Futures 117 (2020) 102510

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Essays

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

Maria Noriega

Univeristy of Oslo, Oslo, Norway

ARTICLE INFO

Keywords:
Artificial intelligence
Bias
Policing
Interrogations
Discrimination
Crime

ABSTRACT

Research presented in this study 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.

This study suggests that the racial and gender bias influencing false confessions may be due to the two fold bias occurring within the interrogator-to-suspect dynamic, referenced in this study as "the Bias-Uncooperative Loop." It argues that applying artificial intelligence within the interrogation room may minimize the two fold bias occurring in the dynamic. It suggests the potential for cooperation between the two parties can be conditioned by programmable similarity; whereby artificial intelligence can mimic the racial, ethnic and/or cultural similarities of the suspect in question. This is reflected in research in different arenas (not inclusive to interrogations) to have an effect on enhanced comfortability and cooperation with AI. This paper assumes similar results within interrogations.

1. Introduction

The research presented in this study examines the potentiality of artificial intelligence as an interrogator within a police interrogation to promote a non-biased[1] 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.[2] Furthermore, from these cases, 70 % of the exonerees were of Black, Latino, or Native American ethnic origins.[3] 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*."[4] Gender bias within interrogations is another contentious

[1] Only racial and gender bias although this paper acknowledges bias can extend to other categories.
[2] Innocence Project (2020) "DNA Exonerations in the United States." https://www.innocenceproject.org/dna-exonerations-in-the-united-states/.
[3] Ibid
[4] Lopez and Zarracina (2017) "Study: black people are 7 times more likely than white people to be wrongly convicted of murder." VOX. https://www.vox.com/policy-and-politics/2017/3/7/14834454/exoneration-innocence-prison-racism.

https://doi.org/10.1016/j.futures.2019.102510
Received 13 August 2019; Received in revised form 27 December 2019; Accepted 30 December 2019
Available online 30 January 2020
0016-3287/

issue, in which women account for 11 % of exonerations based on false confessions, in comparison to 9 % of overall exonerations.[5] 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.”[6] Overall, the percentage statistics suggest that there may be a greater likelihood of conviction based on false confessions for females.[7]

Growing usage of new platforms designed to minimize or ideally eliminate bias within the criminal justice and legal system has shed some light on 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.[8] This study 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 against false confessions.

The focus of this study highlights 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 research 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. I argue 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. I will argue that the model utilized in Pollina and Barretta (2014) can serve as a baseline model for future exportation in the context of guilt admission.

I also will highlight the theoretical cooperation dynamic between the *interrogator-to-suspect* and the *suspect-to-interrogator*. It argues that there is a two-fold bias dynamic occurring, (referred to as the “Bias-Uncooperative Loop” in this study) which is ultimately causing miscommunication of cooperation and willingness to disclose information. This may allow for a more intense, rigorous interrogation, and may lead to the propensity for false confessions. I then argue that the utilization of AI as an interrogator could theoretically halt the Bias-Uncooperative Loop by minimizing or ideally eliminating the two fold bias occurring, and thereby ideally minimizing or eliminating the risk for false confessions. Further, it 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.

2. Background analysis of Pollina and Barretta (2014) for admission of illegal behavior

The research presented in this section shows that artificial intelligence has shown capability of invoking *admissions* of illegal behavior in the context of national security clearance interviews within the United States. In conjunction with the research findings, it 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, it further suggests that manipulating other variables within the design structure could potentially increase the theoretical probability of eliciting self-incriminating confessions of illegal behavior through increasing cooperation and trust.

It is important to note that theoretical assumptions drawn from this section are hypothetical conclusions intended to be tested and explored further through research. Given the infancy of humanoid 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 section points out that humans will *willingly* engage with artificial intelligence, which is a crucial factor in AI assisted interrogations’ realistic applicability and for improvements on the overall relationship.

Though the main component of this theoretical response is based on Pollina and Barretta (2014), which specifically outlines a response detailing criminal behavior, I will outline two other sources Kiesler, Powers, Fussell, and Torrey (2008) and the Ellie virtual interview project at the Institute for Creative Technologies, in Los Angeles, California by Jonathan Gratch. Since disclosing criminal behavior could potentially fall under “sensitive information” or “undesirable behavior” it may support continued research in AI prompted criminal confessions.

2.1. Artificial intelligence and national security clearance interviews

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. 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

[5] Knox (2016) “Amanda Knox: Why Do Innocent Women Confess to Crimes They Didn’t Commit?” *Broadly*. https://broadly.vice.com/en_us/article/9k9wep/amanda-knox-why-do-innocent-women-confess-to-crimes-they-didnt-commit.

[6] Jones (2011). “Under pressure: Women who plead guilty to crimes they have not committed” *Criminology and Criminal Justice* https://doi.org/10.1177/1748895810392193.

[7] Knox.

[8] Electronic Privacy Information Center (2020) “Algorithms in the Criminal Justice System” EPIC https://epic.org/algorithmic-transparency/crim-justice/.

self questionnaires of 120 United States Army basic trainees between the ages of 18 and 40.[9] 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.”[10] In addition, the CG targeted questioning interviewees in the areas of mental health, drug, alcohol, and criminal histories.[11] The CG was programmed to interpret participants’ verbal responses and then proceed to ask subsequent questions based on the answers.[12] The flow chart of the algorithmic design of questioning can be found in Fig. 1. The content of the interviews was similar to sections of Standard Form 86 (Sections 21, 22, 23, and 24).[13] 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.

The results of the study concluded that national security interviews could be conducted by artificial intelligence in the form of a CG agent.[14] They found that information “including prior illegal acts” could be elicited by the agent.[15] The results showed that, most importantly, three participants admitted previous engagement in either criminal alcohol or criminal drug offenses that was never caught.[16] 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.[17] In addition, when able to to speak freely through open ended questions, 14 participants added to their statements (three participants about drug use and one participant on crime).[18] In terms of relating the study to the context of this present study, artificial intelligence has shown the capability of invoking a self-incriminatory *admission* of illegal behavior.

It is important to note that a protection clause was not included in the Pollina and Barretta (2014) p. 45 study which specifically informed participants that, “researchers had an obligation to report serious violations of the law.”[19] Therefore, because participants willingly divulged illegal behavior without the protection clause 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).

In addition, relating to the context of this study, the original research was obtained via one study, and that the sample size was small. Further replication would be needed to substantiate AI’s ability to yield similar results. However, there are many studies that involve human participants interacting with AI, and even disclosures of undesirable behavior or sensitive information (as presented in this study) to AI. With that being said, Pollina and Barretta (2014) 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), I assume their results are repeatable and valid.

2.2. Expanding baseline design for confession elicitation

Section 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. I argue 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.

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 enhance a person’s view on artificial intelligence, which this present study argues could be beneficial in obtaining an incriminating confession.

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, and 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.[20] 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

[9] Pollina and Barretta (2014). "The effectiveness of a national security screening interview conducted by a computer-generated agent.*” Comput. Hum. Behav. 39, C* (September 2014), 39–50. DOI = https://doi.org/10.1016/j.chb.2014.06.010.
[10] Ibid, p. 39.
[11] Ibid, p. 39.
[12] Ibid, p.39.
[13] Ibid, p. 40.
[14] Ibid, p. 47.
[15] Ibid p. 47.
[16] Ibid, Table 1, p. 45.
[17] Ibid, Table 1, p. 45.
[18] Ibid, 45.
[19] Pollina, Barretta, p. 41.
[20] Bainbridge et al. (2008). The effect of presence on human-robot interaction. *Conference: Robot and Human Interactive Communication* 701–706.

