

**INDIVIDUAL DIFFERENCES IN DEEPPAKE DETECTION:  
MINDBLINDNESS AND POLITICAL ORIENTATION**

A Thesis  
Presented to  
The Academic Faculty

by

Zachary R. Tidler

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science in the  
School of Psychology/College of Sciences

Georgia Institute of Technology  
May 2021

**COPYRIGHT © 2021 BY ZACHARY R. TIDLER**

**INDIVIDUAL DIFFERENCES IN DEEPPFAKE DETECTION:  
MINDBLINDNESS AND POLITICAL ORIENTATION**

Approved by:

Dr. Richard Catrambone, Advisor  
School of Psychology  
*Georgia Institute of Technology*

Dr. Bruce N. Walker  
School of Psychology  
*Georgia Institute of Technology*

Dr. Sidni Justus  
*Oglethorpe University*

Date Approved: December 11, 2020

## **ACKNOWLEDGEMENTS**

I would like to thank the efforts of several undergraduate research assistants for their help in developing the study materials: Nidhi Pai, Adam Snoll, Kyle Walker, Sara Ferez, Nilay Mehta, Srikar Sajja, Srane Bayapureddy, and Greg Varghese. I offer thanks to my graduate student colleagues for engaging in brainstorming sessions with me when I know that they really just wanted to enjoy a glass of wine: Sibley Lyndgaard, Corey Tatel, MacKenzie Hughes, Julie Harrison, Jason Tsukahara, and Cody Mashburn. I would also like to thank my advisor, Richard Catrambone, and my committee members, Sidni Justus and Bruce Walker, for their guidance. I am especially grateful to my family for the sacrifices they have made to facilitate my scientific pursuits.

## **TABLE OF CONTENTS**

<b>ACKNOWLEDGEMENTS</b>	<b>iii</b>
<b>LIST OF TABLES</b>	<b>vi</b>
<b>LIST OF FIGURES</b>	<b>vii</b>
<b>LIST OF SYMBOLS AND ABBREVIATIONS</b>	<b>viii</b>
<b>SUMMARY</b>	<b>ix</b>
<b>CHAPTER 1. Introduction</b>	<b>1</b>
1.1 The Uncanny Valley	2
1.2 Bicameralism, Theory of Mind, Mindblindness, and Autism	4
1.3 Mindblindness, Political Orientation, and Susceptibility to Deception	6
1.4 The Present Study	8
1.4.1 Hypotheses	8
1.4.2 Exploratory Analyses	9
<b>CHAPTER 2. Method</b>	<b>11</b>
2.1 Sample	11
2.1.1 Recruitment	11
2.1.2 Compensation	12
2.2 Measures	12
2.2.1 Affect Detection Measures	12
2.2.2 Political Orientation Measures	14
2.2.3 Autism Spectrum Disorder Measure	14
2.2.4 Deepfake Detection Task	15
2.2.5 Exploratory Measures	17
2.2.6 The Assessing Emotions Scale (EQ)	17
2.2.7 Sensory Perception Quotient (SPQ)	17
2.2.8 Conspiratorial Thinking Scale (CT)	17
2.3 Variables Defined	18
2.3.1 Affect Detection Ability (Theory of Mind Ability)	18
2.3.2 Political Orientation	18
2.3.3 Autism Spectrum Traits	19
2.3.4 Deepfake Detection Ability	20
2.4 Procedure	20
<b>CHAPTER 3. Results</b>	<b>22</b>
3.1 Demographic Effects	22
3.1.1 Gender	22
3.1.2 Age	22

<b>3.2 Hypothesis 1: Affect Detection Performance Correlates with Deepfake Detection Performance</b>	<b>22</b>
<b>3.3 Hypothesis 2: Political Orientation Correlates with Affect Detection Performance</b>	<b>23</b>
<b>3.4 Hypothesis 3: Political Orientation Correlates with Deepfake Detection Performance</b>	<b>25</b>
3.4.1 The Incremental Predictivity of Political Orientation	26
<b>3.5 Hypothesis 4: Deepfake Detection Performance Correlates with ASD Quotient</b>	<b>27</b>
<b>3.6 Exploratory Analyses</b>	<b>29</b>
3.6.1 AQ	30
3.6.2 Conspiratorial Thinking	30
3.6.3 Sensory Perception	30
3.6.4 EQ	30
<b>CHAPTER 4. Discussion</b>	<b>34</b>
<b>4.1 Mindblindness, Political Orientation, and Deepfake Detection</b>	<b>34</b>
<b>4.2 ASD and Deepfake Detection</b>	<b>35</b>
<b>4.3 Applications</b>	<b>36</b>
<b>4.4 Limitations and Extending the Work</b>	<b>37</b>
<b>4.5 Conclusions</b>	<b>40</b>
<b>APPENDIX A. Adult Eyes TEst – Example Item</b>	<b>42</b>
<b>APPENDIX B. Adult Faces TEst – Example Item</b>	<b>43</b>
<b>APPENDIX C. Cambridge Mind Reading TEst – Example Item</b>	<b>44</b>
<b>APPENDIX D. Reading the mind in Film TEst – Example Item</b>	<b>45</b>
<b>APPENDIX E. SOCIAL AND ECONOMIC CONSERVATISM SCALE – Example Items</b>	<b>46</b>
<b>APPENDIX F. SAPPLY political compass test – Example Items</b>	<b>47</b>
<b>APPENDIX G. AUTISM Spectrum quotient – Example Items</b>	<b>48</b>
<b>APPENDIX H. Deepfake Detection task – Example Item</b>	<b>49</b>
<b>APPENDIX i. Assessing Emotions Scale – Example Items</b>	<b>50</b>
<b>APPENDIX j. Sensory perception quotient – Example Items</b>	<b>51</b>
<b>APPENDIX K. Conspiratorial Thinking Scale – Example Items</b>	<b>52</b>
<b>REFERENCES</b>	<b>53</b>

## LIST OF TABLES

TABLE 1 - CORRELATIONS AMONG AFFECT DETECTION MEASURES AND THE DEEPFAKE DETECTION TASK	23
TABLE 2 - CORRELATIONS AMONG POLITICAL ORIENTATION MEASURES AND AD ABILITY	24
TABLE 3 - CORRELATIONS AMONG POLITICAL ORIENTATION MEASURES AND DFD PERFORMANCE	25
TABLE 4 - CORRELATIONS AMONG AUTISM SPECTRUM QUOTIENT (AQ) SCALES AND DFD	27
TABLE 5 - CORRELATIONS AMONG AFFECT DETECTION MEASURES, THE AQ, AND THE DEEPFAKE DETECTION TASK	29
TABLE 6 - CORRELATIONS AMONG EXPLORATORY MEASURES AND CRITERION VARIABLES	31
TABLE 7 - DESCRIPTIVE STATISTICS FOR ALL MEASURES	32

## **LIST OF FIGURES**

FIGURE 1 - STATIC EXAMPLE OF A DEEPPFAKE.	1
FIGURE 2 - UNCANNY VALLEY DEMONSTRATION.	3

## **LIST OF SYMBOLS AND ABBREVIATIONS**

ASD    AUTISM SPECTRUM DISORDER

TOM    THEORY OF MIND

DFD    DEEPPAKE DETECTION

AD    AFFECT DETECTION

AQ    AUSTISM SPECTRUM QUOTIENT

AET    ADULT EYES TEST

AFT    ADULT FACES TEST

RTMIF    READING THE MIND IN FILM

CAM    CAMBRIDGE MIND READING

SEC    SOCIAL AND ECONOMIC CONSERVATISM SCALE

EQ    THE ASSESSING EMOTIONS SCALE

SPQ    SENSORY PERCEPTION QUOTIENT

CT    CONSPIRATORIAL THINKING



## SUMMARY

The proliferation of the capability for producing and distributing deepfake videos threatens the integrity of systems of justice, democratic processes, and the general ability to critically assess evidence. This study sought to identify individual differences that meaningfully predict one's ability to detect these forgeries. It was hypothesized that measures of affect detection (theory of mind ability) and political orientation would correlate with performance on a deepfake detection task. Within a sample ( $N = 173$ ) of college undergraduates and participants from Amazon's Mechanical Turk platform, affect detection ability was shown to correlate with deepfake detection ability,  $r(171) = .73, p < .001$ , and general orientation to the political left was shown to correlate with deepfake detection ability,  $r(171) = .42, p < .001$ . Stronger correlations with deepfake detection ability were observed among specific facets of political orientation: economic liberalism,  $r(171) = .40, p < .001$ , and social progressivism,  $r(171) = .57, p < .001$ . Political orientation was shown to add incrementally predictivity in a model that included both, political orientation and affect detection as predictors of deepfake detection ability. The deepfake detection task was also assessed as a predictor of an autism spectrum disorder screening instrument,  $r(171) = -.23, p < .001$ . The results of this study serve to identify populations who are particularly susceptible to deception via deepfake video and to inform the development of interventions that may help defend the vulnerable from nefarious attempts to influence them.

