The Application of Artificial Intelligence in Police Interrogations

An Analysis Addressing the Proposed Effect AI Has on Racial and Gender Bias, Cooperation, and False Confessions

Candidate number/Student Number: 8015

Submission deadline: 15 May, 2019

Number of words: 20,000 words (excl. Preface, Acknowledgements, and References)
Preface

This paper was initially inspired by the 2014 study conducted by Dean A. Pollina and Allison Barretta, “The effectiveness of a national security screening interview conducted by a computer-generated agent.” It showed the potential for the legal futurism of artificial intelligence application in policing. Further, while writing this paper, the Estonian Ministry of Justice made international headlines by launching their decision to develop a robot judge to handle small claims court disputes. Increasing development in artificial intelligences’s functions and abilities continues to contradict previous assumptions. These assumptions restrict these functions to the confines of human performance. In fact, some of the previously “Hollywood-ized” sci-fi accomplishments are now, actually, not so far from reality.

When choosing a master’s thesis topic, I specifically chose a controversial issue in the human rights realm, combining both my interests in technology and my background expertise in criminology and criminal justice. Though this analysis could be applicable internationally, the main focus on this paper focused on the racial and gender inconsistencies in false confessions in the United States criminal justice system, thought to be impacted by police bias. Still today, the international community seems torn by both simultaneously creating and fearing artificial intelligence advancement. Though there are, and should be, many concerns surrounding the ethical developments in the field, artificial intelligence could be a vital asset to use to our advantage.

I hope that the extrapolation of future policing as approached through this paper can positively assist discrimination and bias. In addition, its hope is to be relevant now and in years time for the future implementation of artificial intelligence application in policing.
Acknowledgements

I would like to thank and acknowledge the educators from University of Oslo’s Master’s of Theory and Practice of Human Rights program for all the work that they put in to make this program internationally recognized and regarded. In the Norwegian Center of Human Rights, I would like to thank Knut Asplund, Head, Rule of Law, International Department and Gisle Kvanvig, Director of the ASEAN/Vietnam program, for the opportunity to intern for them and their proactive work on interrogations. I would especially like thank Malcolm Langford, Professor of Public Law at University of Oslo and Co-Director of the Centre on Law and Social Transformation, for being the sole advisor on this paper, who’s legal expertise in the field provided the critical feedback on the context and organization of the new age approach presented in this paper.
# Table of Contents

INTRODUCTION .................................................................................................................................................6

CHAPTER 1: ARTIFICIAL INTELLIGENCE, CONFESSION ELICITATION, AND ITS APPLICATION IN POLICE INTERROGATIONS ..................................................................................................................11

1.1 ELICITING SELF INCRIMINATING CONFESSIONS OF ILLEGAL BEHAVIOR THROUGH AI ..................................................................................................................................................................................11

1.1.1 ARTIFICIAL INTELLIGENCE ASSISTED NATIONAL SECURITY CLEARANCE INTERVIEWS ON SELF-INCRIMINATORY, ILLEGAL BEHAVIOR: .................................................................................................................................11

1.2 DESIGNING AN AI MODEL FOR ENHANCING SELF INCRIMINATION OF ILLEGAL BEHAVIOR ...........................................................................................................................................................................14

1.2.1 PERCEPTION OF HUMAN-LIKE ATTRIBUTES .................................................................................15

1.2.2 PERCEPTION OF HUMAN CONTROL ...........................................................................................15

1.2.3 QUESTION DESIGN ........................................................................................................................16

1.2.2 EXPANDING THE DESIGN OF AN “AI INTERROGATOR” IN CONFESSION ELICITATION. 16

1.2.2.1 INCREASED REALISM ..............................................................................................................16

1.2.2.1.1 PHYSICAL PRESENCE ........................................................................................................17

1.2.2.1.2 AI -TO-HUMAN PHYSICAL PLACEMENT ....................................................................18

1.2.2.1.3 ANTHROPOMORPHISM AND THE UNCANNY VALLEY ...........................................18

1.2.2.2 DECREASED PERCEIVED EXTERNAL HUMAN CONTROL ............................................20

1.2.2.3 DIRECT QUESTIONING FORMAT ...........................................................................................21

1.3 DISCUSSION ...................................................................................................................................................21

CHAPTER 2: ARTIFICIAL INTELLIGENCE AS INTERROGATORS IN POLICE INTERROGATIONS .......................................................................................................................................................................23

2.1 APPLYING PROPOSED AI MODEL DESIGNED FOR POLICE INTERROGATIONS .....................................................................................................................................................................................24

2.1.1 RAPPORT BUILDING THROUGH AI ............................................................................................25

2.1.2 BASELINE QUESTIONS AND RESPONSE ..................................................................................26

2.1.2.1 AI FACIAL-EMOTIONAL DECODING AND RECOGNITION IN HUMANS ....................27

2.1.2.2 AI VERBAL-EMOTIONAL DECODING AND RECOGNITION IN HUMANS ....................27

2.1.2.3 DECEPTION RECOGNITION SOFTWARE ...........................................................................28

2.2 DISCUSSION ...................................................................................................................................................28

2.2.1 APPLICABILITY ....................................................................................................................................30

2.2.2 LIMITATIONS AND FURTHER RESEARCH ..............................................................................31

2.2.2.1 FINANCIAL LIMITATIONS ...............................................................................................31

2.2.2.2 LEGAL LIMITATIONS ..........................................................................................................31

CHAPTER 3: THE INTERROGATOR-TO-SUSPECT INTERFACE: ........................................................................33

3.1 INTERROGATOR BIAS INFLUENCE AND PERCEPTION OF GUILT ....................................................33

3.1.1 DECEPTION DETECTION AND PRESUMPTION OF GUILT ...............................................................34

3.1.2 RACIAL AND GENDER INTERROGATOR BIAS ............................................................................35

3.1.2.1 IMPLICIT RACIAL BIAS INFLUENCE .............................................................................35

3.1.2.2 IMPLICIT GENDER BIAS INFLUENCE .............................................................................36
3.2 DISCUSSION..................................................................................................................38

CHAPTER 4: THE SUSPECT-TO-INTERROGATOR INTERFACE: .................................40
4.1 TYPES OF FALSE CONFESSIONS ..............................................................................40
  4.1.1 VOLUNTARY FALSE CONFESSIONS .................................................................40
  4.1.2 COMPLIANT FALSE CONFESSIONS .................................................................41
  4.1.3 INTERNALIZED FALSE CONFESSIONS ............................................................41
4.2 SUSPECT COMPONENT OF BIAS-UNCOOPERATIVE LOOP ...............................42
  4.2.1 THE BIAS COMPONENT ....................................................................................42
    4.2.1.1 THE RACIAL BIAS STEREOTYPE THREAT .................................................42
    4.2.1.2 THE GENDER BIAS STEREOTYPE THREAT .............................................43
  4.2.2 THE COOPERATION COMPONENT ..................................................................44
4.3 DISCUSSION ............................................................................................................44

CHAPTER 5: NON BIAS ENVIRONMENT AND PROGRAMABLE SIMILARITY ON
HUMAN-TO-AI COOPERATION, TRUST AND SAFEGUARDING AGAINST
FALSE CONFESSIONS.......................................................................................................46
5.1 POSITIVE STUDIES: RACIAL AND GENDER NEUTRAL ARTIFICIAL
INTELLIGENCE MODELS ...............................................................................................46
  5.1.1 RISK ASSESSMENT MODELS AND RACIAL AND GENDER NEUTRALITY ..........47
5.2 MINIMIZING SUSPECT BIAS INFLUENCE, STRESS/ANXIETY, AND FALSE
CONFESSIONS ............................................................................................................49
5.3 HYPOTHETICAL ENHANCEMENT OF SUSPECT AND INTERROGATOR
INTERACTION USING ARTIFICIAL INTELLIGENCE ...............................................50
  5.3.1 PROGRAMMABLE SUSPECT SIMILARITY ON HYPOTHETICAL TRUST AND
    COOPERATION ......................................................................................................50
    5.3.1.2 PROGRAMMABLE PHYSICAL SIMILARITY THROUGH ETHNIC CULTURE AND
    LANGUAGE ...........................................................................................................52
5.4 DISCUSSION ............................................................................................................53

CHAPTER 6: PLATFORM FOR A RACIAL AND GENDER NON BIAS
ENVIRONMENT: DEVELOPMENT, OVERSIGHT, AND TRANSPARENCY 54
6.1 PHASE I: THE PROGRAMMING PHASE: DATA ACCURACY AND EQUALITY
IN PERFORMANCE ........................................................................................................55
6.2 PHASE II: THE APPLICATION PHASE: HUMAN INTERVENTION: VETTING
DATA, BIAS DETECTION SOFTWARE, and Oversight ..................................................56
6.3 PHASE III: THE RESULTS PHASE: TRANSPARENCY AND ACCESS TO DATA
WITHIN THE LAW .......................................................................................................57
6.4 DISCUSSION ............................................................................................................59

CHAPTER 7: CONCLUDING REMARKS AND OUTLOOK ........................................61
Introduction

“If we do it [Artificial Intelligence] right, we might actually be able to evolve a form of work that taps into our uniquely human capabilities and restores our humanity. The ultimate paradox is that this technology may become the powerful catalyst that we need to reclaim our humanity.”

The research presented in this thesis examines the potentiality of artificial intelligence as an interrogator within a police interrogation to promote a non biased environment in an effort to mitigate the ongoing racial and gender divide in statistics regarding false confessions. Ideally, artificial intelligence supplementation may help promote the elicitation of non-coerced, voluntary confessions.

Growing acknowledgment of police-induced false confessions alongside an “in-system” bias within the United States gained public attention in 1992 through the ongoing DNA exonerations of the Innocence Project. From these 364 exonerees, approximately 28% of the cases involved false confessions. Furthermore, from these cases, 70% of the exonerees were of Black, Latino, or Native American ethnic origins. Twenty five years later in 2017, the National Registry of Exonerations’ report Race and Wrongful Convictions in the United States, found a continued pattern in which “law enforcement misconduct and racism also played major roles, such as police deliberately targeting black people for raids, arrests, and false confessions.” Gender bias within interrogations is another contentious issue, in which women account for 11% of exonerations based on false confessions, in comparison to 9% of overall exonerations. These female false confessions are thought to be influenced and targeted by the societal gender bias, especially in the areas of “threats to family responsibilities.”

---

2 Only racial and gender bias although this paper acknowledges bias can extend to other categories.
4 Ibid
the percentage statistics suggest that there may be a greater likelihood of conviction based on false confessions for females.\(^8\)

Growing usage of new platforms designed to minimize or ideally eliminate bias within the criminal justice and legal system has shed some hope for acceptance of human-and-AI cooperative assistance. Beginning as early as 2009, forty-eight States within the United States have utilized some version of a computer-algorithm based risk assessment tool for screening in various stages of the criminal justice system.\(^9\) This progression of artificial intelligence in the criminal justice system has set the premise for this paper. What if we could somehow implement these advancements into the interrogation room, in which these biases and these false confessions often collide? The paper aims to shed light on the proposed benefits artificial intelligence-assisted interrogations could have in the continuously problematic area of racial and gender bias within police interrogations, and its influence on these false confessions.

Critics notoriously point to the “dark side” of artificial intelligence and its impact on humans. The late Steven Hawking once told BBC that “the development of full artificial intelligence could spell the end of the human race….\(^10\) However, collectively, critics seemingly fail to acknowledge the even darker side of human behavior, as illustrated by these current and continuous statistic and research findings on false confessions. In the area of human rights, artificial intelligence has seemingly developed a negative connotation. AccessNow, an international digital rights organization claims that, “As artificial intelligence continues to find its way into our daily lives, its propensity to interfere with human rights only gets more severe…”\(^11\) This analysis intends to challenge those claims, and the stigma itself, arguing that in certain contexts, artificial intelligence advances human rights.

Protections against both racial and gender discrimination are underlying and fundamental principles in the context of international human rights. Police-based bias raises a plethora of discrimination issues in a variety of different avenues from excessive force, to profiling, etc.\(^12\) Given the specificity of racial and gender bias in the interrogation room, the only relevant applicable document concerning police guidelines in the area is discussed in the 2004

---

\(^8\) Knox
\(^12\) American Civil Liberties Union. “Racial Profiling.” ACLU” https://www.aclu.org/issues/racial-justice/race-and-criminal-justice/racial-profiling"
United Nations document *Human Rights Standards and Practice for the Police (Professional Training Series No. 5/Add.3).* However, there was an inability to obtain a list of countries who actively implement or use the guideline manual. In the United States, domestic-based organizations such as the Leadership Conference of Civil and Human Rights and the American Civil Liberties and others increasingly urge awareness and change on the issue. The fact of the matter is, seemingly even with hard or soft law guidelines in place, or protective organizations, the problem of racial and gender bias within the police force exists, and continues to exist within the interrogation room.

This research question this thesis proposes is: “Can artificial intelligence be utilized in an interrogation room to safeguard against police-based discrimination and bias, to effectively promote suspect cooperation, and thereby minimize the risk factors associated with false confessions?