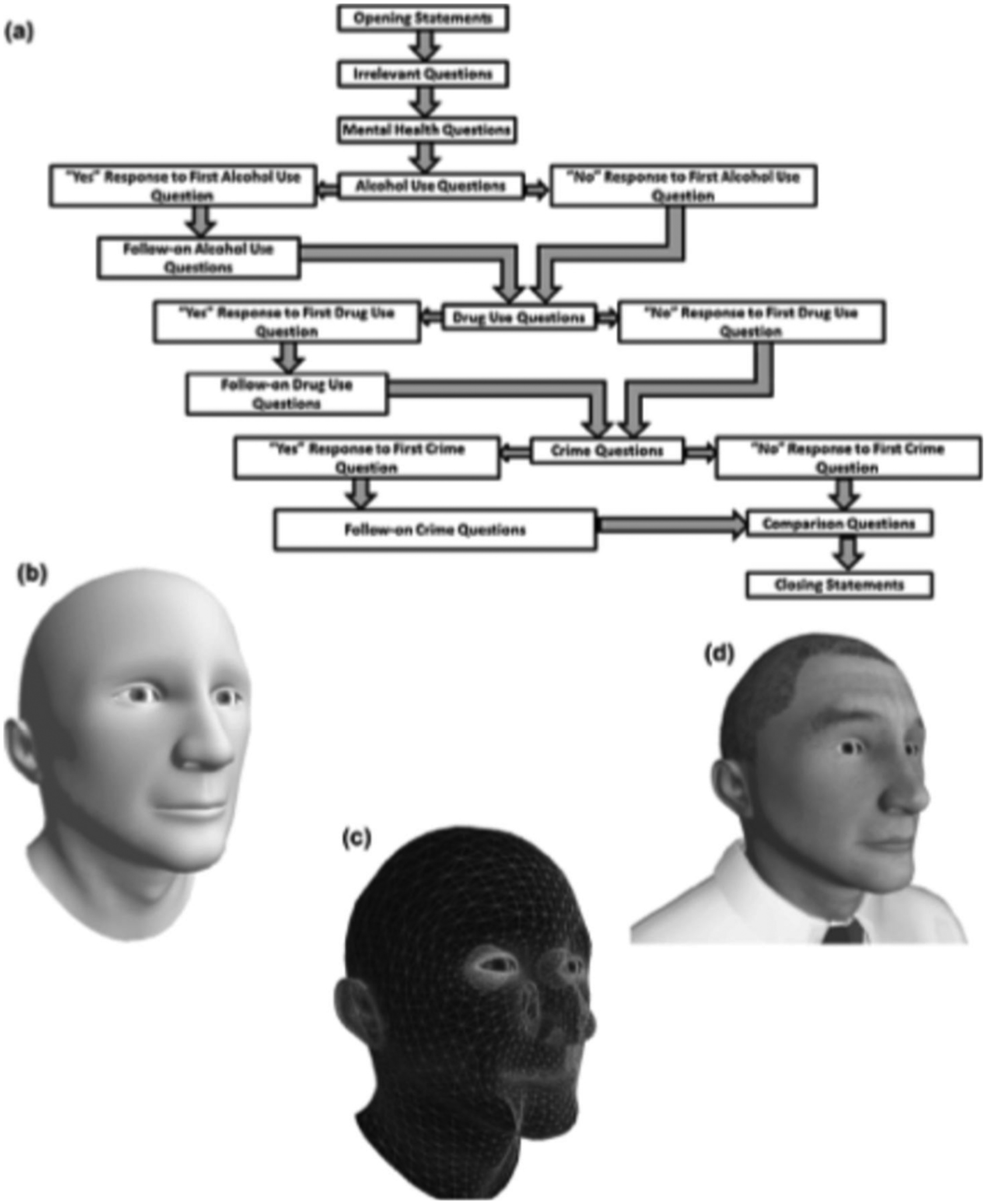

Fig. 1. Pollina and Barretta (2014). "The effectiveness of a national security screening interview conducted by a computer-generated agent." *Comput. Hum. Behav. 39, C* (September 2014), 39–50. https://doi.org/10.1016/j.chb.

Table 1
Participant admissions during the CG interview: alcohol use, drug use, and Criminal History (n = 120).
Source: Pollina and Barretta (2014) p. 45

Category	Admission	Number of participants
Alcohol use	Ever Drank Alcohol	110
	Drank Alcohol While Underage	92
	Still Drink Alcohol	62
	Negative Impact on Work	5
	Negative Impact on Finances	3
	Voluntarily Sought Treatment for Alcohol Use	4
	Advised to Seek Treatment for Alcohol Use	3
	Diagnosed by a Professional with an Alcohol Problem	0
	Drink Alcohol Less than Once Per Week	58
	Drink Alcohol at Least Once Per Week	29
	Drink Alcohol More than Once per Week	16
Drug use	Used Illegal Drugs	55
	Advised to Seek Treatment for Drug Use	2
	Voluntarily Sought Treatment for Drug Use	1
Criminal History	Convicted of a crime	17
	Ever On Probation or Parole	8
	Convicted of Alcohol-Related Offense	5
	Criminal Alcohol-Related Offense	3
	Drug-Related Offense	1
	Engaged in these Activities but Not Caught	3
	Go To Jail	2
	Pay a Fine	9
	Go to Jail and Pay a Fine	5

Nico."[21] In Hancock 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."[22] In regards to respect, Bainbridge et al. (2008) found that the participants afforded the *physical* Nico increased personal space.[23] 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."[24]

In Kiesler et al.'s (2008) "Anthropomorphic Interactions with a Robot and Robot–like Agent" in *Social Cognition*, participant's interactions were studied using either robots or the virtual presence of one on a screen.[25] Participants viewed the robot as more dominant, trustworthy, sociable, responsive, competent, and respectful than the agent and rated it more life-like.[26] 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."[27] 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.[28] However, this particular study may have been significantly flawed, because many of the participants were not "fluent English speakers" and researchers noted significant issues with subject response rates due to this.[29] However, the repeating pattern is the association of trust and respect with physical presence versus virtual presence is noted. The significant take away from this study is that regardless of a flawed participant language component, undesirable behavior was *still* disclosed at some rate.

2.2.1.2. 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 et al. did not evaluate the effect distance has on AI-to-human interaction. Supplementary research conducted by Siegel (2009) on behalf of Massachusetts Institute of Technology, tested AI-to-human distance on compliance.[30] In the study, participants were evaluated for compliance standing 2.5 feet (76 cm) and 5 feet (152 cm) from the robot.[31] The results showed that the decrease in distance had the overall effect of decreasing compliance.[32] The study hypothesized

[21] Ibid, 705.
[22] Hancock et al. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human factors, 53 5*, 517–27.
[23] Bainbridge, Hart, Kim, and Scassellati (2008).
[24] Ibid 706.
[25] Kiesler et al. (2008) "Anthropomorphic Interactions with a Robot and Robot–like Agent." *Social Cognition* 26, 169–181 (2008).
[26] Ibid 177.
[27] Ibid 169.
[28] Ibid 175.
[29] Ibid.
[30] Siegel (2009). "Persuasive Robotics How Robots Change our Minds." *Massachusetts Institute of Technology*. https://core.ac.uk/download/pdf/4411144.pdf.
[31] Ibid, 114.
[32] Ibid, 114.

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."[33] 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."[34] However, what I presently conclude from both previous 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 an suspect-to-interrogator interaction), the distance is also important. Irecommend 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.

2.2.1.3. Physical details. 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."[35] This section aims at understanding the effects of anthropomorphism in AI-to-human interaction in the area of an interrogation. Fink (2012)'s *Anthropomorphism and 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."[36] 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.[37] The author included that "visible design is crucial."[38]

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.[39] 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."*[40]

In other words, the more it indistinguishably resembles human beings, the more of a "trigger [to] our human-to-human empathy levels."[41] 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* which studied perceived likability of the robot.[42] 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."*[43]

In regards to the uncanny phenomenon and human-to-AI trust, "humans appear to infer trustworthiness from affective cues (subtle facial expressions) known to govern human–human social judgments."[44]

Modern technology and robotics considerations are closer to overcoming the uncanny valley 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[45] can display 60 different expressions through material developed to mirror human facial muscles and skin.[46] In 2018, advancements made to Sophia's design included the addition of legs and the ability to walk.[47] 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.

