Moral Decision Making in Adults with Autism: The Role of Capacity in Blame Assignments

Ladan Mohamed

Advised by Fred Volkmar, Irving B. Harris Professor of Child Psychiatry, Pediatrics, and Psychology at the Yale University Child Study Center, School of Medicine

Submitted to the faculty of Cognitive Science in partial fulfillment of the requirements for the degree of Bachelors of Science

Yale University
April 22, 2019

Abstract

This study proposes that theory of mind does not affect ASD (Autism Spectrum Disorder) participants' understanding of intentionality but rather affects their judgment of capacity (whether an event is preventable or not) when evaluating accidental harms. This study assessed the differences between ASD participants and controls on (i) the assignment of blame in unpreventable and preventable cases; (ii) the recognition of capacity in unpreventable and preventable cases; (iii) the likelihood to choose to update blame assignments when presented with an explicit statement of capacity; and (iv) the adjustment of blame assignments after an understanding of capacity is facilitated in participants. The results show that if a situation is perceived as preventable by participants, the ASD group is harsher in their blame assignments. ASD participants are less likely to make correct preventability judgments than controls in both preventable and unpreventable cases. ASD participants are also more likely to choose to update blame assignments than controls in both preventable and unpreventable cases. Finally, among participants who chose to update blame assignments, there was a significant difference between groups on unpreventable questions, where the mean difference in blame adjustment is greater in ASD participants than controls, suggesting that ASD individuals' blame assignments decrease more when preventability is made salient. These findings allow for a better understanding of criminal offenses in adults with ASD, which are often the result of misunderstandings of accidental harms, and can lead to real-world solutions for those who come into contact with the criminal justice system.

1. Introduction

1.1 Moral judgments and theory of mind

Every day, we engage in moral judgments and evaluate others' actions and intentions. We are sensitive to harmful acts and violations of rights (Margoni & Surian, 2016). The development and justification of a moral judgement is a complex socio-cognitive task that relies on mental state reasoning abilities, or theory of mind (Young et al, 2007; Moran et al, 2011). When individuals are asked to evaluate cases of harm, they need to consider the agents' intention and weigh it against the external consequences of the action. This requires a mental state analysis, and neuroscientific evidence confirms an association between moral judgement and theory of mind.

From a developmental perspective, children face difficulties with integrating information about mental states and outcomes. When a moral scenario presents conflicting information about the intention of an agent and the outcome of an action, children tend to rely on the outcome rather than the intention to inform their moral judgement (Young et al, 2007). Older children show greater sensitivity to information about an agent's intentions and this is due to a development of theory of mind and the ability to "integrate this information with information about consequences in the context of moral judgement" (Young et al, 2007). Developmental evidence suggests that mature moral judgments depend on these cognitive processes which are responsible for representing and integrating information about beliefs and outcomes (Young et al, 2007).

1.2 Process of making moral judgments

Much of the literature in moral psychology focuses on the specific inputs to moral judgments, such as causality, intentionality, desires and beliefs, and foreseeability (Monroe & Malle, 2017). However, it is important to consider the process of how moral judgments are made. How do individuals go from perceiving immoral outcomes to considering moral information (e.g. causality, intention) to producing a moral judgment (e.g. blame assignment)? (Monroe & Malle, 2017).

Previous research has attempted to answer this question by distinguishing between two types of judgments. Greene's well-known dual-process model, for example, argues that there are consequentialist judgments and deontological judgments (Greene et al, 2001). While consequentialist judgments are slow to arise, cognitively demanding, and respond to "considerations of outcome and intent," deontological judgments are quick and automatic aversions to harmful outcomes (Cushman, 2015; Monroe & Malle, 2017).

The dual process model is helpful for predicting judgments of moral permissibility, but does not provide predictions for how judgments of blame arise. By itself, the model cannot specify how people process information shown to be integral to moral judgments, such as intentionality, motives, or obligations (Monroe & Malle, 2017). The Path Model of Blame (Malle et al, 2014) tries to fill this gap by specifying the temporal order of processing critical inputs to blame judgments: once it is determined that an agent caused a norm-violating event, perceivers infer information about intentionality, which can send them down one of two paths (Malle et al, 2014). If they determine the agent brought it about intentionally, they infer information about the agent's reasons, and if they determine the agent brought it about unintentionally, they infer information about capacity (whether agent could have prevented the event) or obligation (whether agent should have prevented it) (Malle et al, 2014).

1.3 Autism and moral reasoning

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by impaired social interactions and communication, and a set of restricted and repetitive behaviors (Margoni & Surian, 2016). The American Psychiatric Association's fifth edition of the Diagnostic and Statistical Manual has updated the diagnostic criteria for autism to allow for more accurate assessments of severity level. While there were previously several categories of autism (including

Asperger's syndrome, PDD-NOS, and high functioning autism), the updated manual consolidated these categories into one umbrella diagnosis of Autism Spectrum Disorder.

Note: High-functioning autism and Asperger's syndrome are two forms of autism in which individuals don't display significant motor delays, language delays, and cognitive deficits. For the purposes of this research, we are limiting our analysis to adults with an ASD diagnosis that falls into one of these two categories, as moral judgment tasks require a certain level of cognitive ability that can only be ensured in individuals without intellectual disability and/or cognitive deficits. Throughout this paper, however, both of these categories will be referred to as an Autism Spectrum Disorder (ASD) diagnosis.

Theory of mind has been shown to influence the social-moral abilities of individuals on the autism spectrum. Autism is commonly associated with a delayed maturation of theory of mind, the ability to infer the contents of other people's minds, including beliefs and intentions. Adults with an ASD diagnosis typically succeed on standard tests for theory of mind, yet they remain impaired in real-life social situations (Bowler, 1992).

Traditionally, research on moral judgement and children with ASD has focused on 1) the capacity to distinguish between moral transgressions and conventional transgressions and 2) the ways in which individuals with ASD evaluate the moral permissibility of an action (Margoni & Surian, 2016). We will focus on the second of these topics, particularly on research concerning the relative weight of intention and outcome in the judgments of ASD individuals. In studies involving children, a working hypothesis exists that ASD children perform similarly to typically developing (TD) children when presented with simple and unambiguous moral scenarios, such as a "negative/positive outcome produced by an intentional action with the same valence" (Margoni & Surian, 2016). ASD children appear to develop a basic moral judgement because any difficulties with integrating mental state understanding can be overcome by a reliance on action outcomes. However,

school-aged children with ASD fail to distinguish between accidental harm and intentional harm. While they are able to judge an agent behind intentional harm more harshly than an agent that has caused accidental harm, their justifications do not refer to the agent's intention (Grant et al, 2005). For example, they are unable to distinguish between "failing to come to a planned meeting as a result of canceling the plan without telling or [failing to come] as a result of the bus breaking" (Margoni & Surian, 2016). The same effect is observed in adults with ASD, which suggests that more complex cases of moral judgement, such as accidental harm, require a more substantial contribution of theory of mind.

Neurotypical (control) adults weigh a person's intention more heavily than the outcome of their action when evaluating the moral permissibility of an action. However, adults with ASD and neurotypical adults differ in judgments for accidental harm (intention neutral/outcome negative) and intentional harm (intention negative/outcome negative). While both groups view intentional harm as morally impermissible, adults with ASD view accidental harm as significantly less permissible than neurotypical peers (Moran et al, 2011). ASD participants show an underreliance on information about innocent intentions and an overreliance on negative outcomes (Moran et al, 2011). Making moral judgments about an action based on the analysis of a person's intentions requires theory of mind, especially since exculpation for accidents requires a strong mental state representation to "override a response to the salient information about actual harm" (Young et al, 2007). Moran and colleagues conclude that theory of mind deficits in adults with ASD influence moral judgments by making it difficult to distinguish between neutral and negative intentions (Moran et al, 2011).