To begin this discussion, Chapter 1 presents the hypothetical idea of artificial intelligence eliciting a criminal confession. This chapter will outline the inspiration of this thesis, the 2014 study conducted by Dean A. Pollina and Allison Barretta in the *International Journal of Human Computer Studies*, in which a portion of their study suggested that admissions of illegal behavior (in the context of national security clearance interviews) were and could be made possible in an artificial intelligence environment. This thesis argues, that given the similarity of obtaining an admission of criminal behavior in the national security interview and obtaining a confession of criminal behavior in an interrogation, artificial intelligence could function in an interrogation room setting. Using, their AI design model as a theoretical functional base, this thesis adds a collection of recent developments in human-to-AI interactions to create an enhanced theoretical design model intended to improve upon human and AI cooperation and trust.

Given the evidence presented in Chapter 1, which suggests the ability for culpability admissions, Chapter 2 will continue the focus with a proposed functional artificial intelligence interrogator using the model from Chapter 1. Once this analysis has established a hypothetically functional AI interrogator, with the ability to theoretically aid in eliciting criminal con-
fessions, the analysis will shift focus onto identifying the current and continuous racial and
gender bias problem within police interrogations.

Chapters 3 and 4 both explore the theoretical cooperation dynamic between the interrogator-to-suspect (Chapter 3) and between the suspect-to-interrogator (Chapter 4). This thesis argues both dynamics are influenced by these biases. Research presented in Chapter 3 suggests that an interrogator’s racial and gender bias influences both (1) perception of guilt and (2) overall perception of the suspect. It will also point out that “anti-stereotype conditioning” has shown minimal longterm success in reducing these biases. Moreover, research presented in Chapter 4 suggests that these “perceived” biases by the suspect (even when no bias exists) impacts the perception of suspect cooperation with the interrogator and willingness to disclose information. Chapters 3 and 4 collectively attempt to address the two-fold bias occurring in the interrogation room. Though this notion of an interracial and gender bias dynamic has been acknowledged before in previous research, this thesis suggests both its appearance in the interrogation room, and as a factor of false confessions. Since this phenomena does not have a consistent name amongst scholars, for the purposes of this thesis, it will be referred to as the the Bias-Uncooperative Loop. This simultaneous bias assumption comprises (1) the interrogator’s explicit or implicit racial /gender bias as influencing the interrogator’s perception of suspect cooperation while (2) the suspect perceives the interrogator as stereotypically biased, whereby (a) the suspect believes interrogator assumes guilt (b) the suspect conforms to the stereotypical threat by exhibiting anxiety or stress associated with un-cooperation. This in turn may lead to the interrogator’s confirmation bias of guilt, with the potential to lead to a more intense interrogation, and theoretically elicit a false confession. The common denominator, as this thesis will point out, is the perception of cooperation, in which evidence supports the existence of a stereotype-bias found in interrogation studies. This thesis finds that understanding the Bias-Uncooperative Loop within the dynamic between the two parties requires the separation of how one influences the other in relation to eliciting these false confessions. Ideally, this might help to explain why so many false confessions are racial and gender divided.

Chapter 5 will present how the addition of AI as an interrogator could theoretically halt the Bias-Uncooperative Loop, which is theorized to affect these false confessions by providing a non biased environment and through programmable similarity. Without a biased en-
vironment, the two bias dynamics theoretically cannot exist, since both biases are based on being biased against one another. Further, this thesis acknowledges that cooperation can further be conditioned by programmable similarity, whereby artificial intelligence can mimic the racial, ethnic and/or cultural similarities of the suspect in question, which is thought to have an effect on enhanced comfortability and cooperation with not only AI, but with interrogators as well.

Chapter 6 will then shift focus outward, highlighting how a non-biased environment can and already is achieved within the criminal justice system with the utilization of artificial intelligence. It will highlight its current successes and failures, as well as suggest a three phase validity approach to ensure non-bias maintenance and transparency moving forward.

Chapter 7 will conclude with the findings of this thesis, and provide a proposed outlook on artificial intelligence in policing.
Chapter 1: Artificial Intelligence, Confession Elicitation, and its Application in Police Interrogations

This chapter will examine the potential applicability of AI assistance into police interrogations. In order to address and proactively answer the question proposed in this thesis, it is first necessary to determine the theoretical ability of artificial intelligence-assisted platforms to induce confessions of illegal behavior. The research presented in this chapter shows that artificial intelligence is capable of invoking admissions of illegal behavior in the context of national security clearance interviews within the United States. In conjunction with the research findings, this thesis then suggests that under certain conditions, artificial-intelligence assisted platforms may have the potential capability of eliciting self-incriminatory confessions of illegal behavior and therefore, could be utilized in interrogations. Through the incorporation of recent developments regarding artificial intelligence-human interaction research, this analysis further suggests that manipulating other variables within the design structure could potentially increase the theoretical probability of eliciting self-incriminatory confessions of illegal behavior through increasing cooperation and trust.

1.1 Eliciting Self Incriminating Confessions Of Illegal Behavior Through AI

In evaluating artificial intelligence’s subsequent role in interrogations, it is necessary to first demonstrate a theoretical capability to elicit self incriminating confessions of illegal behavior. This section evaluates self-incriminatory response through artificial intelligence by focusing on United States’ national security clearance procedures and their findings. This section emphasizes that the portion of the national security clearance procedure utilized in this evaluation is based on the admission of both documented and undocumented criminal behavior.

1.1.1 Artificial Intelligence Assisted National Security Clearance Interviews On Self-Incriminatory, Illegal Behavior:

A national security clearance interview (also referred to as federal security clearance interview) is the in-person assessment of the background investigation that targets a variety of

---

14 In stating illegal behavior, the thesis assumes that such illegal behavior admitted is (1) correct information, and not a false confession (2) voluntary
areas such as “personal and professional history, indication of loyalty to the United States, strength of character, trustworthiness, honesty, reliability, discretion, and sound judgment.”

The previous stage of the background investigation includes the written completion of Standard Form 86 Questionnaire for National Security Positions. As per the form, disclosure of illegal behavior is addressed in Section 23 “Illegal Use of Drugs and Drug Activity,” Section 27 “Use of Information Technology Systems”, and Section 29 “Association Record” respectively.

In 2014, the International Journal of Human Computer Studies published “The effectiveness of a national security screening interview conducted by a computer-generated agent” researched by Dean A. Pollina and Allison Barretta, which evaluated the response rate during artificial intelligence assisted national security screening interviews versus self questionnaires of 120 United States Army basic trainees between the ages of 18 and 40. The study was conducted using a computer-generated agent (CG) designed to mimic human attributes including “types of verbal and non verbal behaviors, human like voice characteristics, facial features and facial changes.” In addition, the CG targeted questioning interviewees in the areas of mental health, drug, alcohol, and criminal histories. The CG was programmed to interpret participants’ verbal responses and then proceed to ask subsequent questions based on the answers. The flow chart of the algorithmic design of questioning can be found in Figure 1.1. The content of the interviews was similar to sections of Standard Form 86 (Sections 21, 22, 23, and 24). For the purpose of this present research, it is important to note the study was specifically designed to incorporate Section 23 “Illegal Use of Drugs and Drug Activity” in evaluating self-incriminatory admissions of illegal behavior.

---

15 United States Department of State, All About Security Clearances, DOJ, https://www.state.gov/m/ds/clearances/c10978.htm
16 Ibid
19 Ibid, p. 39
20 Ibid, p. 39
21 Ibid, p.39
22 Ibid, p. 40
The results of the study concluded that national security interviews could be conducted by artificial intelligence in the form of a CG agent. They found that information “including prior illegal acts” could be elicited by the agent. The results showed that, most importantly, 3 participants admitted previous engagement in either criminal alcohol or criminal drug offenses that was never caught. In addition, in terms of illegal behavior, 55 participants admitted to illegal drug use and 17 participants admitted to being previously convicted of a crime. In addition, when able to to speak freely through open ended questions, 14 participants added to their statements (3 participants about drug use and 1 participant on crime).

In terms of relating this study to the context of this paper, artificial intelligence has shown the capability of invoking a self-incriminatory admission of illegal behavior. Even though the study did not focus on Sections 27 or 29, it did focus on Section 23 (among other sections) which substantiated the ability for certain self incriminatory admissions to the CG. In relation to its question design, the Standard Form 86 has a protection clause which states:

"...with respect to Sections 23, 27, and 29, however, neither your truthful responses nor information derived from those responses will be used as evidence against you in a subsequent criminal proceeding.”

---

23 Ibid, p. 47
24 Ibid p. 47
25 Ibid, Table 1, p. 45
26 Ibid, Table 1, p. 45
27 Ibid, 45
28 Questionnaire For National Security Positions
However, it is important to note that this protection clause was not included in the Pollina and Barretta study which specifically informed participants that, “researchers had an obligation to report serious violations of the law.” Therefore, because participants willingly divulged illegal behavior without the protection afforded in the real Standard Form 86, no argument can be made to suggest that the presence of legal protection skewed the presence of illegal admissions. The objective of the study’s research for this present argument intended to establish the ability to achieve such a result. Since the AI-assisted interview showed the ability to elicit self-incriminatory admissions of illegal behavior, incorporating the same approach into police interrogations may also yield similar results (in the area of eliciting a confession of criminal behavior). The following section in this chapter will incorporate the design of this study and additional studies of AI-to-human interaction to show that incorporation of additional AI design manipulation may even improve self-incriminatory response.

It is important to also note that this is one study, and that the sample size was small. Further replication would be needed to substantiate AI’s ability to yield similar results. However, though there are many studies that involve human participants interacting with AI, and even disclosures of undesirable behavior (as presented in this analysis) to AI, the Pollina and Barretta study is the only, current, applicable study in approaching how humans could disclose illegal behavior in this type of format. Since their findings have not yet been reputed (since 2014), this analysis assumes their results are repeatable and valid.

1.2 Designing An AI Model For Enhancing Self Incrimination Of Illegal Behavior

The purpose of Section 1.2 is to address the design of Pollina and Barretta (2014) and discuss other relevant AI-to-human interaction studies. In doing so, this section will first analyze Pollina and Barretta (2014) study’s design that showed the ability to induce a criminal behavior admission. Then, this present research will incorporate other AI-to-human interaction studies to propose an enhanced design structure (adding increased realism and decreased knowledge of perceived human control) stemming from the baseline study.

---

29 Pollina, Barretta, p. 41
1.2.1 Design Of Pollina And Barretta’s Computer-Generated Agent

In Pollina and Barretta’s research, their model was based on previous AI-to-human cooperation research. This section will focus on their design mechanisms including perception of human attributes, perception of human control, and question design.

1.2.1.1 Perception Of Human-Like Attributes

This section will discuss the human like appearance of the CG agent within the study. Pollina and Barretta noted previous research conducted by De Angeli (2009) in PsychNology Journal which indicated that the avatar interviewer must exhibit human like qualities (voice, facial features, expression) in order to “achieve a level of verisimilitude that allows for the CG agent to subsume a level of “intentionality, sociability, and personality.” Subsequent research noted by Pollina and Barretta also indicates that “this increases the degree of presence experienced by the human in the virtual environment.” In this context, perception refers to “the extent to which a person feels immersed in a virtual social environment is related to the person’s perception that there are other sentient beings in that environment.”

As seen from Figure 1.1, the figure is a presumed white male (given the facial feature characteristics and skin tone), although not race nor ethnicity was noted in the design process. In addition, one can see from Figure 1.1 the avatar is also computer generated, so that the interviewee views this entity on a screen versus in a physical presence. Lastly, the figure, though resembling human, is not wholly physically realistic (lack of defining features, lack of full body presence).

1.2.1.2 Perception Of Human Control

The Pollina and Barretta study also mentioned work from Lim & Reeves (2010) and Mohammad & Nishida, (2010), in which suggested that participants who believed that a CG agent was controlled by an external human source, would demonstrate “an increased subjective sense of presence and would therefore be more likely to manifest greater physiological

31 Pollina, Barretta, p. 39-40
32 Gerhard, Moore, & Hobbs, 2004, 2005; Groom et al., 2009; Guadagno, Blascovich, Bailenson, & McCall, 2007; Guadagno, Swinth, & Blascovich, 2011
33 Pollina, Barretta, p. 40
34 Ibid
arousal than participants who believed that the interview was being controlled by a computer.”

The study claims that:

“The participant was informed that they were about to be interviewed by a computer, using a CG character, also known as an avatar, who would instruct them about the acceptable range of responses to different types of questions that would be asked during the interview.”

Pollina and Barretta noted that interviewees had no way of knowing whether the CG agent was acting alone, or being controlled by a computer. They noted that although “interviewees in the present study who believed that a human was controlling the CG interviewer were treating it as a social entity,” though greater physiological arousal was not supported.

1.2.1.3 Question Design
Pollina and Barretta showed that the CG Agent produced questions intended to gather information regarding mental health, drug use, alcohol use, and criminal history in which the previous section in this present research analysis showed success in this area. These were either in the form of direct questions, or open ended questions without context (i.e. “Do you have anything else you wish to add?”).

