approach that could be construed as a real self view. Recently, Angela Smith has developed and defended a view of moral responsibility which places primary importance

on the notion that responsibility is a matter of an agent's actions or attitudes being connected in the appropriate way with her evaluative judgments. (Smith 2005; Smith 2008; Smith 2015) The theory, which she calls the Rational Relations view, draws on a similar view originally presented by T.M. Scanlon (1998) which holds that an action is attributable to an agent for the purpose of moral appraisal if and only if it is connected to the agent's judgment-sensitive attitudes. For each of these theorists, an action or attitude's being attributable to an agent entails that the agent may be appropriately called to *answer for* the action or attitude in question. Thus, a theory of answerability for attitudes and actions, they think, tells us everything we need to know about an agent's status with respect to moral responsibility.[10] In a sense, this sort of theory might be seen as falling somewhere between a reasons-responsive view and a real self view. On the one hand, judgment-sensitive attitudes are an important, deep feature of the agent, and the agent is seen as responsible in virtue of the fact that her attitudes and actions originate from this agential feature. In this way, the view has the makings of a real self account. On the other hand, though, answerability is seen fundamentally as a function of the agent's ability to cite the reasons for which she acted, and, so, there is an obvious sense in which this sort of account contains elements of a reasons-responsive view.

Before turning to Smith's Rational Relations view, I will take some time to summarize some of the basic features of the Scanlonian account on which Smith draws. For Scanlon, moral responsibility is a function of an agent's self-governance and its

[10] Describing this sort of view as an answerability theory is intended to help distinguish it from theories of *attributability* such as Frankfurt's. Frankfurt claims that all that is needed for an agent to be morally responsible is for an action to be attributable to the agent's real self. I will use this terminology in the rest of this section since it is how both Scanlon and Smith refer to their views.

relation to her judgment-sensitive attitudes. Because of this, moral criticism is aimed at critiquing this governance and calling the agent to answer for it. Thus, he writes,

> [M]oral criticism claims that an agent has governed herself in a way that would not be allowed by any principles that no one could reasonably reject. When addressed to the person in question as a fellow participant in a system of co-deliberation, this charge calls for her to explain why this claim is mistaken or to acknowledge that it is valid and that her self-governance has been faulty. (Scanlon 1998, 268)

However, there may be cases in which such a call to answer for one's actions may be seen as unintelligible. If, for example, I were to utter a particularly offensive remark it would be inappropriate for you to call on me to answer for if it turned out that I uttered the remark under post-hypnotic suggestion. In such cases, Scanlon suggests, moral appraisals are inappropriate because these cases "break the usual connection between the action and the judgment-sensitive attitudes of the agent." (277) If such an action is not under the control of the agent's judgment-sensitive attitudes, then "it is not *his* act in the sense required for moral appraisal to make sense." (277) Thus, in order for an agent to be answerable for her actions they must be attributable to her in the sense required for moral appraisal, and this requires that they be connected in the right sort of way to her judgment-sensitive attitudes. These attitudes, Scanlon claims, are "attitudes that an ideally rational person would come to have whenever that person judged there to be sufficient reasons for them and that would, in an ideally rational person, 'extinguish' when that person judged them not to be supported by reasons of the appropriate kind." (20)

These basic features of the Scanlonian view are extended by Smith into the realm of responsibility for attitudes. It is often appropriate, she argues, to hold people to be responsible for attitudes which seem beyond their control (things that we notice or fail to

notice, things that occur to us, spontaneous affective responses, etc.). What explains the appropriateness of maintaining that an individual is responsible for such attitudes is the connection of these attitudes to the agent's evaluative judgments. Thus she writes, "When we praise or criticize someone for an attitude … it seems we are responding to certain judgments of the person which we take to be implicit in that attitude, judgments for which we consider her to be directly morally answerable." (Smith 2005, 251) So, in much the same way that our actions are *ours* by virtue of their being attributable to us, our attitudes are *ours* insofar as they bear the same sort of rational relation to our judgments.[11]

In short, on the Scanlon/Smith view of responsibility, an agent is responsible for her actions or attitudes only to the extent that these attitudes and actions reflect the agent's judgment-sensitive attitudes in such a way that the agent could be, in principle, intelligibly called to answer for them. Attributability and answerability, then, turn out to be coextensive and can tell us everything we need to know about an agent's status as morally responsible or not. If the agent's actions and attitudes are governable by her judgments in the right sort of way, then she is responsible for them. If they are not, then she is not.

David Shoemaker has recently leveled an important challenge against the view that attributability and answerability are both coextensive and able to offer a unified account of moral responsibility and its attendant practices. (Shoemaker 2011) In doing so, he has aimed to show that there are cases in which an action or attitude may be

[11] By "judgments," Smith has in mind something much more expansive than consciously held judgments.

attributable to an agent but that it may nevertheless be unintelligible to claim that she is answerable for it. He offers two types of objections on this point.

The first type of objection relies on an example originally offered by Smith. Suppose, so the example goes, that an individual is afraid of spiders but nevertheless claims sincerely to believe that spiders are not dangerous. In such a case, the individual is seen as holding two conflicting attitudes. On Smith's view, this agent is seen as irrational insofar as these attitudes, i.e. fear and the belief that spiders are not dangerous, are each judgment-sensitive. However, Shoemaker suggests that the Rational Relations view, though it can conclude that the agent is answerable for each of these conflicting attitudes, cannot hold that the agent is answerable for her being in a state of irrationality. Insofar as the irrationality itself is not connected to any judgment-sensitive attitude, it cannot be the case that the agent may be called to answer for it. However, Shoemaker claims, it seems perfectly reasonable to call on an agent to answer for his or her irrationality, so a view which cannot account for this is implausible in an important way.

Shoemaker's second challenge on this point comes by way of an appeal to typical human emotional commitments. Such commitments fall under the domain of "judgment-sensitive" on the Scanlon/Smith view, but this is highly problematic. As Shoemaker puts it,

> Some cares in particular are notoriously independent of reason, for example, our cares for our children or other loved ones. After my child has become a serial killer, for instance, I may arrive at the consciously held propositional belief that he's a worthless human being, that he's dead to me. And yet when I read of his upcoming execution, I may well up with tears or fall into a depression. 'I still care about him,' I may say. 'There are no reasons to do so – he's an awful man – but it still matters to me what happens to him.' (Shoemaker 2011, 610)

What cases like these suggest is that, like the case of irrationality, these cares may be attributable to one, but it may nevertheless be unintelligible to call one to answer for them. Such cares are non-rational, and, as a result, answerability simply does not enter the picture.

In summary, then, Shoemaker's challenge seems to show that answerability and attributability can, in some cases, come apart. An action may be attributable to an agent without it being the case that the agent is answerable for it. If this is true, then answerability as the unifying feature of moral responsibility seems to come up short. However, it seems to me that the problem runs much deeper than this and that the evidence from ASD suggests that these two conceptions of responsibility may come apart in a different way as well, one which calls into question the Scanlon/Smith view of attributability all together.

The basic challenge issued by Shoemaker can be extended in a way that is far more problematic for the Rational Relations view when considered in light of empirical data on alexithymia in ASD. Recall, from Chapter 3, that alexithymia is a set of characteristics which reflect deficits in the cognitive processing of emotions and emotion regulation, and it is commonly seen as being comprised by the following features: "1) difficulty identifying and describing subjective feelings; 2) difficulty distinguishing between feelings and the bodily sensations of emotional arousal; 3) constricted imaginal capacities, as evidenced by a paucity of fantasies; and 4) an externally oriented cognitive style." (Taylor 2000) For the purposes of the present argument, I will be focusing primarily on the inability to identify and distinguish emotions that is experienced by those with alexithymia, i.e. (1) and (2) above.

It seems to me that these are the features which bear directly on an agent's judgment-sensitive attitudes and their relation to the agent. To see why such deficits might be important, consider a revised version of the fear-of-spiders example employed by Smith and described above.

> **Fear of Spiders 2.0:** Suppose that Jones is an individual with an extremely high degree of alexithymia. Jones sincerely believes that spiders are not dangerous, but upon seeing a spider he recoils from it. His heart rate increases, he begins breathing heavily, and he starts to sweat. These responses, he notes, are connected to some general negative affective feeling that he cannot quite discern.

What should we make of Jones' response in this revised example? There are several features of the case that must be pointed out. First, the affective response that Jones is having is, in fact, fear. He is experiencing all of the typical physiological responses that comprise fear, but he is unable to distinguish these from the emotion itself and unable to recognize the emotion as an instance of fear. Second, since Jones is experiencing fear and he believes that spiders are not dangerous, he is simultaneously holding two conflicting judgment-sensitive attitudes. Therefore, Jones is subject to the charge of irrationality in the same way that the agent in Smith's original example would be.

If the Rational Relations view is correct, then it ought to be appropriate for us to call on Jones to answer for his attitudes. However, it seems wholly inappropriate to maintain that he is answerable for his fearful response when he *does not even recognize that he is afraid*. That is, it would be entirely unintelligible to him if we were to demand that he give us an explanation for the emotion that he is feeling because he is incapable of identifying the emotion in question. Nevertheless, the fear certainly does seem to be attributable to him. It is *his* fear, after all. So, this seems to be another case in which

answerability and attributability come apart. The important task, however, is to explain the nature of their severance in this case.[12, 13]

Shoemaker's aim in objecting to the coextension of attributability and answerability was to show that attributability may be a function of something other than an agent's evaluative judgments, specifically, certain non-cognitive features of an agent (e.g. cares). However, the present objection says more than this. My claim is not simply that attributability may depend upon something other than an agent's judgment-sensitive attitudes; it is the stronger claim that appealing to judgment-sensitive attitudes is not in itself sufficient to establish an agent's answerability. Jones' response, in my revised case, arises as a result of his evaluative judgments – or, to use Scanlon's language, it is connected to his judgment-sensitive attitudes. However, what is missing is an appropriate connection between Jones' judgment-sensitive attitudes *and Jones himself*. On the Rational Relations view, an agent's actions are not attributable to her if it is the case that they are divorced from her judgment-sensitive attitudes. So, Smith is positing a two-place relation between actions and judgment-sensitive attitudes. However, the relevant relation

[12] Readers will notice that this objection proceeds in the opposite direction as the objections argued for in the previous two sections. In those sections I began with the intuitive judgment that individuals with ASD are responsible agents and criticized the desire- and value-based views for failing to explain this intuition. Here, however, I am offering a case where autistic agents are intuitively not responsible and criticizing Scanlon/Smith for offering a view which says that they are. Because of this, it might seem as though I'm trying to have my cake and eat it too, but this is not the case. Rather, it is the difference in the respective views which licenses this difference in methodology. The desire- and value-based views are concerned to show that responsibility is a function of a certain *structural* feature of agents whereas the Rational Relations view aims to show that responsibility is a function of the judgments that an agent makes. Thus, it is open to me to object on the basis of intuitions about particular instances of judgment-sensitivity rather than objecting on the basis of agential capacity. Recall that this is how the objection to reasons-responsive views proceeded in the last chapter. Given the similarity of those views to the Scanlon/Smith view (see fn. 21 above) the methodology is, to an extent, mirrored here.

[13] Smith might respond here that the fear that is being experienced in Jones' case is something closer to a phobia, which on her account would be both non-rational and non-attributable to Jones. However, this response seems to me implausible. Jones' fear in this case is not disordered to the point of being non-rational as in the case of a phobia. The fear itself is identical to the fear that any neurotypical agent might feel in Jones' case. The only difference is that Jones can't identify what it is that he's feeling.

is one which contains three terms: actions, judgment-sensitive attitudes, and *the agent*. The presence of alexithymia in ASD seems to suggest that this relation can be compromised by a disconnect between agent and attitude and that there may be actual individuals whose judgment-sensitive attitudes are divorced, in a very real way, from the individual herself. The presence of such individuals seems to constitute powerful evidence against a view which places fundamental importance in determining one's standing as a responsible agent on the connection of actions to evaluative judgments.

Chapter 6: ASD and Quality of Will

In this chapter, we turn to the final approach to moral responsibility commonly pursued in the literature. As we saw in Chapter 4, reasons-responsive views claim that an agent's ability to recognize and act on moral reasons is the central criterion for her being morally responsible. In Chapter 5 we discussed real self theories of responsibility, which claim that agents are responsible for those actions which are attributable to their real selves. Each of these approaches, I argued, has considerable difficulty accounting for the empirical evidence on autism, and in each case counterexamples to representative views from these camps can be readily produced as a result. The goal in this chapter, then, is to determine whether or not a third type of theory, quality of will theory, fares any better in accounting for individuals with ASD. Though there is considerable variation among the views on offer, the basic claim made by quality of will theorists is that an agent is responsible only for those actions or attitudes that reflect either good will, ill will, or insufficiently good will on her part.

The variation among quality of will views arises from the fact that theorists disagree about how best to specify precisely what an agent's will consists in. There are those who claim that the quality of an agent's will is a function of the quality of her judgmental capacities which make her answerable for acting on certain reasons (this is Scanlon's and Smith's construal). There are also some who claim that the quality of an

agent's will is a function of the quality of her character as revealed by her values or her desires or by some other feature of herself as an agent (Frankfurt and Watson take this sort of view). These understandings of quality of will result in views which correspond to what I have called real self views. Having already discussed these approaches in the previous chapter, I will focus here on a set of views that are distinctive due to their construing quality of will terms of the *regard* for others that is revealed in an agent's actions.[1] These views are set apart from those already discussed insofar as they hold that to be a morally responsible agent one must have the capacity to take up a particular type of attitude or stance toward other agents which makes one the fitting target of the reactive attitudes.