A recent study further explored how theory of mind deficits play a role in whether adults with ASD can make appropriate judgments when information relating to the motivations for actions is made salient. Cases involving driver negligence were presented to ASD participants along with the motivation for drivers' actions. ASD participants sympathized less than controls with drivers who

lacked mitigating reasons for their actions, which suggests that "emotionally-driven, rather than cognitively-driven processes may have influenced higher sympathy ratings for the perpetrator in the control group" (Channon et al, 2010). Individuals with ASD, however, may draw upon more deliberate reasoning strategies, or simply use learned social rules, to infer whether an action is right or wrong (Channon et al, 2010). They are able to respond to already-learned morally relevant scenarios but are not able to later generate novel moral distinctions in unfamiliar scenarios (Channon et al, 2010). This finding has implications for social training and interventions in adults with ASD.

1.4 Moral reasoning deficits and criminal behavior in ASD

The deficits adults with ASD face in moral reasoning are particularly relevant when it comes to criminal behavior and violent activity within this population. While an ASD diagnosis is not sufficient to invoke mitigation in cases of violent crime, recent research on the incidence of ASD in criminal settings has suggested that more research is needed to understand the characteristics of individuals with ASD that might contribute to criminal offenses (Lerner et al, 2012). Moral reasoning in particular may be casually relevant to violent criminal activity in adults with ASD (Lerner et al, 2012). In individual cases studies of violent law-breaking in adults with ASD, the common factor was a deficient social understanding, which is attributable to theory of mind deficits (Baron-Cohen, 1998; Kohn et al, 1998). An analysis of the typical motives and triggers of violence in ASD adults shows that more than half of violent acts were motivated by "communicative and social misinterpretations of other person's intentions" (Bjorkly, 2009). The following example illustrates an example of such an offense:

> "The problem began when the family moved to another city. The radio station with which he was fascinated was difficult to access from their new residence. He developed a complex system of aerials, which finally enabled him to tune in successfully. After a year of listening, a local religious radio station set up a new broadcast on the frequency close to his favored station. This interfered with his listening between the hours of 7

> and 10 each evening. He wrote a number of letters to the radio station asking them to stop interfering. He received blessings and Christian tracks [*sic*] in response. Following a further unsuccessful communication, he walked to the radio station carrying a can of petrol, poured the petrol around the station and burnt it down. He had no regrets for his actions and was puzzled what all the fuss was about. He would have successfully avoided detection except that he proudly informed his mother the next morning that he was responsible for the destroyed radio transmitter, a picture of which appeared in the local newspaper" (Barry-Walsh et al, 2004)

From this example and other similar cases, a picture emerges of the situations in which criminal offenses may occur in this population: "an individual with particularly poor theory of mind and emotion regulation is placed in an unfamiliar and overwhelming social scenario. He misperceives the intentions of another individual as hostile (perhaps when such an inference would be justified by observing another agent's actions alone, but would be unjustified when considering intentions, beliefs, and desires), becomes upset and unable to regulate this emotion sufficiently, and is unable to place the situation into his moral rubric; lacking the guidance of internal moral reasoning, and the lodestar of an external moral refrent, he is thus left without any means of directing his actions. He then lashes out, perhaps impulsively and aggressively, to silence the source of his frustration and confusion. Subsequently, during questioning, he lacks the cognitive structure on which to hang his justification (perhaps simply repeating the charges against him to the inquiring officer or lawyer) and is thus less able to make sense of what has happened to himself or to others" (Lerner et al, 2012).

As mentioned in the literature surrounding ASD adults' ability to make correct moral judgments after drawing upon learned social rules, researchers believe that moral reasoning in adults with ASD does not adhere to the same pathway of affective intuition and personal engagement that we observe in the typical population (Greene et al, 2003). Therefore, further research on how adults with ASD understand moral violations is necessary to make stronger claims and consider these

differences in forensic contexts, as it will allow for a better understanding of treatment of individuals with ASD who come into contact with the legal system.

1.5 The present study

This study is interested in the concept of capacity (whether an agent could have prevented an event) that is highlighted in the Path Model of Blame (Monroe & Malle, 2017). We propose that theory of mind does not affect ASD participants' understanding of intentionality but rather affects their judgment of capacity when evaluating moral scenarios with neutral intentions and harmful outcomes. For the rest of this paper, the concept of capacity will be referred to as "preventability," in an effort to make the results and conclusions easier to interpret and understand.

Part 1 of the study is purely experimental and is intended to support a better understanding of moral judgments in people with ASD. Participants are presented with 8 scenarios of accidental harm are asked to make a blame assignment and determine whether the case was preventable or not. Part 2 of the study starts to investigate whether the putative impairment (present in ASD participants) may be ameliorated. In this section, participants are presented with a different 8 scenarios of accidental harm and are asked to make a blame assignment, followed by the opportunity to update that assignment after reading a statement about the preventability of the case. The findings of this study will support a better understanding of the abilities of adults with ASD to refrain from violent reactions and criminal offenses, and how the legal system can be adapted to accommodate for differences in moral reasoning. The following hypotheses are put forward in this paper:

1) Blame assignment based on perceived preventability: If a situation is judged as unpreventable by participants, there will be a significant difference in blame assignment between ASD and controls, as the ASD group will be harsher in their blame assignment. If a situation is judged as preventable by participant, there will be no significant difference in blame assignment between ASD and controls.

2) Accurate judgment of preventable and unpreventable cases: Whether participant judgments of preventability align with the actual question type (preventable vs not preventable) will depend on diagnosis. ASD participants will be more likely to misjudge situations than controls, only on unpreventable questions.

3) Likelihood to choose to update blame assignment when presented with explicit statement about preventability: ASD participants will be more likely to choose to update blame assignments, in both preventable and unpreventable questions.

4) Mean difference in blame adjustment: Of participants who chose to update blame assignments, diagnosis will have an impact on the mean difference in blame adjustment. In unpreventable cases, but not preventable cases, the mean difference in blame adjustment will be greater in ASD participants than in controls.

2. Methods

2.1 Participants

Participants consisted of adults with ASD (self-identified) and neurotypical adult controls (self-identified) over the age of 18. Twenty two participants with ASD (M_{age} = 34.36, 27.27% female) and forty three control participants (M_{age} = 20, 67.44 % female) completed the entire study. One participant with ASD and seven control participants were excluded from analysis because they failed to complete the entire study.

Control participants were recruited within Yale University and New Haven, and through posts on social media. Recruitment of participants with ASD took place through emails to local organizations, schools, and universities serving adults with autism (e.g. Kennedy Krieger Institute, Chapel Haven School), postings on the Community Autism Socials at Yale (CASY) website and Facebook page, autism advocacy websites, and social media outlets. ASD participants had an existing clinical diagnosis of ASD (Asperger's, Autism, ASD, PDD-NOS) and self-reported on the

survey. All participants were given informed consent in accordance with procedures outlined by the Yale University Human Subjects Committee.

2.2 Procedure

Research was conducted through the online Qualtrics™ platform. After the consent process, participants completed the demographics section of the study, which collected gender, age, information about ASD diagnosis (self-report), including age of diagnosis, and geographic location.

Note: The moral scenarios used for the questions in this study were drawn from Monroe & Malle (2017), who had developed a set of prototypically intentional and unintentional events. All scenarios used are provided in the Appendix.