1.2.2 Expanding The Design Of An “AI Interrogator” In Confession Elicitation
Section 1.2.2 intends to expand on the effective design of the CG model in the Pollina, and Barretta study in the areas of increased realism (via physical presence, physical placement, and human likeness) decreased knowledge of perceived external human control, and direct questioning design. This thesis argues that these changes to the original, successful design structure could enhance and perhaps increase the number of self incriminating confessions of illegal behavior by promoting factors such as trust, respect, likability, and compliance.

1.2.2.1 Increased Realism
The following research presented suggests that AI presence has a significant effect on AI-to-human interaction. Research has demonstrated that the effects of increased realism may

---

37 Pollina, Barretta, p. 41
38 Ibid p.43
39 Ibid
40 Ibid, p. 48
41 Ibid
enhance a person’s view on artificial intelligence, which this present analysis argues could be beneficial in obtaining an incriminating confession.

1.2.2.1.1 Physical Presence

In considering human compliance, respect, and trust, the physical presence of an AI entity appears to be a deciding factor. Bainbridge, Hart, Kim, Scassellati (2008) studied AI-to-human interaction in the form of cooperation, trust, and personal space with “Nico-“, either a physically present humanoid robot or a video-displayed robot. They found, in regards to trust, participants in the physical condition were more compliant than those in the virtual and augmented category and that “[the] physical presence afforded higher trust in Nico’s credibility, making subjects more willing to follow through with an unusual request from Nico.” In Hancock, Billings, Schaefer, Chen, et.al (2011)’s “Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction,” they suggest, in robot design, that “type, size, proximity, and behavior of the robot also affect trust.” In regards to respect, Bainbridge, Wilma & Hart, et. al. (2008) found that the participants afforded the physical Nico increased personal space. The researchers suggests that “personal space could be interpreted as a variable of respect; as humans give personal space to those they are unfamiliar with but respect as human.”

Kiesler et al’s (2008) Anthropomorphic Interactions with a Robot and Robot–like Agent. Social Cognition, participant’s interactions were studied using either robots or the virtual presence of one on a screen. Participants viewed the robot as more dominant, trustworthy, sociable, responsive, competent, and respectful than the agent and rated it more lifelike. However, the study found that in regards to disclosure, though “participants were more engaged, [they] disclosed less undesirable behavior, and forgot more with the robot versus the agent.” Though that finding may initially seem contradictory, they noted that "responses to the present and projected robots did not differ;” however, the word count was less. However, this particular study may have been significantly flawed, because many of the participants

43 Ibid, 705
45 A. Bainbridge, Wilma & Hart, et. al.
46 Ibid 706
48 Ibid 177
49 Ibid 169
50 Ibid 175
were not “fluent English speakers” and researchers noted significant issues with subject response rates due to this. However, a repeating pattern is the association of trust and respect with physical presence versus virtual presence is noted.

1.2.2.1.2 AI -To-Human Physical Placement

The placement of an AI entity in relation to an individual is also a critical component of interaction via compliance. It is important to note that Bainbridge, Wilma & Hart, et. al did not evaluate the effect distance has on AI-to-human on interaction. Supplementary research conducted by Siegel (2009) on behalf of Massachusetts Institute of Technology, tested AI-to-human distance on compliance. In the study, participants were evaluated for compliance standing 2.5 feet (76 cm) and 5 feet (152 cm) from the robot. The results showed that the decrease in distance had the overall effect of decreasing compliance. The study hypothesized this occurred because “they had no prior knowledge, relationship, or interactions with the robot, and likely had little experience with robots in general. It is quite possible that at a close distance the subjects were uncomfortable with the robot.” However, the study also suggested that “this negative response to the decrease in interpersonal distance would likely change if the relationship between the robot and human were altered.” The interpersonal distance will be further discussed in AI in interrogations techniques in Chapter 2. However, what this analysis concludes from both studies combined, is that not only is the physical nature of the AI entity more effective at increasing trust and respect (which could be aspects of a suspect-to-interrogator interaction), this thesis recommends that, in regards to best promoting such an interaction, the design should have a 5 feet (152 cm) physical distance between them until rapport is built.

1.2.2.1.3 Anthropomorphism And The Uncanny Valley

Anthropomorphism is another component relevant to increasing interactions between human and AI. Anthropomorphism refers to “an attribution of human characteristics to things that are not human.” This thesis aims at understanding the effects of anthropomorphism in AI-to-human interaction in the area of an interrogation. Fink (2012)’s Anthropomorphism and

---

51 Ibid
53 Ibid, 114
54 Ibid, 114
55 Ibid, 114
56 Ibid, 114
57 Anthropomorphism, Oxford Dictionary


*Human Likeness in the Design of Robots and Human-Robot Interaction,* claims that people's acceptance of AI entities can be strengthened by increasing “a robot’s familiarity by using anthropomorphic (humanlike) design and “human social” characteristics.”

The results indicated that (1) anthropomorphic embodied shaped robots yield positive results and (2) that the physical shape of the agent, specifically targeting the areas of the nose, the eyelids, the mouth and width of the head had an effect on the perception of human-likeness and perception of willingness to interact. The author included that “visible design is crucial.”

However in *How Humans Respond to Robots: Building Public Policy through Good Design,* Knight (2014) suggests that the ultimate design for an AI entity is human likeness without the discomfort caused by distinguishable but semi-realistic human similarity. This discomfort (the “valley”) is interpreted by the uncanny theory, which proposes that:

“A human appearance or behavior can make an artificial figure seem more familiar for viewers — but only up to a point. The sense of viewer familiarity drops sharply into the uncanny valley once the artificial figure tries but fails to mimic a realistic human.”

In other words, the more it indistinguishably resembles human beings, the more of a “trigger [to] our human-to-human empathy levels.” The uncanny theory was further explored in robots in 2016 in *Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley* in which the studied perceived likability of the robot. They found that:

“Uncanny Valley is not the only important factor in determining a robot’s perceived likability: Perceived emotion, for example, may play a central role; predictably, robots showing more positive emotion were perceived as more likable. However, the Uncanny Valley effect persisted among faces perceived as displaying almost no emotion.”

---

59 Ibid, 203
60 Ibid, 203
64 B. Mathura, Maya, Reichling David (2014). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley, Cognition Volume 146, January 2016, Pages 22-32
65 Ibid
In regards to the uncanny theory and human-to-AI trust, “humans appear to infer trustworthiness from affective cues (subtle facial expressions) known to govern human–human social judgments.”

Modern technology and robotics considerations are close to overcoming the uncanny theory in AI-to-human interaction in engineering emotional-expressive hyper-realistic robots like Sophia and Actroid. Sophia, who gained international attention for being granted legal rights can display 60 different expressions through material developed to mirror human facial muscles and skin. This year, in 2018, advancements made to Sophia’s design included the addition of legs and the ability to walk. However, it is important to note that these findings show that this human-like perception extends beyond a physical interpretation but also involves the ability to demonstrate emotion. Both features need to be assessed in designing an effective AI interrogator. In addition, incorporation of certain facial features and dimensions are critical in the design process. Therefore, it may be more functional to focus on the emotional display of the AI entity first before its hyper-human appearance.

1.2.2.2 Decreased Perceived External Human Control

Previous research has indicated and promoted that participant have knowledge (or perceived knowledge) that an AI entity is being controlled by an external human facilitator. Pollina and Barretta (2014) specifically draw upon work from Guadagno (2007) regarding perceived human operators. Guadagno et al. (2007) focused on the effect of AI-to-human interaction regarding persuasive arguments. The results indicated that in “regard to the effect of agency showed that both genders were more influenced by the human controlled avatar [versus the computer controlled].” In similar research regarding computer controlled versus human controlled AI in a gameplay setting, Lim and Reeves found that “whether or not a player believes that game characters are controlled by other humans can significantly alter primitive responses to game play (i.e., physiological arousal), as well as more thoughtful evaluations of the experiences (i.e., greater presence and likability).”

---

66 Ibid
69 Ibid
71 Siegel
72 Ibid
73 Lim, S., & Reeves, B. (2010)
However, recent research is suggesting that the perceived human operator component may significantly hinder willingness to divulge sensitive information. This analysis draws upon research conducted by Jonathan Gratch at the Institute for Creative Technologies, in Los Angeles, California that supports the idea that people may be more willing to talk with an artificial intelligence entity perceived to be non-human influenced.\textsuperscript{74} In Gratch’s study, Ellie is a human-like, full body, avatar psychologist, equipped with both visual and speech recognition. Results indicate that “those who thought they were dealing with a human were indeed less forthcoming, averaging 0.56 compared with the other group’s average score of 1.11.”\textsuperscript{75} In addition, those who perceived Ellie with human influence actually demonstrated “greater fear of disclosing personal information” and seemingly “managed more carefully what they expressed during the session” versus those who perceived Ellie as strictly AI without human control.\textsuperscript{76}

Research has thereby indicated the decreased knowledge of perceived human control shows a greater willingness to disclose sensitive, personal information. This analysis suggests that applying these recent findings to the design structure may increase the willingness to talk and retrieve personal information, and through the same thought process, information of criminality.

1.2.2.3 Direct Questioning Format

Pollina and Barretta (2014) noted that “None of the participants volunteered any information when asked the open-ended response concerning illegal activities that they were involved in.”\textsuperscript{77} Therefore, it is important to note that open-ended questions without context were not as successful in obtaining information within “illegal activities” versus being asked specific, direct questions.

1.3 Discussion

This chapter outlined the findings with Pollina and Barretta, in which this analysis expanded to suggest participants could disclose criminal confessions of illegal behavior with an artificial intelligence presence. It is important to note that theoretical assumptions drawn from

\textsuperscript{75} Ibid
\textsuperscript{76} Ibid
\textsuperscript{77} Pollina, Barretta, p. 45
this chapter are hypothetical conclusions intended to be tested and explored further through research. Given the infancy of humanistic robots, this idea depends and rests on advancements in the field of artificial intelligence, as well as advancements in programming and engineering. However, the U.S. government’s willingness to employ and entertain such a notion suggests a growing acceptability in the political arena.

Aside from a proposed ability to obtain a confession, this chapter pointed out that humans will willingly engage with artificial intelligence, which is a crucial factor in AI assisted interrogations’ realistic applicability. Research implies humans will willingly cooperate with AI, which allows for improvements on the overall relationship. Artificial intelligence interrogations will be further explored in Chapter 2, regarding its proposed ability to perform the same functions as an interrogator.
Chapter 2: Artificial Intelligence As Interrogators In Police Interrogations

The previous chapter suggested research for the basis of this thesis, due to its potential ability to elicit criminal confessions of illegal behavior. This chapter will first demonstrate how artificial intelligence could hypothetically function as an interrogator, in a policing context, performing the same fact finding exercises utilized by American police in the Reid Technique, and potentially with higher accuracy. Although this analysis will discuss the emphasis on the Reid technique in the American system, which places heavier reliance on deception detection, it acknowledges that the practice done by humans is not necessarily consistently accurate nor reliable, and should not be sole indicator of guilt during an interrogation. Given that it is still the corner stone of American interrogations, this thesis has to consider its functionality in that practice. In addition, this chapter will argue that though human based deception detection can yield inconsistencies, and that AI deception detection could yield the more accurate, reliable results as compared to humans, given its success in studies surrounding emotional and verbal decoding. In addition, this analysis proposes that this deception detection should not be the sole indicator of an interrogator assessing guilt, regardless of improved accuracy with the addition of artificial intelligence. It will conclude with discussing its application, noting its limitations, and assessing the future of AI-assisted interrogations.

This thesis assesses the combination of modern artificial intelligence developments in fields ranging from software development to robotics and mechanics. When speaking about an artificial intelligence interrogator, artificial intelligence refers to:

“The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.”

For the purposes of this thesis, when speaking about robotics, it is referring to “a machine resembling a human being and able to replicate certain human movements and functions automatically.” As per this thesis, it suggests that the tasks performed by human interrogators within the interrogation room could be performed by artificial intelligence and that the interrogator itself would be either virtual or robotic.

---

78 Definition of “Artificial Intelligence”, dictionary
79 Definition of “Robot,” dictionary
To begin with, this analysis acknowledges that an artificial intelligence interrogator could be a humanistic, life-like, computer generated avatar (i.e. without a physical presence) or a cyber-physical robot. Though the combinations of designs may theoretically work, this thesis demonstrated in Chapter 1 that human-to-AI interaction is positively increased when the subject is physically interacting with the robot entity. Ideally, the increased interaction would serve beneficial to the overall cooperation between interrogator and suspect.

Secondly, the design process of this artificial interrogation, could be based using expert design of decision tree reasoning, or using predictive machine learning, or could be a combination of both. Expert design (or expert system) in artificial intelligence refers to computer programs that “simulate the judgment and behavior of a human or an organization that has expert knowledge and experience in a particular field.” In reference to decision tree learning, this analysis proposes that the AI design incorporates relevant attributes and relating to value function in how interrogators interact and question subjects. Certain eye movements for example, are used as possible indicators of guilt. Using this method, the predictive measures used in this analysis will use values of that attribute, in which the decision tree model will follow a path corresponding with that value. Another approach could be through machine learning, referring to the computer’s learning process bereft human intervention. Machine learning is a section of artificial intelligence which is “based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.” Human rights and due process issues regarding machine learning will be discussed later in this thesis.