## CHAPTER 1. INTRODUCTION

Imagine you have been granted the power to compel any person you choose to say anything you choose and project that statement to the world. If you were so inclined, what sort of chaos could you cause? You could force a US President to announce a military directive. You could make an Iranian Ayatollah issue a dangerous fatwa. You could ruin a rival co-worker's reputation by forcing them to announce allegiance to a hate-group. You could compel a trusted newscaster to warn of an inbound nuclear missile. A similar power has suddenly been made available to anyone with a personal computer via the proliferation of machine learning software that can produce doctored videos called deepfakes (Figure 1).



**Figure 1 - *Static Example of a Deepfake.*** Which side do you think is doctored? Saturday Night Live fans will recognize the image on the right as comedian Kate McKinnon's portrayal of Senator Elizabeth Warren. The image on the left has been doctored with deepfake technology to put the real Senator Warren's face on McKinnon's body.

The applications of this software class range from innocuous uses, such as novelty mobile apps that can age and de-age a subject, to more sinister uses, like superimposing someone's face into a pornographic film in order to extort them.

This present study was conducted during a moment in history that is particularly vulnerable to disruption by deepfakes, the 2020 US Presidential Race. The 2020 election season is special for several reasons. First, the previous election season was marked with both foreign and domestic attempts to use social media to distribute fallacious information in order to influence the outcome (Mueller, 2019). Second, it comes at a time when a strong social justice movement has primed the national zeitgeist to be particularly intolerant of offensive speech (Lukianoff & Haidt, 2019). Third, the technology to produce these deepfakes has matured to the point that the videos are nearly indistinguishable from their undoctored sources, and production capacity has proliferated to anyone with a moderately powerful computer. The magnitude of this threat is self-evident.

Efforts to mitigate the threat have largely been undertaken by the computer science community in the form of algorithmic instruments that detect various artefacts of the deepfake development process (Albahar & Almalki, 2019). While technological solutions are promising, it is equally important that attention is devoted to understanding individual differences in ability to detect deepfakes and techniques for bolstering this ability.

## **1.1 The Uncanny Valley**

If you have ever been to a theme park whose attractions feature animatronic characters, you may have noticed yourself experiencing an eerie feeling at the sight of characters that emulate human likeness (See Figure 2).



**Figure 2 - Uncanny Valley Demonstration. If this picture makes you uneasy, there could be a number of explanations. Among them is the possibility that you are experiencing the phenomenon known as the uncanny valley. This is the animatronic Donald Trump at Disney World's Hall of Presidents attraction.**

Mori (1970) observed this phenomenon and named it the uncanny valley. The term, uncanny valley, conjures images of a dystopian canyon inhabited by robotic abominations, but it actually refers to a curvilinear depression in the graphic representation of the function that describes people's ratings of affinity for artificial humans (digital or physical) as their likeness approaches high fidelity. Although there are several theories that attempt to explain the cognitive underpinnings of the uncanny valley (Green et al., 2008; MacDorman & Ishiguro, 2006; Moosa & Ud-Dean, 2010; Ramey, 2005) the conflicting perceptual cues explanation (Ferrey et al., 2015) is most relevant to the central hypothesis of the present study. Ferrey and colleagues suggest that when we encounter a simulated human of a certain level of fidelity, we are simultaneously given cues that indicate belongingness to a human category and cues that indicate belongingness to some non-human category. The

conflict of these cues generates a psychological tension that is akin to Festinger's (1957) notion of cognitive dissonance.

It is worth noting that the uncanny valley is a controversial topic, and some have put forth arguments which question the scientific validity of the construct. Perhaps the most prominent of these is that suggestion that the effect is not unique to objects on a spectrum of human likeness but rather a general discomfort that arises when categorizing objects whose appearances are a certain distance from a categorical exemplar (Hanson et al., 2005). However, irrespective of the theoretical explanation for the uncanny valley as a within-person effect, there is room for between-person variance on the magnitude of the effect. Sensitivity to the uncanny valley phenomenon has been shown to vary with an individual's placement on certain personality dimensions (MacDorman & Entezari, 2015) but, more importantly for the present study, sensitivity has been shown to vary significantly between typically developing children and those diagnosed with autism spectrum disorder (ASD) (Feng et al., 2018).

## **1.2 Bicameralism, Theory of Mind, Mindblindness, and Autism**

In Julian Jaynes' 1976 book "The Origin of Consciousness in the Breakdown of the Bicameral Mind", he hypothesizes that consciousness or subjectivity, which Jaynes refers to as "the analogical I", is not a necessary feature of being human, but rather an epiphenomenal consequence of the development of language, specifically that of metaphor. He claims that prior to a certain period in history, the human experience (or perhaps the lack thereof) was that of an automaton, dutifully obeying the hallucinated commands of external masters or gods. In reference to the dual-chambered nature of their mind, Jaynes

called these ancient automatons “bicameral man”. The lack of an analogical I was also extended to observations of others. According to Jaynes, ancient humans did not have an introspective subjective experience, nor did they extrospectively attribute subjectivity to others. The bicameral mind hypothesis was famously critiqued by philosopher, Ned Block (1977), but it was also praised for its audacity and captured the imaginations of many in the field. Two years after the publication of Jaynes’ book, Premack and Woodruff (1978) would come to describe a psychological phenomenon that is similar to that which Jaynes’ bicameral man lacked, called theory of mind (TOM). They described an individual who has TOM as someone who “imputes mental states to himself and others”. In subsequent years TOM has been operationalized through tasks that measure one’s ability to attribute false beliefs to characters in vignettes (e.g., hypothesizing where a character thinks an object is as opposed to where the object actually is) (Wimmer & Perner, 1983) and affect detection tasks, in which participants are asked to identify the emotions felt by the subjects in pictures, movies, and voice recordings. These measures have allowed researchers to rank-order individuals by the quality of their TOMs.

One can imagine the implications of successful and unsuccessful theorizing about the thoughts, feelings, and motivations of others. As is the case with all theories, good ones allow for accurate predictions. TOMs are no exception. Sound TOMs can facilitate prediction and control in one’s social sphere and pathologically deficient TOM ability has been proposed as a significant component of ASD (Baron-Cohen, 1997, 2000, 2001). Baron-Cohen and colleagues called this pathology, mindblindness, and it has been proposed that tests for mindblindness be used as ASD screening instruments. The finding that individuals can be mindblind relates back to claims that Jaynes made regarding modern

humans' relationship to bicameral humans. Jaynes believed that those experiencing schizophrenia were, in a sense, regressing back to a more bicameral state (e.g., obeying hallucinated commands). It is not then all too surprising that, much like individuals diagnosed with ASD, individuals diagnosed with schizophrenia do, in fact, score lower on TOM measures than their psychologically typical counterparts (Bora et al., 2009; Brüne, 2005; Sprong et al., 2007).

### **1.3 Mindblindness, Political Orientation, and Susceptibility to Deception**

Although the cause is in dispute, ASD diagnoses are on the rise (Durkin et al., 2017; Matson & Kozlowski, 2011; Nevison et al., 2018; Smiley et al., 2018). Competing interpretations of this phenomenon are that either something is changing in the world causing people to be more mindblind or improved diagnostic capabilities are allowing the identification of extant mindblindness more readily (It should be noted that I do not mean to imply that TOM deficiency is the *Sine Qua Non* of ASD, although it does appear to be an important component.) The point in noting this prevalence is to suggest that, in a representative democracy, there may be political implications if a substantial proportion of the population is deficient in appropriately attributing mental states to other people. Further support for this possibility comes from a body of literature connecting a particular type of attentional cue (called a *gaze cue*) to ASD and political orientation. Imagine you are speaking with a friend on a street corner. You are lost in conversation, but you suddenly notice that your friend has diverted her gaze from your face to the street behind you. You also notice that her eyes have widened tremendously, indicating fear. In a split second, you are able to guess that she has seen a car veer off-course and head straight for you. The information you gleaned from your friend's eyes was communicated to you through gaze

cues (Moore et al., 2014). It is important to note that, in our thought experiment, you did not only draw information about the location of the car from her gaze. You were able to attribute to your friend an affective state of fear. In other words, gaze cues can inform affective TOMs (Bayliss et al., 2007; Moore et al., 2014). However, there are people who might not have been able to gather the same information that you did from your friend's gaze cues. It has been shown that individuals who score highly on indicators of ASD are particularly insensitive to gaze cueing effects (Bayliss & Tipper, 2005). The connection between sensitivity to gaze cues and political orientation was revealed by Dodd et al (2011). Dodd and colleagues hypothesized that orientation to the political right is associated with a high valuation of social independence (i.e., not being influenced by other people). Their work revealed that those oriented to the political left differed significantly from those oriented to the political right on the magnitude of influence gaze cues produced on reaction time. A continuation of this line of research by Carraro et al (2015) found that the effect is present in a gaze cueing task but not in an arrow cueing task, further highlighting the relationship between attention to social cues and political orientation. The interpretation of the finding is made difficult by a temporal precedence problem. Does the tendency to disattend to gaze cues produce mindblindness and orientation to the political right, or does orientation to the political right cause individuals to actively ignore social influence? Evidence of individual differences in gaze cue effects in early infancy suggest that the former is the more likely the case (Mundy et al., 2007; Simpson et al., 2016).