Part 1 of the study consisted of eight trials, classified by researchers into four preventable cases and four unpreventable cases. The cases were presented in randomized order to participants. In each experimental trial, participants viewed two screens displayed in succession. Participants read a short description of an accidental harm and were asked to make a moral judgment ("How much blame does [agent] deserve?") using a click-and-drag slider bar with endpoints of 0 ("no blame at all") and 100 ("the most blame you would ever give"). In the next screen, participants were presented with a multiple choice yes or no question ("Could [agent] have prevented this?"). Participants were not allowed to revisit previous information throughout all the trials.

Part 2 of the study consisted of another eight trials, classified by researchers into four preventable cases and four unpreventable cases. The cases were presented in randomized order to participants. Each experimental trial in this part consisted of 2-3 screens (minimum of two screens, third screen was dependent on participant response) displayed in succession. Participants read a short description of an unintentional event and were asked to make a moral judgment ("How much blame does [agent] deserve?") using a click-and-drag slider bar with endpoints of 0 ("no blame at all") and 100 ("the most blame you would ever give"). In the next screen, participants were

presented with the same scenario followed by either the statement "[Agent] could have prevented this" or "There is no way [agent] could have prevented this." The same screen asked the following multiple choice (yes/no) question: "Would you like to adjust the amount of blame you gave [agent]." Answering "yes" to this question presented participants with a third screen, which included a new click-and-drag slider bar and asked participants to enter their new answer. Answering "no" to this question took participants to the next trial. Participants were not allowed to revisit previous information throughout all the trials.

2.3 Sample Questions

The following is a sample of a question (preventable case) from Part 1 of the study:

> [Screen 1]
> Sven accidentally broke the Wright's front window. He was playing basketball too close to the house and accidentally hit a ball through the window.
>
> How much blame does Sven deserve?
>
> [Screen 2]
> Sven accidentally broke the Wright's front window. He was playing basketball too close to the house and accidentally hit a ball through the window.
>
> Could Sven have prevented this?

The following is a sample of a question (unpreventable case) from Part 2 of the study:

> [Screen 1]
> Marissa accidentally took the very last handicapped parking space. There was no sign and the handicap marking on the ground had worn away.
>
> How much blame does Marissa deserve?
>
> [Screen 2]
> Marissa accidentally took the very last handicapped parking space. There was no sign and the handicap marking on the ground had worn away.
>
> There is no way Marissa could have prevented this. Would you like to change the amount of blame you gave Marissa?
>
> [Screen 3; if participants selected yes in the previous question]
> Please enter your new answer to the following question: How much blame does Marissa deserve?

3. Results

A two-way ANOVA was conducted to examine the effect of preventability judgment and diagnosis on blame assignment. There was a main effect for preventability, $F(1, 516) = 212.9$, $p < .001$, partial $\eta 2 = .292$, such that the average blame assignment was significantly higher in preventable cases ($M = 54.5$, $SD = 31.6$) than in unpreventable cases ($M = 11.4$, $SD = 14.7$). There was a main effect for diagnosis, $F(1, 516) = 7.519$, $p = .006$, partial $\eta 2 = .014$, such that the average blame assignment was significantly higher in ASD participants ($M = 54.0$, $SD = 36.5$) than in controls ($M = 36.7$, $SD = 31.1$). There was a statistically significant interaction between the effects of preventability selection and diagnosis on blame assignment, $F(1, 516) = 4.76$, $p = .030$, partial $\eta 2 = .009$. If a situation is judged as unpreventable by participants, then there is no significant difference in blame assignment between the two groups, whereas if the situation is perceived as preventable by participants, the ASD group is harsher in their blame assignments.

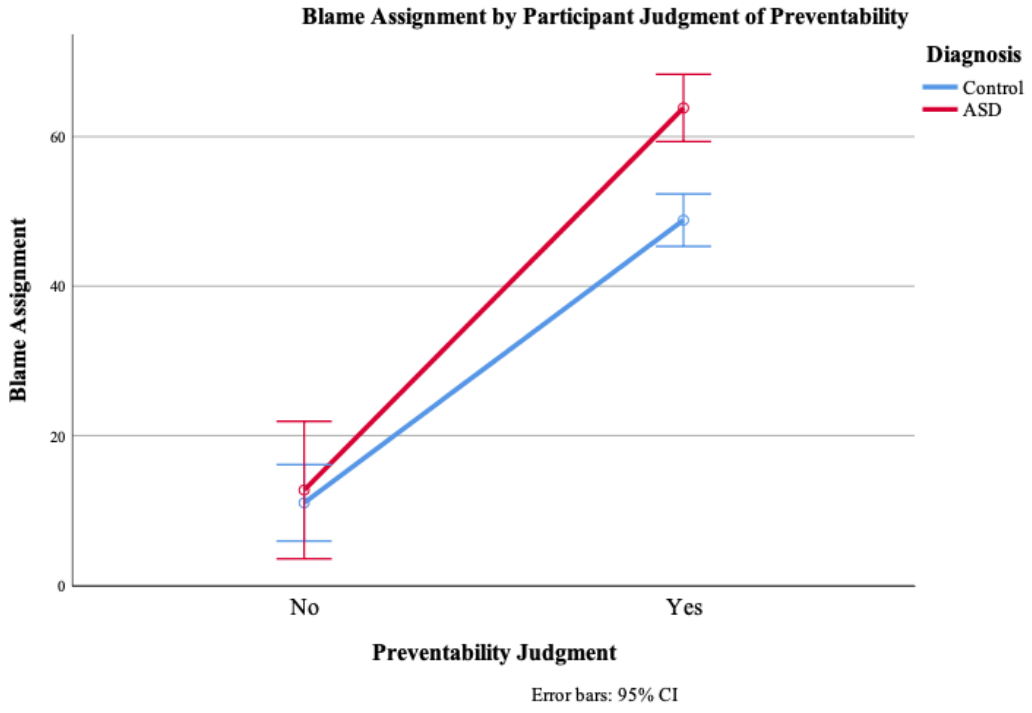

Figure 1. Blame Assignment by Participant Judgment of Preventability

A chi-square test of independence was performed to examine the relation between diagnosis and correctness of preventability judgments in preventable cases. The relation between these variables was significant, X^2 (1, N = 260) = 4.405, p = .036. ASD participants were less likely to make correct preventability judgments than controls in preventable cases.