2.1 Applying Proposed AI Model Designed For Police Interrogations

Section 2.1 will demonstrate that the design detailed in Chapter 1 could be realistically applied to an interrogation setting. This analysis will focus on two particular aspects of the interrogation process: rapport building and deception recognition through baseline question-

---

80 Method suggested by Malcolm Langford, advisor
81 Rouse, Margaret, “DEFINITION expert system,” TechTarget, https://searchenterpriseai.techtarget.com/definition/expert-system
83 Ibid
ing. The dominant method in the American criminal justice system is the Reid-based interrogation, which is “designed to evoke particular behavioral (verbal, nonverbal, and paralinguistic) responses,” with the intent of ascertaining whether the suspect is being truthful or deceitful. This thesis will then argue that the design model presented in Section 1.3 could theoretically execute the same functions, if not with more accuracy, as than a human interrogator.

2.1.1 Rapport Building Through AI

This thesis states that an interrogation can be constituted as “effective’ if it safely and voluntarily elicits the suspect to a confession of guilt of criminal behavior from the suspect. An interrogation normally begins with a rapport building process between the suspect and the interrogator. In this process, the interrogator attempts to appeal to the suspect through conversation, sharing similar likes or preferences in an effort to make (1) the suspect more comfortable and (2) make it more difficult to lie or end the conversation when the topic shifts to criminality. This concept has also been the subject of studies, in which researchers found “cheap talk (i.e., costless, non-binding signals) … to lead to greater human cooperation in repeated interactions.”

Given the algorithmic design of the Pollina and Barretta CG Agent in both asking questions and processing information, it is plausible to create an AI interrogator that can do the same function, in the context of the situation. In essence, this approach is already used in chatbots, company’s self service help chats, etc. In addressing people’s willingness to talk to chatbots, Jeremy Pounder, futures director at Mindshare, remarked that “Surprisingly consumers are very receptive to AI chat technology, and even trust it over human interaction in certain scenarios.”

However, unlike the lack of physical visuality that occurs via chatbot algorithms through computers that can mask the concept of talking to a “robot,” the physical presence (whether virtual or actually physical) of the AI entity would also rely on the comfortability of

---

88 Ibid
the suspect. Pollina and Barretta study noted that “Almost half of the interviewees’ reported that they had at least some preference for the CG interviewer over traditional paper forms, indicating that they felt comfortable with the process.”

What this thesis suggests is we do not have to create that “comfortability” because people have already become used to automated chat bots overtime. However, we indeed have to improve the comfortability level, especially introducing a new scenario and context of interaction. Therefore, regarding the rapport building process, it is essential to draw from the enhanced design model from 1.2.2, especially in the areas of AI-to-human placement and AI physical human likeness.

2.1.2 Baseline Questions And Response

During an interrogation, baseline reactions are established during the initial rapport building conversation and questioning in which the interrogator studies the suspect’s verbal and non-verbal responses. The purpose of these baseline questions is to prompt brain activity from especially both simple memory recall and creativity. This can be observed in the following example scenario:

“When the suspect is remembering something, his eyes will often move to the right. This is just an outward manifestation of his brain activating the memory center. When he's thinking about something, his eyes might move upward or to the left, reflecting activation of the cognitive center. The detective makes a mental note of the suspect's eye activity.”

Once this is established, the theory behind guilt or deception is that truthful answers will target the portion of the brain responsible for memory in which the suspect’s eye moves may show movement towards the right, versus creativity, in which the suspect’s eyes would yield to the left. In addition, non verbal communication would be observed for documented signs of lying including arm, hand, finger, leg, and foot movements. This thesis suggests that, given artificial intelligence advancements, an AI interrogator is (1) able to perform this technique but (2) hypothetically achieve more accurate results than a human interrogator.

---

93 Pollina, Barretta, 47
94 Layton
95 Ibid
96 Ibid
97 Ibid
98 Ulatowska, Joanna (2013). Indicators Of Deception In Different Lie Scenarios. Roczniki Psychologiczne / Annals Of Psychology 2013, Xvi, 1, 127-146
following subsections demonstrate this ability through AI facial emotional recognition, AI verbal emotion recognition and deception recognition software.

2.1.2.1 AI Facial-Emotional Decoding And Recognition In Humans

This thesis uses research regarding the ability of artificial intelligence to recognize and decode human emotions. During these baseline questions, the human interrogator would be looking for facial changes to implicate a suspect. A study conducted by Ohio State University\(^\text{99}\) analyzed “8 facial expressions [from] 184 people from different genders, ethnicities, and overall skin tones” through color patterns of facial blood flow against either computer based algorithm or a human being.\(^{100}\) They found that the computer generated algorithm detected changes in human emotion at a 90% accuracy rate versus being being detected by humans at a 75% accuracy rate.\(^{101}\) It is an essential component that the AI interrogator be able to synthesize emotion, but also in real time. A second study, conducted by Kobayashi and Hara found that through the use of neural network application, based on machine learning, correct facial emotional recognition was established in real time.\(^{102}\) Therefore the a key technique of a human investigator- reading emotional changes, in real time, can be executed by an AI entity.

2.1.2.2 AI Verbal-Emotional Decoding And Recognition In Humans

This section focuses on advancements of verbal-emotional decoding and recognition as assisted through AI. Research has suggested that changes in voice pitch is one common deception indicator that may indicate lying.\(^{103}\) The changes in Hertz are quite small and detectable with “special equipment.” (i.e. not human hearing alone).\(^{104}\) In 2017, a study conducted at Nirma Institute of Technology in Ahmedabad, Gujarat, India, researches designed an “algorithmic approach for detection and analysis of human emotions with the help of voice and speech processing.”\(^{105}\) The study found that through comparison of normal, angry, and panicked voice states, the nervous or panicked state had a “significant increase in the mean values of pitch, time spacing between consecutive words, and increased timbre ascending

---


\(^{100}\) Ohio State University. (2018). At first blush, you look happy—or sad, or angry. Ohio State University. https://news.osu.edu/at-first-blush-you-look-happy—or-sad-or-angry/

\(^{101}\) Ibid


\(^{103}\) Ulatowska

\(^{104}\) Ibid

time.” In general, the study concluded that “with varying emotions, the tonal parameters accordingly change as well.”

This verbal emotional connection could be used in conjunction with the aforementioned facial emotional recognition software to create an AI interrogator with the capacity of analyzing both visual and verbal emotional changes in real time. Subtle changes in these categories rely on the detection skills of a human interrogator; however given biological limitations, the subtleties may not be so easily caught. Modern artificial intelligence advancement has improved the accuracy of these deception cues, which would be beneficial in an interrogation room setting.

2.1.2.3 Deception Recognition Software

Recent developments have led to the advancement of AI deception recognition software. In 2017, the University of Maryland published research studied through the “Deception Analysis and Reasoning Engine (DARE)” in court trial videos to recognize deception cues. By using “multi-modal feature extraction” including motion, audio, and transcript features, feature encoding, facial micro-expression prediction, researchers concluded that:

“Our vision system, which uses both high-level and low level visual features, is significantly better at predicting deception compared to humans. When complementary information from audio and transcripts is provided, deception prediction can be further improved.”

Given the similarity to detection deception in the courtroom, and how it would occur in an interrogation room, this thesis suggests that this same method could be utilized in the interrogation room, and potentially yield similar results.

2.2 Discussion

This chapter has illustrated how artificial intelligence could have the ability to recognize and decode both visual and audio changes in the areas of deception detection. In theory, these abilities, which are already possessed and employed by human interrogators, could also be applied to artificial intelligence. This chapter primarily focused on its integration within the American Reid-based interrogation, which is “designed to evoke particular behavioral
(verbal, nonverbal, and paralinguistic) responses,” with the intent of ascertaining whether the suspect is being truthful or deceitful.\textsuperscript{110} This chapter acknowledges that the Reid Technique itself is not alone responsible for eliciting of false confessions, with the \textit{unlikelihood} of “producing more false than true confessions (simply given the likely base rates associated with innocent vs. guilty suspects).”\textsuperscript{111} However, this thesis suggests that (1) humans are incapable of reliability and accurately assessing guilt or innocence in an interrogation and (2) there is too much reliance on this deception detection as an indicator of guilt.

As for the reliability and accuracy of human based deception detection, Richard A. Leo, Professor of Law and Psychology at the University San Francisco School of Law, noted that”

“Social scientific studies have repeatedly demonstrated across a variety of contexts that people are poor human lie detectors and thus are highly prone to error in their judgment about whether an individual is lying or telling the truth.”\textsuperscript{112}

This chapter has presented relevant studies that support the notion that artificial intelligence can detect physical changes that could be helpful in deception detection. It does not argue that artificial intelligence will be a foolproof means of algorithmically assessing innocence or guilt. However it does suggest that if the Reid Technique continues to be the default method of interrogations, application of artificial intelligence may provide more reliable and accurate indicators of physical change, for the purposes of deception detection. Studies have supported the notion that artificial intelligence can perform better in decoding human emotion both visually and audibly. Ideally, using AI in this decoding might safeguard against false confessions, in which human deception detection may pick up false positives of guilt-suggestive behavior, and thereby subject innocent victims to unfavorable and coercive interrogations.

Secondly, this thesis also acknowledges that deception detection should not be the primary means of assessing innocence or guilt in interrogations. Leo points out that “Once detectives misclassify an innocent person as a guilty suspect, they often subject him to an accusatorial interrogation.”\textsuperscript{113} As previously stated, these behavioral evoking questions by using

\textsuperscript{110} Davis, Deborah and Villalobos, J. Guillermo (2016)
\textsuperscript{113} Ibid
the Reid Technique during an interrogation are thought by the interrogators to be indicators of guilt or innocence, sparking either continuance or discontinuance of an interrogation. Reliance on this method, by either humans or artificial intelligence, should not be the primary indicator of guilt or innocence leading to an intense interrogation. This thesis’ purpose is not to debate other types of interrogation methodology. However, it acknowledges that artificial intelligence could also be implemented within the P.E.A.C.E interrogation method. This method, or Preparation and Planning, Engage and Explain, Account, Closure and Evaluate, is a “less confrontational interview and interrogation” style employed in the United Kingdom. Given its interview style approach, artificial intelligence could have the ability to interpret and store the information provided by the suspect. It also relies on rapport building, which this thesis suggested was a potential capability of artificial intelligence.

2.2.1 Applicability

Police departments have urged technological advancement in policing. In 2013, Damian Green, Minister for policing, criminal justice and victims in the United Kingdom remarked that, "In five years' time we need to look back and see this was the beginning of a technological revolution in policing." In regards to applicability, this thesis is focusing on the functional applicability the model presented in Chapter 1 (in terms of functioning as an interrogator) in an interrogation room. In evaluating the previous studies’ methodological design approaches, Chapter 1 presented an expanded AI design that could theoretically enhance or promote the probability of eliciting confessions by increasing trust, respect and influence. This chapter then demonstrated the applicability of a proposed AI design model that would enhance standard interrogation practices through artificial intelligence advancements in the fields of emotion and deception recognition. Existing research and practice indicate that artificial intelligence may have the ability to not only elicit admissions of illegal behavior but could engage in fact finding techniques at a higher accuracy rate than that of human-interrogators during interrogations. Therefore, given the preliminary findings of modern research, this analysis suggests that this method can be a realistically applicable tool in law enforcement.

115 Ibid
2.2.2 Limitations And Further Research

In addressing this method’s applicability, it is also necessary to draw upon its proposed limitations. This section will address the financial and legal limitations of the proposed design. Technological limitations concerning algorithmic bias will be analyzed in subsequent chapters.

2.2.2.1 Financial Limitations

Firstly, given the financial investment of developing and maintaining highly-detailed artificial intelligence interrogators, these features may pose a challenge to police departments. ASIMO, reported to be “the most advanced humanoid robot” (though its features are stereotypically robotic), is reported to cost ~$2,500,000 USD (~€2,158,629 EUR).117 Police department, in addition, may not allot sufficient funds into technological advancements for police officers. Departmental allocations of technological funding are not necessarily available for all countries, states or providences. For example, in the United Kingdom, it is reported that out of £14bn spent on policing, only 3% [in some forces] is used on on technology.118 In 2014, it was reported that increased use of technology within the West Midland’s police in the United Kingdom had an estimated cost of £25m (~€28,524,366 EUR).119 As for the United States, this thesis could not find technology within police forces as part of the 2018 federal funding.120

2.2.2.2 Legal Limitations

For the purposes of discussing legal limitations, I will only cover proposed legal applicability within the interrogation room. Further detail into legal acceptability will be discussed in Chapter 6 in relation to access and transparency. The first legal limitation is the lack of precedence, being that the extension of artificial intelligence into the interrogation room has not been yet evaluated as a legally permissible format to obtain confessions. Within that same line of reasoning, the mechanical usage of deception detection software utilized by this AI interrogation may be compared to polygraphs, though polygraphs are more invasive than the AI interrogator model and are significantly more restrictive in trial proceedings. Notable

---

118 Anderson, Tim (2013)
countries that legally permit and use polygraphs within the interrogation room are the United States, Canada, the United Kingdom and Australia. However, its legal admissibility as evidence in court proceedings varies. Whether the use of this methodology to 1) obtain the confession or 2) use that confession as evidence would have to still be evaluated in the law.