The thrust of this discussion on the relationship between ASD; mindblindness; disattention to gaze cues; and political orientation, is the proposal that TOM deficiencies are significantly concentrated on the political right and leave people susceptible to



deception/influence via deepfake videos. The argument is further supported by a finding that ASD diagnostic scores (in the direction of psychological typicality) were strongly associated with performance on a lie-detection task (Williams et al., 2018). However, this study also found that TOM (affect detection) tasks were not significantly correlated with the lie detection task. It is important not to confuse this result as evidence contradictory to the present hypotheses. The key differentiator is in the nature of each dependent measure. In Williams et al (2018) participants were asked to spot lies within videos of real people. In the present study participants will be asked to identify computer generated approximations of real people.

#### **1.4 The Present Study**

To reiterate, the aim of the present study is two-fold. The primary aim is to assess the power of TOM (i.e., affect detection) ability and political orientation to predict performance on a deepfake Detection (DFD) task. The secondary aim is to generate preliminary evidence that the DFD task is useful as an ASD screening instrument.

##### *1.4.1 Hypotheses*

##### 1.4.1.1 Hypothesis 1: Affect Detection Performance Correlates with Deepfake Detection Performance

A moderate to high correlation ( $r = .3 - 1.0$ ) (Cohen, 1988) was expected between a factor (composite) score of performance on the 4 affect detection (AD) measures and performance on the DFD task. The factor score was derived by computed a regression

weighted composite of 4 AD measures (DiStefano et al., 2009). This will henceforth be implicit when I refer to factor scores.

#### 1.4.1.2 Hypothesis 2: Political Orientation Correlates with Affect Detection

##### Performance

A moderate to high correlation was expected between political orientation factor score and an AD factor score. Specifically, it was expected that those aligned to the political left will perform better than those aligned to the political right. The operationalization of alignment to the political left is described at length in the “measures” section.

#### 1.4.1.3 Hypothesis 3: Political Orientation Correlates with Deepfake Detection

##### Performance

A moderate to high correlation was expected between political orientation factor scores and performance on the DFD task. Specifically, it was expected that those aligned to the political left would perform better than those aligned to the political right.

#### 1.4.1.4 Hypothesis 4: Deepfake Detection Performance Correlates with ASD Quotient

A moderate negative correlation ( $r = -.3$ ) was expected between performance on the DFD task and a measure of ASD related traits. That is, those who express fewer ASD related traits would perform better on the DFD task. A multiple regression was also performed to determine if DFD ability is incrementally predictive of the ASD quotient in the context of the established affect detection measures.

### 1.4.2 *Exploratory Analyses*

#### 1.4.2.1 Incremental Predictivity of Political Orientation

Hypotheses 1, 2, and 3 address the potential bivariate relationships between AD ability, DFD ability, and political orientation. However, in addition to an assessment of the bivariate relationships it was important to examine these relationships in the context of one another. Given that it seems reasonable to assume that the development of affect detection ability precedes the development of political orientation in one's lifetime, it is conceivable that a relationship between political orientation and DFD ability might become negligible when DFD ability is regressed on political orientation while controlling on AD ability.

#### 1.4.2.2 Conspiratorial Thinking, Sensory Perception Quotient, and Emotional Intelligence

Correlations between DFD, AD and conspiratorial thinking, sensory perception ability, and emotional intelligence were all measured with the intent of generating evidence to support the worthiness of future work.

## CHAPTER 2. METHOD

### 2.1 Sample

#### 2.1.1 Recruitment

An a priori power analysis indicated that the minimum sample size required to achieve  $1 - \beta = .9$  in the presence of moderate correlations was 113. However, given the inclusion of some exploratory measures, a 15 participant-per-variable rule was chosen as a minimum (Cohen, 1988). The final sample (after 5 participants were excluded due to having completed less than 50% of the study) consisted of 173 participants. Of these, 96 were undergraduates at The Georgia Institute of Technology and 77 were recruited from Amazon's Mechanical Turk platform (henceforth referred to as "Mturk"). The former group was recruited through The Georgia Institute of Technology's research participant management platform, *SONA* (*this group of participants will henceforth be referred to as "the Georgia Tech group"*).

##### 2.1.1.1 Georgia Tech Group

The average age of this group ( $n = 96$ ) was 19.39 years old and participants ranged from 18 to 27 years old. There were 52 males, 41 females, 2 gender variant/non-conforming, and 1 participant preferred not to answer. The self-reported political orientation of this group was as follows: 6 conservatives, 38 progressives, 35 moderates, 2 who reported "other", and 15 preferred not to answer. The highest education level obtained by this group was as follows: 90 were working towards their 1<sup>st</sup> college degree, 1 had an associate degree, and 5 already had 1 bachelor's degree.

#### 2.1.1.2 Mturk Group

The average age of this group ( $n = 77$ ) was 34.31 years old and participants ranged from 23 to 64 years old. There were 42 males and 35 females. At the time of data collection, 61 of these participants were located in the United States, 12 were located in India, 3 were located in Cuba, and 1 was located in Brazil. Self-reported political orientation of this group was as follows: 39 conservatives, 16 progressives, 21 moderates, and 1 who reported “other”. The highest education level obtained for this group was as follows: 5 had a high school diploma or GED equivalent, 7 had some college but no degree, 47 had a bachelor’s degree, and 18 had a master’s degree.

The inclusion of both of the recruitment sources served to not only expedite data collection but also to ameliorate attenuated correlations due to a restricted range of talent. The Georgia Institute of Technology student body is subject to strong selection pressures and whenever possible, should be supplemented with participants from a more general population.

#### 2.1.2 *Compensation*

Participants from the Georgia Tech group were compensated with 2 research credits which satisfied requirements for various psychology courses at Georgia Tech. Participants from the Mturk platform were compensated with \$5 USD.

## 2.2 **Measures**

### 2.2.1 *Affect Detection Measures*

#### 2.2.1.1 Adult Eyes Test (AET)

Participants are presented with 36 images of the eyes of people who are experiencing a range of emotions. The participants are asked to respond with the emotion they believe to be represented from 4 options (Baron-Cohen, Jolliffe, et al., 1997). (See Appendix A for sample item)

#### 2.2.1.2 Adult Faces Test (AFT)

Participants are presented with 20 images of the faces of people experiencing a range of emotions. The participants are asked to respond with the emotion they believe to be represented from 2 options (Baron-Cohen, Wheelwright, et al., 1997). (See Appendix B for sample item)

#### 2.2.1.3 Cambridge Mind Reading Test (CMR) (Video Only)

Participants are presented with 50 video clips containing faces of people experiencing a range of emotions. The participants are asked to respond with the emotion they believe to be represented from 4 options (Golan, Baron-Cohen, & Hill, 2006) (See Appendix C for sample item)

#### 2.2.1.4 Reading Minds in Film Test (RMIF)

Participants are presented with 22 scenes from obscure dramatic films in which characters experience a range of emotions. The participants are asked to respond with the emotion they believe to be represented from 4 options (Golan, Baron-Cohen, Hill, et al., 2006) (See Appendix D for sample item)

## 2.2.2 *Political Orientation Measures*

### 2.2.2.1 Social and Economic Conservatism Scale (SEC)

A 12-item self-report assessment that asks participants to respond to political issues on a 100-point “feelings thermometer”(Yilmaz & Saribay, 2017). A high score on this measure is indicative of a conservative political orientation while a low score on this measure is indicative of a progressive political orientation. (See Appendix E for sample items)

### 2.2.2.2 Supply Political Compass Test

A 46-item test that places examinees on 3 dimensions of political orientation: economic liberalism, authoritarianism, and social progressivism. Participants respond on a 5-point Likert-type scale (*Political Compass Project*, n.d.). A high score on the economic liberalism subscale is indicative of alignment to the economic left while a low score is indicative of alignment to the economic right. A high score on the authoritarianism subscale is indicative of a tendency toward authoritarian governance style while a low score on this subscale is indicative of tendency toward libertarian governance style. A high score on the social progressivism subscale is indicative of a socially progressive orientation while a low score on this subscale is indicative of a socially conservative orientation. (See Appendix F for sample items)

## 2.2.3 *Autism Spectrum Disorder Measure*

### 2.2.3.1 The Autism Spectrum Quotient (AQ)

A 50-item questionnaire that places examinees on 5 trait dimensions associated with ASD: social skill, attention switching, attention to detail, communication, and imagination (Baron-Cohen, 2001). (See Appendix G for sample items)