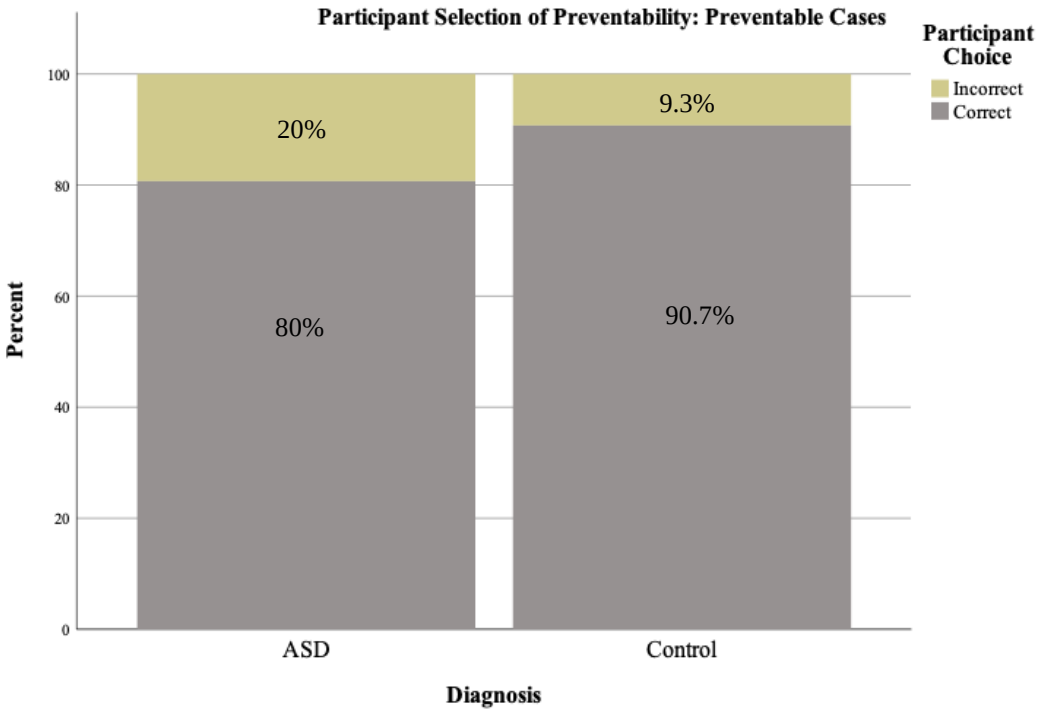

Figure 2. Participant Selection of Preventability: Preventable Cases

A chi-square test of independence was performed to examine the relation between diagnosis and correctness of preventability judgments in unpreventable cases. The relation between these variables was significant, X^2 (1, N = 260) = 4.9998, p = .025. ASD participants were less likely to make correct preventability judgments than controls in unpreventable cases.

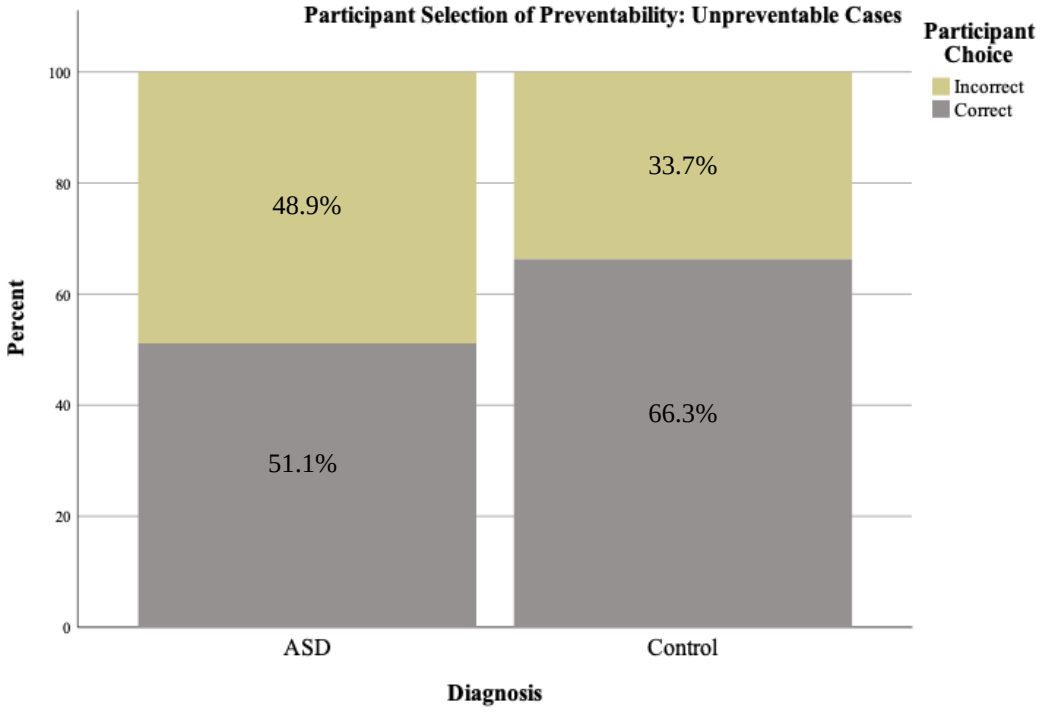

Figure 3. Participant Selection of Preventability: Unpreventable Cases

A chi-square test of independence was performed to examine the relation between diagnosis and choosing to update blame assignments in preventable cases. The relation between these variables was significant, X^2 (1, N = 260) = 10.078, p = .0015. ASD participants were more likely to choose to update blame assignments than controls in preventable cases.

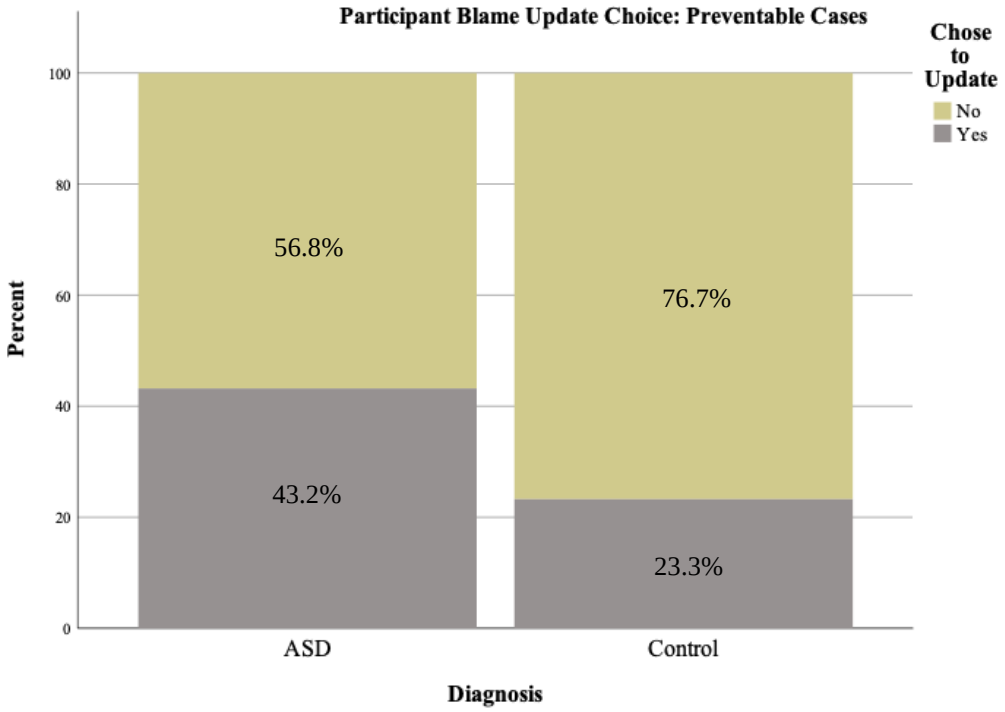

Figure 4. Participant Blame Update Choice: Preventable Cases

A chi-square test of independence was performed to examine the relation between diagnosis and choosing to update blame assignments in unpreventable cases. The relation between these variables was significant, X^2 (1, N = 260) = 34.241, p = 4.868e-09. ASD participants were more likely to choose to update blame assignments than controls in unpreventable cases.