[33] Ibid, 114.
[34] Ibid, 114.
[35] Anthropomorphism Oxford Dictionary.
[36] Fink (2012). "Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction". © Springer-Verlag Berlin Heidelberg. pp. 199–208.
[37] Ibid, 203.
[38] Ibid, 203.
[39] Knight (2014)."How Humans Respond to Robots:Building Public Policy through Good Design." *Carnegie Mellon University*. https://www.brookings.edu/wp-content/uploads/2014/07/HumanRobot-PartnershipsR2.pdf.
[40] Hsu (2012). "Why 'Uncanny Valley' Human Look-Alikes Put Us on Edge." *Scientific American*. https://www.scientificamerican.com/article/why-uncanny-valley-human-look-alikes-put-us-on-edge/.
[41] Melina (2011). "Is 'Mars Needs Moms' Too Realistic?" *Life Science* https://www.livescience.com/33198-is-mars-needs-moms-too-realistic.html.
[42] Mathura and Reichling (2014). "Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley." Cognition Volume 146, January 2016, Pages 22-32.
[43] Ibid.
[44] Ibid.
[45] Reynolds (2018) "The agony of Sophia, the world's first robot citizen condemned to a lifeless career in marketing." *WIRED*, https://www.wired.co.uk/article/sophia-robot-citizen-womens-rights-detriot-become-human-hanson-robotics.
[46] Poulter and Harry (2018) "Sophia takes her first steps: Super-intelligent humanoid robot reveals she now has legs." DailyMail. https://www.dailymail.co.uk/sciencetech/article-5251185/Worlds-robot-citizen-Sophia-gets-legs.html.
[47] Ibid.

2.2.2. Decreased perceived external human control

Previous research has indicated and promoted that participants 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, Blascovich, Bailenson, and McCall (2007) regarding perceived human operators. Guadagno et al. (2007) [48] focused on the effect of AI-to-human interaction regarding persuasive arguments.[49] 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]."[50] 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)."[51]

However, recent research is suggesting that the perceived human operator component may significantly hinder willingness to divulge *sensitive information*. It 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.[52] In Gratch's study, Ellie, a human-like, full body avatar psychologist, is 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."[53] 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.[54]

Research has thereby indicated the decreased knowledge of perceived human control shows a greater willingness to disclose sensitive, personal information. It 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.

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."[55] 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.

3. Artificial intelligence as interrogators in police interrogations

I will 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 it 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 during an interrogation. 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."*[56]

In addition, I propose 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. I also do not encourage that artificial intelligence will be a foolproof means of assessing innocence or guilt. Given that deception detection via the Reid Technique is still the corner stone of American interrogations, it has to consider its functionality in that practice. I also acknowledge 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.[57] 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[58], also a potential capability of artificial intelligence.

[48] Guadagno et al. (2007)." Virtual humans and persuasion: The effects of agency and behavioral realism." *Media Psychology*, 10:1-22, 2007.

[49] Siegel.

[50] Ibid.

[51] Lim and Reeves (2010). "Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player." *International Journal of Human–Computer Studies*, 68(1/2), 57–68.

[52] The Computer Will See You Now (2014). *The Economist.* https://www.economist.com/science-and-technology/2014/08/20/the-computer-will-see-you-now.

[53] Ibid.

[54] Ibid.

[55] Pollina, Barretta, p. 45.

[56] Leo (2009) "False Confessions: Causes, Consequences, and Implications." Journal of the American Academy of Psychiatry and the Law Online September 2009, 37 (3) 332-343; http://jaapl.org/content/37/3/332.

[57] Orlando (2020). "Interrogation Techniques" OLR Research Report. *Connecticut General Assembly.* https://www.cga.ct.gov/olr/.

[58] Ibid.

In practice, I will argue 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. Ideally, using AI 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.

To begin with, it acknowledges that an artificial intelligence interrogator could be a human-like, life-like, computer generated avatar (i.e. without a physical presence) or a cyber-physical robot. Though the combinations of designs may theoretically work, it shows that research assessing 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 of the artificial interrogator, could be based using expert design of decision tree reasoning, or using predictive machine learning, or could be a combination of both.[59] 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."[60] In reference to decision tree learning, it proposes that the AI design incorporates relevant attributes and relating to value function in how interrogators interact and question subjects.[61] Certain eye movements for example, are used as possible indicators of guilt. Using this method, the predictive measures demonstrated in this study will use values of that attribute, in which the decision tree model will follow a path corresponding with that value.[62] Another approach could be through machine learning, referring to the computer's learning process bereft human intervention.[63]

3.1. Applying AI model (1) designed for police interrogations

Section 2.1 will demonstrate that the design detailed in the first portion of this study could be realistically applied to an interrogation setting. I will focus on two particular aspects of the interrogation process: rapport building and deception recognition through baseline questioning. 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.[64] I will then argue that the design model presented in 1 could theoretically execute the same functions, if not with greater accuracy, as/than a human interrogator.

3.1.1. Rapport building through AI

I support 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.[65] 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.[66] 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."[67],[68],[69]

Given the algorithmic design of the Pollina and Barretta (2014) CG Agent in both asking questions and processing information, it is plausible to create an AI interrogator that can perform 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."[70]

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 the suspect. Pollina and Barretta (2014) 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."[71]

What I suggest is we do not have to—- in essence— "create" that "comfortability" because people have already become conditioned to automated chat bots overtime. However, we indeed have to improve the comfortability level, especially introducing a new

[59] Method suggested by Langford (2019), advisor.

[60] Rouse, Margaret. "DEFINITION expert system," *TechTarget*, https://searchenterpriseai.techtarget.com/definition/expert-system.

[61] Sharma, Ishan. "Introduction to Decision Tree Learning." HeartBeat. https://heartbeat.fritz.ai/introduction-to-decision-tree-learning-cd604f85e236.

[62] Ibid.

[63] Balaji and Choubey. "ML | What is Machine Learning?" GeeksforGeeks.com https://www.geeksforgeeks.org/ml-machine-learning/.

[64] Davis and Vllalobos (2016). "Interrogation and the Minority Suspect: Pathways to True and False Confession." *Advances in Psychology and Law: Volume 1*. https://www.researchgate.net/publication/290436784_Interrogation_and_the_Minority_Suspect_Pathways_to_True_and_False_Confession.

[65] Layton, Julia. "How Police Interrogation Works." *Culture* https://people.howstuffworks.com/police-interrogation1.htm.

[66] Ibid.

[67] Sally (1993). "Conversation and cooperation in social dilemmas a meta-analysis of experiments from 1958 to 1992." Ration. Soc. 7, 58–92.

[68] Balliet (2009). "Communication and cooperation in social dilemmas: a meta-analytic review." Ration. Soc. 54, 39–57.

[69] Crandall et al. (2018). "Cooperating with machines." Nature Communications 9.

[70] Shead (2016). "Humans are willing to trust chatbots with some of their most sensitive information." *Business Insider*. https://www.businessinsider.com/humans-are-willing-to-trust-chatbots-with-some-of-their-most-sensitive-information-2016-6?r=UK&IR=T&IR=T.

[71] Pollina, Barretta, 47.

scenario and context of interaction. Therefore, regarding the rapport building process, it is essential to draw from the enhanced design model from 1, especially in the areas of AI-to-human placement and AI physical human likeness.