#### *2.2.4 Deepfake Detection Task*

This is an internally developed task in which participants watch a series of video clips and are asked to indicate whether the human subject of the clip is authentic or the product of deepfake software. The clips (some authentic and some altered) come from a deepfake dataset released publicly by Google, a deepfake generation application called impressions.app, and obscure clips found on Instagram (Rossler et al., n.d.). (See Appendix H for sample item)

Because the measure consisted of extant deepfake videos, there was no opportunity to calibrate the difficulty of each item. This raised the concern that any given video would be either, so obviously inauthentic that participants of all abilities would be able to correctly identify it as such, or so indistinguishable from authentic that participants would be able to correctly identify it at a rate no better than chance. If this were the case, there may not have been sufficient variance in performance upon which to conduct meaningful analyses. To preempt this eventuality, a pilot study was run using participants from Mturk. If the aforementioned pattern of results was observed in the pilot study, the obviously inauthentic videos would be replaced with less obvious versions. Originally three versions of each item were developed, one in which the subject of the clip's entire torso was visible, one which was zoomed in so that only the subject of the clip's face was visible, and one which was zoomed in so that only the subject of the clip's eyes were visible. The logic of developing



these three versions loosely followed that of the development of the Adult Eyes Test in which it was hypothesized that presenting images of eyes alone would offer less affective information and precipitate more response variance in neurotypical participants (Baron-Cohen, Jolliffe, et al., 1997). Stated plainly, it was assumed that difficulty in identifying a deepfake would increase with the extent of zooming-in. Initially, it was planned that three groups of Mturk participants ( $n = 25$ ,  $N = 75$ ) would be recruited and each would complete a different version of the task (torso, face, or eyes). If a particular item was correctly identified above a certain frequency threshold (above 90%), that item would be replaced with its corresponding zoomed-in version. However, after the first group of Mturk participants ( $n = 25$ ) completed the torso version of the task (which was thought to be the most difficult version) it was determined that no further piloting would be necessary. The task was internally consistent ( $\alpha = .83$ ) and any items which were answered correctly above the 90% threshold were retained as attention checks.

The final task consisted of 62 items. Of these, 32 were deepfakes and 30 were authentic. For each item, the participant watched a 5 to 8 second video clip. The participant was then asked to indicate whether they believed the video was authentic or a deepfake and how confident they were in their response on a 4-level Likert-type scale (not confident at all, slightly confident, somewhat confident, very confident). The task was scored by coding correct responses as 1 and incorrect responses as -1. The response codes were then multiplied by the confidence rating (coded as 1-4) and scores on all items were summed. Based on the Mturk pilot, items were categorized as easy (items in the lower 3<sup>rd</sup> of the pilot sample's aggregated score), medium (items in the middle 3<sup>rd</sup> of the pilot sample's aggregated score), or hard (items in upper 3<sup>rd</sup> of the pilot sample's aggregated score). To

avoid order effects, the items were divided into 4 blocks that each contained a roughly equal number of easy, medium, and hard items (the blocks were not exactly the same size because 62 is not divisible by 4). Blocks were presented in random order and items within each block were also randomized. At the end of the task, participants were offered the opportunity to explain what features of the videos they may have been focusing on as clues to reveal the authenticity of the video. This question was answered in an open-ended format. The final task exhibited similar internal consistency to that of the pilot administration ( $\alpha = .83$ ).

#### *2.2.5 Exploratory Measures*

The following measures were administered in an exploratory fashion with the intent of generating evidence to support the worthiness of future work.

#### *2.2.6 The Assessing Emotions Scale (EQ)*

A 32-item self-report assessment that taps participants' impressions of their own emotional intelligence (Schutte et al., 2009). (See Appendix I for sample Items)

#### *2.2.7 Sensory Perception Quotient (SPQ)*

A 40-item self-report assessment that asks participants to indicate the degree to which they agree with statements about their sensory and perceptive experience on a 4-point Likert-type scale (Tavassoli et al., 2014). (See Appendix J for sample items)

#### *2.2.8 Conspiratorial Thinking Scale (CT)*

A 15-item questionnaire that asks participants to rate the truthfulness of statements regarding conspiracies on a 5-point Likert-type scale (Brotherton et al., 2013). Example statement: The power held by heads of state is second to that of small unknown groups who really control world politics. (See Appendix K for sample items)

## **2.3 Variables Defined**

### *2.3.1 Affect Detection Ability (Theory of Mind Ability)*

Affect detection (AD) ability is operationally defined as the participant's performances on the four AD measures aggregated into a regression weighted factor score (DiStefano et al., 2009) (Individual measures are described in detail in the "Measures" section below). To derive the factor score, scores on the four component measures were factor analyzed (Promax rotated, maximum likelihood extraction method, retaining eigenvalues greater than 1) to establish the contribution weight of each toward a latent AD factor. The single factor solution explained 80% of the variance in the component measures. The component loadings were as follows: AET, .9; AFT, .71; CMR, .96; RTMIF, .85. Then, each participant's scores on the component measures (Adult Eyes Test, Adult Faces Test, Cambridge Mind Reading Test, and the Reading the Mind in Film Test) were weighted based upon regression coefficients in an equation that treats the component measures as predictor variables and the latent factor as the criterion variable. Weighted scores were then summed into a single value, representing an AD factor score. In addition to the regression weighted factor score, scores on the component measures were also considered as unique predictor variables.

### *2.3.2 Political Orientation*

Political orientation was operationalized in a fashion similar to that of AD. In addition to serving as unique predictor variables, the SEC scale and the 3 subscales of the Supply Political Compass test (economic liberalism, authoritarianism, and social progressivism) were treated as components of a regression weighted factor score. This dimension reduction should not be taken as an assertion on behalf of the author that it is necessarily appropriate to treat political orientation as a unidimensional construct. This procedure is only a way to simplify (perhaps oversimplify) the correlational analyses that would follow. In several cases the relationships between the components of political orientation and the criterion variables may be more illuminating than the political orientation factor score. Scores on the component measures were factor analyzed (unrotated principle axis factoring, retaining eigenvalues greater than 1) and a two-factor solution was extracted that accounted for 79% of the variance in the component measures. However, due to the pattern of loadings on the first factor (SEC, -.72; economic liberalism, .81; authoritarianism, .12; and social progressivism, .58), only the first factor was used to define the political orientation factor score. Only authoritarianism loaded substantially (.47) on the second factor so it was decided that the manifest authoritarianism score was sufficient to capture the construct. The political orientation factor (the first factor) accounted for 49% of the variance in the component measures.

### *2.3.3 Autism Spectrum Traits*

Autism spectrum traits are defined as the participants score on the Autism Spectrum Quotient (AQ) test (Baron-Cohen, 2001). At a more granular level, the AQ contains 5 trait subscales. These include social skill, attention switching, attention to detail,

communication, and imagination. Total AQ and the subscales will each be considered as predictor variables.

#### *2.3.4 Deepfake Detection Ability*

Deepfake detection (DFD) ability was operationalized as the participant's performance score on the DFD task.

### **2.4 Procedure**

Participants were simultaneously recruited via announcements in general psychology courses at The Georgia Institute of Technology and HIIT (Mturk's name for a work opportunity) request on Mturk. Interested participants signed up for the study via Mturk or the SONA participant management platform and were immediately provided a link that directed them to the study tasks and measures. The tasks and measures were implemented on the Qualtrics survey platform. The survey began with a briefing statement and a request for consent. Consenting participants then began the demographic portion of the survey. After responding to basic demographic items (age, gender, location, self-report political orientation, education level) all the remaining measures and tasks were administered. The order of these measures and tasks was randomized but all items within a given measure/task were administered in the order commensurate with their original development. Three attention check items were randomly distributed throughout the measures. These items explicitly instructed the participant to select a certain response. If the response was not selected, that participants survey immediately ended and their data were not collected. The study took 85.04 minutes to complete on average with the Georgia Tech group taking slightly longer (89.09 minutes on average) than the Mturk group (80 minutes). Three

participants were excluded from the calculation of average study duration as it was clear that they finished the study and left the window open without submitting their results, causing their completion time to be unrealistically large.

## CHAPTER 3. RESULTS

### 3.1 Demographic Effects

#### 3.1.1 Gender

A significant gender effect was observed only in conspiratorial thinking,  $t(168) = -4.130$ ,  $p < .001$ ,  $d = .63$ . Women ( $M = 49.88$ ,  $SD = 10.5$ ) scored significantly higher than men ( $M = 42.35$ ,  $SD = 13.28$ ), indicating that women in the sample displayed a stronger tendency toward conspiratorial thought. The sample contained an insufficient number of gender variant/non-confirming participants ( $n = 2$ ) to draw meaningful comparisons.

#### 3.1.2 Age

Significant correlations were observed between age and affect detection ability,  $r(171) = -.33$ ,  $p < .001$ , deepfake detection performance,  $r(171) = -.33$ ,  $p < .001$ , and progressivism,  $r(171) = -.39$ ,  $p < .001$ .