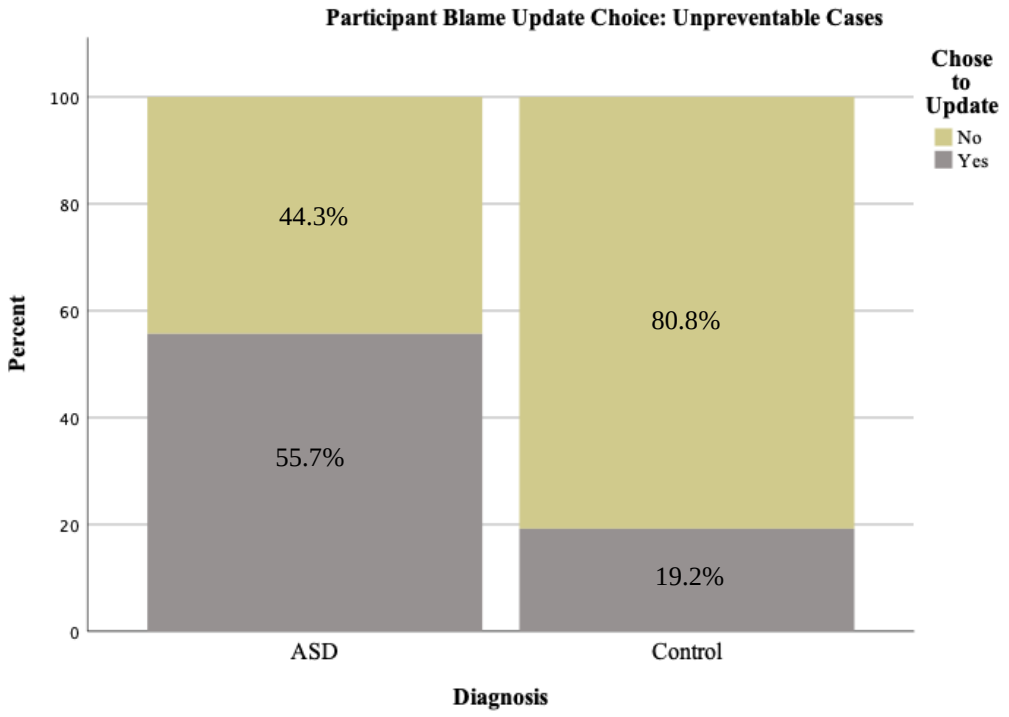

Figure 5. Participant Blame Update Choice: Unpreventable Cases

An independent samples t-test was conducted to compare mean difference in initial and new blame assignments between ASD and controls in preventable cases. Among participants who chose to update blame assignments, there was no significant difference between mean difference in blame assignments between ASD (M = 10.2, SD = 19.6) and controls (M = 12.6, SD = 21.8) in preventable cases, t(75.780) = .515, p = .608.

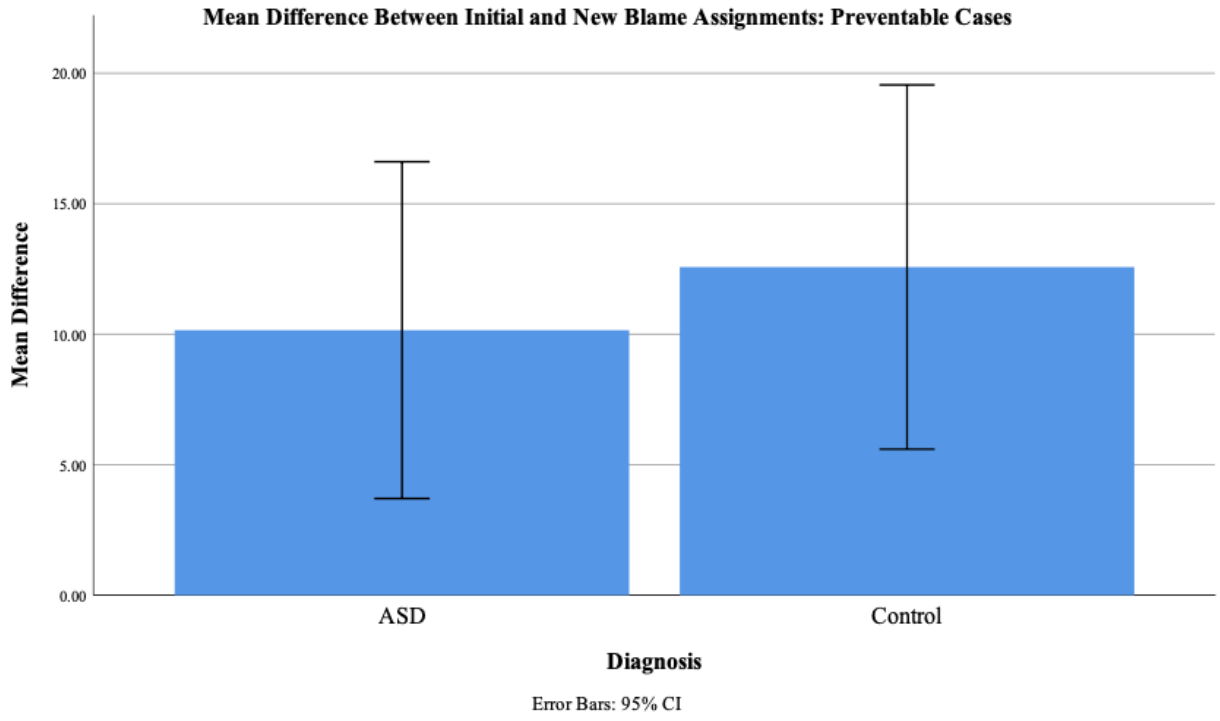

Figure 6. Mean Difference Between Initial and New Blame Assignments: Preventable Cases

An independent samples t-test was conducted to compare mean difference in initial and new blame assignments between ASD and controls in unpreventable cases. Among participants who chose to update blame assignments, there was a significant difference between mean difference in blame assignments between ASD participants (M = -41.8, SD = 33.8) and controls (M = -22.8, SD = 20.5) in unpreventable cases, t(79.254) = 3.165, p = .002. The mean difference in blame adjustment was greater in ASD participants than controls, suggesting that their blame assignments decrease more.

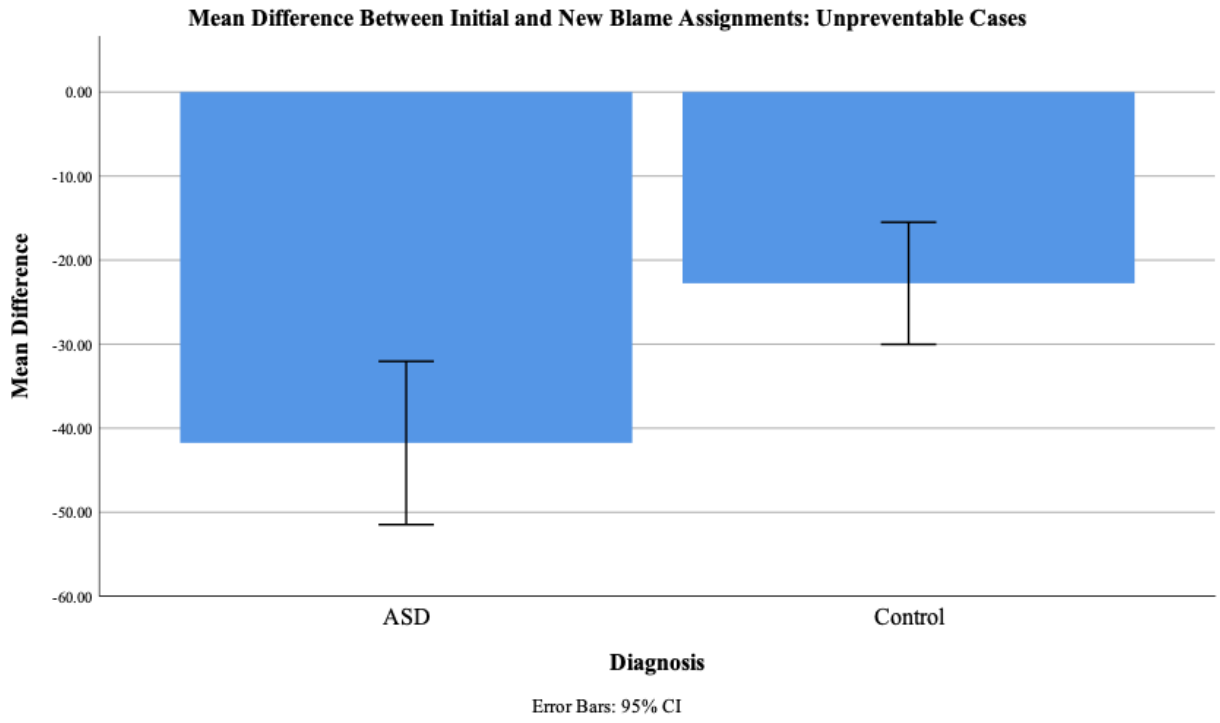

Figure 7. Mean Difference Between Initial and New Blame Assignments: Unpreventable Cases

