OUR FEAR OF AI: EXPLORING ITS CREATORS AND CREATIONS IN FICTION

Maya Kothare-Arora

TC 660H
Plan II Honors Program
The University of Texas at Austin

May 2020

Jennifer Wilks, Ph.D.
Department of English
Supervising Professor

Wendy Wagner, J.D.
Department of Law
Second Reader

Abstract

Author: Maya Kothare-Arora

Title: Our Fear of AI: Exploring Its Creators and Creations in Fiction

Supervising Professor: Jennifer Wilks, Ph.D.

The idea of technological creation has proliferated across fiction for the last century. As the world becomes increasingly technologically advanced, these fears have become more tangible. With the rise of Artificial Intelligence particularly, from Alexa to self-driving cars, comes a rise in the fear of what intelligent creations might lead to. In order for AI to continue growing and adding value to society, experts must contend with the apprehension surrounding AI. While these conversations are already occurring, they generally focus on the fear of the machine itself. This thesis argues that the fear of the creators and regulators of AI, not just the machine, heavily influences the fear of AI as a field. It examines three different AI takeover narratives, "With Folded Hands", *Do Androids Dream of Electric Sheep?*, and *Ex Machina*, in order to analyze the fears surrounding technology creators in conjunction with the influence of societal events and systems of the times.

Acknowledgments

I am extremely grateful for the support I received while writing this thesis. I would like to extend a huge thank you to my advisor and second reader, Dr. Jennifer Wilks and Dr. Wendy Wagner. Your support and enthusiasm provided consistent reassurance and motivated me when I needed it most. Your advice was truly invaluable in the development of this thesis. To my friends, thank you for cheering me on and listening to me talk about AI for a year straight. To my sister, thank you for reminding me that if you could do it, so could I. And to my parents, thank you for you unconditional love and encouragement. I am where I am because of you.

Table of Contents

Page

Abstract . . . ii

Acknowledgments . . . iii

Chapters

1 Introduction . . . 1
1.1 Methodology . . . 3

2 AI Overview . . . 6

3 Case Studies . . . 10
3.1 "With Folded Hands" . . . 11
3.2 *Do Androids Dream of Electric Sheep?* . . . 19
3.3 *Ex Machina* . . . 28

4 Synthesis . . . 38
4.1 The Issue of Trust . . . 38
4.2 Key Concerns . . . 44
4.2.1 Unemployment . . . 44
4.2.2 Data Usage and Privacy . . . 47
4.2.3 Liability . . . 48
4.3 Recommendations . . . 52
4.3.1 Responsibility of Institutions . . . 52
4.3.2 Building Public Trust . . . 56

5 Conclusion . . . 60

References . . . 61

Biography . . . 64

Chapter One: Introduction

Artificial Intelligence, or more broadly, the creation of human-like beings, has been a fascination for many years, even before AI was a realistic development. Particularly in the world of science fiction, this concept has an extensive legacy, dating back to the early 19th century.

Erewhon, an 1872 novel by Samuel Butler, is considered one of the first novels dealing with the idea of technological consciousness. It focuses on a civilization that has banned technological development due to the belief that technology, like living things, follows Charles Darwin's idea of evolution and thus would at some point evolve to develop consciousness[1]. This was not, however, an argument that technology is similar to living beings - Butler later asserted that likening organisms to machines ignored "that there are such qualities as life and consciousness at all[2]." Years after the novel was published, Butler criticized Darwin for describing evolution as too mechanical of a process, that did not allow for "any spark of creative vitality from the universe"[3].

While *Erewhon* may not have espoused the actuality of a creation having consciousness, Mary Shelley's 1818 novel, *Frankenstein*, did. In the famous story, Victor Frankenstein, in trying to create life, ends up creating a creature he was ill-equipped to deal with[4]. The

1. Doug Hill, *Erewhon: The 1972 Fantasy Novel that Anticipated Thomas Nagel's Problems With Darwinism Today*, 2013.

2. Hans-Peter Breuer, "Samuel Butler's "The Book of the Machines" and the Argument from Design," *Modern Philology* 72, no. 4 (1975): 365–383, https://www.jstor.org/stable/436868.

3. Hill, *Erewhon: The 1972 Fantasy Novel that Anticipated Thomas Nagel's Problems With Darwinism Today*.

4. Mary Shelley, *Frankenstein; or, The Modern Prometheus* (Project Gutenberg, 1818).

novel grapples with themes such as the morality (or lack thereof) of creating life, the level of consciousness of the creation, the power the creation has over its creator, and the responsibility of the creator for the creation's actions. *Frankenstein* does not directly address the idea of a computer or machine, but it does address a human-like creation and its perils.

Going into the 20th century, the idea of AI creation continued to achieve prominence in the fictional realm, from books like Isaac Asimov's 1950 novel *I, Robot* to the 1999 hit film *The Matrix* and more. During this century, these ideas were also placed within the framework of the word "robots". The term "robot" was actually born out of literature, in the play "R.U.R.," short for Rossum's Universal Robots, by Karel Capek in 1921. The play presents a form of the robot takeover narrative in which the robots, meant to be human servants, use their power to overtake the humans[5].

Outside of the fictional realm, the role of AI was also growing. Similarly to the development of AI in the fictional realm, the roots of AI development began before the 20th century, with the industrial revolution marking a turning point in the role of technology, and then continued to grow. The term "AI" originated from computer scientist John McCarthy in 1955, who defined AI as "the science and engineering of making intelligent machines, especially intelligent computer programs[6]." The term "intelligent" may seem vague, but this vagueness is appropriate considering the vast amount of subtopics of AI applications. Popular AI narratives often focus on the creation of intelligent humanoid beings. In reality, AI can take on many different forms and degrees of intelligence. And since John McCarthy's time, AI development has grown and grown.