### 3.2 Hypothesis 1: Affect Detection Performance Correlates with Deepfake Detection Performance

Demonstrated in Table 1, strong correlations (Cohen, 1988) were observed between each of the predictor variables (the four AD measures and the AD factor score) and the criterion measure (DFD task). These results support hypothesis 1 by demonstrating a strong relationship between affect detection ability and performance on the deepfake detection task at both the manifest and latent level.

**Table 1 - Correlations Among Affect Detection Measures and the Deepfake Detection Task**

Measures	1	2	3	4	5	6
1. Adult Eyes Test	—					
2. Adult Faces Test	.62***	—				
3. Cambridge Mind Reading	.85***	.68***	—			
4. Reading the Mind in Film	.75***	.60***	.80***	—		
5. Affect Detection Composite	.92***	.73***	.98***	.86***	—	
6. Deepfake Detection Ability	.68***	.54***	.71***	.62***	.73***	—

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

### **3.3 Hypothesis 2: Political Orientation Correlates with Affect Detection Performance**

As displayed in Table 2, the AD factor score correlated moderately with economic liberalism and the political orientation factor score; and correlated strongly with social progressivism.



**Table 2 - Correlations Among Political Orientation Measures and AD ability**

Measures	1	2	3	4	5	6
1. SEC Scale <sup>a</sup>	—					
2. Economic Liberalism	-.56***	—				
3. Authoritarianism	.05	.14	—			
4. Social Progressivism	-.40***	.54***	.08	—		
5. Political Orientation <sup>b</sup>	-.82***	.91***	.11	.64***	—	
6. Affect Detection	-.19*	.42***	.22***	.70***	.44***	—

*Note:* <sup>a</sup>SEC Scale = *Social and Economic Conservatism Scale*. <sup>b</sup>Political Orientation = *Latent Political Orientation Factor Score*.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

A significant main effect of self-reported political orientation on affect detection ability was also observed,  $F(4,168) = 14.83$ ,  $p < .001$ ,  $\eta_p^2 = .26$ . Pairwise comparisons at an alpha level of .005 (after Bonferroni correction) showed significant mean differences between conservatives ( $M = -.81$ ,  $SD = .8$ ) and progressives ( $M = .37$ ,  $SD = .93$ ), moderates ( $M = .17$ ,  $SD = .89$ ), those who reported “other” ( $M = 1.06$ ,  $SD = .07$ ), and those who

preferred not to answer ( $M = .24$ ,  $SD = .45$ ). No significant mean differences were observed in any of the other pairwise comparisons.

The results support the prediction that orientation to the political right is predictive of deficiency in affect detection ability. However, at a more granular level, the relatively weak predictivity of the SEC scale and trait authoritarianism was not expected.

### 3.4 Hypothesis 3: Political Orientation Correlates with Deepfake Detection Performance

As displayed in Table 3, performance on the deepfake detection task correlated moderately with economic liberalism and the political orientation factor score; and correlated strongly with social progressivism.

**Table 3 - Correlations Among Political Orientation Measures and DFD performance**

Measures	1	2	3	4	5	6
1. SEC Scale <sup>a</sup>	—					
2. Economic Liberalism	-.56***	—				
3. Authoritarianism	.05	.14	—			
4. Social Progressivism	-.40***	.54***	.08	—		

Measures	1	2	3	4	5	6
5. Political Orientation <sup>b</sup>	-.82***	.91***	.11	.64***	–	
6. Deepfake Detection	-.21***	.40***	.14	.57***	.42***	–

*Note:* <sup>a</sup>SEC Scale = *Social and Economic Conservatism Scale*. <sup>b</sup>Political Orientation = *Latent Political Orientation Factor Score*.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

A significant main effect of self-reported political orientation on deepfake detection ability was also observed,  $F(4,168) = 8.24, p < .001, \eta_p^2 = .164$ . Pairwise comparisons at an alpha level of .005 (after Bonferroni correction) showed significant mean differences between conservatives ( $M = 26.44, SD = 40.63$ ) and progressives ( $M = 74.13, SD = 51.23$ ), moderates ( $M = 63.98, SD = 48.5$ ), and those who preferred not to answer ( $M = 78, SD = 41.28$ ). No significant mean differences were observed in any of the other pairwise comparisons.

The results support the prediction that orientation to the political left is predictive of deficiency in deepfake detection performance. However, at a more granular level, the relatively weak predictivity of the SEC scale and trait authoritarianism was not expected.

#### 3.4.1 *The Incremental Predictivity of Political Orientation*

When comparing a model that contained only affect detection as a predictor of DFD to a model that contained both affect detection and political orientation as predictors of DFD, the inclusion of political orientation significantly improved the overall predictivity of the model ( $R^2_{\text{null model}} = .535$ ;  $R^2_{\text{full model}} = .547$ ;  $R^2_{\text{change}} = .012$ ,  $p = .04$ ). However, it should be noted that the regression weight of political orientation in a model that controls on affect detection ( $\beta_{\text{std}} = .125$ ,  $p = .04$ ) is considerably less than the regression weight of political orientation in a model in which it is the sole predictor of DFD ( $\beta_{\text{std}} = .42$ ,  $p < .001$ ). This suggests something akin to a mediation effect, although the suggestion of a causal mechanism is not made.

### 3.5 Hypothesis 4: Deepfake Detection Performance Correlates with ASD Quotient

As displayed in Table 4, performance on the deepfake detection task correlated moderately negatively with the communication and imagination subscales of the AQ; and correlated weakly negatively with the total score on the AQ.

**Table 4 - Correlations Among Autism Spectrum Quotient (AQ) Scales and DFD**

Measures	1	2	3	4	5	6	7
1. AQ Social Skill	—						
2. AQ Attention Switching	.23***	—					
3. AQ Attention to Detail	-.06	-.08	—				

Measures	1	2	3	4	5	6	7
4. AQ Communication	.41***	.26***	-.13	–			
5. AQ Imagination	.24***	.15	-.10	.35***	–		
6. AQ (Total)	.69*	.53***	.22***	.73***	.57***	–	
7. Deepfake Detection	-.05	.08	.08	-.46***	-.32***	-.26***	–

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

The negative correlation between deepfake detection performance and the AQ total score is supportive of hypothesis 4, although a stronger correlation was expected.

No specific hypotheses were offered regarding the AQ subscales, but the strengths of the correlations between DFD and the imagination subscale, and to a lesser degree the imagination subscale, is worthy of note.

The data were further probed to seek evidence that DFD might offer incremental predictivity as an ASD screening instrument. Table 5 shows that the bivariate correlation between DFD and the AQ was of a similar magnitude to the bivariate correlations between the AQ and the various affect detection tasks. Stronger relationships were observed between CAM and the affect detection factor score. However, when the AQ was regressed on DFD while controlling on the affect detection factor score (model significance:  $F(2,170) = 8.33, p < .001, R^2 = .09$ ), no significant influence of DFD ( $\beta_{\text{std}} = -.1, p = .35$ ) was detected

(Influence of the affect detection score:  $\beta_{\text{std}} = -.22, p = .04$ ). The same result was observed when the AQ was regressed on DFD while controlling on CMR score (model significance:  $F(2,170) = 8.79, p < .001, R^2 = .09$ ) (Influence of DFD: ( $\beta_{\text{std}} = -.09, p = .37$ ); Influence of CMR score: ( $\beta_{\text{std}} = -.23, p = .03$ ). While inconclusive, these results offer some evidence that DFD may not contribute incremental predictivity as an ASD screening instrument.

**Table 5 - Correlations Among Affect Detection Measures, the AQ, and the Deepfake Detection Task**

Measures	1	2	3	4	5	6	7
1. Adult Eyes Test	—						
2. Adult Faces Test	.62***	—					
3. Cambridge Mind Reading	.85***	.68***	—				
4. Reading the Mind in Film	.75***	.60***	.80***	—			
5. Affect Detection Composite	.92***	.73***	.98***	.86***	—		
6. AQ (total)	-.26***	-.22***	-.30***	-.21***	-.30***	—	
7. Deepfake Detection	.68***	.54***	.71***	.62***	.73***	-.26***	—

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

### 3.6 Exploratory Analyses

(See Table 6 for Correlation Matrix)

### *3.6.1 AQ*

A moderate negative correlation was observed between AQ and political orientation factor score. At a more granular level, economic liberalism and social progressivism were each moderately negatively correlated with AQ. These results suggest that orientation to the political right is predictive of autistic traits. No significant correlation was observed between authoritarianism and AQ.

### *3.6.2 Conspiratorial Thinking*

A strong negative correlation was observed between conspiratorial thinking and affect detection factor score. Moderate negative correlations were observed between conspiratorial thinking and political orientation factor score, social progressivism and DFD. A weak positive correlation was observed between conspiratorial thinking and EQ. Weak negative correlations were observed between conspiratorial thinking and economic liberalism, authoritarianism, and social progressivism.