4. Discussion

4.1 Discussion of Part 1:

The findings of the analysis of variance revealed that there was a significant interaction effect between preventability judgment and diagnosis on blame assignment. This analysis did not separate responses based on researcher-determined preventability (as discussed in the methods, each question is set in the experiment as either preventable or unpreventable), but rather on whether participants perceived, and subsequently judged, cases to be either preventable or unpreventable. As shown in Figure 1, when moral scenarios were perceived as unpreventable by participants, there was no significant difference in blame assignments between the two groups. When moral scenarios were perceived as preventable by participants, there was a significant difference between ASD participants and controls (ASD participants assigned higher blame on these cases). These findings do not support the previously stated hypothesis, which predicted the opposite trend. We hypothesized that there would be a significant difference in blame assignments between ASD and controls in scenarios perceived as unpreventable, but not in scenarios perceived as preventable. Control participants are expected to assign higher blame when they perceive something as preventable and lower blame when they perceive something as unpreventable. Further, previous research has found that adults with ASD assign high blame in cases of accidental harm. We therefore predicted that perceived preventability would not have any impact on the ASD group's blame assignments, and that ASD participants would assign high blame across all cases, whether they were perceived as preventable or unpreventable.

However, these findings show that ASD participants do in fact differentiate between preventable and unpreventable cases, and this is evident in the difference in their blame assignments across these two types of judgments, which is larger than the difference within controls. While the general behavior pattern between ASD participants and controls is similar, when a scenario is judged

to be preventable, controls are more liberal in their blame assignments. Inherently in perceiving and judging a case to be preventable, an individual understands that the agent in question should be held accountable for their actions. An interesting finding here is that ASD participants are less forgiving than controls when assigning blame in these cases, but are as understanding as controls when assigning blame in cases perceived as unpreventable. One potential explanation for this, which will be further explored in this section, is that ASD participants simply judged more cases to be preventable than controls, which might lead to the large difference observed in blame assignments between ASD participants and controls. It might be that ASD participants actually assigned higher blame in both cases they perceived to be preventable and cases they perceived to be unpreventable, but also happened to judge more cases to be preventable (sometimes incorrectly) than controls.

The first set of chi-square analyses further explore the question of ASD participants' ability to make judgments about preventability compared to controls. For these analyses, responses were separated according to whether questions were actually preventable or not preventable (researcher-determined). Accordingly, each participant had four judgments included in the analysis for preventable cases and four judgments in the analysis for unpreventable cases. As shown in Figure 2, there was a significant difference between ASD participants and controls in making correct preventability judgments in preventable cases. 71 of 88 (80%) judgments made by ASD participants on preventable questions were correct, compared to 156 of 172 (90.7%) judgments made by controls. This suggests that ASD participants are less likely to recognize preventable cases as preventable. Similarly, as shown in Figure 3, there was a significant difference between ASD participants and controls in making correct preventability judgments in unpreventable cases. 45 of 88 (51.1%) judgments made by ASD participants on unpreventable questions were correct, compared to 114 of 172 (66.3%) judgments made by controls. This suggests that ASD participants are also less likely to recognize unpreventable cases as unpreventable. These findings only partially

support the previously stated hypothesis, which predicted that ASD participants would be more likely to misjudge preventability when compared to controls, but only on unpreventable questions. Our hypothesis was based on the assumption that ASD participants' high blame assignments on accidental harms (in previous literature) was due to a failure to recognize unpreventable cases. In other words, we predicted that ASD participants viewed every case as preventable, and expected to see a difference in judgment accuracy between controls and ASD in only the unpreventable condition, which would suggest that ASD participants have trouble identifying a case as unpreventable. However, the observed effect shows that diagnosis and accuracy of preventability judgment are not independent in either preventable or unpreventable cases. Control preventability judgments were more aligned with the actual nature of the question, which suggests that diagnosis matters when it comes to whether an individual can recognize preventable and unpreventable cases.

This finding, combined with the interpretation of the ANOVA analysis, tells us that while ASD participants are more likely to incorrectly judge preventability, when they perceive something as either preventable or unpreventable, their behavior closely mimics that of controls, or how judgments "should" be made. That is to say, when ASD participants perceive something as preventable, they are assigning higher levels of blame accordingly (albeit more strongly/harshly compared to controls) than when they perceive something as unpreventable. This behavior is expected from the average neurotypical behavior. It seems then, that the largest problem facing individuals with ASD is correctly judging whether something is preventable or not. If this could be facilitated, one would expect that their blame assignments would more closely resemble that of controls. However, such a conclusion would assume that preventability is largely influencing blame assignments, as opposed to other aspects of these moral scenarios. One limitation of this study is that it focuses solely on preventability judgments and fails to account for other aspects of moral judgments that might be causing the difference in blame assignment, such as judgments of moral

wrongness, which could be measured on a continuous scale. For example, both the case of accidentally bumping into someone in a hallway and accidentally killing someone after forgetting their allergy fall into the preventable category. Although these cases are both due to carelessness, and are both preventable, this study design fails to account for the fact that determining one case to be more morally wrong or worse than another may influence blame assignments, both in controls and ASD participants. Despite this, in part 2 we will discuss the findings of chi-square analyses and how blame assignment adjustments differ between controls and participants with ASD, which still provides us with valuable information about how individuals with ASD update blame assignments after they are provided with the "correct" judgment of preventability.

A final concern regarding the design of part 1 of this study is raised by the results presented in Figure 3. As mentioned previously, 45 of 88 (51.1%) judgments made by ASD participants on unpreventable questions were correct, compared to 114 of 172 (66.3%) judgments made by controls. While the significant difference between groups suggests that ASD participants are less likely to accurately recognize that something is unpreventable, this finding might raise questions about the behavior of controls. One might expect the control success rate in unpreventable cases to be closer to 100%, just as in the preventable condition controls correctly judged preventability 90.7% of the time. A success rate of 66.3% in the control group may lead one to question whether the unpreventable questions themselves are confusing or more open to interpretation than researchers believed them to be. Future studies or changes to the current design may consider re-evaluating scenarios to ensure that they can be clearly recognized as either preventable or unpreventable, at least within the general control population.

4.2 Discussion of Part 2:

The findings of tests run on participant data from Part 1 of this study provided us with valuable information about how ASD participants differ from controls in their blame assignments

and whether they correctly judge preventability. Although the findings do not entirely align with our hypotheses, the results suggested that ASD participants do not necessarily have trouble distinguishing between preventable and unpreventable cases and making blame assignments accordingly. Rather, they show a deficit in accurately recognizing preventability when compared to neurotypical peers. The aim of this part of the study was to see if this deficit could be ameliorated by presenting participants with an explicit statement about the true nature of the case ("[Agent] could have prevented this" or "There is no way [agent] could have prevented this") and providing them with the opportunity to update their blame assignment.