I will focus on two representatives of such a view in the remainder of this chapter each of which holds that the capacity for having quality of regard requires the taking up of a particular stance toward other agents. The first, advanced by Michael McKenna, claims that in order to be a responsible agent one must be able to take a communicative stance toward other responsible agents and that this stance reveals one's quality of regard by allowing one to engage in a "moral responsibility exchange" with others. The second, defended by David Shoemaker, holds that in order to be a responsible agent[2] one must be able to take up a robust empathic stance toward other agents and that this stance enables one to have a particular quality of regard. Each of these views takes its cue from P.F. Strawson's seminal essay, "Freedom and Resentment," and, given the significance of

[1] The terminology for these distinctions in quality of will theories is drawn from Shoemaker 2015b.

[2] It should be noted here that Shoemaker's view is a pluralistic one which holds that there are three distinct types of moral responsibility. My remarks in this chapter apply only to Shoemaker's discussion of accountability-responsibility.

Strawson's basic insight and his influence on quality of will theories more generally, I will begin with a discussion of his work in section 1. Then, I will offer fuller exposition (as well as criticism) of McKenna's and Shoemaker's view in sections 2 and 3, respectively. Finally, in section 4, I will attempt to show that, with some modification, a version of Shoemaker's view might be able to give an adequate account of the responsibility of individuals with ASD.[3]

§1. Strawson and the Reactive Attitudes

Strawson's aim in "Freedom and Resentment" was to offer a way toward a resolution of the longstanding debate over the compatibility of our moral responsibility practices with the truth of determinism. In doing so, however, he presents the groundwork for quality of will theories of responsibility. He begins with the commonplace observation of "the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions." (Strawson 1962, 48)

By placing the focus on the types of attitudes and intentions that exist within our interpersonal relationships, Strawson's arguments served to locate *responsibility* within these as well. He writes,

[3] There are many theorists who classify their views as "quality of will" views. For example, Scanlon takes his view (presented in the appendix of chapter 5) to be a quality of will view. Additionally Nomy Arpaly, for example, in her book, *Unprincipled Virtue* (2006) makes reference to the quality of the agent's will and the role that it plays in moral responsibility as a function of the agent's responsiveness to reasons. Many others take the quality of an agent's will to be important for responsibility as well (see e.g. Vargas 2015), but I have chosen to focus on McKenna's and Shoemaker's views here for two reasons. First, as I noted in the text above, there is a great deal of variation in what theorists take "quality of will" to refer to, and offering a comprehensive classification of these would, I think, detract from the argument of this project. Second, McKenna's and Shoemaker's views share the important characteristic of construing regard as being revealed or enabled by the agent's ability to take up a unique stance or set of attitudes toward others, and, in my opinion, considering this sort of capacity and its presence or absence in autism will help to reveal something important both about the nature of responsibility and about individuals with ASD.

> We should think of the many different kinds of relationship which we can have with other people … Then we should think … of the kind of importance we attach to the attitudes and intentions towards us of those who stand in these relationships to us, and of the kinds of *reactive* attitudes and feelings to which we ourselves are prone. In general, we demand some degree of goodwill or regard on the part of those who stand in these relationships to us, though the forms we require it to take vary widely in different connections. (Strawson 1962, 49-50)

Thus, for Strawson, the demand for goodwill is of central importance, and being responsible consists in being the appropriate target of the reactive attitudes when one violates this demand. This feature of Strawson's view is a revolutionary shift in the literature on moral responsibility as it construes responsibility as a response-dependent phenomenon.[4] That is, on his view, to be a responsible agent *just is* to be the appropriate target of the reactive attitudes, and these attitudes (which may include "such things as gratitude, resentment, forgiveness, love, and hurt feelings") (Strawson 1962, 48) are "essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in *their* attitudes and actions." (53)

The notion of appropriateness here is important as it leads Strawson to consider two types of cases in which it would be inappropriate for us to hold reactive attitudes such as resentment or indignation. The first type of case deals with actions that can be characterized as excused or justified in some respect. According to Strawson, these sorts of cases leave the agent with the option to use such pleas as, 'He didn't mean to', 'He didn't know', or 'They left him no alternative' (50). Strawson writes,

> None of [these pleas] invites us to suspend towards the agent, either at the time of his action or in general, our ordinary reactive attitudes. They do not invite us to view the *agent* as one in respect of whom these attitudes are in any way inappropriate. They invite us to view the *injury* as one in respect of which a

[4] I am borrowing the phrase "response-dependent" from Watson 2014. Watson's is an insightful discussion of Strawson's view from which I learned a great deal.

> particular one of these attitudes is inappropriate. They do not invite us to see the *agent* as other than a fully responsible agent. (50)

The more interesting cases – i.e. those that point toward what it means to be a moral agent – are those latter cases that show an *agent* to be an inappropriate target of the morally reactive attitudes. These cases prompt pleas such as 'He wasn't himself', 'He's only a child,' or 'He's a hopeless schizophrenic'. (51) These, Strawson says, invite us to suspend our reactive attitudes toward the agent in all cases, rather than doing so only in cases where the circumstances of an injury provide the agent with an excuse. They allow us to look on an agent with an objectivity of attitude and to view the agent as something to be managed or as an object of social policy rather than one toward whom we can direct our participant reactive attitudes. To adopt the objective attitude toward one, then, is to withdraw from personal relations with her. Those who can legitimately offer pleas of the second type lack, as Gary Watson puts the point, "the capacity to have and respond to normative expectations and hence the recognition of something like the general demand for good will and the like." (Watson 2014, 21)

Central to both pleas, according to Strawson, is the extent to which the attitudes or actions of the agent reflect good or ill will. With this groundwork in place, I will now turn to a discussion of two contemporary versions of this sort of theory, both of which rely heavily on Strawson's construal.

§2. McKenna's Conversational Theory[5]

There are several important features of the conversational theory that McKenna ably defends. Understanding these features and how they fit together will be the focus of this section. A complete analysis of the theory would require far more attention than I can

[5] The material in this section is drawn from Stout (forthcoming).

give it here, and so, my goal in this section will simply be to give the reader a general sense of the theory where a detailed treatment is not necessary.

As I see it, McKenna's view consists of three theses. The first is a metaphysical claim about the relationship between being responsible and holding responsible. The second is a Strawsonian quality of will thesis, and the third is the claim that we should understand responsible agency as being modeled on an analogy with a linguistic conversation. I will discuss each of these in turn.

2.1: The Modest Metaphysical Thesis

McKenna's first substantive claim is one about the relationship between being responsible and holding responsible, and in defending it, he takes up the debate between those who, on the one hand, maintain that holding others morally responsible is metaphysically more basic than being responsible – the normative interpretation – and those who, on the other hand, hold that there are facts about moral responsibility which are metaphysically prior to facts about holding responsible – the extreme metaphysical interpretation. Those who take the normative interpretation claim that our practices of holding others responsible take explanatory priority over the facts that obtain with respect to any given agent and are metaphysically more basic in grounding morally responsible agency. So, those who take this broadly Strawsonian view of responsibility claim that the facts about the appropriateness of *holding* responsible and the norms that govern the practices of doing so serve to fix the relevant facts about *being* morally responsible. Thus, the norms involved in holding others responsible – or the conditions that make it appropriate to hold others responsible – are metaphysically more basic than any independent facts about responsible agents.

In contrast with this, those who endorse what McKenna refers to as the "extreme metaphysical interpretation" take the converse of the above approach. The extreme metaphysical interpretation says that there is a set of facts about responsible agency, and these facts are what determine our practices of holding one another responsible. Or, to put it another way, our practices of holding others responsible simply track the facts of responsible agency. On this view, then, being responsible is wholly independent of the norms which govern our practices of holding others responsible. In other words, being responsible is metaphysically more basic than both the practices involving holding responsible and the norms which govern such practices. As one proponent of this view, Michael Zimmerman, writes, "Someone is blameworthy … if it is correct, or true to the facts, to judge that there is a 'debit' in his 'ledger' (etc.). It is important to note that, in the context of inward moral praise or blame, worthiness of such praise or blame is a strictly nonmoral type of worthiness; it is a matter of the truth or accuracy of judgments." (Zimmerman 1988, 38, as quoted in McKenna 2012, 44)

Rather than taking sides in this debate, McKenna attempts to argue for a compromise position which he calls the "modest metaphysical interpretation." He claims that, "Holding responsible should in the first place answer to the facts about what it is to be responsible. Yet … I contend … that being morally responsible presupposes considerations employed from the standpoint of holding responsible." (McKenna 2012, 50) In other words, McKenna sees the facts about an agent's being responsible and the practices of holding her responsible as interdependent. As a result, he bases his theory around the notion that being responsible and holding responsible are fundamentally integrated with one another. Or, as McKenna puts it, "the metaphysical thesis proposed is

that there is an irreducible relation of interdependence between being and (pertinent norms pertaining to) holding responsible." (54) So, according to the modest metaphysical interpretation neither being responsible nor holding responsible is metaphysically prior to the other. Instead, McKenna proposes that we understand being responsible by reference to the quality of an agent's will and that we then understand the nature of quality of will in virtue of the capacities that an agent might have relative to a kind of moral responsibility conversation that is analogous to a successful linguistic conversation between competent speakers.

2.2: Quality of Will

In order to better understand McKenna's modest metaphysical interpretation, it will be important to understand the role that quality of will plays in his work. He proposes introducing a quality of will thesis which states, "Being morally responsible for an action is to be settled in terms of the moral quality of the will with which an agent acts." (58) However, given his modest metaphysical interpretation, this is to be partially understood with respect to the practices of holding responsible. Thus, he writes, "the moral quality of an agent's will, as well as a person's standing as a morally responsible agent, is dependent upon that agent's appreciation of the expectations of due regard for others as revealed in the practices of those holding responsible." (58) Given this, McKenna attempts to reorient the classical Strawsonian position around this quality of will thesis so that being responsible is a function of one's acting with a certain quality of will and being morally blameworthy is dependent upon how that will is revealed in an agent's actions. He then gives an account of the reactive moral emotions that is sensitive to this reorientation. On his view, the morally reactive emotions are responses to the

quality of an agent's will when she acts, and this means that they are fundamentally cognitive insofar as they are have as their propositional object that one has acted with an objectionable quality of will. Moreover, McKenna argues that the reactive attitudes are best understood in terms of their overt manifestation rather than any private experience of them. This means that in order for an individual to understand the reactive emotion that he or she experiences, he or she must understand what sorts of practices would be fitting manifestations of the experienced emotion.

So, on McKenna's view, what is of primary importance for determinations of responsible agency is a fact about the agent in question, namely the quality of the agent's will. Because of this, reactive attitudes are responses to this quality of will which are to be understood in cognitive terms and by reference to their public manifestations. It follows from this, McKenna argues, that agents who are exempted from being objects of the reactive emotions are those who cannot understand the practices of holding responsible when these practices are directed toward the agent in response to her being perceived as evincing an objectionable quality of will. Moreover, if the agent cannot understand these practices, then she cannot engage in them either, and this notion leads naturally to a discussion of the second portion of McKenna's modest metaphysical interpretation to which I will now turn.

2.3: The Moral Responsibility Exchange

On McKenna's view, the ability to hold others responsible is necessary in order for an agent to be responsible. In justifying this thesis, McKenna appeals to Paul Russell's notion of "moral sense." For Russell, the condition of moral sense is just the ability to "feel and understand moral sentiments or reactive attitudes," (Russell 2004,

293) and, he argues, it is a necessary condition for morally responsible agency. “Agents who lack moral sense,” Russell writes, “are missing something that is vitally important and that ‘normal’ agents do and must possess. More specifically, when an agent lacks moral sense, her *sensitivity* to moral considerations is diminished and her motivation to be guided by these considerations is impoverished and limited.” (294) Taking up this line of thought, McKenna claims that agents who are unable to hold other agents responsible by feeling and understanding the moral emotions lose access to a specific type of reason. One example of the reasons he has in mind are second-personal reasons like those proposed by Stephen Darwall. Of these, Darwall writes, “A second-personal reason is one whose validity depends on presupposed authority and accountability relations between persons and, therefore, on the possibility of the reason’s being addressed person-to-person.” (Darwall 2006, 8) In further discussing the second-personal nature of responsibility, Darwall says, “I claim that reactive attitudes are always implicitly second-personal and that they therefore invariably carry presuppositions of second-personal address about the competence and authority of the individuals who are their targets, as well as about those who have them.” (67) So, by making use of Russell’s notion of moral sense, and Darwall’s treatment of the second-personal features of responsible agency, McKenna argues that being responsible is dependent on an agent’s ability to hold others responsible because without this ability an agent would lack the capacity to recognize and respond to second personal reasons as these reasons require an ability to hold others to second-personal demands. To this end, McKenna writes,

> If Darwall is correct about the very existence, nature, and pervasiveness of these reasons, and to my mind he is, we have a straightforward source of support for my claim about moral responsibility’s dependence on the nature of holding morally responsible. The practices by which others hold one morally responsible are

> themselves expressions of demands that as a competent agent one must be able to grasp and treat as reasons that apply to one. In the absence of this ability, a person would be unable to recognize and respond to a vast array of reasons presented to morally responsible agents. (McKenna 2012, 84)

Given all of the above, McKenna proposes a theory of moral responsibility which gives full treatment to the communicative nature of our moral responsibility practices. In doing so, he builds a theory which presents responsibility as modeled on an analogy with a linguistic conversation saying,

> By acting as she does, the morally responsible agent *opens up the possibility of a conversation* about the moral value of her action, and most notably, what it reveals about the quality of her will. This initial contribution provides a basis for members of a moral community responding to her by holding her morally responsible, thereby engaging in a dialogue with her. Given the unfolding conversation, it is now the agent's place – her conversational role – to extend the conversation by offering some account of her conduct either by appeal to some excusing or justifying consideration or instead by way of an acknowledgment of a wrong done, perhaps an apology offered. (88-89)

So, there are three stages to the conversation, or, as he calls it, the "moral responsibility exchange." The first is the stage of moral contribution in which the responsible agent makes an opening salvo, followed by the stage of moral address in which her peers hold her responsible, and, finally, the stage of moral account in which the responsible agent gives some account of her actions.