### *3.6.3 Sensory Perception*

The sensory perception quotient did not correlate significantly with any variables of interest.

### *3.6.4 EQ*

A moderate negative correlation was observed between EQ and AQ. A weak positive correlation was observed between EQ and conspiratorial thinking.

**Table 6 - Correlations Among Exploratory Measures and Criterion Variables**

Measures	1	2	3	4	5	6	7	8	9	10
1. EQ <sup>a</sup>	–									
2. CT <sup>b</sup>	.21***	–								
3. SPQ <sup>c</sup>	.10	.05	–							
4. AQ <sup>d</sup>	-.31***	.05	-.10	–						
5. AD Composite <sup>e</sup>	.08	-.53***	.02	-.29***	–					
6. Economic Liberalism	.02	-.16*	.05	-.23***	.42***	–				
7. Authoritarian <sup>f</sup>	.14	-.17*	.001	-.13	.22***	.14	–			
8. Social Progressivism	-.05	-.49***	-.01	-.26***	.7***	.54***	.08	–		
9. Political Orientation	-.13	-.32***	.05	-.23***	.44***	.91***	.11	.64***	–	
10. DFD <sup>g</sup>	.07	-.40***	.14	-.26***	.73***	.40***	.14	.57***	.42***	–

*Note:* <sup>a</sup>EQ = *The Assessing Emotions Scale*. <sup>b</sup>CT = *Conspiratorial Thinking Scale*. <sup>c</sup>SPQ = *Sensory Perception Quotient*. <sup>d</sup>AQ = *Autism Spectrum Quotient*. <sup>e</sup>AD Composite = *Affect Detection Factor Score*. <sup>f</sup>Authoritarian = *Authoritarianism Subscale of Supply*. <sup>g</sup>DFD = *Deepfake Detection Performance*.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .



**Table 7 - Descriptive Statistics for All Measures**

Measures	Minimum Score	Maximum Score	Possible Range	Mean	Standard Deviation
EQ <sup>a</sup>	84	224	32-224	162.43	22.25
CT <sup>b</sup>	16	69	15-75	45.68	12.62
SPQ <sup>c</sup>	12	26	0-35	18.87	2.33
AQ <sup>d</sup>	6	37	0-50	22.48	5.47
Economic Liberalism	21	74	15-75	47.76	7.79
Authoritarian <sup>e</sup>	29	62	15-75	48.10	5.06
Social Progressivism	40	73	16-80	54.44	8.85
Conservatism	-91	974	-200-1200	481.34	184.57
Reading the Mind in Film	3	22	0-22	11.31	4.26
Cambridge Mind Reading	6	44	0-50	27.97	10.92

Measures	Minimum Score	Maximum Score	Possible Range	Mean	Standard Deviation
Adult Eyes Test	3	25	0-36	15.80	6.20
Adult Faces Test	6	20	0-20	16.79	2.58
DFD <sup>g</sup>	-35	166	-248-248	59.27	50.69

*Note:* <sup>a</sup>EQ = *The Assessing Emotions Scale*. <sup>b</sup>CT = *Conspiratorial Thinking Scale*. <sup>c</sup>SPQ = *Sensory Perception Quotient*. <sup>d</sup>AQ = *Autism Spectrum Quotient*. <sup>e</sup>Authoritarian = *Authoritarianism Subscale of Supply*. <sup>g</sup>DFD = *Deepfake Detection Task*.

## **CHAPTER 4. DISCUSSION**

Although propaganda-wielding and the weaponization of information is not new, the threat from nefarious utilizations of deepfake technology is especially grave. The capacity to produce these videos is increasingly widespread, the social and political zeitgeist predisposes us to be particularly reactive to compromising videos (in some cases, justifiably so), and bad actors on the international and domestic stage are already deploying this technology to wreak havoc. Even the President of the United States, Donald J. Trump, has used this technology against his political opponents (Frum, 2020).

### **4.1 Mindblindness, Political Orientation, and Deepfake Detection**

This study set out to investigate whether certain individual differences could be identified that would predict susceptibility to being fooled by deepfake videos. The primary traits of interest were theory of mind ability (operationalized as affect detection ability) and political orientation. The results of the study suggest that while both of these traits can be separately predictive of one's ability to spot deepfake videos, theory of mind deficiency appears to be the more salient of the two. Political orientation (specifically orientation to the political right) was shown to predict poor deepfake detection ability but this relationship is mediated by theory of mind ability. However, the results of this study also suggest that theory mind ability is not distributed evenly across the political spectrum. Deficiency in one's ability to theorize about the minds of others does appear to be predictive of alignment with the political right, a finding that is in concordance with previous work which established the relationship between orientation to the political right and inattention to gaze (social) cueing (Carraro et al., 2015; Dodd et al., 2011). This is concerning in any climate

but given the widely publicized attempts by belligerent states to interfere in the domestic affairs of opponent nations, the unequal distribution of susceptibility to misinformation across the political spectrum is worthy of society's collective attention (Badawy et al., 2018; Broniatowski et al., 2018; Davis, 2018; Mueller, 2019).

## **4.2 ASD and Deepfake Detection**

A secondary aim of the study was to generate preliminary evidence that deepfake detection might be a viable method by which to screen for autism spectrum disorder. While the observed negative correlation between the Autism Spectrum Quotient and performance on the deepfake detection task is in accordance with the prediction stated in hypothesis 4, performance on the deepfake detection task did not offer incremental predictivity when controlling on the more tried and true affect detection measures. While this does not necessarily preclude using deepfake detection to screen for ASD generally, the measure would, at the very least, need to be refined. However, a deepfake detection task does come with the advantage of being more language independent than many ASD diagnostic instruments. In the adult version of the Autism Spectrum Quotient for instance, participants must be able to read and respond to complete sentences. In the child version of the Autism Spectrum Quotient, the patient's parent or guardian must complete the questionnaire. In contrast, the deepfake detection task requires only that a person be able to indicate authenticity or inauthenticity of a video and then report response confidence. By the same logic, one could even imagine implementing an ASD screening instrument which replaced deepfake detection with a task which measured insensitivity to the uncanny valley effect. Such a task could be implemented with very little reliance on verbal ability.

While this study did not generate convincing evidence that the current instantiation of the deepfake detection task has utility as an ASD screening instrument, it is worthy of note that two subscales of the Autism Spectrum Quotient showed a markedly stronger relationship with deepfake detection performance than did the other three. The communication subscale, for which a positive score indicates that the examinee has difficulty expressing themselves to others, correlated negatively with deepfake detection performance. This is not necessarily a surprising result as it seems reasonable to assume that an individual who has difficulty recognizing affect conveyed by the facial expressions of others would also have difficulty communicating with others. The imagination subscale, for which a positive score indicates that the examinee has difficulty producing mental imagery, also correlated negatively with deepfake detection performance, although to a lesser degree than did the communication subscale. Perhaps it is the case that the ability to produce mental imagery is integral to the ability to affect categorizations or deepfake vs. authentic distinctions. In the case of affect detection, it is plausible that categorization is facilitated by comparing the observed facial expression to mental images of exemplars. Future work should investigate this possibility.

### **4.3 Applications**

The results of this study have the potential to inform several practical applications. Among these is the development and implementation of educational/training interventions intended to bolster one's credulity during media consumption. The relationship between theory of mind ability (affect detection ability) and deepfake detection ability, revealed in this study suggests that interventions which improve one's ability to recognize affect in others could additionally reduce one's susceptibility to being fooled by deepfakes. Several

extant interventions purport to do exactly that (Begeer, 2014; Fletcher-Watson et al., 2014; Hadwin & Kovshoff, 2013; Kuoch & Mirenda, 2003; Lantz, 2002; Noel & Westby, 2014; Vass et al., 2018). Perhaps these interventions could be retooled to help defend the most susceptible among us from deception via deepfakes. Future work should investigate the efficacy of these interventions for this purpose.

The identification of individual differences that predict one's deepfake detection ability might also inform the provident allocation of resources intended to defend against their influence. As discussed in the introduction, the computer science community has begun to develop tools that help identify videos that have been digitally altered (Albahar & Almalki, 2019). The situation may eventually require that all videos are subject to this scrutiny and that all viewers rely on the recommendation of such software. In the meantime, however, it may be prudent to prioritize distributing these tools to the most vulnerable among us.

There is additionally the possibility that the identification of individual differences that predict one's deepfake detection ability could facilitate self-selection for adoption of the aforementioned defensive tools or even simply heightened personal vigilance. That is to say, if individuals are made aware of their own susceptibility to deception via deepfake video, they may wish to seek out and add some of these tools to their media consumption habits. They may also wish to alter their personal criteria of credibility.