The second legal limitation of the use of this model is whether or not the AI entity itself could be legally recognized as an interrogator. Again, there is currently no legal precedent concerning this issue. However, several AI legal advancements suggest a potential resolution of the issue. In 2017, the previously mentioned Sophia robot was granted legal personhood and citizenship in Saudi Arabia. However, artificial intelligence accountability is highly debated, regarding the issue of programmer responsibility and the future autonomous decisions that an AI entity can make.

---


125 Reynolds

Chapter 3: The Interrogator-To-Suspect Interface:

Thus far, this thesis has presented the hypothetical AI functionality in interrogations. Now, this thesis will shift forward to address the racial and gender bias thought to be impacting false confessions. This chapter will introduce the Bias-Uncooperative Loop, which this thesis argues may be factor contributing to the racial and gender statistic anomalies in false confessions. This chapter will subsequently highlight the critical effect racial and gender bias has in a police interrogation in eliciting confessions. It suggests that understanding the bias dynamic within the interrogation room requires analyzing the effect of bias from both (1) the interrogator-to-suspect interface and (2) suspect-to-interrogator interface. The review of research in following chapters seeks to identify current bias issues in both interfaces, before discussing of how artificial intelligence could theoretically eliminate the Bias-Uncooperative Loop thought to affect false confessions, which will be covered later in this thesis.

The focus of Chapter 3 will concentrate on the position of the interrogator in reference to the suspect, and how the his or her bias (both racial and/or gender) is further exacerbated by standard interrogatory techniques including deception detection and presumption of guilt. This analysis contends that though these techniques may prove helpful, the underlying fallibility may be due to the preconceived notions of the interrogator. This chapter will present evidence to suggest that racial and gender stereotypes do influence (1) suspect perception of guilt and (2) cooperation and overall perception of the suspect.

3.1 Interrogator Bias Influence And Perception Of Guilt

This section will begin with understanding interrogatory techniques and the dangers surrounding both human deception detection and presumption of guilt. In combination with those findings, this analysis will highlight how these fallible methods intensify these biases (racial and gender in the context of this thesis) which may be a crucial component of police induced false confessions.

127 Redlich, Meissner,
3.1.1 Deception Detection And Presumption Of Guilt

First and foremost, it is important to note the lack of supporting scientific research and evidence between human interrogators’ “hunches” and the accuracy in assessing guilt. In Redlich and Meissner’s *Techniques and Controversies in the Interrogation of Suspects: The Artful Practice versus the Scientific Study*, they point out both a lack of training and usage of the scientific method in regards to employing these interrogation practices. They note that current scientific research does not bolster the notion that “behaviors or response styles” are necessarily reliable indicators of distinguishing between innocence or guilt, nor does it support the foolproof skill or expertise of the interrogator. Richard A. Leo Ph.d and J.D. points out that, “American police are taught, falsely, that they can become human lie detectors capable of distinguishing truth from deception at high, if not near perfect, rates of accuracy.” Numerous research studies cast strong doubt on the Reid technique's purported 85% accuracy rate, with research suggesting an approximate 53% accuracy in deception detection (Bond and DePaulo (2006)). Redlich and Meissner further suggest that this “hunch” “pseudo” scientific method is worsened with a spiraling effect from the primary deception detection, which enables the presumption of guilt, and from there, the following techniques and practices to gain a confession.

The deception detection is a critical component of investigator bias, which is suggested by Redlich and Meissner to contribute to confirmation bias. Confirmation bias refers to “the phenomenon in which information that is consistent with one’s hypothesis or expectations is given credence, whereas information that is inconsistent is discounted, ignored, or actively re-interpreted to be consistent with the hypothesis (Darley & Fazio, 1980; Nickerson, 1998).” In Meissner & Kassin 2004, research suggested a link between this investigator bias and false confessions. Redlich and Meissner found that when together, “the use of questionable deception detection techniques and a strong presumption of guilt on the part of

---

128 Ibid
129 Ibid
130 Leo, Richard A. (2009)
131 Redlich, Meissner
132 Ibid
133 Ibid
134 Ibid
135 Ibid
the investigator can be dangerous to innocent suspects, placing them at risk for the pressures of interrogation (Kassin 2005; Meissner & Kassin, 2004).”

3.1.2 Racial And Gender Interrogator Bias

Research and statistics support the notion that race/ethnicity and gender bias may play a significant role in contributing to police bias, and thus confirmation bias, and the outcome of an interrogation. This thesis acknowledges that racial and gender bias occurs outside of police custody though committed by police, but for the purposes of this analysis (with the focus on interrogations), evidence on racial/gender treatment will be within police custody and an interrogation/interview setting.

3.1.2.1 Implicit Racial Bias Influence

In the United States, social media and technology (i.e. car-cams and body-cams) have given rise to a nationwide outrage concerning police maltreatment of black or ethnic civilians. Phillip Atiba Goff, Ph.D, social psychologist at the University of California, Los Angeles and co-founder of the Center for Policing Equity found that in surveying 12 different police departments, “black residents were more often subjected to police force than white residents.”

A large portion of debate regarding bias within the United States regarding racial injustice concerns a both conscious and unconscious racial bias of police officers. John Dovidio, Ph.D, a social psychologist at Yale University, noted in regards to implicit biases, “a large proportion of white Americans have these [implicit] biases, and it's hard to expect police officers to be any different.” However, such implicit racial bias is not racially divided, with research supporting that the black community, including black police officers also exhibit negative racial stereotypes towards their own group. Weir (2016) points out that “when officers' training and experiences confirm racial stereotypes, those biases appear to hold more sway over their behavior.” She specifically points out the “unconscious association between black individuals and crime,” in which she suggests that this implicit notion can affect police behavior. Furthermore she notes that it is not necessarily anchored to obvious or explicit racist beliefs of the individuals.

136 Ibid
138 Ibid
140 Ibid
141 Ibid
142 Ibid
In the context of the interrogation, there appear to be a few research studies focusing on the implicit bias. In Appleby (2015), *Guilty Stereotypes: The Social Psychology Of Race And Suspicion In Police Interviews And Interrogations*, the author found that the interrogation room was also subject to race-based police bias. Appleby reports that “innocent Black suspects are at a greater risk of being erroneously judged as guilty during police interviews and interrogations.” Further, the study found that “police officers judged Black suspects to be less cooperative and less forthcoming than White suspects.” This “perceived” cooperation, or lack thereof, may also be a factor in the deception detection and presumption of guilt. The same study found racial differences in selection of “different strategies” within the interrogation, noting that Black suspects were not as forth coming as White suspects, for reasons not answered in the study. The author theorized that “Police may be exploiting this lack of information when making lie judgments,” while noting that this assumption needed further analysis and research before definitively coming to that conclusion. This thesis will point out that this cooperation dynamic is key to both the suspect and interrogator interface, which will be discussed in Chapter 4.

3.1.2.2 Implicit Gender Bias Influence

Gender bias may also play a significant role in how interrogators go about eliciting a confession from a suspect. As previously noted in this thesis, false confessions made by females may be influenced on the gender-biased, societal expectations of the interrogator, especially in the areas of maternal and family expectations.” In 2011, a study conducted by the University of Bristol found that in a survey of 50 incarcerated British females that women were both more susceptible to coercion and “threats to family responsibilities.”

Again similar to race in the interrogation room, this type of bias is also under-researched as to how and why this occurs. In understanding this phenomenon, this thesis draws upon the 2014 study findings of Barry C. Feld at the University of Minnesota Law School,

---


144 Ibid

145 Ibid

146 Ibid

147 Jones, Stephen (2011)

who found in juvenile justice based interviews, substantial differences in suspect perception on the basis on gender.\textsuperscript{149} The study reported that:

\textit{“Despite their objective similarities, interviews with juvenile justice personnel report substantial differences in how they perceive boys and girls. While some described girls as more cooperative than boys, most offered much more negative characterizations of females, describing them as emotional, confrontational, or verbally aggressive.”}\textsuperscript{150}

Stereotypical association with females is also found in cases dealing with children. Andrea Lewis writing on behalf of University's Center on Wrongful Convictions, in the Albany Law Review commented that, "When a woman is accused of child abuse or murder after the death of a child she is judged both societally and legally through stereotyped ideals of womanhood and motherhood.”\textsuperscript{151} Lastly, in combination with their false confession statistics, women may be even more likely to be accused of “mistaken crimes” (i.e. homicides found to be suicides) at a rate of approximately 66\%\textsuperscript{152}

Assuming that interrogation practices are consistently the same regardless of gender, and consistent with the findings of Feld's juvenile delinquent study, gender-sensitivity training may be necessary to avoid false confession dilemmas.\textsuperscript{153} The stereotypical view of females (emotional, etc) appears to directly influence the perception of the suspect, which as previous research via this thesis has noted, could hypothetically impact the deception detection and perception of guilt. In the context of women’s perception of interrogation dominance/control and police force’s majority of male officers, Cleary and Bull (2018) suggest that:

\textit{“These findings speak to the manner in which interrogating officers relate to suspects within a specific social interaction, it is perhaps useful to consider them in the context of the extremely high rates of prior victimization reported by incarcerated American females.”}\textsuperscript{154}

Therefore, this thesis suggests that gender bias plays a significant role within the interrogation room.

\textsuperscript{150} Ibid
\textsuperscript{151} Knox, Amanda (2016)
\textsuperscript{152} Ibid
\textsuperscript{153} Feld
3.2 Discussion

The issue this thesis highlights is the racial and gender bias prevalence in both the deception detection and the presumption of guilt affecting *innocent* suspects. Redlich, Meissner point out that “the interrogation techniques…often produce true confessions by guilty suspects, but simultaneously increase the risk of false confessions by innocent individuals who are subjected to these same procedures.”\(^{155}\)

Research presented here has shown that the already criticized deception detection and presumption of guilt is inherently skewed by human fallibility and worsened by both conscious or unconscious biases that infiltrate the fact finding process. This thesis argues that the interrogator’s biased view of a suspect enables the Bias-Uncooperative Loop. The bias infiltrates both the deception detection and presumption of guilt, leading the interrogator to perceive the suspect as uncooperative, which in turn leads the interrogator to perceive the suspect as guilty. Leo writes that “the subject who is guarded, uncooperative, and offers broad denials and qualified responses is also believed to be deceptive and therefore guilty.”\(^{156}\) The common denominator of “uncooperative” found in both racial and gender studies suggests a link to the loop process. This might help to explain the racial and gender statistical divide in false confessions within the criminal justice system.

Through research has continued to highlight the police-based bias dilemma, less research, (so it seems), has been done to target and effectively minimize bias altogether. Current research attempts to pin point unconscious racial bias suggests that racial stereotype workshops or “interventions” have been unsuccessful in long term racial bias reduction.\(^{157}\) Weir commented that:

“In two studies with more than 6,300 participants, all of the interventions reduced implicit prejudice in the short term. But none of those changes lasted more than a couple of days following the intervention—and in some cases, the effects vanished within a few hours (Journal of Experimental Psychology, 2016).”\(^{158}\)

Thus, intentional stereotype reworking of our brains may work in the short term, but given that bias is likely deeply rooted in our subconscious, such techniques may not provide an easy fix.

\(^{155}\) Redlich, Meissner
\(^{156}\) Leo
\(^{157}\) Weir
\(^{158}\) Ibid
This chapter showed that both racial and gender stereotype bias (when held by the interrogator) seemingly influence both (1)suspect perception of guilt and (2)overall suspect cooperation perception. It introduced the first portion of the Bias-Uncooperative loop which, in Chapter 4 will be discussed in further detail with regard to suspect perceptions of the interrogator.
Chapter 4: The Suspect-To-Interrogator Interface:

In Chapter 3, this thesis began the two-part interaction dynamic between the interrogator and the suspect. The focus of Chapter 4 will consider the position of the suspect in reference to the interrogator within the context of eliciting false confessions. This thesis utilizes the findings from current research, which suggest that suspect cooperation is a reflection of an attributed interrogator bias. The second portion of the Bias-Uncooperative Loop concerns the suspect in this interaction dynamic, specifically how the suspect perceives the interrogator as racially or gender biased and, thereby, internalizes this projected bias into assuming the interrogator finds him or her already guilty by default. This theory is thought to lead to a “perceived” uncooperative suspect. This reinforces the perception of the suspect by the interrogator, whereby the underlying interrogator bias can affect the interrogator’s impression of cooperation. This loop may represent a “chicken or egg” dilemma. Furthermore, the suspect perceiving this stereotype threat and police bias may cause the suspect to be less forthcoming with information, which potentially could heighten the intensity of an interrogation, resulting in a false confession.

The focus of Chapter 4 will first consider the theories behind false confessions and how and why they occur with innocent suspects. From there, this thesis will focus on the Bias-Uncooperative Loop using Appleby’s (2015) suggestions as to why suspects may hold this underlying assumption. Similar to the previous areas of interrogation discussed, the number of studies reflecting interactions between suspect-to-interrogator perception is minimal and also needs further research.