In order to add legitimacy to the conversational model, McKenna proposes a further feature of the analogy. One obvious and important aspect of a linguistic conversation is that the speaker who makes an utterance typically intends for that utterance to be understood by her audience, that is, she intends her utterance to have *meaning*. In the same way, McKenna claims, "a morally responsible agent acts with the knowledge that her conduct is always a *potential* object of interpretation by members of the (or a) moral community. She therefore is able to understand her own actions as

having meaning." (94) And that meaning, on this view, is found in what is revealed about the agent's quality of will through her actions. In summary, McKenna offers the following description of the theory he is proposing:

> [W]hen a morally responsible agent acts, she understands the standpoint of holding morally responsible, and so understands that others in a moral community are liable to take her actions as indicative of the quality of her will. In acting, she must be able to appreciate how her conduct could be interpreted as reflective of the quality of will with which she does act. And she must be able to adjust her behavior in ways sensitive to these considerations – or so I have argued. This is irrespective of whether as a matter of fact she allows those considerations to have any bearing upon her intentions or whether these matters even occur to her when she does act. *A morally responsible agent must be able to act as if her actions were morally significant bearers of meaning among communicating agents. And because of this, her actions thereby take on meaning merely by virtue of her being such an agent.* (99, emphasis added)

In short, McKenna's view is characterized by three basic theses. It is one in which being responsible and holding responsible are held to be interdependent rather than either having metaphysical priority. It is oriented around a quality of will thesis so that any judgment of responsibility must be made in reference to the quality of an agent's will. Most importantly it highlights the expressive nature of responsibility and its attendant practices by assigning a great deal of importance to the ability of an agent to take part in the moral responsibility exchange. Having understood this, we are now in a position to understand the challenge that autism spectrum disorder and its distinctive moral psychology poses for such a theory.

§2.4. The Challenge from ASD

The challenge that the evidence from ASD poses for McKenna's view is this: it may be the case that some individuals with ASD can engage in McKenna's responsibility exchange in exactly the way that his view requires yet are not responsible despite this ability. The reason for this possibility is that it is not difficult to imagine an individual

with ASD who engages in the responsibility exchange in a purely functional manner. To see how this may be and why it would be problematic for McKenna, let's return to the case of Adam presented at the beginning of Part II. Recall that Adam was meant to be the paradigm case of an autistic person who is, at least intuitively, morally responsible and that he was largely typical in terms of his emotional profile. For the purposes of evaluating McKenna's account, however, let's stipulate instead that Adam has a high degree of alexithymia. Call this person Adam*, and suppose that his alexithymia is such that he generally does not understand his own emotional responses to others in the context of morally charged scenarios. He has strong emotional responses in these cases, but he simply cannot make sense of them. Nevertheless, he still has a complex and highly effective set of compensatory strategies and, as a result, is able to navigate social situations quite well but only in virtue of his ability to "go through the motions," so to speak. Now, according to McKenna's conversational theory, the moral responsibility exchange is underpinned by both the modest metaphysical interpretation of the relationship between being responsible and holding responsible as well as a broadly Strawsonian quality of will thesis. This being the case, I would now like to examine whether or not someone like Adam* can rightly be said to engage in the exchange that McKenna describes.

To begin, consider the passage from McKenna quoted above. He begins by saying, "when a morally responsible agent acts, she understands the standpoint of holding morally responsible…" (McKenna 2012, 99) Can this understanding be attributed to Adam*? It certainly seems so. From the fact that Adam* is able to make reliable moral judgments and distinguish them from judgments of convention, it seems to follow that he

is able to recognize when others have violated moral norms. Moreover, what seems to be suggested by the empirical literature on autism is that the capacity to make reliable moral judgments stems from the capacity to experience emotional empathy, a capacity which, I have stipulated, Adam* possesses. If this is true, then it would suggest that Adam* is, as Russell requires, able to "feel and understand moral sentiments or reactive attitudes." That is, his ability to have an affective response to others suggests that Adam* is in fact capable of experiencing at least some of the reactive attitudes.[6] Now, it might be objected at this point, simply experiencing an affective response is not sufficient for holding morally responsible and instead the agent must be feeling a particular reactive emotion. This strikes me as entirely correct, and I will return to this objection below.

In addition to being able to hold others responsible, McKenna claims that, "A morally responsible agent must be able to *act as if* her actions were morally significant bearers of meaning among communicating agents." (99) (Recall also, that agent meaning, on McKenna's account, involves reference to the quality of the agent's will.) Can this ability be attributed to someone like Adam*? Some might claim that the answer is "no," and that this is where McKenna's theory is able to account for autistic individuals. The reason for this is simple. Being able to express meaning, it might be argued, requires the ability to understand how others will interpret one's actions, which requires the ability to represent the mental states of others, and this is precisely the ability that autistic individuals lack. However, I want to resist this characterization. More specifically, I want

[6] I've left it an open question whether or not his experience of these emotions is in any way standard or whether he experiences reactive emotions in a way that is phenomenologically like the way in which neurotypical individuals experience them. However, on a view like McKenna's which is concerned with a primarily functional capacity (as I argue his is), it may not be important that autistic individuals experience the same emotions as the rest of us, so long as they understand the behavioral manifestations that their emotions call for in a given context. I return to this below.

to deny that an understanding of another individual's mental state is necessary for one to act as if one's actions are bearers of meaning.

I have cited some recent research in previous chapters which suggests that autistic individuals are capable, to a certain degree, of learning how to engage in felicitous social interaction. That is, they are capable of recognizing certain features of others which will allow them to successfully engage in social transactions, and I have ascribed this ability to our fictional agent, Adam*, as well. This, so far as we know, does not constitute learning cognitive empathy. Rather, it seems to be a type of functional learning which eases social facility. I have also pointed to some evidence which suggests that these individuals are capable, either consciously or unconsciously, of developing compensatory methods that lead to a greater degree of social awareness and appropriate social interaction. The upshot of these facts for the present discussion is that such individuals seem to be perfect examples of agents who can perform meaningful actions but who nevertheless lack an important capacity that is directly related to responsibility. This is partially because the compensatory system that someone like Adam* develops is a purely functional system. That is, it is a system which allows him to engage successfully and prudentially with the outside world, but it is not such that it negates the incapacities associated with ASD, nor is it such that it can succeed in all morally relevant situations. If it were, then we might be tempted to say that autistic individuals who have developed compensatory strategies are fully responsible. However, this strikes me as the wrong conclusion to draw for reasons that I will make clear presently.

Astute readers will notice a tension in what I have so far written. I have claimed that the ability of autistic individuals to make the moral/conventional distinction along

with their capacity to have emotional responses to the affective states of others suggests that these individuals are able to hold others responsible. However, I have also just claimed that the compensatory heuristics developed by high-functioning autistic individuals are such that they allow for the possibility of engaging in a moral conversation like the one McKenna has in mind. If the former is true, then it may imply that the actions of autistic individuals do not actually implicate their quality of will because the conversational theorist may simply argue that having an emotional response and recognizing moral transgressions – even jointly – do not constitute understanding the standpoint of holding responsible. On the other hand, if the latter claim is true, then it may imply that autistic individuals are indeed responsible and that their autism is simply a mitigating factor in their responsibility.[7] In short, it may be that I am either describing a class of agents who lack precisely those capacities that McKenna identifies or that I am describing a class of people who are responsible agents, but whose disorder may, in some cases, provide them an excuse. That being said, I take myself to be doing neither of these. Rather, what I am after is a description of an agent who occupies a middle ground. That is, I am looking for a case of someone who is functional in such a way that her quality of will is implicated in her actions but lacks, at least intuitively, vital capacities for responsibility – someone who is *merely* functional in this way.

To answer these objections, we must consider how McKenna proposes that we understand the nature of emotions – by first understanding their fitting behavioral manifestations. He writes, "I want to resist [the] picture of the emotions … in which private episodes and subjective states are more fundamental. In my estimation … the

[7] I'm grateful to both David Shoemaker and Michael McKenna for raising this worry.

order of conceptual priority is reversed … What makes a private episode of, say, resentment intelligible to its subject *as* resentment is precisely an appreciation of the criteria indicators that would be manifested in a public display of *that* emotion." (McKenna 2012, 69) To see why someone like Adam* might prove difficult for McKenna's view to explain in this regard, consider the response that Adam* might have to being held responsible by a peer. I have stipulated that Adam* is alexithymic, so let us suppose that upon being held to account he experiences a strong emotional response (as a result of his capacity for emotional empathy), but he simply cannot identify the emotion that he feels. However, Adam* has been able to compensate for his emotional experiences by tracking past outward manifestations of them (to help make sense of this, we might think of Adam* as making model-free judgments rather than model-based judgments as I suggested in chapter 4), and these outward manifestations are precisely what McKenna sees as important, at least insofar as they are typically associated with the relevant internal states. My contention, then, is that it is plausible to hold that someone with Adam*'s moral psychology could make sense of the appropriate social practices governing a situation and not only understand but also *identify* the emotion that he is experiencing in light of an understanding of the actions that are called for in a given situation and that these could lead him to identify the internal state accordingly (imagine this as a sort of emotional inverse spectrum problem). Thus, for Adam*, even though he may be unable to identify the emotional response that he has in a given situation, his functional capability to recognize socially appropriate transactions would allow him to do so and would thus entail that he *actually does feel the relevant emotion,* and so he has a moral sense in the way that McKenna requires. If this is true, and I think that it must be

on McKenna's view, then we have an agent who meets all of the requirements for fully responsible agency set out by the conversational theory but who seems not to be fully responsible, and, therefore, an agent who occupies the middle ground that I proposed above. He lacks access to a range of facts about other people which may be necessary to meet certain epistemic conditions of responsibility, and more importantly, he lacks the ability to relate to his own emotions in important ways. He can have appropriate emotional responses, but he simply cannot make sense of them in the way required for being fully morally responsible. Instead, Adam* is an agent who responds to the world around him in ways that allow him to make sense of his environment and of his relationships with others. However, there is a sense in which his actions are a part of what he takes to be a grand moral choreography. He is functional enough to take part in the choreographed moves, and carrying on moral "conversations" is an important part of this. However, the ability to function in these conversations fails to explain important factors outside of this functionality, and these factors are such that they make a difference for how we understand responsible agency.

I take it that this challenge points toward a significant problem for McKenna's account. Insofar as certain autistic persons might meet the necessary conditions that the theory proposes but fail to be responsible. I will offer a tentative solution to this problem in section 4, but first I will consider whether Shoemaker's account can give an adequate account of the moral responsibility of those with ASD where McKenna's could not.

§3. Shoemaker's Empathic Account

Providing an adequate exposition of Shoemaker's regard-based quality of will approach to moral responsibility will require situating it within his broader project, and that will be the task to begin this section.

§3.1. The Tripartite Theory

Most generally, Shoemaker aims to construct a theory of responsibility which can account for all of the facts our moral responsibility practices. Key among these, Shoemaker claims, is the fact that there are a number of agents to whom we respond with ambivalence. This fact motivates Shoemaker's central claim which is that moral responsibility is not unified but that there are, instead, three distinct types of responsibility – attributability, answerability, and accountability. This observation leads him to construct what he calls a "quality of *will*s" approach in which he argues that the three types of responsibility are grounded on three distinct understandings of quality of will. So, on his view, attributability-responsibility is characterized by responses to an agent's quality of character, answerability-responsibility is characterized by responses to an agent's quality of judgment, and accountability-responsibility (the sort which is of interest to us in this chapter) is characterized by responses to an agent's quality of regard. Each of these types of responsibility, he claims, is needed to account for all that we mean when we talk of moral responsibility.

Importantly, Shoemaker employs a Strawsonian strategy in that he understands each type of responsibility to be characterized by the sentimental responses that would be fittingly directed at agents. Sentiments, on Shoemaker's account, have a number of crucial features which distinguish them from emotions more generally. The defining

characteristics of sentiments are that they are both pan-cultural and encapsulated from judgment.[8] Elaborating on the former, Shoemaker writes, "Sentiments … are generally taken to refer to pan-cultural emotional syndromes, those emotions instantiated universally with roughly equivalent feelings, thoughts, and action tendencies." (Shoemaker 2015b, 21) With respect to the latter feature, he claims that sentiments are encapsulated from judgment in at least two respects. First, he writes, "[the sentiments] are *stably recalcitrant*, that is they sometimes persist even in the face of contrary judgments." (22) Second, he suggests, sentiments tend to motivate unthinking action independently of any judgments on the part of the agent regarding the merits of those actions which is further evidence of their encapsulation from judgment. Finally, Shoemaker also claims that the sentiments are encapsulated from judgment in a deliberative sense. He writes, "Sometimes when we are unsure about what to do, and where the counsel of judgment is indeterminate or simply runs out, we projectively imagine ourselves into alternative scenarios as robustly as we can in order to discover how we will *feel* about them." (23)

With this understanding of the sentiments in hand, Shoemaker goes on to propose that different sentimental pairs can help to carve out the boundaries of the three different types of responsibility. So, on his view, the contours of attributability-responsibility can be seen by examining the fittingness of the sentiments of admiration and disdain. We can come to an understanding of answerability by examining the appropriateness of the sentimental pair of regret and pride. And, finally, we can get at the nature of accountability-responsibility by looking at the fittingness conditions of what Shoemaker calls "agential anger" and gratitude. Since our interest in this chapter is in the regard

[8] In his account of the sentiments, Shoemaker draws heavily on D'Arms and Jacobson 2006; D'Arms and Jacobson 2000; D'Arms and Jacobson 2003; D'Arms 2013.

sense of quality of will, I will leave out any discussion of Shoemaker's accounts of attributability and answerability here and will turn instead to a detailed discussion of his theory of accountability.