#### **4.4 Limitations and Extending the Work**

There are also several ways in which the design of this study limited the extent certain conclusions could be drawn. First, with the exception of those from the Google dataset, most of the deepfake videos contained celebrity subjects. Because of this, the concern arises that examinees ability to detect the inauthenticity of a particular video was not due to a recognition of affective inconsistencies but rather a mismatch between the appearance of a celebrity they are familiar with and the subject of the video. For example, the face of actor Tom Cruise is used in one of the videos. If an examinee recognized this video as being inauthentic by noting that Tom Cruise has never sported the hairstyle he does in the video, he/she would not be engaging the cognitive mechanism central to the original research question (affect detection). An effort was made to forestall this objection by asking participants to report the cues they relied on to judge the authenticity of the videos. In this open-ended question only 13 of 173 participants input the strings, “celeb”, “fam”, or “recog”. These strings were chosen because jointly they envelope most variations of the words, “celebrity”, “famous”, “familiar”, “recognize”, etc.... The celebrity issue was further addressed by computing a DFD score that included only items containing deepfakes from the Google dataset (which did not contain celebrities) and reanalyzing the relationships between DFD ability and affect detection ability ( $r = .49, p < .001$ ); and the relationship between DFD and political orientation ( $r = .29, p < .001$ ). While these correlations are weaker than those observed with the full DFD measure, it is important to note that the deepfakes from the Google dataset were all of a higher quality than the others and thus considerably more difficult to detect. It is therefore suspected that the attenuation of the correlations is due more to a restricted range of difficulty than it is due to the inclusion of celebrity faces. Future researchers could further address this issue by

developing original deepfake videos that range in quality and do not utilize celebrity models.

Another potential shortcoming of this study lies in the fact that the measure chosen to operationalize traits associated with ASD is not necessarily a commonly used ASD screening/diagnostic instruments. Results from a diagnostic instrument like the Autism Diagnostic Observation Schedule (Lord et al., 2002) would be a more valid proxy for ASD, but such a measure is considerably more onerous to administer and interpret than is the Autism Spectrum Quotient. Future research should assess the relationship to ASD in a more quasi-experimental fashion by comparing participants who are diagnosed with ASD with neurotypical participants.

It is also important to note that the author makes no claim that the variables included in this study represent an exhaustive model of the cognitive traits and abilities that predict one's deepfake detection performance. In fact, the present data offer evidence that so-called third variables may be moderating the strength of the relationships at the core of this study. For instance, when data from the Georgia Tech portion of the sample ( $n = 96$ ) are isolated and analyzed, the correlation between the affect detection factor score and performance on the deepfake detection task ( $r = .30, p < .001$ ) is smaller than when data from the Mturk portion of the sample ( $n = 77$ ) are isolated ( $r = .78, p < .001$ ). This attenuation is suspected to be due to a restricted range of talent among the Georgia Tech group as the admissions process selects for students with high cognitive ability. The natural extension of this line of reasoning is that general cognitive ability may be moderating the relationship between theory of mind ability and deepfake detection ability. In other words, it may be the case that the intellectually gifted are able to deploy some heuristic with which



to augment deficiencies in theory of mind ability. Future research should investigate this possibility by testing the moderating effect of cognitive ability.

Another issue is the ecological invalidity of the deepfake detection task. The video clips in this task had no sound and were not associated with any source or other context. Under the conditions in which one might ordinarily encounter a deepfake video, the video would likely be accompanied by sound and source information. These additional sources of information may be instrumental in facilitating the evaluation of the video's authenticity. Future work should present deepfake videos in a more ecologically valid context. The technology also exists for digitally altering voice recording in a way that is similar to the altered video of a deepfake. Future work should include voice forgeries to determine if the same individual differences that predict deepfake detection ability also predict one's ability to detect altered voice information.

Future research should also further probe implications of the relationships between the exploratory measures included in this study and deepfake detection ability. Arguably the most interesting of these relationships is that which is observed between conspiratorial thinking, theory of mind, and deepfake detection ability. One would expect conspiratorial thinking to be more effect than cause, but this is not necessarily the case and should be investigated.

## **4.5 Conclusions**

1. Those who are deficient in their ability to correctly identify the emotions of others are similarly deficient in their ability to spot deepfake videos.

2. Those oriented to the political right tend to be more deficient in theory of mind ability than those oriented to the political left.
3. Those oriented to the political right tend to be less able to detect deepfake videos than those oriented to the political left. However, political orientation is not incrementally predictive of deepfake detection ability when controlling on affect detection ability.
4. The deepfake detection task in its current state does not appear to demonstrate utility as a screening instrument for autism spectrum disorder.

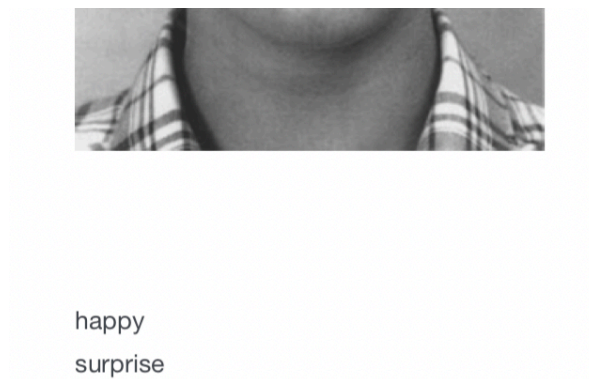
## APPENDIX A. ADULT EYES TEST – EXAMPLE ITEM



playful  
comforting  
irritated  
bored

*Note.* In this 37-item task participants select one of the four emotions that they believe is representative of the eyes in the image. The image of the stimulus has been cropped in the attempt to not reveal any proprietary measures.

## APPENDIX B. ADULT FACES TEST – EXAMPLE ITEM




*Note.* In this 40-item task participants select one of the two emotions that they believe is representative of the faces in the image. The image of the stimulus has been cropped in the attempt to not reveal any proprietary measures.

## APPENDIX C. CAMBRIDGE MIND READING TEST – EXAMPLE

### ITEM

choose just one word, the word which you consider to be most suitable. Before making your choice, make sure that you have read all 4 words.



convinced - being persuaded by an argument

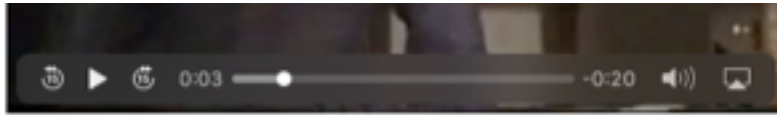
sociable - to enjoy being in other people's company and talking and listening to them

resentful - feeling annoyed about something or being ill-used

luring - tempting, enticing others to do something or go somewhere

*Note.* In this 50-item task participants select one of the four emotions that they believe is representative of the subject in the video. The image of the stimulus has been cropped in the attempt to not reveal any proprietary measures.

## APPENDIX D. READING THE MIND IN FILM TEST – EXAMPLE ITEM



At the end of the scene, how is the woman feeling?

bothered - to feel disturbed or worried about something

embarrassed - to be made to feel self-conscious and shy, often in a social situation

surprised - to feel unexpectedly amazed or shocked

interested - to have one's attention held by something or someone

*Note.* In this 22-item task participants select one of the four emotions that they believe is representative of the subject in the video. The image of the stimulus has been cropped in the attempt to not reveal any proprietary measures.

## APPENDIX E. SOCIAL AND ECONOMIC CONSERVATISM SCALE

### – EXAMPLE ITEMS

---

Indicate your feelings about the following issues on a scale from 0 to 100.

0 = completely negative and 100 = completely positive.

	Negative feelings						Positive Feelings				
	0	10	20	30	40	50	60	70	80	90	100
Abortion											
Patriotism											
Traditional marriage											
Religion											

**APPENDIX F. SAPPLY POLITICAL COMPASS TEST – EXAMPLE**

**ITEMS**

The government should be less involved in the day to day life of its citizens.

Strongly agree    Somewhat agree    Neither agree nor disagree    Somewhat disagree    Strongly disagree



## APPENDIX G. AUTISM SPECTRUM QUOTIENT – EXAMPLE

### ITEMS

---

When I'm reading a story, I can easily imagine what the characters might look like.



## APPENDIX H. DEEPPFAKE DETECTION TASK – EXAMPLE ITEM

In this section you'll be presented with a series of videos. Some of the people in the videos have been digitally altered. Please do your best to identify the videos that have been digitally altered.

Has the person in this video been digitally altered?



Altered

Not Altered

*Note.* In this 62-item task participants are asked whether the video is altered or not and are then asked to indicate their confidence level (not shown).

## APPENDIX I. ASSESSING EMOTIONS SCALE – EXAMPLE ITEMS

Below is a list of statements. Please read each statement very carefully and rate how strongly you agree or disagree by selecting the appropriate option opposite each statement.