4.1 Types Of False Confessions

For the purpose of this analysis, Section 4.1 will first discuss three types of false confessions: voluntary false confessions, compliant false confessions, and internalized false confessions. It will then proceed to discuss current psychological theories which may play a role in why innocent suspects confess to crimes they have not committed.

4.1.1 Voluntary False Confessions

Voluntary false confessions are when an innocent individual “voluntarily” confesses to a crime he or she has not committed. Though there are several subcategories of potential

---

theories as to why these particular false confessions occur, for the purpose of this analysis, the influence this thesis focuses on will be due to “feelings of guilt over past transgressions.”

This thesis suggests that this type of confession may have a connection within the gender bias category. Due to the usage of guilt in this particular false confession, this thesis suggests that there may be a link between the “emotional” stereotypical categorization via interrogators of women suspects that may put them at risk of this type of false confession. Further research however should be done to examine this proposition.

4.1.2 Compliant False Confessions

In compliant false confessions, the suspect feels that a confession is the only way out of the situation. Reasons as to why suspects elicit these types of confessions include: “To escape a bad situation, To avoid a real or implied threat, To gain some kind of reward.” For the purposes of this analysis, the “real or implied threat” portion may be linked to the suspect-to-interrogator perception of guilt, and perhaps a more intense interrogation, which will be discussed in this chapter.

4.1.3 Internalized False Confessions

Internalized false confessions are the third type of false confession in which suspects “believe they are in fact guilty, even though they have no recollection of the crime.” This type of false confession may also be influenced by a gender bias. It could be argued that women are more susceptible to these types of false confessions because they are reported to be more receptive to false memory syndrome. A statistical study indicated, when comparing occurrence of false memory syndrome between the two separate genders, approximately 92 percent were female. Again, multiple factors and numerous unanswered questions still exist for this false confession phenomena, and false confessions in general, but this analysis urges consideration of a stereotypical influence as a factor for this gender divide.

---

160 Ibid
161 Ibid
162 Ibid
163 Ibid
164 Knox
165 Ibid
4.2 Suspect Component Of Bias-Uncooperative Loop

4.2.1 The Bias Component

Previous research suggests that a suspect’s perception of being perceived as guilty may impact how he or she is perceived as uncooperative by the interrogator. Again it is important to acknowledge that this perception as “uncooperative” may be due to increased stress and anxiety levels, as suggested in the Appleby study. Thus, it requires further research and studies for a more definite correlation. The following two sections will discuss how the suspect hypothetically responds to assuming interrogator bias is prevalent in the interrogation room.

4.2.1.1 The Racial Bias Stereotype Threat

In regards to how race effects perception in the interrogation room, Appleby suggested that the higher rate of innocent black suspects found guilty could be due to the “concerns about being stereotyped as criminals.” The study pointed out that:

“Research on stereotype threat shows that when Black participants are asked to identify their race on a demographics questionnaire prior to the task, they perform worse than participants asked to complete the demographic questionnaire after the task (Steele & Aronson, 1995). That is, when race is not primed – as it had not been in the pre-interrogation interview – participants do not yet have the concern that they will confirm cultural stereotypes.”

Returning to Chapter 3 and the discussion of the unconscious or conscious association between the black community and crime, it is possible that this guilty perception by interrogators is misunderstood by Black individuals fearing confirmation of the stereotype. This analysis assumes that the racial stereotype threat evident in the Black community could extend to other ethnic groups as well who also have a social stereotype of related crime. In supporting that notion, this thesis cites Davis and Vllaalobos (2016) published in *Advances in Psychology and Law*, who found that:

“Stereotypes linking minority groups to crime in general (such as race) or to specific crimes...can lead minorities to feel increased arousal and anxiety during interrogation. In turn, this can lead to enhanced risk of being perceived as deceptive... and therefore of being subjected to a coercive interrogation.”

---

166 Appleby
167 Ibid
168 Ibid
169 Davis, Deborah and Vllaalobos, J. Guillermo (2016)
Appleby, did not distinguish by race in her study, in regards to anxiety and stress detection, but suggested that more research would have had to have been conducted in this area before a definitive difference in levels could be made.

4.2.1.2 The Gender Bias Stereotype Threat

Women, on the other hand, may also suffer from this stereotype threat, causing them to behave and respond accordingly like the stereotype. Minimal statistics and studies exist regarding understanding gender division within the interrogation space. In 2016, Amanda Knox, the American women who caught international headlines for her imprisonment in Italy following the death of her roommate, published an opinionated article covering her experience with a woman’s treatment within an interrogation room. She suggests that a societal hard-wired stereotypical view of women impacts women’s self view in an interrogation.170 She points out that:

“Women are raised under a different social incentive structure than men, where attitudes of compliance and deference to authority are more encouraged. This finds its most damning realization in the interrogation room, a situation designed to amplify the absolute control and authority of investigators.”171

This thought process is consistent with current research conclusions. In Cleary and Bull (2018), they found that, “Women did report a greater sensitivity to Dominance/Control and a slight preference for Humanity/Integrity approaches compared to men.”172 Though they did not notate whether the interrogating officer was male or female during the study, they were inclined to believe it was in reference to a male interrogator, due to the percentage of male to female officers173 in the police department.174 In addition, Northwestern psychology and gender studies profession Alice H. Eagly, suggests that, “Women are commonly in roles that involve caring and cooperation. Expectations are formed for men to be more influential and women more easily influenced.”175 Could this power struggle and stereotypical adherence

170 Knox
171 Ibid
173 “Statistically likely, as only 10% of supervisory officers in local police departments (Reaves, 2015) and 12% in local sheriff's offices (Burch, 2016) are female.”
174 Cleary
175 Knox
to gender roles play a part of women and the likelihood for false confessions? Again, this analysis urges further study and consideration.

4.2.2 The Cooperation Component

This thesis also argues that in addition to the suspect being biased (perceived as uncooperative), this “un-cooperation” extends to the suspect’s willingness to disclose information. The Appleby study found that in interpreting the interviews, “Black suspects did withhold more information about their actions from the interviewer than the White suspects.”\(^{176}\) Theories as to why such occurred are debatable and need further research. Appleby, as well as this thesis, suggest that it may be the influence of both underlying bias and police trust.\(^{177}\) Thereby, this thesis suggests that this interpretation as “un-cooperation” could potentially lead to more intense interrogations, and produce a compliant false confession.

4.3 Discussion

This chapter presented the second component of the Bias-Uncooperative Loop demonstrating that the suspect’s perception of a biased environment can contribute to the overall perception of guilt. This chapter presented theories and scenarios in which these types of false confessions may occur. Again, the current research examining the false confession phenomena as well as racial and gender differentiation in interrogations are for the most part hypothetical and conclusions drawn from them are not absolute. The minimal studies regarding race and gender behavior within an interrogation room make it more difficult to have any definitive conclusions regarding reasonings behind false confessions.

This chapter intended to point out the bias perceived through the suspect and its proposed effect on the interrogation, the interrogator, and false confessions. Though a variety of factors could play a role in this perception, this analysis drew upon the findings of the Appleby study suggesting a racial stereotype threat in combination with police trust in willingness to disclose information. In addition, it took a theoretical/philosophical approach when it came to how societal stereotypes of women also may have the same effect on this uncooperative perception. Ideally, we would want to discourage the perception of suspect un-cooperation but enhance the actual cooperation by means of willingness to disclose information. This thesis

\(^{176}\) Appleby

\(^{177}\) Ibid
suggests this may be a reason for false confessions, as the interrogation may intensify and then result in the suspect falsely confessing. Besides minimizing the bias component, this thesis suggests trust and identifying with the interrogator could be important factors in suspect willingness to cooperate.
Chapter 5: Non Bias Environment And Programmable Similarity On Human-To-Ai Cooperation, Trust And Safeguarding Against False Confessions

In the previous two chapters, this analysis has demonstrated the problematic bias influence of both the suspect and the interrogator within the interrogation room. This analysis holds the opinion that the Bias-Uncooperative Loop may help to explain the racial and gender divide in statistics. This chapter will expand even further on the theoretical artificial intelligence interrogator discussed in Chapters 1 and 2, which could hypothetically (1) minimize (or ideally eliminate interrogator bias) through artificial intelligence’s ability to be racial/gender neutral, and therefore (2) minimize (or ideally eliminate) the suspect’s anxiety, stress and stereotype confirmation by removing the interrogator’s implied bias, ideally to (3) to enhance the suspect and interrogator interactions via programmable suspect similarity, which in turn could (4) minimize the risk of the Bias-Uncooperative Loop and false confessions. In regards to the theoretical non-bias interrogator portion of this chapter, it is important to note that bias has been evident in past artificial intelligence decisions, and any artificial intelligence platform cannot be completely exempt from some programmer bias. This chapter intends to present past evidence to suggest that artificial intelligence can be racial and gender neutral within a policing atmosphere, and therefore, could potentially exert the same success in an interrogation room setting. However, this thesis requires that Chapter 5 bias assumptions be considered in conjunction with Chapter 6 regarding weaknesses, maintenance, and transparency needs.

5.1 Positive Studies: Racial And Gender Neutral Artificial Intelligence Models

This thesis will argue that artificial intelligence exhibits the capability of being racial and gender neutral in decision making, therefore, making it an idealistic approach for a non-biased interrogator. This thesis will be using current policing racial and gender neutral algorithmic use as to be hypothetically employed in this exploration of artificial intelligence assisted interrogations. This section attempts to present findings that show that artificial intelligence has shown racial and gender non-bias in risk assessment tools both in the United States and the United Kingdom. This analysis is using risk assessment tools as a comparison for hy-
pothetical racial and gender neutrality because other factors (i.e. past convictions etc. which may also be evident during an interrogation) and conclusions drawn from such factors need not be biased by race or gender.

5.1.1 Risk Assessment Models And Racial And Gender Neutrality

Though the achievement of a completely racially/gender non-biased artificial intelligence program is not yet obtainable, this analysis will argue that in the context of criminal justice, some employed algorithms have been successful in being racially and gender non-biased. Again, this analysis acknowledges that both racially and gender unbiased artificial intelligence, especially in the policing area, is a work in progress. This section will use risk assessment algorithms employed by law enforcement within the United States and the United Kingdom as an indicator of the ability for an unbiased racial environment. In addition it is important to note that many of the studies focusing on an algorithmic bias effect are only published within the last one to three years, therefore, research regarding it is also in its infancy. Further, this analysis acknowledges that results regarding bias/non-bias using risk assessment tools will not necessarily predict results using artificial intelligence in an interrogation environment.

Within the United States, the Public Safety Assessment risk assessment tool is aimed at enhancing a judge’s pretrial decision process178 and currently used in some capacity in approximately 29 jurisdictions across the United States.179 In its assessment, the PSA “does not include direct measures of ascribed status related to race, class, or gender, and …[nor] does not include arrests or charges as risk factors.”180 In DeMichele et. al’s The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky, they found that, though race and gender seemed to be factors in analyzing Failure to Appear (FTA), “no significant differences found for new crimes and new violent crimes between black and white defendants.”181 In addition, they pointed out that the findings “do not… exacerbate disparate treatment by race and gender.”182 They also in-

---

178 Public Safety Assessment, “Risk Factors and Formula” https://www.psapretrial.org/about/factors
181 Ibid
182 Ibid
clude the findings from Stevenson (2017), who in analyzing Failure to Appear and New Criminal Arrest (NCA), “did not find any increase in racial bias due to the use of the PSA.” As for gender, they concluded that the Public Safety Assessment “to be free of predictive bias for FTAs and NCAs” Lastly, the Arnold Foundation found that “Black and White defendants assessed with the PSA succeed at virtually identical rates.” The Virginia Pretrial Risk Assessment Instrument (VPRAI) is a similar risk assessment tool intended to measure FTA and NA amongst other pretrial indicators. Danner, VanNostrand, and Spruance (2016), found that:

“[The] VPRAI-Revised... is free of race and gender predictive bias. Specifically, the VPRAI-Revised risk levels have the same meaning for People of Color and Whites and for females and males. Thus, concerns about predictive bias are successfully addressed and the VPRAI-Revised may be considered race and gender neutral.”

Gender-neutrality through risk assessment tools also comes with proactive responsibility, in which the Pretrial Justice Institute commented it required “thought[ful] design, test[ing] and objectively appli[cation].”

Racial/gender neutral artificial intelligence algorithms have also received attention on an international level in the United Kingdom, although its usage is not as prevalent as in the American System. The United Kingdom uses HART, or “Harm Assessment Risk Tool” which is similar to the United States based assessments in the areas of prediction and risk of future offending and eligibility for the Constabulary’s Checkpoint programme. Though it does not directly mention gender neutrality, the Durham Constabulary police force specifically stated that race was not a factor in the algorithms decisions. Though the United States still appears on the forefront is this issue, steps towards a similar algorithmic system are in earlier stages elsewhere, showing AI usage growth.