§3.2. Agential Anger and Accountability

In exploring the nature of accountability responsibility, Shoemaker begins by giving an account of the sentiment of agential anger.[9] As he understands it, the syndrome of agential anger "consists in (a) feelings of heat and aggression … (b) thoughts about slights, and (c) action tendencies to revenge or retribution (communicated as such)." (Shoemaker 2015b, 90) Anger, on Shoemaker's view, is important because it is the sentiment which evaluates one's regard as poor. The reason that anger is best understood as evaluating poor quality of regard is to be found in its association with thoughts about being slighted, and being slighted, Shoemaker claims, consists in a failure to take another seriously.[10] On this score, he writes,

> …my taking you seriously is a matter of the extent to which I take your specific normative perspective to bear a weight in my own deliberative perspective in the generally valenced way it does for you. Your normative perspective is your take on what things are good or bad, and so includes your attitudes toward your treatment by others. For me to take you seriously, your view of the goodness or badness of such treatment, for example, must count for me too *qua* practical agent, and in sync with how it counts for you. (Shoemaker 2015b, 97)

This conception of what it means to fail to take one seriously points the way, he thinks, toward two distinct types of regard.

---

[9] His focus on anger rather than on resentment is an important departure from how theorists typically proceed. See, e.g. Wallace 1994.

[10] On slights, Shoemaker approvingly cites Aristotle's claim that "slighting is the actively entertained opinion of something as obviously of no importance." Ibid., citing Aristotle/Roberts 1954, 92/1378a-b.

The first of these, evaluational regard, "obtains when facts about what you take to be good or bad are also taken by me to then count sufficiently in favor of, roughly, promoting or respecting what you take to be good, or avoiding or thwarting what you take to be bad." (97) One can fail to demonstrate regard for others in this sense in at least three ways. First, one might fail to even take notice that facts about another's normative perspective are putative reasons. Second, one might actively judge that facts about another's normative perspective are not reasons, and, third, one might fail to give reasons about another's normative perspective their proper weight or authority. In addition to evaluational regard, Shoemaker argues that taking another person seriously requires a sufficient degree of emotional regard. As he puts it,

> This occurs when I *care* about your normative perspective, such that I am disposed to respond emotionally to its fortunes. In other words, when the things you view as good are respected or promoted, I am disposed to respond emotionally in sync with you, feeling happy, proud, or gratified, say; and when the things you view as good are undermined, or the things you view as bad are respected or promoted, I am disposed to feeling dismay, frustration, embarrassment, or shame, say, right alongside you. (99-100)

On his view, then, a failure to respond emotionally in tune with the up and down fortunes of certain others constitutes a slight.

Importantly, regarding others in these ways requires that one be able to take up two distinct empathic stances. In order to treat another with sufficient evaluational regard, one must have a robust capacity for what Shoemaker calls evaluational empathy. This stance requires that one be able to "take up the normative perspective of others to perceive what facts about their normative perspectives, if any, both seem to be, and actually count as, reasons." (100) Likewise, treating others with emotional regard requires that we "take up others' normative perspectives in opening ourselves up to

feeling what they do or would feel in various circumstances." (101) Emotional empathy, in the sense that Shoemaker specifies, involves a number of things. He writes, "What this means is that, in order to feel what the object of my perspective-taking feels, I must project my *feeling* self into his perspective, and to be emotionally affected in sync with him, I have to imagine caring about the things he does … in a way that is roughly akin to the way he cares about them." (101) So, central to the capacity for showing regard for others, on Shoemaker's account, is the ability to take up a robust empathic stance toward other agents.

From all of these claims, Shoemaker proposes the following definition of what it is for an agent to be accountability responsible.

> **Accountability**: One is an accountable agent just in case one is liable for being a fitting target of a subset of responsibility responses to one – a subset organized around the paradigm sentimental syndrome pair of agential anger/gratitude – in virtue of one's quality of regard. To have quality of regard an agent must be capable of either (a) coming to see facts about others' (or the agent's own) normative perspectives as putative reasons in the agent's own normative deliberations, as a function of evaluational empathy, or (b) coming to feel what others feel in a simpatico fashion as a function of emotional empathy. An agent is accountable for some specific attitude or action just in case it accurately displays either or both of these features of the agent's quality of regard. (113)

The important question for our purposes, then, is whether or not Shoemaker's theory of accountability-responsibility can give an adequate characterization of individuals with ASD. Fortunately, unlike the other theorists that have been discussed up to this point, Shoemaker addresses this question explicitly. So, I will turn now to a discussion and evaluation of his account of the responsibility of individuals with autism.

§3.3 Accountability in ASD

Shoemaker's basic position with respect to individuals with ASD is that, while they might be responsible agents on some conceptions of responsibility, they are not

responsible in the accountability sense. His starting point for this claim is the dilemma proposed by Kennett and discussed in the introductory chapter of this work. As he describes it,

> either [individuals with ASD] are … exempt/mitigated from accountability, despite their general moral concern for doing the right thing, their general conformance with morality, and the accountability responses both they and others may level at them, or they are not exempt/mitigated, in which case empathic impairments do not have the importance for regard and accountability I have taken them to have. (168)

His approach to the dilemma is to grasp the first horn and undermine the notion that individuals with ASD being exempt/mitigated is implausible. In order to show this, he relies on the claim that individuals with ASD show considerable impairments in identifying empathy.[11] Importantly, Shoemaker's conception of accountability requires that an agent be capable of coming to see reasons about another's normative perspective via evaluational empathy *or* that the agent be capable of feeling what others feel via emotional empathy in the way that the specific circumstances might demand. So, in order to defend the claim that autistic persons are not accountable, he needs to show that they fail to meet both of these disjuncts.

In arguing that individuals with ASD lack sufficient capacity for evaluational regard, Shoemaker addresses Kennett's suggestion that they can come to have a moral concern by acknowledgment of the moral law. The claim here is just that by evaluating the worth of their own ends, individuals with ASD can come to see that others have a normative perspective which is worth pursuing also. This, Shoemaker concedes, could be

[11] Shoemaker's notion of "identifying empathy" includes both evaluational empathy and emotional empathy as described above. He contrasts identifying empathy with detached empathy which he suggests consists merely in perspective taking. This latter type corresponds to what I called cognitive empathy in earlier chapters.

an alternate route to evaluational regard, but it is unlikely, he thinks, that this is what is going on in the case of autism. Instead, he adopts McGeer's interpretation which, recall, was to suggest that what those with ASD actually display is a concern for order rather than a concern for the moral law or for the normative perspective of others, and, in doing so, he advances an argument against Kennett's suggestion much like the one that I have presented contra McKenna's view here, namely that meeting certain functional demands is not sufficient for moral responsibility. To my mind, Shoemaker is correct on this count. Indeed, given the Counterfactual Thinking Hypothesis that I presented and defended in Part I, a severe difficulty in evaluational empathy is precisely what we should expect to observe in individuals with ASD.

In addition to arguing that individuals with ASD lack the capacity for evaluational empathy, Shoemaker claims that autistic persons also lack emotional empathy. In arguing for this, he focuses on the role that agential anger plays in accountability. The fundamental action tendency of agential anger, he thinks, is to communicate one's anger to the one who has done the slighting. "[W]hat anger motivates," Shoemaker argues, "is *making the slighter fully aware of what he has done*. It is a demand to get him to appreciate, to acknowledge, the emotional havoc (and worse) that he has wreaked." (107) This demand, then, is fundamentally a demand for a kind of emotional identification, and it is this feature which gives rise to his primary argument that individuals with ASD are not accountable agents. He writes,

> [G]iven their deficits of emotional empathy, those with high-functioning autism would seem unable to adhere to the demand for acknowledgment. This point is buttressed by the fact that, according to both parental reports and experimental results, those with high-functioning autism tend to experience neither guilt nor pride (or experience them only in rare cases) … Now if being accountable is just a function of being eligible for being appropriately held accountable, and the

> fundamental aim of holding accountable – expressing the accountability demand – is for acknowledgment and a certain emotional experience and transformation as a result, because those with high-functioning autism tend to be unable to do either, it would not be appropriate to hold them accountable and consequently they would not *be* accountable (or at least their accountability would be significantly mitigated). (171)

In light of the evidence that I presented in chapter 3 of this work, it should be clear that I think there are considerable problems with this argument. I will turn to these now.

Shoemaker's argument requires the truth of two empirical claims: (1) individuals with ASD have impaired emotional empathy, and (2) individuals with ASD are unable to experience the appropriate acknowledgment responses to the accountability demand. It seems to me, however, that neither of these claims is warranted by the available empirical evidence. With respect to (1), I have presented a great deal of evidence which suggests a robust emotional life in individuals with ASD. I argued at length in chapter 3 that autistic individuals are able to experience a wide range of emotions, understand their own emotions, recognize emotions in others, and, importantly, engage in robust forms of emotional empathy. In support of (1), Shoemaker cites the work of Hobson et al. which I discussed at length previously. His central claim is that what is lacking in ASD is an engaged, intersubjective element of empathy which "rest[s] on their inability to take up the perspectives of others in any real way." (168) However, when one looks beyond the Hobson et al. studies, as we saw, the evidence seems to tell in favor of intact emotional empathy in ASD. Evidence exists which seems to show that individuals with ASD experience personal distress in response to negative emotions in others, that they "catch" emotions from others by way of emotional contagion, that they show empathic concern or sympathy toward others to a degree equivalent to controls, and that they are capable of

experiencing affective empathy in the sense specified in chapter 3.[12] The presence of these capacities in ASD is more apparent in light of the pervasiveness of alexithymia in autism. I presented evidence previously, which shows that alexithymia causes empathy deficits in a number of populations (most notably those with anorexia nervosa and those with traumatic brain injury) and that when researchers control for alexithymia in studies involving ASD populations the emotional impairments in ASD virtually disappear. The studies that Shoemaker cites do not account for this feature of ASD, and, so, they do not seem to me to warrant the claims he makes about emotional empathy in autism, or, at the very least, it makes those claims speculative and unable to bear the theoretical weight that he puts on them.

Nevertheless, it may be open to Shoemaker to claim that even if I am correct that the evidence tells in favor of the presence of emotional empathy in ASD, agential anger toward autistic individuals would continue to be unfitting due to their purported inability to experience guilt or pride and, therefore, to acknowledge the accountability demand. This response, however, does not withstand the empirical evidence presented in chapter 3 either. Shoemaker bases his claim that these emotions are unavailable to those with ASD, once again, on the work of Hobson et al., but there are problems with this interpretation. First, pride seems to be an emotion that individuals with ASD clearly do experience. What was borne out in the Hobson et al. studies was not that individuals with ASD lack feelings of pride but that they lack a motivation to share their achievements or their pride

[12]"[Affective] empathy presupposes the ability to differentiate between oneself and the other. It requires that one is minimally aware of the fact that one is having an emotional experience due to the perception of the other's emotion, or more generally due to attending to his situation … In order for my happiness or unhappiness to be genuinely empathic it has to be happiness or unhappiness about what makes the other person happy." (Stueber 2014)

with others. Second, I devoted a good deal of time to discussing guilt earlier on in this work, and, though its role in ASD was considerably more complex than that of other emotions, the evidence suggests that those with ASD do, in fact, experience guilt but that it is experienced (or, more accurately, expressed) in atypical ways.

One reason for the perception that guilt is lacking in ASD, I argued, was that studies of guilt experiences relied on the ability to describe events or episodes in which the participant felt guilty and that such tasks stack the deck against individuals with ASD because they require intact episodic memory which individuals with ASD lack. However, in one experiment from the Hobson et al. study the authors aimed to study the experience of guilt directly, rather than by way of self-report, by attempting to induce a guilty response in participants. In these tasks, researchers would hand the participant a toy or a pair of spectacles that was designed to break when moved in a particular way. The study was constructed so that the participant was lead to think that he or she had broken the item and his or her response was then observed after the break occurred. The level of guilt induced was inferred by the tendency of the participants to (1) express negative affect, (2) attempt to repair the broken item, and to (3) show signs of bodily retreat by physically withdrawing from the tester. Importantly, Hobson et al. found no significant difference between the ASD and neurotypical groups along any of these dimensions. The only way in which ASD differed significantly from controls was in their tendency not to have a "guilty look" after breaking the item.[13] This condition, however, is far too subjective to be rigorously applied, and, in my opinion, tells us little about whether the

[13] In the study, they define a "guilty look" as "an active gaze pattern towards and/or away from the tester involving some combination of anxiety and reassurance-seeking, and then relief when the tester explained it was not the participant's fault. The quality of the exchange involves the rater emotionally, so that there is sympathy with the child's personal distress at harm done." (Hobson et al. 2006, 102).

participants actually experience the emotion in question. The fact that "looking guilty" was dependent to some extent on the emotional response of the rater could tell us just as much about the neurotypical bias of raters as it does about the emotional life of those with ASD. Additionally, within the Hobson et al. study at least one of the autistic participants performed at control levels on this feature of the guilt task as well. It seems to me that what these responses show is that guilt is atypically manifested in ASD, but that it is very likely present. The participants with ASD showed an appropriately valenced emotional response, and they made an effort to rectify the wrong. Both of these are important sentimental markers on Shoemaker's view, and they seem to be present in ASD. Given this, it seems to me that there is little ground to claim that they do not experience genuine guilt responses.[14] If Shoemaker is to claim that autistic individuals "tend to experience neither guilt nor pride," then more evidence is needed. What the evidence he cites seems to show is not that guilt and pride are absent but that they are manifested in atypical, idiosyncratic ways. However, this idiosyncrasy is not sufficient ground to deny that individuals with ASD are accountable agents.