The second set of chi-square analyses explored whether diagnosis was related to likelihood of participants updating blame assignments. For these analyses, responses were once again separated according to whether questions were actually preventable or not preventable (researcher-determined). Accordingly, each participant had four judgments included in the analysis for preventable cases and four judgments in the analysis for unpreventable cases. We hypothesized that ASD participants would be more likely than controls to choose to update blame assignments in both preventable and unpreventable cases; this prediction was confirmed by the results. As shown in Figure 4, there was a significant difference between ASD participants and controls in choosing to update blame assignments in preventable cases. ASD participants were more likely to update their blame assignments, with 43.2% of these participants selecting the update option, compared to 23.3% of controls who chose to update. This finding aligns with the analyses from part 1, as ASD participants were less likely to correctly identify preventable cases. This finding, however, suggests that facilitating ASD participants' understanding of preventability causes them to update blame assignments, rather than sticking with their initial choice. Similarly, as shown in Figure 5, there was a significant difference between ASD participants and controls in choosing to update blame assignments in unpreventable cases. ASD participants were more likely to update their blame

assignments, with 55.7% of these participants selecting the update option, compared to 19.2% of controls who chose to update. This finding also aligns with the analyses from part 1, as ASD participants were less likely to correctly identify unpreventable cases. Further, it makes sense that the percentage of ASD participants who chose to update is higher in unpreventable cases than preventable cases (55. 7% vs 43.2%) because fewer ASD participants were able to correctly identify unpreventable cases than preventable cases. The findings of these two analyses are especially useful because they confirm that an understanding of preventability can be facilitated. In response to this, one might argue that it is not so obvious that an actual understanding is being facilitated in ASD participants. Perhaps these participants are exhibiting some form of response bias and choosing the option to update their blame assignments simply because they have the option to do so. To see if these results are still observed without the existence of this bias, a revised version of this study might consider including a third condition, in addition to preventable and unpreventable cases, where the second question asking participants "Would you like to change the amount of blame you gave [agent]?" is not preceded by an explicit statement about preventability. Rather, the statement preceding this question would be unrelated to any concept about morality or the judgment at hand. For example, participants might read a question that follows a format such as the following, where the preventability statement is replaced by an irrelevant statement.

> [Screen 1]
> Marissa accidentally took the very last handicapped parking space. There was no sign and the handicap marking on the ground had worn away.
>
> How much blame does Marissa deserve?
>
> [Screen 2]
> Marissa accidentally took the very last handicapped parking space. There was no sign and the handicap marking on the ground had worn away.
>
> **The word "apple" has five letters.** Would you like to change the amount of blame you gave Marissa?

> [Screen 3; if participants selected yes in the previous question]
> Please enter your new answer to the following question: How much blame does Marissa deserve?

The expectation is that participants, both ASD and controls, would not choose to update their blame assignment in the new condition involving such statements. Including this condition would help differentiate between whether the increased likelihood to update in ASD participants is due to an actual understanding of preventability being facilitated or due to a response bias.

The final test (independent samples t-test) aimed to complement the previous finding that ASD participants chose to update their assignments more often than controls and see whether they also differ from controls in their mean difference between initial and new blame assignments. In other words, in preventable and unpreventable cases, of participants who chose to update their blame assignments, we are interested in whether one group experienced a significantly greater mean difference in scores. We hypothesized that diagnosis would have an impact on the mean difference in blame assignment in unpreventable cases only, more specifically that the mean difference in blame assignment would be greater in ASD participants than control. This was supported by the data: there was no significant difference between mean difference in blame assignments between ASD ($M = 10.2$, $SD = 19.6$) and controls ($M = 12.6$, $SD = 21.8$) in preventable cases, but there was a significant difference between mean difference in blame assignments between ASD participants ($M = -41.8$, $SD = 33.8$) and controls ($M = -22.8$, $SD = 20.5$) in unpreventable cases. These findings, shown in Figures 6 and 7, work well with our earlier analyses. Given that ASD participants are likely already assigning higher levels of blame in preventable cases, we don't expect that participants who choose to update their assignment would do so by a large amount. On the other hand, we know that ASD participants have more trouble correctly identifying unpreventable cases. The significant difference here is particularly important because it suggests that the mean difference in blame adjustment is greater in ASD participants, so when ASD participants choose to update in unpreventable cases,

they are decreasing blame by an average of 41.8 points, as compared to controls, who decrease by an average of 22.8 points. These findings do not necessarily suggest that facilitating an understanding of preventability in ASD participants makes them more forgiving than the general population. Since we know that ASD participants assign low blame (like controls do) when they perceive a case to be unpreventable, a more legitimate interpretation of this result might be that ASD participants are already determining these cases to be preventable when they are not, thus assigning higher blame, so when they choose to update based on the "new" information presented about how a case was unpreventable, they immediately decrease their assignment by a large number. Knowing that individuals with ASD tend to significantly decrease blame assignment after learning that a case was unpreventable has implications for understanding offenses committed because of a misunderstanding of preventability or intentionality. For example, if a person with ASD realized that an altercation was not preventable, might that prevent him or her from engaging in violent or criminal activity as a result? Alternatively, having a better sense of how we can facilitate an understanding of preventability, or related concepts in the realm of theory of mind, might have implications for social training and interventions. Both of these avenues will be further discussed in section 5.

4.3 Additional limitations:

It is important to note a few other limitations of this study. First, the small sample size of ASD participants could be viewed as a potential limitation and the results could be difficult to relate to the population of ASD adults. This study could be expanded to include a larger sample of these individuals. Second, as a result of the methods used to recruit control participants (primarily in New Haven, and through social media posts that reached the researcher's circles), a majority of participants were likely Yale University undergraduates and university community members. These controls are not representative of the entire population on many factors, particularly age and

education level. A revised version of this study might consider running control participants through Amazon Mechanical Turk (MTurk), which would generate a more representative sample of the general population. Lastly, due to IRB exemption status and restrictions on compensating participants, this study was kept as short as possible to ensure that participants would complete the entire survey. With more time for data collection, a revised version of this study might consider compensating participants or offering an incentive. This would allow for a third part of the study, which might include measures of sympathy, theory of mind, and/or alexithymia. These measures are typically lengthy (between 30-60 items) but would be useful to explore how these scores predict judgments and blame assignments.

5. Implications & Future Directions

The two most relevant findings from this study in the context of understanding criminal offenses in individuals with ASD are the following: 1) diagnosis matters when it comes to whether an individual can recognize preventable and unpreventable cases, and ASD participants are less likely than controls to correctly identify preventability, and 2) ASD participants tend to significantly decrease blame assignment after they are explicitly told that a case is unpreventable. These deficits, or differences, in moral reasoning, paired with conditions of conflict or ambiguity, may decrease the capacity of individuals with ASD to refrain from violent and criminal offenses (Lerner et al, 2012). For example, an adult with ASD may engage in violent crime because he misunderstands the preventability of an accidental harm directed towards him. The results of our study further our understanding of situations like these, as we can conclude that if the ASD adult was aware of the unpreventable nature of the situation, he would assign less blame to the perpetrator, and as a result (potentially) not engage in any violence.

The ultimate aim of this type of research is to make policy recommendations for changes in our legal system. As our society becomes more neurodiverse, we need to be aware of the differences in

moral reasoning and thinking in adults with ASD, and accommodate for these differences in our evaluation of their culpability in criminal cases. From an evaluative view of law and emotion, our ability to make an informed evaluation of an individual's culpability "rests almost entirely on the degree to which his motives are comprehensible and evident" (Lerner et al, 2012). Adults with ASD, however, are unable to present motives through a normative process of moral reasoning. This raises concerns because this inability to present motives also influences an individual's ability to "communicate effectively with law enforcement officials, to consult with defense attorneys, and, critically, to self-advocate during legal proceedings" (Lerner et al, 2012). Surely, this inability to communicate effectively may lead to officials in the criminal justice system (officers, judges, etc.) to have a hard time making intuitive sense of the ASD defendant's odd motives, which introduces a systematic bias into their evaluation of guilt (Lerner et al, 2012). In order to solve this issue, legal researchers and psychologists should work together to answer the following questions: Is an evaluative view of emotion in criminal law fair for adults with ASD? How can the prosecution and jury erase biases that exist simply because an adult with ASD does not fit into their existing moral schema (biases that ultimately influence their evaluation of guilt)? (Lerner et al, 2012). The future of the legal system can be reformed if there are policies put in place which help prevent adults with ASD from engaging in future criminal activity, consider assigning more lenient punishments, or provide a platform that allows for these adults to self-advocate more effectively.