3.1.2. Baseline question 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.[72] The purpose of these baseline questions is to prompt brain activity from especially both simple memory recall and creativity.[73] 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.*"[74]

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.[75] In addition, non verbal communication would be observed for documented signs of lying including arm, hand, finger, leg, and foot movements.[76] It 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. The following subsections demonstrate this ability through AI facial emotional recognition, AI verbal emotion recognition and deception recognition software.

3.1.2.1. AI Facial-emotional decoding and recognition in humans. During these baseline questions, the human interrogator would be looking for facial changes to implicate a suspect. A study conducted by Ohio State University[77] 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.[78] 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.[79] 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 (1997) found that through the use of neural network application, based on machine learning, correct facial emotional recognition was established in real time.[80] Therefore the a key technique of a human investigator- reading emotional changes, in real time, can be executed by AI.

3.1.2.2. AI verbal-emotional decoding and recognition in humans. Research has suggested that changes in voice pitch are one common deception indicator that may indicate lying.[81] The changes in Hertz are quite small and detectable with "special equipment." (i.e. not human hearing alone).[82] In 2017, a study conducted at Nirma Institute of Technology in Ahmedabad, Gujarat, India, proposed research for an "algorithmic approach for detection and analysis of human emotions with the help of voice and speech processing."[83] 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 time."[84] In general, the study concluded that "with varying emotions, the tonal parameters accordingly change as well."[85]

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 or even possible. Modern artificial intelligence advancement has improved the *accuracy* of these deception cues, which would be beneficial in an interrogation room setting.

[72] Layton, Julia. "How Police Interrogation Works." *Culture* https://people.howstuffworks.com/police-interrogation1.htm.
[73] Ibid.
[74] Ibid.
[75] Ibid.
[76] Ulatowska (2013). "Indicators of deception in different lie scenarios." Roczniki Psychologiczne / Annals Of Psychology 2013, XVI, 1, 127-146.
[77] Benitez-Quiroz, Srinivasan, and Martinez (2018). "Facial color is an efficient mechanism to visually transmit emotion." Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1716084115.
[78] 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/.
[79] Ibid.
[80] Kobayashi and Hara (1997). "Facial interaction between animated 3D face robot and human beings." *IEEE **Xplore*** 2002. https://doi.org/10.1109/ICSMC.1997.633250.
[81] Ulatowska.
[82] Ibid.
[83] Dasgupta (2017). "Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing." *International Journal of Computer Trends and Technology (IJCTT)* – Volume 52 Number 1 October 2017.
[84] Ibid.
[85] Ibid.

3.1.2.3. Deception recognition software. Recent developments have led to the advancement of AI deception recognition software. In 2017, the University of Maryland published a study regarding research analyzed through the “Deception Analysis and Reasoning Engine (DARE)” in court trial videos to recognize deception cues.[86] 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.”*[87]

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

4. The interrogator-to-suspect interface

This portion of the study will shift forward to address the racial and gender bias thought to be impacting false confessions. This section will introduce the Bias-Uncooperative Loop, which argues may be factor contributing to the racial and gender statistical anomalies in false confessions. It will subsequently highlight the hypothetical effect of Bias-Uncooperative Loop on racial and gender bias 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 to follow 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.

This study will first 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.[88] I contend that though these techniques may prove helpful, the underlying fallibility may be due to the preconceived notions of the interrogator. This study 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.

4.1. 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, it 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.

4.1.1. 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.[89] They note that current scientific research does not bolster the notion that “behaviors or response styles” are necessarily reliable indicators of distinguishing factors between innocence or guilt, nor does it support the foolproof skill or expertise of the interrogator.[90] 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.”[91] 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 & DePaulo, 2006).[92] 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.[93]

The deception detection is a critical component of investigator bias, which is suggested by Redlich and Meissner to contribute to confirmation bias.[94] 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).”[95] In Meissner and Kassin (2004), research suggested a link

[86] Galeon (2018) “A New AI That Detects ‘Deception’ May Bring an End to Lying as We Know It.” *FUTURISM*. https://futurism.com/new-ai-detects-deception-bring-end-lying-know-it.

[87] Wu, Singh, Davis, and Subrahmanian (2017). “Deception Detection in Videos.” *Cornell University* Library. https://arxiv.org/abs/1712.04415.

[88] Redlich and Meissner (n/a). “Techniques and Controversies in the Interrogation of Suspects: The Artful Practice versus the Scientific Study.” (to appear in) *J. L. Skeem, K. Douglas, & S. Lilienfeld (Eds.), Psychological science in the courtroom: Controversies and consensus.* https://digitalcommons.utep.edu/cgi/viewcontent.cgi?article=1036&context=christian_meissner.

[89] Ibid.

[90] Ibid.

[91] Leo (2009).

[92] Redlich and Meissner.

[93] Ibid.

[94] Ibid.

[95] Ibid.

between this investigator bias and false confessions.[96] Redlich and Meissner found that when together, “the use of questionable deception detection techniques and a strong presumption of guilt on the part of the investigator can be dangerous to innocent suspects, placing them at risk for the pressures of interrogation (Kassin 2005; Meissner & Kassin, 2004).”[97]

4.1.2. 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.

4.1.2.1. 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.”[98]

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.”[99] 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.[100] Weir (2016) points out that “when officers' training and experiences confirm racial stereotypes, those biases appear to hold more sway over their behavior.”[101] She specifically points out the “unconscious association between black individuals and crime,” in which she suggests this implicit notion can affect police behavior.[102] Furthermore she notes that it is not necessarily anchored to obvious or explicit racist beliefs of the individuals.[103]

In the context of the interrogation, there appear to be 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.”[104] Further, the study found that “police officers judged Black suspects to be less cooperative and less forthcoming than White suspects.”[105] 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.[106] 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.[107] I will point out that this cooperation dynamic is key to both the suspect and interrogator interface.

4.1.2.2. 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, 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.”[108] 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 susceptive to coercion and “threats to family responsibilities.”[109]

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, I draw upon the 2014 study findings of Barry C. Feld at the University of Minnesota Law School, who found in juvenile justice based interviews, substantial differences in suspect perception on the basis on gender.[110] The study reported that:

“Despite their objective similarities, interviews with juvenile justice personnel report substantial differences in how they perceive boys and

[96] Ibid.
[97] Ibid.
[98] Weir (2016). “Policing in black & white,” *Monitor on Psychology: Vol 47, No. 11*, https://www.apa.org/monitor/2016/12/cover-policing.
[99] Ibid.
[100] Carbado and Richardson (2018) Book Review “The Black Police: Policing Our Own.” 131 *Harv. L. Rev*. https://harvardlawreview.org/2018/05/the-black-police-policing-our-own/.
[101] Ibid.
[102] Ibid.
[103] Ibid.
[104] Appleby (2015) ““Guilty Stereotypes: The Social Psychology Of Race And Suspicion In Police Interviews And Interrogations” *City University of New York (CUNY) CUNY Academic Works*. https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1518&context=gc_etds113.
[105] Ibid.
[106] Ibid.
[107] Ibid.
[108] Jones (2011).
[109] Knox (2016) “Amanda Knox: Why Do Innocent Women Confess to Crimes They Didn’t Commit?” *Broadly*. https://broadly.vice.com/en_us/article/9k9wep/amanda-knox-why-do-innocent-women-confess-to-crimes-they-didnt-commit.
[110] Feld (2014) “Questioning Gender: Police Interrogation of Delinquent Girls” *University of Minnesota Law School.* https://scholarship.law.umn.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1301&context=faculty_articles.