One of my primary objectives in chapter 3 was to show that individuals with ASD have rich, and largely typical, emotional capacities. Where these are absent, I argued, their absence is to be attributed not to the emotional profile of ASD but to alexithymia or to the counterfactual cognitive demands of their context. I think that picture is correct and is bolstered by what I have said so far in this chapter. However, it is important to keep in

[14] It should be noted that there are important problems with this study. First, while it is *likely* that what the study tested was the experience of guilt, this is not obviously the case. In the study, the breaking of the item was accidental for all participants involved. So, in breaking it, the participants did nothing *wrong*, and as a result we might expect them not to feel guilty. Second, and astonishingly, the researchers make no reference to apology anywhere in the study. Insofar as being moved to apologize is one of the key action tendencies of guilt, it is surprising that it does not come up in the observation of either participant group.

mind that Shoemaker is working with an incredibly strong conception of empathy, one that requires that the agent be able to imaginatively project oneself into another's position and take on her cares. This sort of imaginative projection is something that, given their impairments in counterfactual representation, autistic persons are unlikely to accomplish, and, so, it remains open to Shoemaker to claim that they are not accountable due to this more specific empathic impairment. Indeed, given his strong conception of empathy, this is the conclusion to which he is committed, and this is what leads him to take the first horn of Kennett's dilemma, described above. However, this way of proceeding seems to me to be a consolation prize of sorts for Shoemaker. What would be better would be for him to have a way of rejecting Kennett's claim that moral concern is arrived at by something other than an emotional capacity (thus preserving the claim that empathy is a necessary condition for accountability) *and* to have the means to preserve the strong intuition that high-functioning autistic persons are, in fact, accountable agents (thus allowing him to differentiate them from psychopaths in the way that he wants) all while explaining our sometimes ambivalent responses to them. I don't think that Shoemaker needs to settle for second best here, and in the next section I will propose some revisions that will, I think, give him a better way of accounting for the responsibility of autistic persons.

§4. Revisions

I am quite sympathetic to the quality of will approach, generally, and I think that such an approach gives us the best hope for dealing adequately with the empirical evidence regarding ASD. So, I want to consider now some ways in which we might

rescue the views discussed so far from the challenges that I have leveled against them while preserving what makes them appealing.

§4.1. Empathy Revisited

The way forward for Shoemaker's view seems to me rather obvious. We can sidestep the challenge from ASD by simply weakening his conception of emotional empathy. As he presents it, emotional empathy requires a great deal from agents. It requires that we take up the perspective of another person, that we share their cares, that we project our emotional selves into their position and that we imagine what our emotional response would be, etc. I have serious doubts about whether the majority of moral agents are ever capable of engaging in such a process at all. Moreover, such a robust process seems unnecessary to preserve Shoemaker's theory. Consider the example of a failure of emotional regard that he offers:

> Suppose my spouse has been terribly mistreated at work one day, and she comes home very upset. Once she tells me the story, if I am not upset as well – upset along the same dimensions as her, and with respect to what was done to her – this will likely occasion her anger, and it seems to do so fittingly; my lack of emotionally in-sync response is, in other words, a slight. (Shoemaker 2015b, 100)

Surely Shoemaker is correct that anger is fitting in this case. However, on his view, what happens here is this: when my spouse comes home and tells me her story, I must imagine myself caring about the things that she cares about, project my feeling self into her situation, imagine how I would feel, and then respond with the appropriate emotion. If I fail on any of these counts, then my spouse can be fittingly angry with me for slighting her. These requirements, however, strike me as altogether too demanding.

To see why, imagine, instead, that when my spouse comes home and tells me her story that I become sad (emotional contagion), and I have a sinking feeling in my gut

(personal distress). Now, if my response consisted solely of these two responses, then anger may be fitting after all. This is because this minimal response could leave open the possibility that I am merely responding to her expression of some negative emotion. So, perhaps it could be the case that sadness and my physical discomfort merely arise due to the fact that I have an aversion to crying, say, and that I express concern solely for the purpose of making this aversive action stop. If my spouse were to discover this, then her anger would seem fitting enough. However, the weaker conception of emotional empathy that I have been defending throughout this project requires more than this. So, imagine that, in addition to feeling sadness and experiencing physical discomfort in response to her story, I then express an overt concern *for her well-being* (empathic concern) while being fully aware that my sadness is about what makes her sad, i.e. it is about her being mistreated (affective empathy). This is a genuinely empathic response, and it seems to me that it would be entirely unfitting for her to become angry with me were she to find out, say, that all of this took place without my ever imaginatively putting myself in her place and considering how I would feel were her cares to have been mine.

The more robust conception of empathy that Shoemaker has in mind certainly may have a place in our moral relationships. The ability to identify with others so completely seems to me to be a quality that we should find admirable in others. However, as the case above shows, it seems not to be a quality the absence of which should occasion anger. So, this more robust and imaginative form of empathy is better construed, I think, as something supererogatory. When another agent identifies with us in this way, it would be appropriate to be exceedingly grateful to that person. A deep appreciation for their ability to be emotionally engaged with us would be fitting, but its status as

supererogatory makes it unfitting to hold strong negative reactive sentiments when others do not engage in this deeper form of identification with us.

Weakening the emotional empathy requirement would make it such that Shoemaker would no longer need to come up with a plausible explanation for why individuals with ASD are not accountable agents. Rather, he could simply claim that they *are* accountable agents by virtue of their capacity for emotional empathy (a claim which separates them from the psychopath and, thus, gets Shoemaker out of Kennett's dilemma) and that we respond to them ambivalently, when we do, because they are often *excused from* certain sentimental responses due to their pervasive difficulties with counterfactual representations. On this picture, the story to be told about the accountability of individuals with ASD would go something like this: individuals with ASD (1) are accountable agents by virtue of their capacity for emotional empathy. However, (2) given their impairments in counterfactual thinking they are often excused from accountability responses because they have difficulties identifying the emotions of others, or of understanding the circumstances under which certain emotions are elicited, etc. Nevertheless, (3) under some circumstances they *are* able to think counterfactually – namely, as I argued in chapter 2, when the counterfactual elements are made explicit to them – so there are times when it is appropriate to respond to them with agential anger. Knowing when one is faced with an instance of (2) or an instance of (3) is a difficult empirical matter (and one which Shoemaker is fully prepared to accept), but this story gives us an explanation for our ambivalence toward individuals with ASD, offers an empirically adequate account of the emotional profile of such persons, and offers a way

out of the dilemma proposed by Kennett. Moreover, and importantly, it accomplishes all of this without compromising any additional feature of Shoemaker's view.

§4.2. Conversation Revisited

So, there may be a way that we can revise Shoemaker's empathic account in order to save it from the challenge from ASD. Can the same be said of McKenna's account? Recall that the trouble for McKenna's view arose from the fact that we were able to imagine an individual with ASD who could "go through the motions" of McKenna's responsibility exchange yet seemed not to be fully responsible. It seems to me that what enables this counterexample is McKenna's moderate metaphysical thesis, and if we jettison this thesis and adopt the normative interpretation of the relationship between being and holding responsible, then we can retain what is appealing about his view while avoiding the challenge advanced in §2 of this chapter.

The reasons that the modest metaphysical thesis opens the door to the challenge from the functional autistic agent are twofold. First, it makes the susceptibility to the moral responsibility exchange a function of some actual fact about the agent's quality of will, and, second, this results in the reactive attitudes being defined primarily by their appropriate outward manifestation. That is, on his view, facts about the agent's will make certain outward manifestations of reactive emotions appropriate. However, the modest metaphysical thesis makes it the case that the outward manifestation of reactive emotions serves as a means of understanding the content of those emotions as well, and this is precisely what enables the functional agent to engage successfully in the moral responsibility exchange because the outward manifestations of emotions are what are being tracked by the compensatory system of our imaginary autistic agent. So, if we

revise McKenna's view such that what is crucial is that an agent be able to understand the perspective of holding responsible by way of an understanding of the reactive emotions themselves (and not simply by tracking the situations in which their manifestation would be appropriate), then the functional compensatory strategy will no longer give one genuine access to the moral responsibility exchange in which genuine participation is a necessary condition for moral responsibility. To put the point another way, the modest metaphysical thesis allows the functional agent to achieve a "moral sense" simply by understanding when certain practices are appropriate because it results in the view that to understand the practices of holding responsible is essentially to understand the reactive emotions themselves. Doing away with this thesis, then, would require that the agent genuinely understand the reactive emotions independently of the practices that they involve, and this is what the merely functional agent lacks.

This revision preserves what is genuinely interesting and insightful about McKenna's account insofar as it allows him to retain the conversational approach to responsibility. He is correct that there is something essential about the ability to engage with one another on moral terms, and his view brings this out nicely. The revision suggested here allows him to do so without running afoul of the empirical evidence regarding autism.

Interestingly, this revision of McKenna's view, along with my proposed revision of Shoemaker's picture of accountability, may actually give provide the groundwork for a synthesis of the two theories. Fully developing such an account would be beyond the scope of this chapter, but it is worth making a first (speculative) pass at it here. Recall that Shoemaker's account is concerned primarily with the communicative element of

agential anger and its calling to account those toward whom it is directed, and it is the fittingness of anger which determines responsibility on his view. McKenna's view applies this same communicative component more broadly in that it gives us a fuller picture of how agents engage in the communicative practice together. Additionally, the revision that I have proposed for McKenna's view would lead to the adoption of Shoemaker's preferred explanatory direction[15] (i.e. from fitting attitudes to responsibility rather than the interdependence of the two) and would offer a way of explaining why the purely functional agent, Adam*, is not responsible. On the other side, the revision that I have proposed for Shoemaker's view would allow him to maintain that individuals with ASD can genuinely engage in the communicative practices of moral responsibility, thereby preserving the strong intuition that we have in the paradigm case of Adam that he is responsible. In short, without the revisions proposed here, McKenna's view was too inclusive in that it counted more severely impaired autistic persons like Adam* responsible, and Shoemaker's view was too exclusive in that it counted paradigm autistic agents like Adam not responsible. A synthesis of their views, however, may come closer to getting things right. So, a first attempt at such a synthesis might look something like the following:

> **The Sentimental Responsibility Exchange**: To be morally responsible one must be capable of taking part in a moral responsibility exchange which consists of three stages. First, the agent makes a moral contribution whereby she performs some action or holds some attitude which she understands makes her the fitting target of agential anger, an understanding which she achieves through her familiarity with anger's emotional syndrome and through her ability to engage in emotional empathy with others thereby understanding the emotional impact (i.e. agent meaning) of her contribution. Second, at the stage of moral address, other agents communicate their agential anger to the agent and call her to account for

[15] Though Shoemaker is technically neutral on the issue of explanatory direction, he has clear leanings toward this way of conceiving it.

> the slight caused by her actions. Finally, at the stage of moral account, the agent either explains why her actions do not make her a fitting target of agential anger, or she responds by experiencing guilt and its action tendency toward apology and reparation.

Surely this picture requires further analysis before it can stand up to much extended scrutiny. However, as a first pass it seems to me to be a plausible way of bringing these two quality of will views closer together.[16] So, in the end, the two may not end up that far from one another. However, if either view is to get things right in the case of autism each will need to be revised, and the revisions that I have proposed here help to accomplish that.

§5. Conclusion

This chapter concludes our discussion of the general theoretical approaches to moral responsibility on offer in the literature. I have presented two prominent quality of will theories and argued that each is open to an important challenge from the empirical data on autism. McKenna's view seems to render the verdict that some individuals with ASD are responsible where they seem clearly not to be, and Shoemaker's view seems to be too strong in that it claims that individuals with ASD are never (or perhaps rarely) accountable for their actions which strikes many (myself included) as implausible. However, I have also suggested that, with some revision, these theories might be saved from the challenges that I have advanced. The reason for this latter effort is that I think that a quality of will approach to moral responsibility is likely our best hope for giving an account of the responsibility of individuals with ASD. While I can't offer a positive view of my own in this work, there is much to like about the views of both McKenna and

[16] This, of course, applies only to Shoemaker's conception of accountability and not to the whole of his tripartite view.

Shoemaker, and if one of these can account for the hard case of autism, then we will be better off for it.

Chapter 7: Moving Forward

§1. The Road So Far

In this project, I have made a number of bold claims (some surely more plausible than others), so it will be helpful to begin this chapter by recapitulating what has been argued thus far. I began by framing this dissertation around Jeanette Kennett's argument in favor of the Kantian basis for moral agency. Roughly, Kennett's argument was this: Psychopaths, it is often claimed, are not moral agents due to their lack of empathy. So, any person who lacks empathy is not a moral agent. However, people with autism also lack empathy, and it is generally agreed that people with autism are moral agents. Therefore, moral agency does not require empathy. My claim at the outset was that Kennett's argument is an important one in that it proposed that we could learn something important about agency and responsibility by attending to the facts about autism but that her account of autism did not give an accurate picture of the empirical data. So, the motivation for the project was to take what was good about Kennett's paper (as well as Victoria McGeer's subsequent response) by trying to give an account of the cognitive and affective psychology of ASD that was both accurate and comprehensive enough to be of use to philosophers interested in agency and responsibility. That task occupied Part I of this project.

In chapter 2, I set out to give a thorough treatment of the cognitive features of ASD. I began by canvassing the three dominant cognitive theories of the disorder: the Theory of Mind hypothesis, the Executive Dysfunction hypothesis, and the Weak Central Coherence hypothesis. During the course of discussing these, I noted that each of them seemed to involve, in some manner or other, a reliance on counterfactual thinking, and, after presenting evidence that individuals with ASD show signs of impairment in their ability to represent counterfactual states, I proposed my own unifying hypothesis, the Counterfactual Thinking hypothesis. My central claim was that ASD can be best characterized by an impairment in the ability to represent counterfactual states. I showed how this hypothesis could underwrite the three dominant theories on offer in the literature, and I argued that its explanatory reach went beyond these other approaches in that it offered a ready explanation of deficits in pretense and episodic cognition as well. In the end, I suggested that even if this hypothesis did not succeed in offering a fully unifying explanation of ASD it makes a strong case for the view that counterfactual thinking deficits are a central feature of autism, and this weaker hypothesis is sufficient to generate the theoretical conclusions that I have defended.