Lastly, while the most direct application of this research is in the context of the legal system, the findings of this study can also contribute to our more general understanding of social and interpersonal relationships in adults with ASD. The idea that we can teach concepts such as theory of mind in social skills training groups has been previously researched in children with ASD. The results are significant and show that theory of mind training, which involves repetitive instruction about emotions/beliefs and uses positive feedback, improves the social skills of children with ASD

and has an impact on their daily interactions with peers (Adibsereshki et al, 2014). We suggest that extending this sort of training to the moral realm may be beneficial for adults with ASD, as adults encounter and are expected to understand moral situations more often than children are. Training for adults with ASD can include similarly repetitive instruction about moral situations and preventability. Combining this type of intervention training research with the design of our current study, and observing whether blame assignments and recognition of preventability change as a result of this training, may connect back to the legal implications of our research by providing a powerful tool in helping reduce criminal activity in this population, especially if this training is introduced in early adult years (i.e. high school).

Author Contributions

Ladan Mohamed developed the concept for this thesis based on discussions with Fred Volkmar. The study was designed, conducted, and analyzed by Ladan Mohamed under the advisement of Fred Volkmar. The writing was done by Ladan Mohamed, and a draft was read and reviewed by Fred Volkmar, who provided comments and suggestions.

Acknowledgements

I am especially grateful to my advisor, Fred Volkmar, for his constant support and efforts over the past year in helping to make this thesis possible. I would also like to thank Mark Sheskin and Joshua Knobe for their advice and assistance throughout.

References

Adibsereshki, Narges & Nesayan, Abbas & Asadi Gandomani, Roghayeh & Karimlou, Masood. (2015). The Effectiveness of Theory of Mind Training On the Social Skills of Children with High Functioning Autism Spectrum Disorders. Iranian journal of child neurology. 9. 40-9.

Baron-Cohen, S. (1988), AN ASSESSMENT OF VIOLENCE IN A YOUNG MAN WITH ASPERGER'S SYNDROME. Journal of Child Psychology and Psychiatry, 29: 351-360.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. Journal of Autism and Developmental Disorders, 31(1), 5-17.

Barry-Walsh JB, Mullen PE. (2004). Forensic aspects of Asperger's Syndrome, The Journal of Forensic Psychiatry & Psychology, 15:1, 96-107

Bjørkly, S. (2009). Risk and dynamics of violence in Asperger's syndrome: A systematic review of the literature. Aggression and Violent Behavior, 14(5), 306-312.

Bowler DM (1992) "Theory of mind" in Asperger's syndrome. J Child Psychol Psychiatry 33:877–893. Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 2105–2108.

Channon, Shelley & Fitzpatrick, Sian & Drury, Helena & Taylor, Isabelle & Lagnado, David. (2010). Punishment and Sympathy Judgments: Is the Quality of Mercy Strained in Asperger's Syndrome?. Journal of autism and developmental disorders. 40. 1219-26. 10.1007/s10803-010-0980-4.

Cushman, F., Sheketoff, R., Wharton, S., and Carey, S. (2013). The development of intent-based moral judgment. Cognition 127, 6–21. doi: 10.1016/j.cognition. 2012.11.008

Grant, C., Boucher, J., Riggs, K., and Grayson, A. (2005). Moral understanding in children with autism. Autism 9, 317–331. doi: 10.1177/1362361305055418

Greene, Joshua & Haidt, Jonathan. (2003). How (and Where) Does Moral Judgment Work?. Trends in cognitive sciences. 6. 517-523. 10.1016/S1364-6613(02)02011-9.

Kohn, Y & Fahum, T & Ratzoni, G & Apter, Alan. (1998). Aggression and sexual offense in Asperger's Syndrome. The Israel journal of psychiatry and related sciences. 35. 293-9.

Lerner, M. D., Haque, O. S., Northrup, E. C., Lawer, L., & Bursztajn, H. J. (2012). Emerging perspectives on adolescents and young adults with high-functioning autism spectrum disorders, violence, and criminal law. The Journal of the American Academy of Psychiatry and the Law, 40(2), 177–190.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. Psychological Inquiry, 25, 147– 186.

Margoni, F., & Surian, L. (2016). Mental State Understanding and Moral Judgment in Children with Autistic Spectrum Disorder. Frontiers in psychology, 7, 1478. doi:10.3389/fpsyg.2016.01478

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. Journal of Experimental Psychology: General, 146(1), 123-133.

Moran, J. M., Young, Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., et al. (2011). Impaired theory of mind for moral judgment in high-functioning autism. Proc. Natl. Acad. Sci. U.S.A. 108, 2688–2692.

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O., and Convit, A. (2007). Who cares? Revisiting empathy in Asperger syndrome. J. Autism Dev. Disord. 37, 709–715. doi: 10.1007/s10803-006-0197-8

Wechsler, D. (2011).WASI-II: Wechsler Abbreviated Scale of Intelligence (2nd Ed.). Oxford: Pearson Assessment.

Young L, Cushman F, Hauser M, Saxe R (2007). The neural basis of the interaction between theory of mind and moral judgment. Proc Natl Acad Sci USA 104:8235–8240.

Appendix

Part 1 Preventable Cases:
1. Sven accidentally broke the Wright's front window. He was playing baseball too close to the house and accidentally hit a ball through the window.
2. Ryan accidentally set the Andersons' house on fire. He was burning leaves too close to the house and not paying very close attention to the fire which spread to the house.
3. Amanda accidentally misreported her income on her taxes. She did not carefully review her tax information.
4. David accidentally bumped his classmate. David was not watching where he was walking

Part 1 Unpreventable Cases:
1. Matt accidentally killed Frank. Matt unwittingly gave Frank expired medicine because the box did not have an expiration label.
2. Jim accidentally kicked Aaron. Jim had a migraine and could not see clearly in front of him.
3. Katherine accidentally spray-painted what looked like graffiti on Jane's door. The spray-can was mislabeled, and Katherine thought she was spraying sealant on the door.
4. Tommy accidentally left the restaurant without leaving the waiter a tip. The bill indicated that the tip was added to the bill, but it was not.

Part 2 Preventable Cases
1. Brianna accidentally made the infant sick. She forgot that the food contained peanuts, which she knew the baby was allergic to.
2. Bob accidentally damaged a local wildlife preserve. He forgot to read the environmental impact report describing potential damages.
3. Lisa accidentally shot Tom in the arm. She was cleaning the gun, but she forgot to check whether it was loaded.
4. Fred accidentally told his friend the wrong wedding date because he was distracted and wrote down the wrong date.

Part 2 Unpreventable Cases
1. Ted accidentally hit a man with his car. Ted's brakes failed to work.
2. Marissa accidentally took the very last handicapped parking space. There was no sign and the handicap marking on the ground had worn away.
3. Randy accidentally pulled out some of Eve's hair. He ran his hand through her hair, not knowing she recently started chemo treatments.
4. Liz accidentally took a t-shirt out of the store without paying. Someone slipped the shirt in her bag with other clothes that she already paid for.