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."[111]

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."[112] 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 %.[113]

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.[114] The stereotypical view of females (emotional, etc) appears to directly influence the perception of the suspect, which as previous research via this study 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:

"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."[115]

4.2. Discussion of findings

The issue highlights the racial and gender bias prevalent in both the deception detection and the presumption of guilt affecting innocent suspects. Redlich and 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."[116]

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. It 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."[117] The common denominator of "uncooperative" found in both racial and gender studies suggests a link to the loop process. This might help to start 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.[118] Weir commented that:

"*In two studies with more than* 6300 *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 hoursJournal of Experimental Psychology, 2016."*[119]

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.

5. The suspect-to-interrogator interface

This section will consider the position of the suspect in reference to the interrogator within the context of eliciting false confessions. 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 guilt by default. This theory is thought to lead to a "perceived" uncooperative suspect. This reinforces the perception of the suspect by the interrogator. This loop may represent a "chicken or egg" dilemma, meaning who's bias is influencing who's. Furthermore, the suspect perceiving this stereotype threat and police bias may cause the suspect to be less forthcoming with information, which potentially

[111] Ibid.
[112] Knox (2016).
[113] Ibid.
[114] Feld.
[115] Cleary and Bull (2018). "Jail Inmates' Perspectives on Police Interrogation."Psychology Crime and Law. https://www.researchgate.net/publication/326531981_Jail_Inmates'_Perspectives_on_Police_Interrogation.
[116] Redlich and Meissner.
[117] Leo.
[118] Weir.
[119] Ibid.

could heighten the intensity of an interrogation, resulting in a false confession. 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.

5.1. *Suspect component of the Bias-Uncooperative Loop*

5.1.1. *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 (2015). Thus, it requires further research and studies for a more definite correlation.[120] The following two sections will discuss how the suspect hypothetically responds to assuming interrogator bias is prevalent in the interrogation room.

5.1.1.1. The racial bias stereotype threat. In regards to how race effects perception in the interrogation room, Appleby (2015) suggested that the higher rate of innocent black suspects found guilty could be due to the "concerns about being stereotyped as criminals."[121] 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.*"[122]

Returning to 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 study 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, Idy cite Davis and Vllalobos (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.*"[123]

Appleby (2015), did not distinguish by race 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.

5.1.1.2. 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 Op-Ed covering her experience with a woman's treatment within an interrogation room. She suggests that a societal hardwired stereotypical view of women impacts women's self view in an interrogation.[124] 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.*"[125]

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."[126] 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 officers[127] in the police department.[128] In addition, Northwestern psychology and gender studies professor 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."[129] Could this

[120] Appleby.
[121] Ibid.
[122] Ibid.
[123] Davis and Vllalobos (2016).
[124] Knox.
[125] Ibid.
[126] Cleary and Bull (2018).
[127] "Statistically likely, as only 10% of supervisory officers in local police departments (Reaves, 2015) and 12% in local sheriff ;'soffices (Burch, 2016) are female."
[128] Cleary.
[129] Knox.

power struggle and stereotypical adherence to gender roles play a part of women and the likelihood for false confessions? Again, I urge for further research and consideration.

5.1.2. The cooperation component

I argue that in addition to the suspect being biased (perceived as uncooperative), this "un-cooperation" extends to the suspect's willingness to disclose information. Appleby (2015) found that in interpreting the interviews, "Black suspects did withhold more information about their actions from the interviewer than the White suspects."[130] Theories as to why such a pattern of results occurred are debatable and need further research. Appleby, as well as this study, suggest that it may be the influence of both underlying bias and police trust.[131] Thereby, it suggests that this interpretation as "un-cooperation" could potentially lead to more intense interrogations, and produce a compliant false confession.

5.2. Discussion of findings

This section 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. It 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.

I intended to point out both 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 study drew upon the findings of Appleby (2015), suggesting a racial stereotype threat in combination with police trust in willingness to disclose information. In addition, I took a theoretical/philosophical approach when it came to how societal stereotypes of women also may have the same effect on this uncooperative perception. It 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, it suggests trust and identifying with the interrogator could be important factors in suspect willingness to cooperate, which will be discussed next. Therefore, it 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. Stanford University published Empirically Validated Strategies to Reduce Stereotype Threat, in which the first suggestion was to "(1)Remove Cues That Trigger Worries About Stereotypes."[132] This study suggests that Blacks/ethnic minorities and females will continue to exhibit false positives of deception and be at risk for falsely confessing because of this underlying stereotype threat unless the bias is removed.

6. Programable similarity in human-to-AI trust and cooperation

First, an important consideration is to note that variables in studying human interrogation regarding bias, such as the influence of subtle cues, make both the study and replication incredibly difficult or nearly impossible. It is likely for this reason that few studies focus on interrogator bias and suspect perception. The ability to control and standardize an AI interrogator during interrogation (versus a human interrogator) can help to equalize interrogator response, allowing for greater ease in results interpretation. Body language on part of the interrogator is as integral of a part of this process as the subtle cues the suspect presents.

Ideally, we would want to discourage the perception of suspect un-cooperation but enhance the actual cooperation by means of willingness to disclose information. I propose the hypothetical enhancement of suspect and interrogator interaction using artificial intelligence could accomplish such. It 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; therefore it could yield safer[133] interrogations. 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.

6.1. Programmable physical suspect similarity through race and gender

I argue 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."[134] Fink argues that both influence and perception

[130] Appleby.
[131] Ibid.
[132] Stanford University."Empirically Validated Strategies to Reduce Stereotype Threat" Standford University. https://ed.stanford.edu/sites/default/files/interventionshandout.pdf.
[133] By means against pressure of false confessions.
[134] Fink.

depend upon physical components of the AI, and therefore are critical to the overall design.[135] She also urged the consideration for "demographic, cultural factors [22] [41], individual preferences, and the context of use."[136]

One dilemma in this 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."[137] The study— primarily focusing on understanding of human 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 study explored 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)."*[138]

This study notes that even though these findings were based on a virtual form of artificial intelligence, the evidence of AI gender 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."[139] They found that "participants were more swayed by the persuasive arguments of a virtual human that matched their gender than one who did not."[140] 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 study suggests that same gender interrogators should be utilized.

6.2. *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 study 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."[141] In addition, John E. Reid & Associates, commented that interpreters should also be familiar with the "cultural background, religious beliefs, and value system."[142] It suggests that the physical manipulation regarding the areas of ethnic and cultural similarity of artificial intelligence could have the same effect of increased 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 overall may be perceived as guilt by the interrogator.[143] Davis and Vllalobos (2016) speculated that "language difficulties can increase vulnerability to confession."[144] In the original basis of this study, Pollina and Barretta (2014), noted their study should be expanded to include "CG agents with different phys- ical characteristics or culture-specific utterances."[145] 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 Time-kettle located in Pasadena, California in the United States.[146]

[135] Ibid.
[136] Ibid.
[137] Ackerman (2018). "Humans Show Racial Bias Towards Robots of Different Colors: Study" IEEE Spectrum. https://spectrum.ieee.org/automaton/robotics/humanoids/robots-and-racism.
[138] Guadagno et al. (2007).
[139] Ibid.
[140] Ibid.
[141] Andrew and Tobia (2014). "Interrogation in War and Conflict: A Comparative and Interdisciplinary Analysis." London, England: Routledge. p. 239.
[142] Reid, John E. "The Use of an Interpreter During an Interview." PoliceLink http://policelink.monster.com/training/articles/1963-the-use-of-an-interpreter-during-an-interview.
[143] Davis and Vllalobos (2016).
[144] Ibid.
[145] Pollina and Barretta (2014).
[146] Timekettle, "WT2 Real-time Wearable Translator" Indiegogo. https://www.indiegogo.com/projects/wt2-real-time-wearable-translator#/.

6.3. Discussion of findings

Davis and Vllalobos (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.*"[147]

This section 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. Though human-to-AI cooperation needs substantial further research, past research suggests a working relationship between the two. It was important for this section 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, the 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 study 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.