---

183 Ibid
184 Ibid
185 Pretrial Justice Institute,
186 Ibid
187 ~14% of the U.K. police applying computerized/algorithmic data analysis or decision-making
189 Ibid
5.2 MINIMIZING SUSPECT BIAS INFLUENCE, STRESS/ANXIETY, 
AND FALSE CONFESSIONS

Given the research presented in 5.1, in combination with the Bias-Uncooperative Loop theory, eliminating the interrogator bias could hypothetically eliminate the perceived interrogator bias via the suspect, and theoretically decrease the anxiety and stress caused false positive of guilt associated with conforming to the interrogator’s biased stereotype. It will not necessarily make an interrogation any less stressful, noting Appleby’s findings of stress levels detected regardless of race.\textsuperscript{190} This stereotype bias presumption is also thought to impact suspect willingness to disclose information, referring back to 4.2.2 “The Cooperation Component,” where Black suspects did withhold more information than White suspects. This analysis is going to assume that given the similarities of Black and ethnic minorities in interrogation behavior, ethnic minorities would likely show similar results, though the participation rates of willingness to disclosure information in ethnic minorities needs to be further studied. Much of the minimal research linking stereotype bias to interrogations does not provide a realistic strategy on how to fix the problem, other than acknowledge its existence.

Unfortunately, even with the collection of research in this thesis, there will not be one concrete answer to why the false interrogations continue to arise amongst these two groups. This analysis suggests that the hypothetical “removal” of interrogator bias will simultaneously remove the suspect’s preconceived bias against themselves. When studying stereotype threat, much of the research that surfaces focuses on the educational setting, but the same approach can be applicable in this scenario as well. Standford University published \textit{Empirically Validated Strategies to Reduce Stereotype Threat}, in which the first suggestion was to “(1) Remove Cues That Trigger Worries About Stereotypes.”\textsuperscript{191} Given the previous research discussed in Chapter 3 regarding the difficulties if not an impossibility of long term removal of bias, this analysis suggests that Blacks/ethnic minorities and females will continue to exhibit false positives of deception because of this underlying stereotype threat unless the bias is removed. As we saw from 5.1, criminal justice policing algorithms show potential in being racial and gender neutral, suggesting the same non bias environment could be instituted within the interrogation room.

\textsuperscript{190} Appleby
5.3 Hypothetical Enhancement Of Suspect And Interrogator Interaction Using Artificial Intelligence

Building on of 5.1 and 5.2, this portion of the thesis aims to show how advances in artificial intelligence could theoretically improve the suspect and interrogation interaction beyond the non-bias component, arguing that programmable similarity may affect a suspect’s view of the interrogator, and his or her willingness to trust, and therefore disclose information. This in turn would improve the overall cooperation. Again, I stress that is a purely hypothetical scenario on the basis of both racial-interrogation studies and continuance of human-to-AI cooperation studies. Further research and actual testing of this theory would be needed.

5.3.1 Programmable Suspect Similarity On Hypothetical Trust And Cooperation

This section is dedicated to employing developments regarding programmable similarity to the suspect to hypothetically enhance not only the human-to-AI cooperation, but also the underlying trust and cooperation between the suspect and the interrogator. It argues that this buildup of theoretically increased trust and cooperation could yield safer interrogations, by means of promoting voluntary increased disclosure of information. The subsections of 5.3.1. will consist of (1) programmable physical similarity to the suspect including both race and gender and (2) cultural and language programmability of the suspect.

5.3.1.1 Programmable Physical Similarity Through Race And Gender

This thesis argues that a manipulation of the artificial intelligence interrogator to appear physically similar to the suspect may help with human-to-AI cooperation. Fink (2012) found that factors influencing artificial intelligence acceptance included its “familiarity by using anthropomorphic (humanlike) design and ‘human social’ characteristics.” Fink argues that both influence and perception depend upon physical components of the AI, and therefore are critical to the overall design. She also urged the consideration for “demographic, cultural factors [22] [41], individual preferences, and the context of use.”

The first dilemma this analysis will study is whether race could extend to artificial intelligence without the human component. The findings of the 2018 study conducted by University of Canterbury suggest that even though not human, people will categorized robot color...
into race as if they were human and “adapt their behavior accordingly.” The study, primarily focused on understanding of humans tendency to also be racially biased towards robots, supports the notion that humans will racialize robots. Being that this can have both positive and negative results, attribution of race would therefore play a crucial role in cooperation. Given these findings, this study suggests that interrogators should appear racially similar to the suspect, as to not revert back to racial bias and thus the Bias-Uncooperative Loop.

In terms of gender, this thesis will explore the same dilemma as race, considering whether humans will perceive the gender of artificial intelligence, again without the human component and whether gender stereotypes will be present. Again, assessing human-to-AI perception of gender is critical as not to engage the suspect’s stereotype threat. Guadagno et. al (2007) found stereotypes consistent with human-to-human perception were also found in human-to-AI perception. They found that:

“...Differences may be due to participants’ expectations for interacting with a computer being more consistent with masculine stereotypes (e.g., competent), whereas expectations for interacting with a human are more consistent with feminine stereotypes (e.g., warm).”

This thesis notes that even though these findings were based on a virtual form of artificial intelligence, the evidence of AI stereotyping was still evident, suggesting similar results would occur using physical models, and theoretically lead back to the Bias-Uncooperative loop.

In the same study, exposing participants to same-gendered artificial intelligence, they found that behavioral realism, perception, and social influence depended on a “gender match” between the “gender of the virtual human and the research participant.” They found that "participants were more swayed by the persuasive arguments of a virtual human that matched their gender than one who did not.” Therefore, gender in physical AI interrogators could yield similar results. However, before conclusions can be drawn in this arena, human female suspects interacting to human female interrogators needs further discussion and study. Thus, this thesis suggests that same gender interrogators should be utilized.

---

196 Guadagno et.al
197 Ibid
198 Ibid
5.3.1.2 Programmable Physical Similarity Through Ethnic Culture And Language

The section assumes that similarity through ethnic culture and language may also positively impact the suspect-to-interrogator interaction. This thesis bases such an assumption on findings of interpreters used during police interrogations. In *Interrogation in War and Conflict: A Comparative and Interdisciplinary Analysis*, “The ethnic identity is also a factor when it comes to trust that a witness or suspect has in an interpreter…An interpreter from the same ethnic group engenders confidence, as the witness more easily identifies with the interpreter and therefore feels more comfortable with them.”199 In addition, John E. Reid & Associates, commented that interpreters should also be familiar with the "cultural background, religious beliefs, and value system.”200 This thesis suggests that the physical manipulation regarding the areas of ethnic and cultural similarity of artificial intelligence could have the same effect of increase comfortability and trust with the interrogator.

On a similar note, language and “utterances” also should be considered when designing an artificial intelligence interrogator. Language barriers in minorities is also thought to be an influence in false confessions both in failure of understanding and ability to communicate, and the stress and anxiety caused by these language barriers, which again may be perceived as guilt by the interrogator.201 Davis and Villalobos (2016) speculated that “language difficulties can increase vulnerability to confession.”202 In the original basis of this analysis, Pollina and Barretta (2014), noted their study should be expanded to include “CG agents with different physical characteristics or culture-specific utterances.”203 Therefore, artificial intelligence could have the ability to be multilingual and respond in real time to the suspect, as well as translate and communicate in the preferred language to human police. This technological concept is already achieved through the design prototype of the WT2 Real-time Translating Earphone by Timekettle located in Pasadena, California in the United States.204

---

201 Davis and Villalobos
202 Ibid
203 Pollina and Barretta
5.4 Discussion

Davis and Villalobos (2016) point out that:

“In an increasingly multicultural society, it is imperative to take into consideration the role of social category membership and cultural differences when designing and enforcing the law and its procedures. Among many needed reforms, interrogators need much more and better training regarding ethnic and cultural differences impacting performance in interrogation.”

This chapter intended to present a hypothetical solution to the racial and gender bias dilemma in interrogations by suggesting that artificial intelligence supplementation may minimize the risk of false confessions made by blacks, minorities, and women. This chapter showed that not only can artificial intelligence become both racially and gender neutral in its decision making, it can also be physically programmable to mimic the racial, gender, cultural and language similarities of the suspect. This is thought to have an impact on trust and cooperation. Though human-to-AI cooperation needs substantial further research, past research suggests a working relationship between the two. It was important for this chapter to point out the ability of humans to “humanify” artificial intelligence in both race and gender. This is a critical piece of research, suggesting that if not careful, artificial intelligence design process may reconstruct the same biases they were created to tackle.

It also points out the effect of both cultural reference and language in an interrogation. This analysis theorizes that artificial intelligence’s ability to appeal to culture visually, audibly, and through general reference and understanding could also impact a suspect’s ability to trust and or cooperate with the interrogator. Further research needs to be conducted regarding perception of culture through artificial intelligence, assuming the results are likely similar to that of University of Canterbury’s racial findings.

---

205 Davis and Villalobos
Chapter 6: Platform For A Racial And Gender Non Bias Environment:  
Development, Oversight, And Transparency

All five chapters presented thus far have acknowledged the potential for artificial intelligence in theoretical terms for usage in the interrogation room. This chapter expands on this notion, addressing questions like “How can we control the biases artificial intelligence might develop?” and “Will this notion be a socially accepted solution?” “Will this be legally acceptable? and “Will this be a realistic approach in the American criminal justice system?”

This chapter theorizes the achievement of a gender and racial non bias environment through artificial intelligence needs to be separated into three distinct phases: the programming, the application, and the results. This thesis argues that in order to control artificial intelligence bias creation tendencies, the developers in the programming phase would be responsible for data accuracy and equality in performance. The second phase, the application phase, would require the vetting of data and the introduction of procedural bias detection software, maintenance and oversight once in usage. The third phase, the results phase, would require transparency within the law, giving access to the reasoning of the critical intelligence to suspects/defendants if warranted during trial proceedings. This analysis argues that failures in all three categories have allowed for popularized AI platforms to develop a biased tendency, and thereby cause a mass societal critique and reluctancy towards artificial intelligence usage. This analysis stresses that this methodology is taken from the point of criminological perspective, and that by no means do I, as the author, possess the data programming and technological expertise to conclusively resolve such an issue. It merely highlights the seemingly overlapping issues within the context of bias. This thesis will argue that addressing the confidence in artificial intelligence non-bias may require systematic change in all of these categories in order for greater acceptance moving forward. This methodology, this thesis argues, could also apply with AI used in the interrogation room. Since this concept is theoretical, there is no basis to test its actual acceptance and usage by police forces. Theoretically speaking, criminal justice system’s usage of artificial intelligence, coupled with its upholding in the American legal system suggests that this method may be accepted in both avenues.
6.1 Phase I: The Programming Phase: Data Accuracy And Equality In Performance

In the development of non-biased artificial intelligence, this analysis argues that both data accuracy and equality in performance are necessary factors in its design. The data used is an integral part of the design process. Rashida Richardson, director of policy research at the AI Now Institute at New York University, noted that:

“A lot of the existing AI machine learning systems are using what data is available… You’re either getting it from government information — public open data — or other data sets that maybe are just easily accessible,”\(^{206}\)

The problem with the easy accessible data, is that it is often “incomplete,” which may lead to negatively impacting group diversity, and therefore, eventually triggering a bias in the analytics.\(^{207}\) In addition, the data sets themselves may be biased; citing Henry Hinnefeld, data scientist at Civis Analytics, who noted that: “Failing to consider the bias in your datasets at best can lead to poorly performing models, and at worst can perpetuate a biased decision making process.”\(^{208}\) On that same note, Sarah Jeong, American journalist specializing in artificial intelligence, candidly writes: “…Although machine learning has huge potential, data sets with ingrained biases will produce biased results - garbage in, garbage out.”\(^{209}\) Therefore, the data sets used within the creation of these algorithms should be looked at more closely.

As for equality in performance, Richardson argues that besides the data, training humans to be more conscious of their unconscious biases in software development is essential for the design process.\(^{210}\) It may not be as easy as just removing “race” or “gender” from the parameters, but rather how other factors that may be linked to those categories can cause machines to develop a bias seemingly against those two categories.\(^{211}\) Factors which affected the parameters race and gender also need consideration. COMPAS, an American risk assessment scale, came under fire in 2016 when ProPublica found that reoffending prediction within the software was biased towards black defendants.\(^{212}\) Furthermore, in Loomis v. Wisconsin, the

---


\(^{207}\) Ibid


\(^{210}\) Gale

\(^{211}\) Ibid

\(^{212}\) Electronic Privacy Information Center
defendant argued predictions made by the software were inherently gender biased because of the separate scale differentiated by gender, in which “assessment results will differ based on gender alone.” Therefore, equality in performance requires minimizing human bias within the design process, as well as ensuring equality in both its application and performance.

6.2 Phase II: The Application Phase: Human Intervention: Vetting Data, Bias Detection Software, And Oversight

At MIT’s AeroAstro Centennial Symposium, Elon Musk remarked that:

“I’m increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don’t do something very foolish. I mean with artificial intelligence we’re summoning the demon.”

This introduces the second phase in considering artificial intelligence’s non-bias which suggests the need for human intervention once the algorithmic programs are in use.