After presenting an account of the cognitive features of autism, I turned my attention to the affective profile of the disorder. My primary goal in that discussion (chapter 3) was to press against the sort of view that Kennett offers of individuals with ASD as being profoundly impaired in their emotional capacities. The empirical literature actually suggests the opposite, I argued, and it turns out that individuals with ASD have rich emotional lives. I argued in that chapter that individuals are unimpaired in their ability to understand and to experience emotions and that they have intact capacities for

emotion recognition and for empathy. I spent a great deal of time showing evidence that suggests that empathy, in particular, is largely unaffected in ASD. I also presented evidence which suggests that where emotional deficits are observed, these are not a feature of autism but are instead a feature of alexithymia, a disorder which has extremely high rates of comorbidity with ASD. Additionally, I argued that certain atypical features of the social emotions in ASD were attributable to certain counterfactual elements of the emotions themselves or of the context in which they are studied. Part I of this project, then, gives a strikingly different picture of the psychology of ASD than that which most philosophers appeal to.

After offering this detailed, empirically supported picture of the cognitive and affective psychology of ASD, I then turned my attention, in Part II, to bringing this psychological picture to bear on a number of prominent theoretical approaches to moral responsibility. I began, in chapter 4 by considering reasons-responsive theories of responsibility. More specifically, I looked closely at the theory proposed by John Martin Fischer and Mark Ravizza, as theirs is widely considered the gold standard of such theories. I argued, however, that the empirical data on ASD casts doubt on the reasons-responsive approach for several reasons. Chief among these was that Fischer and Ravizza's conception of moderate reasons-responsiveness gave an incorrect characterization of the receptivity and reactivity to reasons of individuals with ASD. Moreover, I argued that correcting this would require Fischer and Ravizza to adopt a view which they explicitly, and correctly, reject. Finally, I claimed that the data on ASD gives us reason to think that the so-called subjective features of responsibility that Fischer and Ravizza propose are inadequate as well. These considerations, at the very least, give

us reason to abandon Fischer and Ravizza's classical conceptual analysis in favor of some other methodology, and, I think, they present a (defeasible) presumption against the reasons-responsiveness view.

These challenges to Fischer and Ravizza's view, I claimed were sufficient to motivate us to look elsewhere for a theory of moral responsibility and, so, I turned in the following chapter to a discussion of real self theories. On these views, an agent is responsible for actions or attitudes which originate in, or are attributable to, her real self. I gave considerable attention to the most famous of these views, Harry Frankfurt's hierarchical view and Gary Watson's dual system view. Each of these accounts, I argued, required that an agent be able to reflect on her desires in important ways. So, on Frankfurt's view, the agent must have second-order volitions which require her to reflect on and identify with her first order desires, and, on Watson's view, the agent must have the ability to reflect on her desires through the filter of her values. I cited some evidence that desire understanding is impaired in ASD and that these impairments, along with the pervasive counterfactual deficits, preclude this sort of self-reflection. Having offered these challenges, I then sought a characterization of the real self which could go some way toward overcoming them. I offered a sketch of an account which was similar to a view defended by David Shoemaker, and which suggested that the real self may be best understood in terms of the agent's emotional dispositions and her historical pattern of moral judgments. While this account is admittedly underdeveloped, I argued that it comes closer to giving an adequate account of individuals with ASD and could, potentially, plausibly characterize the real self of neurotypical agents as well.

This brought me, finally, to a discussion of quality of will theories. These theories find their philosophical ancestry in P.F. Strawson's essay, "Freedom and Resentment," and I discussed two such theories at length, namely, Michael McKenna's conversational theory and David Shoemaker's empathy-based theory of accountability. I argued that each of these views runs up against problems when trying to account for ASD, but that these problems were not as severe as the problems encountered by other theories. I then proposed revisions to each account in an effort to help them to overcome the respective challenges from ASD. These revisions, I claimed, may point the way toward a plausible synthesis of Shoemaker's and McKenna's view, and I offered an outline of what such a view would look like. In the end, I suggested that quality of will theories give us the best hope for arriving at a theoretical account of responsibility which is also up to the task of giving an adequate account of the responsible agency of autistic persons.

One important upshot of the discussion of theoretical approaches to responsibility in Part II is that it seems to be the case that the approaches best able to handle the empirical evidence on ASD are those that give a primary place to the emotional capacities (especially empathy) of the agent. As we saw, the most adequate conception of the real self seems to be one that involves, necessarily, the agent's emotional dispositions or cares. Additionally, one of the key takeaways from the discussion both quality of will views addressed here is that they succeed in accounting for people with ASD, when they do, in virtue of the fact that they place a great deal of importance on the emotions for responsible agency. This is an important conclusion given where this project started. Recall that we wanted to know why autistic persons are moral agents and psychopaths are not, and that Kennett's solution to this problem was to argue that autistic persons are able

to reason their way to a sort of moral concern. What we have now, however, is a great deal of evidence and argumentation that tells directly against Kennett's claim that people with autism take a rationalist route to moral agency. In fact, their route seems to go through their emotional abilities.

§2. Looking Ahead

I have covered a considerable amount of empirical and theoretical work in both philosophy and psychology in the six chapters leading up to this point. However, there is still much work to be done that falls outside the scope of this project which warrants some attention here. First, and most importantly, the primary theoretical contribution that I have offered to autism research is, at this point, speculative. Though I am confident that the counterfactual thinking hypothesis is well-grounded in the available empirical literature, much more research is needed to substantiate it. Future research on this issue could help to further our understanding of the cognitive underpinnings of ASD and, as a result, help to develop effective interventions. Each of the following questions is deserving of further examination in light of the evidence that I proposed in Part I, and answering them may prove helpful in confirming or disconfirming the counterfactual thinking hypothesis: (a) is there a causal connection between counterfactual thinking and weak central coherence? (b) Does performance on counterfactual thinking tasks predict patterns of pretense? (c) Is counterfactual thinking related to social emotional competence? Additionally, examining the use of the subjunctive mood in autistic speech may offer insight into the presence or pervasiveness of counterfactual thinking deficits associated with the disorder. To my knowledge, no such studies currently exist.

Irrespective of its role in autism, I think I have indirectly shown throughout the course of this project that counterfactual thinking is an important cognitive feature for moral responsibility and for moral judgment more generally. In chapter 4, I argued that counterfactual thinking deficits could make individuals with ASD more prone to adopting a model-free rather than model-based system of moral judgment. To my knowledge no one has studied the role of counterfactual thinking in moral judgment, and it seems to me that there is the potential for significant progress to be made by doing so.

In addition to avenues for future empirical work, there is still much theoretical work to be done in light of what I have presented here. I have been focused on addressing general approaches to moral responsibility and examining how well they are able to account for the facts of autism. However, this means that I have passed over, in many cases, discussion of particular views, and, so, it is an open question as to whether or not specific versions of some of these approaches can withstand the empirical test of autism. Using the empirical data as a test case for individual views will be an important undertaking going forward as new views are continually being offered.[1]

Finally, to this point I have been focused on using the empirical data from ASD as a way of testing existing theories, but I have yet to present a positive view of my own. Doing so is well beyond the scope of this project, but it will be the focus of future work moving forward. I have already suggested that a quality of will view is the most promising approach for an empirically adequate account of moral responsibility, and I have suggested that some revisions to current theories might bring us closer to realizing this. However, it remains to be seen whether a better account is in the offing.

[1] See, for example, Vargas 2015.

The presence of persisting questions such as these could be either troubling or gratifying. In one sense, it could be taken as a sign that a project is incomplete in important ways. A project of this scope and size that sets out to answer questions but leaves a preponderance of them open would be of only very questionable success. On the other hand, persisting questions such as those that I've outlined here could be taken as a sign that the project is rich enough and sufficiently interesting to generate new and important questions that are left lying in its wake after it has cut through those it originally set out to answer. Ultimately, I hope that the questions that remain here are of the latter sort.

BIBLIOGRAPHY

Accardo, Pasquale J. and Barbara Y. Whitman. *Dictionary of Developmental Disabilities Terminology, 2nd Edition*. Baltimore, MD: Paul H. Brookes Publishing Co., 2002.

Adams, Marcus. "Explaining the Theory of Mind Deficit in Autism Spectrum Disorder." *Philosophical Studies*. Vol. 163. (2013): 233-249.

Adshead, Gwen. "Commentary on 'Psychopathy, Other-Regarding Moral Beliefs and Responsibility.'" *Philosophy, Psychiatry, and Psychology*. Vol. 3. (1996): 279-281.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: Fifth Edition.*

Aristotle. *The Rhetoric and the Poetics of Aristotle*. Translated by W. Rhys Roberts. New York: The Modern Library, 1954.

Arpaly, Nomy. *Unprincipled Virtue*. New York: Oxford University Press, 2003.

Atwood, Anthony, Uta Frith, Beate Hermelin. "The Understanding and Use of Interpersonal Gestures by Autistic and Down's Syndrome Children." *Journal of Autism and Developmental Disorders*. Vol. 18, No. 2. (1988): 241-257.

Baron-Cohen, Simon, Alan M. Leslie, and Uta Frith. "Does the autistic child have a 'theory of mind'?" *Cognition*. Vol. 21. (1985): 37-46.

Baron-Cohen, Simon, Michelle O'Riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. "Recognition of Faux Pas by Normally Developing Children and

Children with Asperger Syndrome or High-Functioning Autism." *Journal of Autism and Developmental Disorders*. Vol. 29, No. 5. (1999): 407-418.

Baron-Cohen, Simon and Sally Wheelwright. "The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences." *Journal of Autism and Developmental Disorders*. Vol. 34, No. 2. (2004): 163-175.

Bartlett, Frederic C. *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press, 1932.

Beadle, Janelle N., Sergio Paradiso, Alexandria Salerno, and Laurie M. McCormick. "Alexithymia, emotional empathy, and self-regulation in anorexia nervosa." *Annals of Clinical Psychiatry*. Vol. 25, No. 2. (2013): 107-120.

Begeer, Sander, Marc De Rosnay, Patty Lunenburg, Hedy Stegge, and Mark Meerum Terwogt. "Understanding of emotions based on counterfactual reasoning in children with autism spectrum disorders." *Autism*. Vol. 18, No. 3. (2014): 301-310.

Begeer, Sander, Hans M. Koot, Carolien Rieffe, Mark Meerum Terwogt, and Hedy Stegge. "Emotional competence in children with autism: Diagnostic criteria and empirical evidence." *Developmental Review*. Vol. 28. (2008): 342-369.

Begeer, Sander, Mark Meerum Terwogt, Patty Lunenburg, and Hedy Stegge. "Brief Report: Additive and Subtractive Counterfactual Reasoning of Children with High-Functioning Autism Spectrum Disorders." *Journal of Autism and Developmental Disorders*. Vol. 39. (2009): 1593-1597.

Belcher, Hannah Louise. 2015. “Watch: Women Diagnosed with Autism Late Tell Their Stories.” *Autism Speaks*. Web. Accessed 25 February 2016. https://www.autismspeaks.org/blog/2015/09/17/watch-women-diagnosed-autism-late-tell-their-stories

Beldarrain, Marian Gomez, J. Carlos Garcia-Monco, Elena Astigarraga, Ainara Gonzalez, and Jordan Grafman. “Only spontaneous counterfactual thinking is impaired in patients with prefrontal cortex lesions.” *Cognitive Brain Research*. Vol. 24. (2005): 723-726.

Bird, Geoffrey and Richard Cook. “Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism.” *Translational Psychiatry.* No. 3. (2013): 1-8.

Bird, Geoffrey, Clare Press, Daniel C. Richardson. “The Role of Alexithymia in Reduced Eye-Fixation in Autism Spectrum Conditions.” *Journal of Autism and Developmental Disorders*. Vol. 41. (2011): 1556-1564.

Bird, Geoffrey, Giorgia Silani, Rachel Brindley, Sarah White, Uta Frith, and Tania Singer. “Empathic brain responses in insula are modulated by levels of alexithymia but not autism.” *Brain*. Vol. 133. (2010): 1515-1525.

Blair, R. James R. "Brief Report: Morality in the Autistic Child." *Journal of Autism and Developmental Disorders.* Vol. 26, No. 5. (1996): 571-579.

_____. “A Cognitive Developmental Approach to Morality: Investigating the Psychopath.” *Cognition*. Vol. 57. (1995): 1-29.

_____. "Psychophysiological responsiveness to the distress of others in children with autism." *Personality and Individual Differences*. Vol. 26. (1999): 477-485.

Blair, R. J. R., Lawrence Jones, Fiona Clark, and Margaret Smith. "The psychopathic individual: a lack of responsiveness to distress cues?" *Psychophysiology.* Vol. 34, No. 2 (1997): 192-198.

Bowler, Dermot M., John M. Gardiner, and Sebastian B. Gaigg. "Factors affecting conscious awareness in the recollective of adults with Asperger's syndrome." *Consciousness and Cognition*. Vol. 16. (2007): 124-143.

Bratman, Michael. "Responsibility and Planning." *The Journal of Ethics*. Vol. 1, No. 1. (1997): 27-43.

_____. *Structures of Agency.* New York: Oxford University Press, 2007.

Broekhof, Evelien, Lizet Ketelaar, Lex Stockmann, Annette van Zijp, Marieke G. N. Bos, and Carolien Reiffe. "The Understanding of Intentions, Desires and Beliefs in Young Children with Autism Spectrum Disorder." *Journal of Autism and Developmental Disorders.* Vol. 45. (2015): 2035-2045.

Bruck, Maggie, Kamala London, Rebecca Landa, and June Goodman. "Autobiographical memory and suggestibility in children with autism spectrum disorder. *Development and Psychopathology*. Vol. 19. (2007): 73-95.

Brundson, Victoria and Francesca Happé. "Exploring the 'fractionation' of autism at the cognitive level." *Autism*. Vol. 18, No. 1. (2014): 17-30.

Capps, Lisa, Connie Kasari, Nurit Yirmiya, and Marian Sigman. "Parental Perception of Emotional Expressiveness in Children With Autism." *Journal of Consulting and Clinical Psychology*. Vol. 61, No. 3. (1993): 475-484.