7. Potential for achievement, acceptance, and consistency

In regards to hypothetical non bias achievement, acceptance, and consistency, this study looks at current policing racial and gender neutral algorithmic use in this exploration into artificial intelligence assisted interrogations. This study is specifically using risk assessment tools as a comparison for hypothetical racial and gender neutrality. The following section will demonstrate that the idea of a "non-biased" machine is achievable.

7.1. 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, I will argue that in the context of criminal justice, some employed algorithms have been successful in being racially and gender non biased. Again, it acknowledges that both racially and gender unbiased artificial intelligence, especially in the policing area, is a work in progress.

Within the United States, the Public Safety Assessment (PSA) risk assessment tool is aimed at enhancing a judge's pretrial decision process[148] and currently used in some capacity in approximately 29 jurisdictions across the United States.[149] 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."[150] 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."[151] In addition, they pointed out that the findings "do not… exacerbate disparate treatment by race and gender."[152] They also include 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."[153]

As for gender, they concluded that the Public Safety Assessment "to be free of predictive bias for FTAs and NCAs"[154] Lastly, the Arnold Foundation found that "Black and White defendants assessed with the PSA succeed at virtually identical rates."[155] 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]."[156]

Racial/gender neutral artificial intelligence algorithms have also received attention on an international level in the United

[147] Davis and Vllalobos (2016).
[148] Public Safety Assessment, "Risk Factors and Formula" https://www.psapretrial.org/about/factors.
[149] Pretrial Justice Institute (2017). PRETRIAL RISK ASSESSMENT CAN PRODUCE RACE-NEUTRAL RESULTS, https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=5cebc2e7-dfa4-65b2-13cd-300b81a6ad7a&forceDialog=0.
[150] DeMichele et al. (2018). "The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168452.
[151] Ibid.
[152] Ibid.
[153] Ibid.
[154] Ibid.
[155] Pretrial Justice Institute.
[156] Ibid.

Kingdom, although its usage is not as prevalent[157] as in the American System.[158] 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 of future offending and eligibility for the Constabulary's Checkpoint programme.[159] Though it does not directly mention gender neutrality, the Durham Constabulary police force specifically stated that race was not a factor in the algorithm's 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.

Therefore, racial and gender non-bias can be achieved through AI, and in addition, it suggests acceptance of additional AI utilization, given its prevalence in the United States, as well as the United Kingdom.

7.2. Programming and software maintenance, vetting, and regulation

I argue that failures have allowed for popularized AI platforms to develop a biased tendency, and thereby cause a mass societal critique and reluctancy towards artificial intelligence. In the development of non-biased artificial intelligence, I argue for 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."*[160]

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.[161] 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."[162] 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."[163] 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.[164] 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.[165] 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.[166] Furthermore, in Loomis v. Wisconsin, the 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."[167] Therefore, equality in performance requires minimizing human bias within the design process, as well as ensuring equality in both its application and performance.

Further, 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."[168] Hinnefeld also argues for such, noting an algorithmic potential impact on people's lives.[169] 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."[170]

[157] ~14% of the U.K. police applying computerized/ algorithmic data analysis or decision-making.

[158] Oswald, Grace, Urwin, and Barnes (2018), "Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality."Information & Communications Technology Law 27:2, 223-250. https://www.tandfonline.com/doi/pdf/10.1080/13600834.2018.1458455.

[159] Ibid.

[160] Gale (2018). "Can Artificial Intelligence Be Unbiased? PCMA. https://www.pcma.org/can-artificial-intelligence-be-unbiased/.

[161] Ibid.

[162] Hinnefeld (2018). "Reducing bias and ensuring fairness in data science." The Civis Journal. https://medium.com/civis-analytics/reducing-bias-and-ensuring-fairness-in-data-science-424ded3badbb.

[163] Jeong (2018). "AI is an excuse for Facebook to keep messing up" The Verge https://www.theverge.com/2018/4/13/17235042/facebook-mark-zuckerberg-ai-artificial-intelligence-excuse-congress-hearings.

[164] Gale.

[165] Ibid.

[166] Electronic Privacy Information Center (2020).

[167] Ibid.

[168] Baer and Kamalnath (2017), "Controlling machine-learning algorithms and their biases." McKinsey and Company. https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases.

[169] Hinnefeld (2018). "Reducing bias and ensuring fairness in data science." The Civis Journal. https://medium.com/civis-analytics/reducing-bias-and-ensuring-fairness-in-data-science-424ded3badbb.

[170] Gale.

She argues that, though AI can develop their own biases in the data, consistency in data vetting will, in time, eliminate the algorithmic biases.[171]

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 a 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."[172] 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.*"[173]

In conclusion, bias-free programming is absolutely achievable, but requires extensive work and expertise to do so. It is imperative that companies invest in ensuring proper data sets, expert programmers and non bias education, consistent vetting, and maintenance. Unlike humans, who have shown that racial biases are nearly impossible to change regardless of education or intervention, machines allow us to control these variables that we cannot control or change in humans. Hopefully, this may aid in maintaining bias-free artificial intelligence programs. However given its novelty, time would be needed to consider its actual aid.

8. Limitations & outlook

8.1. Legal limitations

There are a few legal limitations this study will recognize for further consideration. 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 countries that legally permit and use polygraphs within the interrogation room are the United States[174], Canada,[175] the United Kingdom[176] and Australia.[177] 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.[178] However, artificial intelligence accountability is highly debated, regarding the issue of programmer responsibility and the future autonomous decisions that an AI entity can make.[179]

Thirdly, in terms of tech innovation in general, laws recognizing such against the protections afforded by the People are lagging behind. Currently, laws within the United States favor protecting algorithmic design over transparency. If AI is going to be used in an interrogation room, addressing rights of the suspects or defendants may be an issue. For example, in 2016 the defendant in Loomis v. Wisconsin 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.[180] 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."[181] Aside from the concerning regarding the assessments decision factors of race and gender, there are two notable problems regarding the

[171] Ibid.

[172] Hale (2018). "2,531 viewsSep 25, 2018, 10:30pm IBM's Unbiased Approach To AI Discrimination" Forbes. https://www.forbes.com/sites/korihale/2018/09/25/ibms-unbiased-approach-to-ai-discrimination/#8007c3971185.

[173] Ibid.

[174] Stromberg (2014) "Lie detectors: Why they don't work, and why police use them anyway." Vox. https://www.vox.com/2014/8/14/5999119/polygraphs-lie-detectors-do-they-work.

[175] Yuen (2015). "Fear the examiner, not the polygraph." Toronto Sun, https://torontosun.com/2015/04/12/fear-the-examiner-not-the-polygraph/wcm/020cf4de-e22b-4efa-b3db-812fd303d128.

[176] Police trial lie detectors on sex offender suspects (2011). BBC News. https://www.bbc.com/news/uk-16371043.

[177] How the Truth Comes Out (2004). The Sydney Morning Herald. https://www.smh.com.au/national/how-the-truth-comes-out-20040820-gdjkzp.html.

[178] Reynolds.

[179] Roach (2016). "Holding Killer Robots Accountable? The New Moral Challenge of 21st Century Warfare." Columbia University Journal of International Affairs. https://jia.sipa.columbia.edu/online-articles/holding-killer-robots-accountable.

[180] Loomis and Wisconsin (2016) Docket No. 16-6387, "BRIEF FOR THE UNITED STATES AS AMICUS CURIAE" Wisconsin, United States https://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf.

[181] Loomis and Wisconsin (2016).

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. 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.

8.2. 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).[182] Police departments, 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 technology.[183] 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 £25 m (~€28,524,366 EUR).[184] As for the United States, I could not find technology within police forces as part of the 2018 federal funding.[185]

8.3. Outlook

The exploration into how artificial intelligence advancements in different fields can be used and their subsequent impact on social change is an essential piece of discussion complimentary to innovation. Future-oriented research needs to be encouraged if any innovation intends to bring out social benefits and consequences. While this study 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.[186] 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.[187] Current research and implementation are clearly serving as building blocks for more artificial intelligence within the criminal justice sector.