Firstly, the data used in these algorithms needs to be (1) maintained by humans and (2) consistently vetted. Comparing data maintenance to that of an experienced gardener encouraging plant growth, Baer and Kamalnath note that: “The most dangerous myths about machine learning is that it needs no ongoing human intervention... Much human oversight is needed.” Hinnefeld also argues for such, noting an algorithmic potential impact on people’s lives. Therefore, private companies responsible for the development of these algorithms used within the criminal justice system need to consider the human maintenance required for non-biased performance. Secondly, these algorithms need to be consistently data vetted. Marie-Eve Piche, CFO at Pymetrics, stated that, “As much as we are scared sometimes of the technology introducing bias, we can control it, we can educate it, we need to audit it.” She argues that, though AI can develop their own biases in the data, consistency in data vetting will, in time, eliminate the algorithmic biases.

---

213 Ibid
216 Hinnefeld
217 Gale
218 Ibid
In addition, recent developments in bias detection software would help the non-bias maintenance in these algorithms. In late 2018, IBM targeted fixing artificial intelligence bias, through the development of an easily available rating system, AI Fairness 360 toolkit software, reported to be able to: “rank the relative fairness of an AI platform and explains how decisions are made.”\textsuperscript{219} IBM commented that:

“The fully automated software service explains decision-making and detects bias in AI models at runtime – as decisions are being made – capturing potentially unfair outcomes as they occur. Importantly, it also automatically recommends data to add to the model to help mitigate any bias it has detected.”\textsuperscript{220}

Hopefully, this may aid in maintaining bias-free artificial intelligence programs. However given its novelty, time would be needed to consider its actual aid.

As for major innovators push for oversight, the majority of the research was attributed to American national security concerns and warfare or holding the AI directly accountable, which are not entirely applicable concepts for this analysis. This thesis points out that various 2019 articles did mention the need for oversight and regulation, however, no federal plans are seemingly within place within the United States. The closest oversight is within New York City with the 2018 formation of the “Automated Decision Systems Task Force” which is “a cross-disciplinary group of city officials and experts in artificial intelligence (AI), ethics, privacy, and law.”\textsuperscript{221} However, the oversight has not made any significant progress as of April 2019, and are in debate over the working definition of “automated decision system.”\textsuperscript{222}

6.3 Phase IIi: The Results Phase: Transparency And Access To Data Within The Law

The last phase of consideration is the results phase, in which this thesis contends that the law needs to address and recognize transparency and access to data. This section will show that, currently, laws within the United States favor protecting algorithmic design over transparency. It will also show that arguments made suggesting the protection of algorithmic design exploits a person’s due process rights. In 2016, the defendant of Loomis v. Wisconsin

\textsuperscript{220} Ibid
\textsuperscript{222} Ibid
noted two significant dilemmas of artificial intelligence within due process. The case challenged both court reliance of COMPAS, one of the risk assessment tools utilized in the United States courts, without the ability of the defendant to question or challenge the results, and whether such a result was a violation on the basis that the factors of race and gender were decisional factors within the assessment.\textsuperscript{223} This particular case considered the risk assessment tool in the content of judicial consideration to determine sentencing.\textsuperscript{224} The case reached the Wisconsin Supreme Court, in which Court ruled against Loomis, specifically noting that the results indicating he was “high risk” by the COMPAS report were based on the “usual factors” of his crime and criminal history.\textsuperscript{225} The Court rejected the defendant’s argument, which insisted that gender was a “criminological factor” contributing to a higher risk assessment score in men. The Court argued that “gender is used only for statistical “norming,” or comparing an offender to a group of the same gender for purposes of determining the offender’s relative level of risk.”\textsuperscript{226}

Aside from the concerning regarding the assessments decision factors of race and gender, there are two notable problems regarding the transparency of AI within the law: (1) decisions generated by these algorithms are not subject to defendant questioning or examination and (2) those algorithms are legally allowed to be trade-secret, even to the defendant whose sentencing it directly reflects.

In regards to the first issue, this analysis pointed out that nearly all States within the United States use risk assessment tools within the criminal justice system, therefore, nearly all defendants are subject to these algorithm predictions. Depending on the State, and the discretion of the judge, reliance on these prediction scores may vary. If the defendant and legal counsel are not allowed access to the system on which such decisions are based, it opens the door for complete legality in itself. In defending against the movement of this case to the United States Supreme Court, Wisconsin’s attorney general, Brad D. Schimel, stated that, “The use of risk assessments by sentencing courts is a novel issue, which needs time for further percolation.”\textsuperscript{227}

\textsuperscript{225} Ibid
\textsuperscript{226} Loomis v. Wisconsin,
\textsuperscript{227} Ibid
In regards to the second issue regarding trade-secrecy, the Court’s favoring of protection of algorithm trade secrets could be problematic moving forward. COMPAS stated that:

“The key to our product is the algorithms, and they’re proprietary,” one of its executives said last year. “We’ve created them, and we don’t release them because it’s certainly a core piece of our business.”

There indeed needs to be a balance between a defendant’s rights to access to his or her results and a company’s right to withhold certain information. Again the novelty of such a dilemma doesn’t yet allow for an answer. Further, AI itself could also be held responsible in further law making. All these legal factors need to be considered when applying artificial intelligence into an interrogation room as well. AI induced confessions could be subjected to legal scrutiny and applicability during a trial.

6.4 Discussion

This chapter has raised awareness in areas that could benefit from further consideration in the design, the maintenance, the transparency and access of artificial intelligence in the criminal justice system. This chapter attempted to shed light on the questions posed at the beginning of this section, and further research would be needed to effectively answer those questions. This chapter was included to draw attention to where general improvements could be made, and where further developments need consideration. In regards to the design, another factor to consider is that since an AI interrogator algorithm is not yet in place, prototypes would be to be assessed on whether it could equate values equally to determine a result (assessing guilty tendencies etc.), as compared to risk assessment tools. As this chapter showed with risk assessment tools, and in the Loomis v. Wisconsin case, issues within risk assessment tools are based on the design’s assessment to be racially or gender influenced by including them as factors. Legally speaking, the United States seems to disproportionately weigh AI design protection over the defendants’ rights of due process.

Again, this thesis is not an expert in computer science, nor a legal expert, and therefore this analysis urges further consideration of topics brought up in this chapter to be carefully examined by experts within the field. Though bias will consistently be an issue within algorithm programs, there are ways to effectively minimize bias, and the risks associated with a flawed program. Human intervention and oversight are key factors in moving forward with

228 Ibid
this new technological development. As for its place in law, the application of artificial intelligence in the legal system is still developing alongside the developing technology.

This analysis points out that in applying artificial intelligence into new areas, like interrogations, it would ultimately require examination and consideration of its usage into already utilized arenas. It suggests that the concept of AI interrogations could be welcomed as a police force supplementary tool given artificial intelligence’s usage, reliance, and legal backing. As stated in Chapter 2, laws regarding the usage of an AI as an interrogator are still inexistent, and the closest comparison is polygraphs, which are utilized in interrogations. Therefore, in true sense of acceptance, conclusions drawn are purely hypothetical. In essence, unbiassed artificial intelligence is a work in process, and as society, we recognize that. Over time, improvements in algorithmic design and bias detection, alongside legal transparency will hopefully pave the way for future usage. However, society is now attempting to address these concerns through recognition, systematic reconsiderations, improvements bias detection, major figures calling for oversight and regulation, and laws upholding its usage. This is showing that we are displaying confidence in using and continuing artificial intelligence.
Chapter 7: Concluding Remarks And Outlook

Richard A. Leo Ph.D and J.D. commented in the Journal of the American Academy of Psychiatry and the Law that, “[False Confessions] are consistently one of the leading, yet most misunderstood, causes of error in the American legal system and thus remain one of the most prejudicial sources of false evidence that lead to wrongful convictions.” This thesis urged further consideration as to why the false confessions seem to be statistically high in the areas of race and gender and how the addition of artificial intelligence may minimize these biases. Again, there is no definite answer as to why this occurs, and also there could be a variety of different factors influencing this phenomenon. This thesis suggested the Bias-Uncooperative loop could be the common denominator of perceived cooperation, seemingly suggesting that it may be an influential factor in both interrogator and suspect perceptions of one another.

All of the chapters presented in this thesis collectively intended to support the notion of a possible utilization of artificial intelligence where it has not yet been used before. In expanding insight on this problem, this thesis suggested using artificial intelligence to hypothetically eliminate both sides of bias, which would not only improve the perception of cooperation, but theoretically improve the actual suspect-to-interrogator interaction. It suggests current biased perception may influence more intense interrogations, and thereby put innocent suspects at risk of making false confessions. Promoting a bias free environment and effective cooperation hopefully could yield the potential for minimized false confessions. Furthermore, this thesis also presented procedural safeguards on how society should be responsible for this new technology.

Feasability

Given the abstract approach in utilizing artificial intelligence in the interrogation realm, main questions to be drawn from this thesis are whether this concept is a feasible solution now or in the future?, when these processes would be expected to be both societally accepted and integrated into the system?, and the types of cases/crimes it would be best suited
for?. While this thesis was in the drafting stages, the Estonian Ministry of Justice launched its proposition for the design of a “robot judge” which would handle small claims court disputes.\(^\text{230}\) In addition, new research conducted by forensic psychologist Francesco Pompedda from Åbo Akademi University in Åbo, Finland implemented computer-generated avatar interviewing to train interviewers and improve interview quality in alleged child sexual abuse cases.\(^\text{231}\) Current research and implementation are clearly serving as building blocks for more artificial intelligence within the criminal justice sector. Given that this thesis took roughly 9 months of research, and given the consistent release of related new research and press within that time period, I guesstimate that certain aspects of this idea are feasible now, but within the next five years we are likely to be seeing much more of it, and likely in greater detail.

Another feasibility question is what types of crimes this type of approach would be best suited for. There are several different approaches given the conclusions drawn from this analysis. One, artificial intelligence interrogator models could function only on the basis of assessing guilt via deception detection, by means of the physical human manifestation of guilt related predictors discussed in Chapter 2, given the ability and increased accuracy in deception detection software. Another possible route for this is the actual obtainment of information, where the artificial intelligence interrogator appeals to the cultural/ethnic/linguistic similarity of the suspect in order to gain insight or information. Stemming from these, investigators of certain types of crimes could also narrow down the usage of artificial interrogators. This analysis suggests that, if the design encompasses a machine learning function, it may be best to separate both crimes and function (guilt versus information) in order to study both the accuracy, and improvements in specific functions, given specific crimes. However, moving forward, an under-researched area is the national discrepancy of drug related crime and African Americans, who, according to NAACP, use drugs at similar rates, but are convicted 6 times more than their white counterparts.\(^\text{232}\) Again, ideally this artificial intelligence approach should be able to be applicable in all cases in the future.


\(^{231}\) Pompedda, Franseco, (2018) “Training in Investigative Interviews of Children: Serious Gaming Paired with Feedback Improves Interview Quality” Åbo Akademi University https://pdfs.semanticscholar.org/14d9/45841824f7a003eece3bb38cf1dac0a26a1ed.pdf

OUTLOOK

When statistics show year after year that these stereotypes have seemingly remained factors in police injustice, and studies seemingly focus on existence without an effective remedy, does that suggest that human behavior maybe incapable of change? This analysis then returns to the beginning quote of this paper, which boldly states “The ultimate paradox is that this technology may become the powerful catalyst that we need to reclaim our humanity.” This thesis is intended to promote a nouveau unorthodox approach to addressing biased interrogations and false confessions. Though many might argue that this paper may spark a “Robot Takeover” or comically compare it to the Terminator or RoboCop series, this paper intended to provoke the concept of how artificial intelligence does not become a human replacement, but rather help us become more human ourselves. Hopefully, my thesis, current theses being written, or theses to come explore the collaboration of both human rights and artificial intelligence. I hope that artificial intelligence can compliment international human rights and be used to tackle discrimination and bias.

Given artificial intelligence advancement, its increasing application, not only in the criminal justice system, but exponentially in various other fields, and its legal and societal uphold and acceptance, police forces outsourcing to artificial intelligence is not a far fetched idea within the next five years. We, as human rights advocates, criminologists, as the developers, as the programmers, as the law makers, as the police, and as members of society all need to be proactive in this new age of artificial intelligence justice. This thesis urges that before any type of artificial intelligence is either considered or implemented in a new platform, it requires multi-angular and multi-disciplinary considerations, commentary, and suggestions.

References

Articles


Public Safety Assessment, “Risk Factors and Formula” https://www.psapretrial.org/about/factors


Books

Definitions
Anthropomorphism, Oxford Dictionary
Definition of Artificial Intelligence, dictionary
Definition of “Robot,” dictionary

Rouse, Margaret, “DEFINITION expert system,” TechTarget, https://searchenterpriseai.techtarget.com/definition/expert-system

External Studies Cited Collectively Within Studies
(As cited in Pollina and Baretta (2014), page 12) Gerhard, Moore, & Hobbs, 2004, 2005; Groom et al., 2009;
Guadagno, Blascovich, Bailenson, & McCall, 2007; Guadagno, Swinth, & Blascovich, 2011

**Federal Government Documents**


**International Documents**


**Literature**


Legal

Misc
Langford, Malcolm(2019), collection of suggestions for AI interrogator functionality

Studies


Ohio State University. (2018). At first blush, you look happy—or sad, or angry. Ohio State University. https://news.osu.edu/at-first-blush-you-look-happy--or-sad-or-angry/