Carlson, Stephanie and Louis Moses. “Individual Differences in Inhibitory Control and Children’s Theory of Mind.” *Child Development,* Vol. 72, No. 4. (2001): 1032-1053.

Cook, Richard, Rebecca Brewer, Punit Shah, and Geoffrey Bird. "Alexithymia, Not Autism, Predicts Poor Recognition of Emotional Facial Expressions." *Psychological Science*. Vol. 24, No. 5. (2013): 723-732.

Crane, Laura and Lorna Goddard. “Episodic and Semantic Autobiographical Memory in Adults with Autism Spectrum Disorders.” *Journal of Autism and Developmental Disorders.* Vol. 38. (2008): 498-506.

Crockett, Molly. “Models of Morality.” *Trends in Cognitive Sciences.* Vol. 17, No. 8. (2013): 363-366.

Cushman, Fiery. “Action, Outcome, and Value: A Dual-System Framework for Morality.” *Personality and Social Psychology Review*. Vol. 17, No. 3. (2013): 273-292.

_____. “From Moral Concern to Moral Constraint.” *Current Opinion in Behavioral Sciences*. Vol. 3. (2015): 58-62.

Cushman, Fiery, Liane Young, and Joshua Greene. “Our Multi-System Moral Psychology: Towards a Consensus View.” In *Oxford Handbook of Moral Psychology.* 47-71. Edited by John Doris. New York: Oxford University Press, 2010.

D’Arms, Justin. “Value and the Regulation of Sentiments.” *Philosophical Studies,* Vol. 163. (2013): 3–13.

D'Arms, Justin and Daniel Jacobson. 2006. "Anthropocentric Constraints on Human Value." In *Oxford Studies in Metaethics*, vol. 1. 99-126. Edited by Russ Shafer-Landau. New York: Oxford University Press.

_____. "The Moralistic Fallacy: On the 'Appropriateness' of the Emotions." *Philosophy and Phenomenological Research*, Vol. 61. (2000): 65–90.

_____. "The Significance of Recalcitrant Emotions (or, Anti-Quasijudgmentalism)." Reprinted in *Philosophy and the Emotions*. Edited by Anthony Hatzimoysis Cambridge: Cambridge University Press, 2003.

Darwall, Stephen. *The Second-Person Standpoint: Morality, Respect, and Accountability.* Cambridge, Massachusetts: Harvard University Press, 2006.

Davis, Mark H. *Empathy: A social psychological approach.* Oxford, Westview Press, 1994.

Dawson, Geraldine, Deborah Hill, Art Spencer, Larry Galpert, and Linda Watson. "Affective Exchanges Between Young Autistic Children and Their Mothers." *Journal of Abnormal Child Psychology*. Vol. 18, No. 3. (1990): 335-345.

De Brigard, F., D. R. Addis, J. H. Ford, D. L. Schacter, and K. S. Giovanello. "Remembering what could have happened: Neural correlates of episodic counterfactual thinking." *Neuropsychologia*. Vol. 51. (2013): 2401-2414.

Dolan, Ray and Peter Dayan. "Goals and Habits in the Brain." *Neuron*. Vol. 80. (2013): 312-325.

Doris, John. *Talking to Our Selves: Reflection, Ignorance, and Agency*. New York: Oxford Unviersity Press, 2015.

Drayton, Stefane, Kandi Turley-Ames, and Nicole Guajardo. “Counterfactual thinking and false belief: the role of executive function.” *Journal of Experimental Child Psychology*. Vol. 108. (2011): 532-548.

Duff, Antony. “Psychopathy and Moral Understanding.” *American Philosophical Quarterly*. Vol. 14, No. 3. (1977): 189-200.

Dziobek, Isabel, Kimberly Rogers, Stefan Fleck, Markus Bahnemann, Hauke R. Heekeren, Oliver T. Wolf, and Antonio Convit. "Dissociation of Cognitive and Emotional Empathy in Adults with Asperger Syndrome Using the Multifaceted Empathy Test (MET)." *Journal of Autism and Developmental Disorders*. Vol. 38. (2008): 464-473.

Elliott, Carl. “Diagnosing Blame: Responsibility and the Psychopath,” *Journal of Medicine and Philosophy*, Vol. 17. (1992): 200-214.

Feshbach, N. D. “Sex differences in empathy and social behavior in children.” In *The Development of Prosocial Behavior*, 315-338. Edited by N. Eisenberg. New York: Academic Press, 1982.

Fischer, John Martin. “The Free Will Revolution (Continued).” *Journal of Ethics*. Vol. 10, No. 3. (2006): 315-345.

Fischer, John Martin and Mark Ravizza. *Responsibility and Control: A Theory of Moral Resonsibility.* New York: Cambridge University Press, 1998.

Frankfurt, Harry G. "Alternate Possibilities and Moral Responsibility." 1969. In *Free Will, 2nd Ed.*, 167-176. Edited by Gary Watson. New York: Oxford University Press, 2003.

_____. “Freedom of the Will and the Concept of a Person.” 1971. In *The Importance of What We Care About,* 11-25. New York: Cambridge University Press, 1998.

_____. “Identification and Wholeheartedness.” 1987. In *The Importance of What We Care About,* 159-176. New York: Cambridge University Press, 1998.

_____. "The Faintest Passion." *Proceedings and Addresses of the American Philosophical Association.* Vol. 66, No. 3. (1992): 5-16.

Frith, Uta. *Autism: Explaining the Enigma*. Oxford: Basil Blackwell, 1989.

Frith, Uta and Francesca Happé. “Autism: beyond ‘theory of mind.’” *Cognition*. Vol. 50. (1994): 115-132.

Frith, Uta, John Morton, and Alan Leslie. “The cognitive basis of a biological disorder: autism.” *Trends in Neurosciences.* Vol. 14, No. 10. (1991): 433–438.

Gallagher, Helen and Christopher Frith. “Functional Imaging of ‘Theory of Mind.’” *TRENDS in Cognitive Science.* Vol. 7, No. 2. (2003): 77-83.

German, Tim and Shaun Nichols. “Children’s counterfactual inferences about long and short causal chains.” *Developmental Science.* Vol. 6. (2003): 514-523.

Gleichgerrcht, Ezequiel, Teresa Torralva, Alexia Rattazzi, Victoria Marenco, Maria Roca, and Facundo Manes. "Selective impairment of cognitive empathy for moral judgment in adults with high functioning autism." *Social, Cognitive, and Affective Neuroscience*. (2012): 1-9.

Golan, Ofer, Emma Ashwin, Yael Granader, Suzy McClintock, Kate Day, Victoria Leggett, and Simon Baron-Cohen. "Enhancing Emotion Recognition in Children with Autism Spectrum Conditions: An Intervention Using Animated Vehicles

with Real Emotional Faces." *Journal of Autism and Developmental Disorders*. Vol. 40. (2010): 269-279.

Goldman, Alvin. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford: Oxford University Press, 2006.

Grandin, Temple. *Thinking in Pictures: My Life with Autism Expanded Edition*. New York: Vintage Books, 2006. Accessed at http://www.grandin.com

Grant, Cathy M., Kevin J. Riggs, and Jill Boucher. "Counterfactual and Mental State Reasoning in Children with Autism." *Journal of Autism and Developmental Disorders*. Vol. 34, No. 2. (2004): 177-188.

Green, Joshua. "The Secret Joke of Kant's Soul." In *Moral Psychology Volume 3. The Neuroscience of Morality: Emotion, Brain Disorders, and Development.* 35-79. Edited by Walter Sinnott-Armstrong. Massachussetts: MIT Press, 2008.

Guajardo, Nicole, Jessica Parker, and Kandi Turley-Ames. "Associations among false belief understanding, counterfactual reasoning, and executive function." *British Journal of Developmental Psychology*. Vol. 27. (2009): 681-702.

Hadjikhani, N., N. R. Zurcher, O. Rogier, L. Hippolyte, E. Lemonnier, R. Ruest, N. Ward, A. Lassalle, N. Gillberg, E. Billstedt, A. Helles, C. Gillberg, P. Solomon, K. M. Prkachin, and C. Gillberg. "Emotional contagion for pain is intact in autism spectrum disorders." *Translational Psychiatry*. Vol. 4. (2014): 1-9.

Haidt, Jonathan. "The Emotional Dog and Its Rational Tail: A Social-Intuitionist Approach to Moral Judgment." *Psychological Review*. Vol. 108, No. 4. (2001): 814-834.

Haji, Ishtiyaque. *Moral Appraisability: Puzzles, Proposals, and Perplexities.* New York: Oxford University Press, 1998.

Happé, Francesca and Angelica Ronald. "The 'Fractionable Autism Triad': a Review of Evidence from Behavioral, Genetic, Cognitive, and Neural Research. *Neuropsychology Review*. Vol. 18, No. 4. (2008): 287-304.

Harris, Paul, Tim German, and Patrick Mills. "Children's use of counterfactual thinking in causal reasoning." *Cognition*. Vol. 61. (1996): 233-259.

Hill, Elisabeth L. "Evaluating the theory of executive dysfunction in autism." *Developmental Review*. Vol. 24. (2004a): 189-233.

_____. "Executive dysfunction in autism." *TRENDS in Cognitive Sciences.* Vol. 8, No. 1. (2004b): 26-32.

Hill, Elisabeth, Sylvie Berthoz, and Uta Frith. "Brief Report: Cognitive Processing of Own Emotions in Individuals with Autistic Spectrum Disorder and Their Relatives." *Journal of Autism and Developmental Disorders*. Vol. 34, No. 2. (2004): 229-235.

Hirvelä, Shari and Klaus Helkama. "Empathy, values, morality and Asperger's syndrome." *Scandinavian Journal of Psychology.* Vol. 52. (2011): 560-572.

Hobson, R. Peter. "The coherence of autism." *Autism*. Vol. 18, No. 1. (2014): 6-16.

Hobson, R. Peter, Gayathri Chidambi, Anthony Lee, Jessica Meyer, U. Muller, J. I. M. Carpendale, M. Bibok, and T.P. Racine. "Foundations for Self-Awareness: An Exploration through Autism." *Monographs of the Society for Reasearch in Child Development*. Vol. 71, No. 2. (2006): 1-166.

Hoffman, Martin. *Empathy and Moral Development*. Cambridge: Cambridge University Press, 2000.

Hopkins, Ingrid Maria, Michael W. Gower, Trista A. Perez, Dana S. Smith, Franklin R. Amthor, F. Casey Wimsatt, and Fred J. Biasini. "Avatar Assistant: Improving Social Skills in Students with an ASD Through a Computer-Based Intervention." *Journal of Autism and Developmental Disorders*. Vol. 41. (2011): 1543-1555.

Hume, David. *A Treatise of Human Nature, 2nd Edition*. 1740. Edited by L. A. Selby-Bigge. New York: Oxford University Press, 2009.

Jacobson, Daniel. "Regret, Agency, and Error." In *Oxford Studies in Agency and Responsibility,* Vol. 1, 95-125. Edited by David Shoemaker. Oxford: Oxford University Press, 2013.

Jarrold, Christopher. "A Review of Research into Pretend Play in Autism." *Autism.* Vol. 7, No. 4. (2003): 379-390.

Jurado, Maria Beatriz, and Monica Roselli. "The Elusive Nature of Executive Functions: A Review of our Current Understanding." *Neuropsychology Review*. Vol. 17, No. 3. (2007): 213-233.

Kanner, Leo. "Autistic Disturbances of Affective Contact." *The Nervous Child*. Vol. 2. (1943): 217-250.

Kennedy, Daniel P. and Ralph Adolphs. "Perception of emotions from facial expressions in high-functioning adults with autism." *Neuropsychologia*. Vol. 50. (2012): 3313-3319.

Kennett, Jeanette. "Autism, Empathy, and Moral Agency." *The Philosophical Quarterly*. Vol. 52, No. 208. (2002): 340-357.

_____. "Reasons, Reverence, and Value." In *Moral Psychology*, Vol. 3, 259-264. Edited by Walter Sinnott-Armstrong. Cambridge, Massachusetts: MIT Press, 2008.

Kenworty, Lauren, Benjamin E. Yerys, Laura Gutermuth Anthony, and Gregory L. Wallace. “Understaning Executive Control in Autism Spectrum Disorders in the Lab and in the Real World.” *Neuropsychological Review*. Vol. 18. (2008): 320-338.

Koster-Hale, Jorie, Rebecca Saxe, James Dungan, and Liane L. Young. "Decoding moral judgments from neural representations of intentions." *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 110, No. 14. (2013): 5648-5653.

Lane, Richard D., Geoffrey L. Ahern, Gary E. Schwartz, and Alfred W. Kaszniak. “Is Alexithymia the Emotional Equivalent of Blindsight?” *Biological Psychiatry*. Vol. 42. (1997): 834-844.

Lane, Richard D. and Schwartz, Gary E. “Levels of Emotional Awareness: A Cognitive-Developmental Theory and Its Application to Psychopathology.” *American Journal of Psychiatry*. Vol. 144. (1987): 133-143.

Leevers, Hilary J. and Paul L. Harris. “Counterfactual Syllogistic Reasoning in Normal 4-Year-Olds, Children with Learning Disabilities, and Children with Autism.” *Journal of Experimental Child Psychology*. Vol. 76. (2000): 64-87.

Leslie, Alan M., Ron Mallon, and Jennifer A. DiCorcia. "Transgressors, victims, and cry babies: Is basic moral judgment spared in autism?" *Social Neuroscience*. Vol. 1. (2006): 270-283.

Lind, Sophie E. and Dermot M. Bowler. “Episodic Memory and Episodic Future Thinking in Adults with Autism.” *Journal of Abnormal Psychology*. Vol. 119, No. 4. (2010): 896-905.

Losh, Molly and Lisa Capps. “Understanding of Emotional Experience in Autism: Insights from the Personal Accounts of High-Functioning Children with Autism.” *Developmental Psychology*. Vol. 42, No. 5. (2006): 809-818.