Another area of outlook 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 study. 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. 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, investigations of certain types of crimes could also narrow down the usage of artificial interrogators. A good starting point is the under-researched area of 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.[188] Again, ideally this artificial intelligence approach should be able to be applicable in all cases in the future.

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 Centre for 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 study, who's legal expertise in the field provided the critical feedback on the context and organization of the new age approach presented.

[182] Thirteen Advanced Humanoid Robots for Sale Today (2016). Smashing Robotics. https://www.smashingrobotics.com/thirteen-advanced-humanoid-robots-for-sale-today/

[183] Anderson (2013). "Technology on the beat: how IT can enhance policing." The Guardian. https://www.theguardian.com/public-leaders-network/2013/dec/18/criminal-justice-technology.

[184] Dodd (2014), Police force spends £25m on switch to technology-led crime-fighting. The Guardian. https://www.theguardian.com/uk-news/2014/jul/21/west-midlands-police-technology-led-crime-fighting

[185] Dorm (2017). "Police grants: What's being federally funded in 2018?"Police Grants Helphttps://www.policegrantshelp.com/columnists/samantha-dorm/articles/459658006-police-grants-whats-being-federally-funded-in-2018/.

[186] Tangermann (2019) "ESTONIA IS BUILDING A "ROBOT JUDGE" TO HELP CLEAR LEGAL BACKLOG" Futurism, https://futurism.com/the-byte/estonia-robot-judge.

[187] Pompedda (2018). "Training in Investigative Interviews of Children: Serious Gaming Paired with Feedback Improves Interview Quality" Åbo Akademi Universityhttps://pdfs.semanticscholar.org/14d9/45841824f7a003eee3bb38cf1dac0a26a1ed.pdf.

[188] NAACP (2020), "CRIMINAL JUSTICE FACT SHEET" NAACP, https://www.naacp.org/criminal-justice-fact-sheet/.

References

Ackerman, E. (2018). Humans show racial bias towards robots of different colors: Study. *IEEE spectrum*. https://spectrum.ieee.org/automaton/robotics/humanoids/robots-and-racism.

Anderson, T. (2013). *Technology on the beat: How IT can enhance policing.* The Guardianhttps://www.theguardian.com/public-leaders-network/2013/dec/18/criminal-justice-technology.

Andrew, C., & Tobia, S. (2014). *Interrogation in war and conflict: A comparative and interdisciplinary analysis.* London, England: Routledge239.

Anthropomorphism, Oxford Dictionary.

Appleby, S. C. (2015). *Guilty stereotypes: The social psychology of race and suspicion in police interviews and interrogations.* City University of New York (CUNY) CUNY Academic Workshttps://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1518&context=gc_etds.

Baer, T., & Kamalnath, V. (2017). *Controlling machine-learning algorithms and their biases.* McKinsey and Companyhttps://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases.

Bainbridge, A. W., Hart, J., Kim, E., & Scassellati, B. (2008). The effect of presence on human-robot interaction. *Conference: Robot and Human Interactive Communication,* 701–706.

Balaji, Abjishek and Choubey Amal Kumar. "ML | What is Machine Learning?" GeeksforGeeks.com https://www.geeksforgeeks.org/ml-machine-learning/.

Balliet, D. (2009). Communication and cooperation in social dilemmas: a meta-analytic review. *Rationality and Society, 54*, 39–57.

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2018). Facial color is an efficient mechanism to visually transmit emotion. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1716084115.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Bulletin, 10*, 214–234.

Carbado, D. W., & Richardson, S. L. (2018). *(Book review) "The black police: Policing our own".* 131 Harv. L. Rev. 1979https://harvardlawreview.org/2018/05/the-black-police-policing-our-own/.

Cleary, H., & Bull, R. (2018). Jail Inmates' Perspectives on Police Interrogation. *Psychology Crime and Law* https://www.researchgate.net/publication/326531981_Jail_Inmates'_Perspectives_on_Police_Interrogation.

Crandall, J. W., Oudah, M. T., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*. https://www.nature.com/articles/s41467-017-02597-8.

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist, 35*, 867–881.

Dasgupta, P. B. (2017). Detection and analysis of human emotions through voice and speech pattern processing. *International Journal of Computer Trends and Technology (IJCTT), 52*(October (1)) 2017.

Davis, D., & Vllalobos, J. G. (2016). Interrogation and the minority suspect: Pathways to true and false confession. *Advances in Psychology and Law, 1*https://www.researchgate.net/publication/290436784_Interrogation_and_the_Minority_Suspect_Pathways_to_True_and_False_Confession.

DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., & Misra, S. (2018). *The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in Kentucky*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168452.

Dodd, V. (2014). *Police force spends £25m on switch to technology-led crime-fighting.* The Guardianhttps://www.theguardian.com/uk-news/2014/jul/21/west-midlands-police-technology-led-crime-fighting.

Dorm, S. (2017). *Police grants: What's being federally funded in 2018?* Police Grants Helphttps://www.policegrantshelp.com/columnists/samantha-dorm/articles/459658006-police-grants-whats-being-federally-funded-in-2018/.

Electronic Privacy Information Center (2020). *Algorithms in the criminal justice system.* EPIChttps://epic.org/algorithmic-transparency/crim-justice/.

Feld, B. C. (2014). *Questioning gender: Police interrogation of delinquent girls.* University of Minnesota Law Schoolhttps://scholarship.law.umn.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1301&context=faculty_articles.

Fink, J. (2012). *Anthropomorphism and human likeness in the design of robots and human-robot interaction.* Berlin Heidelberg: © Springer-Verlag199–208.

Gale, C. (2018). Can artificial intelligence be unbiased? *PCMA*. https://www.pcma.org/can-artificial-intelligence-be-unbiased/.

Galeon, D. (2018). *A new AI that detects "Deception" may bring an end to lying as we know it.* FUTURISMhttps://futurism.com/new-ai-detects-deception-bring-end-lying-know-it.

Guadagno, R. E., Blascovich, J., Bailenson, J. N., & McCall, C. (2007). Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology, 10*, 1–22 2007.

Hale, K. (2018). *2,531 views Sep 25, 2018, 10:30pm IBM's unbiased approach to AI discrimination.* Forbeshttps://www.forbes.com/sites/korihale/2018/09/25/ibms-unbiased-approach-to-ai-discrimination/#8007c3971185.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., Visser, E. D., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517–527.

Hinnefeld, H. (2018). Reducing bias and ensuring fairness in data science. *The Civis Journal*. https://medium.com/civis-analytics/reducing-bias-and-ensuring-fairness-in-data-science-424ded3badbb.

How the Truth Comes Out (2004). *The Sydney morning herald.* https://www.smh.com.au/national/how-the-truth-comes-out-20040820-gdjkzp.html.

Hsu, J. (2012). Why "Uncanny valley" human look-alikes put us on edge. *Scientific American*. https://www.scientificamerican.com/article/why-uncanny-valley-human-look-alikes-put-us-on-edge/.

Innocence Project (2020). *DNA exonerations in the United States.* https://www.innocenceproject.org/dna-exonerations-in-the-united-states/.

Jeong, S. (2018). *AI is an excuse for Facebook to keep messing up.* The Vergehttps://www.theverge.com/2018/4/13/17235042/facebook-mark-zuckerberg-ai-artificial-intelligence-excuse-congress-hearings.

Jones, S. (2011). Under pressure: Women who plead guilty to crimes they have not committed. *Criminology and Criminal Justice*. https://doi.org/10.1177/1748895810392193.