	Strongly Disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I am aware of my emotions when I express them.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

## APPENDIX J. SENSORY PERCEPTION QUOTIENT – EXAMPLE

### ITEMS

---

	Strongly Agree	Agree	Disagree	Strongly Disagree
I would be able to feel a one millimetre cut in my skin.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

## APPENDIX K. CONSPIRATORIAL THINKING SCALE –

### EXAMPLE ITEMS

---

There is often debate about whether or not the public is told the whole truth about various important issues. This brief survey is designed to assess your beliefs about some of these subjects. Please indicate the degree to which you believe each statement is likely to be true on the following scale: Definitely not true; Probably not true; Not sure/cannot decide; Probably true; Definitely true

The government is involved in the murder of innocent citizens and/or well-known public figures, and keeps this a secret

Definitely not true	Probably not true	Not sure/cannot decide	Probably true	Definitely true
---------------------	-------------------	---------------------------	---------------	-----------------

## REFERENCES

- Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and Countermeasures Systematic Review. *Journal of Theoretical and Applied Information Technology*, 97(22).
- Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173.
- Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.
- Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*, 2, 3–20.
- Baron-Cohen, S. (2001). Theory of mind and autism: A review. *International Review of Research in Mental Retardation*, 23(23), 169–184.
- Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, 38(7), 813–822.
- Baron-Cohen, S., Wheelwright, S., Jolliffe, & Therese. (1997). Is there a " language of the eyes"? Evidence from normal adults, and adults with autism or Asperger syndrome. *Visual Cognition*, 4(3), 311–331.
- Bayliss, A. P., Frischen, A., Fenske, M. J., & Tipper, S. P. (2007). Affective evaluations of objects are influenced by observed gaze direction and emotional expression. *Cognition*, 104(3), 644–653.
- Bayliss, A. P., & Tipper, S. P. (2005). Gaze and arrow cueing of attention reveals individual differences along the autism spectrum as a function of target context. *British Journal of Psychology*, 96(1), 95–114.
- Begeer, S. (2014). Theory of mind interventions can be effective in treating autism, although long-term success remains unproven. *Evidence-Based Mental Health*, 17(4), 120–120.
- Block, N. (1977). *Review of Julian Jaynes, Origin of Consciousness in the Breakdown of the Bicameral Mind*.
- Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: Meta-analysis. *Schizophrenia Research*, 109(1–3), 1–9.

- Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health, 108*(10), 1378–1384.
- Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology, 4*, 279.
- Brüne, M. (2005). “Theory of mind” in schizophrenia: A review of the literature. *Schizophrenia Bulletin, 31*(1), 21–42.
- Carraro, L., Dalmaso, M., Castelli, L., & Galfano, G. (2015). The politics of attention contextualized: Gaze but not arrow cuing of attention is moderated by political temperament. *Cognitive Processing, 16*(3), 309–314.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L Erlbaum Associates.
- Davis, S. (2018). *Russian Meddling in Elections and Referenda in the Alliance*. NATO Parliamentary Assembly.
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation, 14*(1), 20.
- Dodd, M. D., Hibbing, J. R., & Smith, K. B. (2011). The politics of attention: Gaze-cuing effects are moderated by political temperament. *Attention, Perception, & Psychophysics, 73*(1), 24–29.
- Durkin, M. S., Maenner, M. J., Baio, J., Christensen, D., Daniels, J., Fitzgerald, R., Imm, P., Lee, L.-C., Schieve, L. A., & Van Naarden Braun, K. (2017). Autism spectrum disorder among US children (2002–2010): Socioeconomic, racial, and ethnic disparities. *American Journal of Public Health, 107*(11), 1818–1826.
- Feng, S., Wang, X., Wang, Q., Fang, J., Wu, Y., Yi, L., & Wei, K. (2018). The uncanny valley effect in typically developing children and its absence in children with autism spectrum disorders. *PloS One, 13*(11), e0206343.
- Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: A novel account of the uncanny valley. *Frontiers in Psychology, 6*, 249.
- Festinger, L. (1957). *A Theory of Cognitive Dissonance* (Vol. 2). Stanford university press.
- Fletcher-Watson, S., McConnell, F., Manola, E., & McConachie, H. (2014). Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD). *Cochrane Database of Systematic Reviews, 3*.

- Frum, D. (2020, April 27). *The Very Real Threat of Trump's Deepfake*. The Atlantic. <https://www.theatlantic.com/ideas/archive/2020/04/trumps-first-deepfake/610750/>
- Golan, O., Baron-Cohen, S., & Hill, J. (2006). The Cambridge mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *Journal of Autism and Developmental Disorders*, 36(2), 169–183.
- Golan, O., Baron-Cohen, S., Hill, J. J., & Golan, Y. (2006). The “reading the mind in films” task: Complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience*, 1(2), 111–123.
- Green, R. D., MacDorman, K. F., Ho, C.-C., & Vasudevan, S. (2008). Sensitivity to the proportions of faces that vary in human likeness. *Computers in Human Behavior*, 24(5), 2456–2474. <https://doi.org/10.1016/j.chb.2008.02.019>
- Hadwin, J. A., & Kovshoff, H. (2013). A review of theory of mind interventions for children and adolescents with autism spectrum conditions. *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*, 413.
- Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., Fernandez, R., & Stephanou, H. (2005). Upending the uncanny valley. *AAAI*, 5, 1728–1729.
- Kuoeh, H., & Mirenda, P. (2003). Social story interventions for young children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, 18(4), 219–227.
- Lantz, J. (2002). *Theory of mind in autism: Development, implications, and interventions*.
- Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2002). Autism diagnostic observation (ADOS) manual. *Los Angeles: Western Psychological Services*.
- Lukianoff, G., & Haidt, J. (2019). *The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure*. Penguin Books.
- MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, 16(2), 141–172.
- MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3), 297–337.
- Matson, J. L., & Kozlowski, A. M. (2011). The increasing prevalence of autism spectrum disorders. *Research in Autism Spectrum Disorders*, 5(1), 418–425.
- Moore, C., Dunham, P. J., & Dunham, P. (2014). *Joint attention: Its origins and role in development*. Psychology Press.
- Moosa, M. M., & Ud-Dean, S. M. (2010). Danger avoidance: An evolutionary explanation of uncanny valley. *Biological Theory*, 5(1), 12–14.



- Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33–35.
- Mueller, R. S. (2019). *Report on the investigation into Russian interference in the 2016 presidential election* (Vol. 1). US Department of Justice Washington, DC.
- Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., & Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, 78(3), 938–954.
- Nevison, C., Blaxill, M., & Zahorodny, W. (2018). California autism prevalence trends from 1931 to 2014 and comparison to national ASD data from IDEA and ADDM. *Journal of Autism and Developmental Disorders*, 48(12), 4103–4117.
- Noel, K. K., & Westby, C. (2014). Applying theory of mind concepts when designing interventions targeting social cognition among youth offenders. *Topics in Language Disorders*, 34(4), 344–361.
- Political Compass Project*. (n.d.). Retrieved July 6, 2020, from <http://sapplypoliticalcompass.com/>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526.
- Ramey, C. H. (2005). The uncanny valley of similarities concerning abortion, baldness, heaps of sand, and humanlike robots. *Proceedings of Views of the Uncanny Valley Workshop: IEEE-RAS International Conference on Humanoid Robots*, 8–13.
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. F. (n.d.). Learning to detect manipulated facial images. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea*, 27.
- Schutte, N. S., Malouff, J. M., & Bhullar, N. (2009). The assessing emotions scale. In *Assessing emotional intelligence* (pp. 119–134). Springer.
- Simpson, E. A., Miller, G. M., Ferrari, P. F., Suomi, S. J., & Paukner, A. (2016). Neonatal imitation and early social experience predict gaze following abilities in infant monkeys. *Scientific Reports*, 6, 20233.
- Smiley, K., Gerstein, B., & Nelson, S. (2018). Unveiling the autism epidemic. *Journal of Neurology and Clinical Neuroscience*, 2(2), 1.
- Sprong, M., Schothorst, P., Vos, E., Hox, J., & Van Engeland, H. (2007). Theory of mind in schizophrenia: Meta-analysis. *The British Journal of Psychiatry*, 191(1), 5–13.
- Tavassoli, T., Hoekstra, R. A., & Baron-Cohen, S. (2014). The Sensory Perception Quotient (SPQ): Development and validation of a new sensory questionnaire for adults with and without autism. *Molecular Autism*, 5(1), 29.

- Vass, E., Fekete, Z., Simon, V., & Simon, L. (2018). Interventions for the treatment of theory of mind deficits in schizophrenia: Systematic literature review. *Psychiatry Research*, 267, 37–47.
- Williams, D. M., Nicholson, T., Grainger, C., Lind, S. E., & Carruthers, P. (2018). Can you spot a liar? Deception, mindreading, and the case of autism spectrum disorder. *Autism Research*, 11(8), 1129–1137.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128.
- Yilmaz, O., & Saribay, S. A. (2017). The relationship between cognitive style and political orientation depends on the measures used. *Judgment and Decision Making*, 8.