Loveland, Katherine A., Deborah A. Pearson, Belgin Tunali-Kotoski, Juliana Ortegon, and M. Cullen Gibbs. "Judgments of Social Appropriateness by Children and Adolescents with Autism." *Journal of Autism and Developmental Disorders*. Vol. 31, No. 4. (2001): 367-376.

McGeer, Victoria. "The Varieties of Moral Agency: Lessons from Autism (and Psychopathy)." In *Moral Psychology Volume 3. The Neuroscience of Morality: Emotion, Brain Disorders, and Development.* 227-257. Edited by Walter Sinnott-Armstrong. Massachussetts: MIT Press, 2008.

McKenna, Michael. “Contemporary Compatibilism: Mesh Theories and Reasons-Responsive Theories.” In *Oxford Handbook of Free Will, 2nd Edition*. 175-198. Edited by Robert Kane. New York: Oxford University Press, 2011.

_____. *Conversation and Responsibility.* New York: Oxford University Press, 2012.

_____. “Reasons Reactivity and Incompatibilist Intuitions.” *Philosophical Explorations*. Vol. 8, No. 2. (2005): 131-143.

McKenna, Michael and D. Justin Coates. “Compatibilism.” *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Summer 2015. Available at http://plato.stanford.edu/archives/sum2015/entries/compatibilism.

_____. “Compatibilism: The State of the Art.” *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Summer 2015. Available at http://plato.stanford.edu/entries/compatibilism/supplement.html

Mele, Alfred. “Fischer and Ravizza on Moral Responsibility.” *Journal of Ethics.* Vol. 10, No. 3. (2006): 283-294.

Montebarocci, Ornella, Paola Surcinelli, Nicola Rossi, and Bruno Baldaro. “Alexithymia, Verbal Ability and Emotion Recognition.” *Psychiatric Quarterly*. Vol. 82. (2011): 245-252.

Moran, Joseph M., Liane L. Young, Rebecca Saxe, Su Mei Lee, Daniel O'Young, Penelope L. Mavros, and John D. Gabrieli. "Impaired theory of mind for moral judgment in high-functioning autism." *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 108, No. 7. (2011): 2688-2692.

Morris, Robin, Jessica Bramham, Emma Smith, and Kate Tchanturia. “Empathy and social functioning in anorexia nervosa before and after recovery.” *Cognitive Neuropsychiatry*. Vol. 19, No. 1. (2014): 47-57.

Nozick, Robert. *Philosophical Explanations*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1981.

Nuske, Heather J., Giacomo Vivanti, and Cheryl Dissanayake. “Are emotion impairments unique to, universal, or specific in autism spectrum disorder? A comprehensive review.” *Cognition and Emotion*. Vol. 27, No. 6. (2013): 1042-1061.

Ozonoff, Sally, Bruce F. Pennington, and Sally J. Rogers. “Executive Function Deficits in High-Functioning Autistic Individuals: Relationship to Theory of Mind.” *Journal of Child Psychology and Psychiatry.* Vol. 32, No. 7. (1991): 1081-1105.

Peterson, Donald M. and Dermot M. Bowler. “Counterfactual Reasoning and False Belief Understanding in Children with Autism.” *Autism*. Vol. 4, No. 4. (2000): 391-405.

Phillips, Wendy, Simon Baron-Cohen, and Michael Rutter. “To what extent can children with autism understand desire?” *Development and Psychopathology*. Vol. 7. (1995): 151-169.

Rajendran, Gnanathusharan and Peter Mitchell. “Cognitive Theories of Autism.” *Developmental Review*. Vol. 27. (2007): 224-260.

Rieffe, Carolien, Mark Meerum Terwogt, and Katerina Kotronopoulou. “Awareness of Single and Multiple Emotions in High-functioning Children with Autism.” *Journal of Autism and Developmental Disorders.* Vol. 37. (2007): 455-465.

Riggs, Kevin J., Donald M. Peterson, Elizabeth J. Robinson, and Peter Mitchell. “Are Errors in False Belief Tasks Symptomatic of a Broader Difficulty with Counterfactuality?” *Cognitive Development.* Vol. 13. (1998): 73-90.

Rogers, Kimberly, Isabel Dziobek, Jason Hassenstab, Olver T. Wolf, and Antonio Convit. “Who Cares? Revisiting Empathy in Asperger Syndrome.” *Journal of Autism and Developmental Disorders.* Vol. 37. (2007): 709-715.

Ronson, Jon. *The Psychopath Test: A Journey Through the Madness Industry*. New York: Riverhead Books, 2011.

Russell, Paul. "Responsibility and the Condition of Moral Sense." *Philosophical Topics.* Vol. 32, Nos. 1 & 2. (2004): 287-305.

Ryan, Lee, Siobhan Hoscheidt, and Lynn Nadel. “Perspectives on episodic and semantic memory retrieval.” In *The Handbook of Episodic Memory, Vol. 18,* 5-18. Edited by E. Dere, A. Easton, L. Nadel, and J.P. Huston. Elsevier, 2008

Saxe, R. and N. Kanwisher. “People Thinking about Thinking People: The Role of the Temporo-Parietal Junction in ‘Theory of Mind.’” *NeuroImage.* Vol. 19. (2003): 1835-1842.

Scambler, D. J., S. Hepburn, M. D. Rutherford, E. A. Wehner, and S. J. Rogers. “Emotional Responsivity in Children with Autism, Children with Other Developmental Disabilities, and Children with Typical Development.” *Journal of Autism and Developmental Disorders*. Vol. 37. (2007): 553-563.

Scanlon, T. M. "The Significance of Choice." 1988. In *Free Will, 2nd Ed.*, 352-371. Edited by Gary Watson. New York: Oxford University Press, 2003.

_____. *What We Owe to Each Other*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1998.

Schacter, Daniel L., Roland G. Benoit, Filipe De Brigard, and Karl K. Szpunar. “Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions.” *Neurobiology of Learning and Memory*. Vol. 117. (2015): 14-21.

Scott, Fiona J., Simon Baron-Cohen, and Alan Leslie. “If Pigs Could Fly: A Test of Counterfactual Reasoning and Pretense in Children with Autism.” *British Journal of Developmental Psychology*. Vol. 17. (1999): 349-362.

Searle, John. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences*. Vol. 3. (1980): 417-457.

Shulman, Cory, Ainat Guberman, Noa Shiling, and Nirit Bauminger. "Moral ad Social Reasinging in Autism Spectrum Disorders." *Journal of Autism and Developmental Disorders*. Vol 42. (2012): 1364-1376.

Shoemaker, David. “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility.” *Ethics.* Vol. 121, No. 3. (2011): 602-632.

_____. “Ecumenical Attributability.” In *The Nature of Moral Responsibility: New Essays*, 115-140. Edited by Randolph Clarke, Michael McKenna, and Angela Smith. New York: Oxford University Press, 2015a.

_____. “Psychopathy, Responsibility, and the Moral/Conventional Distinction.” *The Southern Journal of Philosophy*. Vol. 49, Spindel Supplement. (2011): 99-124.

_____. *Responsibility from the Margins*. New York: Oxford University Press, 2015b.

Smith, Angela. “Control, Responsibility, and Moral Assessment.” *Philosophical Studies*. Vol 138. (2008): 367-392.

_____. “On Being Responsible and Holding Responsible.” *The Journal of Ethics.* Vol. 11. (2007): 465-484.

_____. “Responsibility as Answerability.” *Inquiry*. Vol. 58, No. 2. (2015): 99-126.

_____. “Responsibility for Attitudes: Activity and Passivity in Mental Life.” *Ethics*. Vol. 115, No. 2. (2005): 236-271.

Sripada, Chandra. “Moral Responsibility, Reasons, and the Self.” In *Oxford Studies in Agency and Responsibility, Vol. 3*. 242-264. Edited by David Shoemaker. New York: Oxford University Press, 2015.

Strawson, P. F. "Freedom and Resentment." 1962. In *Perspectives on Moral Responsibility,* 45-66. Edited by John Martin Fischer and Mark Ravizza. New York: Cornell University Press, 1993.

Stout, Nathan. “Conversation, Responsibility, and Autism Spectrum Disorder.” *Philosophical Psychology*. (Forthcoming).

Stueber, Karsten. “Empathy.” In *The Stanford Encyclopedia of Philosophy*. Edited by Edward Zalta. Spring 2014. Available at http://plato.stanford.edu/archives/spr2014/entries/empathy

Taylor, Graeme. “Recent Developments in Alexithymia Theory and Research.” *Canadian Journal of Psychiatry*. Vol. 45, No. 2. (2000): 134-142.

Terrett, Gill, Peter G. Rendell, Sandra Raponi-Saunders, Julie D. Henry, Phoebe E. Bailey, and Mareike Altgassen. “Episodic Future Thinking in Children with Autism Spectrum Disorder.” *Journal of Autism and Developmental Disorders.* Vol. 43. (2013): 2558-2568.

Thomson, Judith Jarvis. “The Trolley Problem.” *The Yale Law Journal*. Vol. 94, No. 6. (1985): 1395-1415.

Tilghman-Osborne, Carlos, David A. Cole, and Julia W. Felton. “Definition and measurement of guilt: Implications for clinical research and practice.” *Clinical Psychology Review*. Vol. 30. (2010): 536-546.

Todd, Patrick and Neal Tognazzini. “A Problem for Guidance Control.” *The Philosophical Quarterly*. Vol. 58, No. 233. (2008): 685-692.

Uljarevic, Mirko and Antonia Hamilton. "Recognition of Emotions in Autism: A Formal Meta-Analysis." *Journal of Autism and Developmental Disorders*. Vol. 43. (2013): 1517-1526.

Van Hoeck, Nicole, Elizabet Begtas, Johan Steen, Jenny Kestemont, Marie Vandekerckhove, and Frank Van Overwalle. (2014). “False belief and counterfactual reasoning in a social environment.” *Neuroimage*. Vol. 90. (2014): 315-325.

Van Hoeck, Nicole, Ning Ma, Lisa Ampe, Kris Baetens, Marie Vandekerckhove, and Frank Van Overwalle. "Counterfactual thinking: an fMRI study on changing the past for a better future." *Social Cognitive & Affective Neuroscience*. Vol. 8, No. 5. (2013): 556-564.

Van Hoeck, Nicole, Ning Ma, Frank Van Overwalle, and Marie Vandekerckhove. "Counterfactual thinking and the episodic system." *Behavioral Neurology*. Vol. 23. (2010): 225-227.

Vargas, Manuel. *Building Better Beings: A Theory of Moral Responsibility*. New York: Oxford University Press, 2015.

Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Cambridge Massachusetts: Harvard University Press, 1994.

Watson, Gary. "Free Agency." 1975. In *Free Will, 2nd Ed.*, 337-351. Edited by Gary Watson. New York: Oxford University Press, 2003.

_____. "Peter Strawson on Responsibility and Sociality." In *Oxford Studies in Agency and Responsibility, Vol. 2,* 15-32. Edited by David Shoemaker and Neal Tognazzini. New York: Oxford University Press, 2014.

_____. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Agency and Answerability: Selected Essays.* New York: Oxford University Press, 2004.

Williams, Bernard. "Moral Luck." 1976. Reprinted in *Moral Luck: Philosophical Papers 1973-1980*. Cambridge: Cambridge University Press, 1981.

Williams, Claire and Rodger Ll. Wood. “Alexithymia and emotional empathy following traumatic brain injury.” *Journal of Clinical and Experimental Neuropsychology*. Vol 32, No. 3. (2010): 259-267.

Williams, Donna. *Autism: An Inside-Out Approach*. London: Jessica Kingsley Publishers Ltd., 1996.

Wing, Lorna, and Judith Gould. “Severe Impairments of Social Interaction and Associated Abnormalities in Children: Epidemiology and Classification.” *Journal of Autism and Developmental Disorders*. Vol. 9. (1979): 11-29.

Wolf, Susan. *Freedom within Reason.* New York: Oxford University Press, 1990.

Wolff, Sula. “The History of Autism.” *European Child & Adolescent Psychiatry*. Vol. 13, No. 4. (2004): 201-208.

Wood, Rodger Ll. and Claire Williams. "Inability to empathize following traumatic brain injury.” *Journal of the International Neuropsychological Society*. Vol. 14. (2008): 289-296.

_____. “Neuropsychological correlates of organic alexithymia.” *Journal of the International Neuropsychological Society*. Vol. 13. (2007): 471-479.

Yirmiya Nurit, Marian D. Sigman, Connie Kasari, and Peter Mundy. “Empathy and Cognition in High-Functioning Children with Autism.” *Child Development*. Vol. 63. (1992): 150-160.

Zalla, Tiziana, Luca Barlassina, Marine Buon, and Marion Leboyer. "Moral judgment in adults with autism spectrum disorders." *Cognition*. Vol. 121. (2011): 115-126.

Zelazo, Philip, Alice Carter, Steven Reznick, and Douglas Frye. “Early Development of Executive Function: A Problem-Solving Framework.” *Review of General Psychology*, Vol. 1, No. 2. (1997): 198-226.

Zimmerman, Michael. *An Essay on Moral Responsibility*. New York: Rowman and Littlefield Press, 1988.

BIOGRAPHY

Nathan Stout was born in Goshen, Indiana on January 21, 1985 to Phillip and Carol Stout and spent most of his childhood and adolescence in his hometown of Jackson, Michigan. He and his wife, Jennifer, were married on May 20, 2006 in Bourbonnais, Illinois. The two earned their Bachelor's degrees together in 2007 from Olivet Nazarene University. Soon thereafter, they moved to Kalamazoo, Michigan where Nathan earned an MA in Philosophy from Western Michigan University in 2009. They currently reside in New Orleans, Louisiana. Nathan's primary research interests are in ethics and moral psychology, and his work has appeared in a number of philosophy journals and edited volumes.