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition, 26*, 169–181.

Knight, H. (2014). *How humans respond to robots: Building public policy through good design.* Carnegie Mellon Universityhttps://www.brookings.edu/wp-content/uploads/2014/07/HumanRobot-PartnershipsR2.pdf.

Knox, A. (2016). *Amanda Knox: Why do innocent women confess to crimes they didn't commit?* Broadlyhttps://broadly.vice.com/en_us/article/9k9wep/amanda-knox-why-do-innocent-women-confess-to-crimes-they-didnt-commit.

Kobayashi, H., & Hara, F. (1997). Facial interaction between animated 3D face robot and human beings. *IEEE Xplore 2002*. https://doi.org/10.1109/ICSMC.1997.633250.

Langford, M. (2019). *Collection of suggestions for AI interrogator functionality*.

Layton, J. How police interrogation works. Culture, https://people.howstuffworks.com/police-interrogation1.htm.

Leo, R. A. (2009). False Confessions: Causes, Consequences, and Implications. *Journal of the American Academy of Psychiatry and the Law Online, 37*(September (3)), 332–343. http://jaapl.org/content/37/3/332.

Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies, 68*(1/2), 57–68.

Loomis, & Wisconsin (2016). *Docket No. 16-6387, "Brief for the United States as amicus curiae" Wisconsin, United States.* https://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf.

Lopez, G., & Zarracina, J. (2017). *Study: Black people are 7 times more likely than white people to be wrongly convicted of murder.* VOXhttps://www.vox.com/policy-and-politics/2017/3/7/14834454/exoneration-innocence-prison-racism.

Mathura, M., & Reichling, D. B. (2014). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition, 146*(January), 22–32.

Meissner, C. A., & Kassin, S. M. (2004). "You're guilty, so just confess!" Cognitive and behavioral confirmation biases in the interrogation room. In G. D. Lassiter (Ed.).

Interrogations, confessions, and entrapment (pp. 85–106). New York: Kluwer Academic/Plenum Publishers.
Melina, R. (2011). Is' mars needs moms' too realistic? *Life Science*. https://www.livescience.com/33198-is-mars-needs-moms-too-realistic.html.
NAACP (2020). *Criminal justice fact sheet.* NAACPhttps://www.naacp.org/criminal-justice-fact-sheet/.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220.
Ohio State University (2018). *At first blush, you look happy—Or sad, or angry.* Ohio State Universityhttps://news.osu.edu/at-first-blush-you-look-happy–or-sad-or-angry/.
Orlando, J. (2020). *"Interrogation Techniques." OLR Research Report*Connecticut General Assemblyhttps://www.cga.ct.gov/olr/.
Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and "Experimental" proportionality. *Information & Communications Technology Law, 27*(2), 223–2250. https://doi.org/10.1080/13600834.2018.1458455.
Police trial lie detectors on sex offender suspects (2011). *Police trial lie detectors on sex offender suspects.* BBC Newshttps://www.bbc.com/news/uk-16371043.
Pollina, D. A., & Barretta, A. (2014). The effectiveness of a national security screening interview conducted by a computer-generated agent. *Computers in Human Behavior, 39*, 39–50. https://doi.org/10.1016/j.chb.2014.06.010 C (September 2014).
Pompedda, F. (2018). *Training in investigative interviews of children: Serious gaming paired with feedback improves interview quality.* Åbo Akademi Universityhttps://pdfs.semanticscholar.org/14d9/45841824f7a003eee3bb38cf1dac0a26a1ed.pdf.
Poulter, S., & Harry, P. (2018). *Sophia takes her first steps: Super-intelligent humanoid robot reveals she now has legs.* DailyMailhttps://www.dailymail.co.uk/sciencetech/article-5251185/Worlds-robot-citizen-Sophia-gets-legs.htm.
Pretrial Justice Institute (2017). *Pretrial risk assessment can produce race-neutral results.* https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey = 5cebc2e7-dfa4-65b2-13cd-300b81a6ad7a&forceDialog = 0.
Public Safety Assessment. Risk factors and formula, https://www.psapretrial.org/about/factors.
Redlich, A.D. & Meissner, C. (n/a) Techniques and controversies in the interrogation of suspects: The artful practice versus the scientific study (to appear in) D.L. Skeem, K. Douglas, & S. Lilienfeld (Eds.), Psychological science in the courtroom: Controversies and consensus. https://digitalcommons.utep.edu/cgi/viewcontent.cgi?article = 1036&context = christian_meissner.
Reid, J.E. The use of an interpreter during an interview, PoliceLink http://policelink.monster.com/training/articles/1963-the-use-of-an-interpreter-during-an-interview.
Reynolds, E. (2018). *The agony of Sophia, the world's first robot citizen condemned to a lifeless career in marketing.* WIREDhttps://www.wired.co.uk/article/sophia-robot-citizen-womens-rights-detriot-become-human-hanson-robotics.
Roach, S. C. (2016). Holding killer robots accountable? The new moral challenge of 21st century warfare. *Columbia University Journal of International Affairs*. https://jia.sipa.columbia.edu/online-articles/holding-killer-robots-accountable.
Rouse, M. DEFINITION expert system, TechTarget https://searchenterpriseai.techtarget.com/definition/expert-system.
Sally, D. (1993). Conversation and cooperation in social dilemmas a meta-analysis of experiments from 1958 to 1992. *Rationality and Society, 7*, 58–92.
Sharma, I. Introduction to decision tree learning, HeartBeat. https://heartbeat.fritz.ai/introduction-to-decision-tree-learning-cd604f85e236.
Shead, S. (2016). *Humans are willing to trust chatbots with some of their most sensitive information.* Business Insiderhttps://www.businessinsider.com/humans-are-willing-to-trust-chatbots-with-some-of-their-most-sensitive-information-2016-6?r = UK&IR = T&IR = T.
Siegel, M. S. (2009). *Persuasive robotics how robots change our minds.* Massachusetts Institute of Technologyhttps://core.ac.uk/download/pdf/4411144.pdf.
Stanford University. Empirically validated strategies to reduce stereotype threat, Stanford University. https://ed.stanford.edu/sites/default/files/interventionshandout.pdf.
Stromberg, J. (2014). *Lie detectors: "Why they don't work, and why police use them anyway".* Voxhttps://www.vox.com/2014/8/14/5999119/polygraphs-lie-detectors-do-they-work.
Tangermann, V. (2019). *Estonia is building a "robot judge" to help clear legal backlog.* Futurismhttps://futurism.com/the-byte/estonia-robot-judge.
The Computer Will See You Now (2014). *The economist.* https://www.economist.com/science-and-technology/2014/08/20/the-computer-will-see-you-now.
Thirteen Advanced Humanoid Robots for Sale Today (2016). *Smashing robotics.* https://www.smashingrobotics.com/thirteen-advanced-humanoid-robots-for-sale-today/.
Timekettle, WT2 real-time wearable translator. Indiegogo. https://www.indiegogo.com/projects/wt2-real-time-wearable-translator#/.
Ulatowska, Joanna (2013). Indicators of deception in different lie scenarios. *Roczniki Psychologiczne/Annals of Psychology, XVI*(1), 127–146.
Weir, K. (2016). Policing in black & white. *Monitor on Psychology, 47*(11)https://www.apa.org/monitor/2016/12/cover-policing.
Wu, Z., Singh, B., Davis, L. S., & Subrahmanian, V. S. (2017). *Deception detection in videos.* Cornell University Libraryhttps://arxiv.org/abs/1712.04415.
Yuen, J. (2015). *Fear the examiner, not the polygraph.* Toronto Sunhttps://torontosun.com/2015/04/12/fear-the-examiner-not-the-polygraph/wcm/020cf4de-e22b-4efa-b3db-812fd303d128.