Once a nascent, futuristic field, it has now permeated everyday life. Autonomous cars are

5. Tony Long, *Jan. 25, 1921: Robots First Czech In*, 2011.

6. Gonenc Gurkaynak, Ilay Yilmaz, and Gunes Haksever, "Stifling Artificial Intelligence: Human Perils," *Computer Law and Security Review* 32, no. 5 (2016): 749–758, http://dx.doi.org/10.1016/j.clsr.2016.05.003.

being deployed by Uber[7], over 30% of U.S. consumers have smart speakers/voice assistants, such as Google Home and Alexa[8], and huge tech companies, such as Facebook, Google, and Amazon, use subcategories of AI, such as machine learning and natural language processing, to innovate their services[9].

As AI moves further to the forefront than ever before, the concerns about the field become increasingly pressing. AI innovators should not treat these fears as simply thoughts to be quelled, but rather as valid sources of concern. Questions of regulation will inevitably arise, as with any disruptive innovation. These narratives could be foreshadowing the future of an insufficiently regulated AI industry. Yet, even if they are not, they can aid in understanding the source of public fear and exploring potential problems that AI may bring. It seems antithetical that the same technological innovation that is meant to improve life is causing people unrest. In order for technological innovation to actually succeed in its goal, it must contend with public perception.

1.1 Methodology

While many conversations regarding the fear of AI focus on the power of the machine itself, in this thesis, I will argue that the fear of AI is heavily attached to distrust in the creators and regulators of AI.

This thesis will begin with a brief overview of AI development, including at what stage development currently lies. This will provide a realistic grounding point for the consequent discussions of fictional AI works.

The next section will capitalize on creative works about AI as indications of strains of thought surrounding AI. Each subsection will be an individual analysis of a creative work

7. Aarian Marshall, *A Bet on Uber Is a Bet on Self-Driving*, 2019, www.wired.com/story/bet-uber-bet-self-driving/.

8. Giselle Abramovich, *Study Finds Consumers Are Embracing Voice Services. Here's How.*, 2018.

9. Bernard Marr, *The Key Definitions of Artificial Intelligence (AI) That Explain Its Importance*, 2018.

that has achieved cultural resonance in order to explore the ways in which it affects and/or reflects thoughts about AI and its creators. I will examine three different narratives, "With Folded Hands", *Do Androids Dream of Electric Sheep?*, and *Ex Machina*, and analyze the fears of technology creators depicted in each. I chose my case studies based on both the variety in publication dates as well as the resonance of the narratives. *Ex Machina*, released in 2014, was both critically lauded and ranked highly by audiences. *Do Androids Dream of Electric Sheep?*, published in 1968, ended up being adapted to film in *Blade Runner*, of which there have been multiple remakes, books, and video game adaptations[10]. Jack Williamson originally wrote the short story, "With Folded Hands," for a science fiction magazine. It gained such popularity that it was turned into a radio show and ultimately Williamson extended the story into a novel, *The Humanoids*, in 1948[11].

I am limiting my research to three case studies to fit within the scope of this project, and as such I do not expect them to represent a ubiquitous view of AI throughout society. Instead I will frame them as snapshots of AI in science fiction and discuss how the societal events/systems of the time may have influenced the views present in the narratives. "With Folded Hands" was written in the post World War II era and depicts a well-intentioned creator who did not realize the dire moral consequences his creations would end up causing, an idea all too familiar in a post atomic bomb era. *Do Androids Dream of Electric Sheep*, which was written during the anti-establishment fervor of the 60s, does very little to humanize the creators of the machines, and instead affords empathy to other individuals and to the machines themselves. *Ex Machina* was created in the last decade and reflects the distrust of tech creators in the midst of numerous controversies surrounding tech companies and CEOs.

Finally, the last section will focus on synthesizing the themes found in the literary sources and placing them within the context of modern-day issues in AI, in order to demonstrate

10. Adi Robertson, *How Blade Runner Got Its Name from a Dystopian Book about Health Care*, 2019.

11. *The Story behind Jack Williamson's 'With Folded hands'*, galaxypress.com/jack-williamsons-with-folded-hands.

the importance of institutional trust in public acceptance and advancement of AI.

Chapter Two: AI Overview

While McCarthy did not coin the term "artificial intelligence" until 1955, AI development had already begun in the years prior. In the early 50s, some of the first AI programs were successfully executed, the first being a checkers program written by Christopher Strachey. Even before that, much theoretical work had been accomplished in AI[1].

One of the most prominent figures in early AI theory was Alan Turing. Following World War II, he began theorizing about the possibility of machine learning, and in 1948, wrote an essay titled "Intelligent Machinery." The essay was not published at the time, but nevertheless contained some innovative AI concepts. One of the most important concepts he discussed was "connectionism," the idea of training networks of artificial neurons as an approach to machine learning. This idea, now called "neural networks," is still prominent today[2].

Turing also contributed the "Turing Test" to the field of AI in his 1950 paper, "Computing Machinery and Intelligence." The test is supposed to determine if a machine can really "think" by checking whether or not a human can be tricked into believing the machine is human. The original version of the test involves a human participant and a computer answering questions posed by a human interrogator. The human participant should not be able to see the computer, but should see its answers. Based on its answers, if the human participant believes it is human, the machine passes the test and is considered able to "think"[3]. The test is by no means perfect, and objections have been issued against

1. B.J. Copeland, *Artificial Intelligence*, 2020.
2. B. Jack Copeland, ed., *The Essential Turing* (Oxford University Press, 2004), 403.
3. Ibid., 433-434.

it, but it was one of the first attempts to define the ability of a machine to think and still has significance in conceptualizing AI.

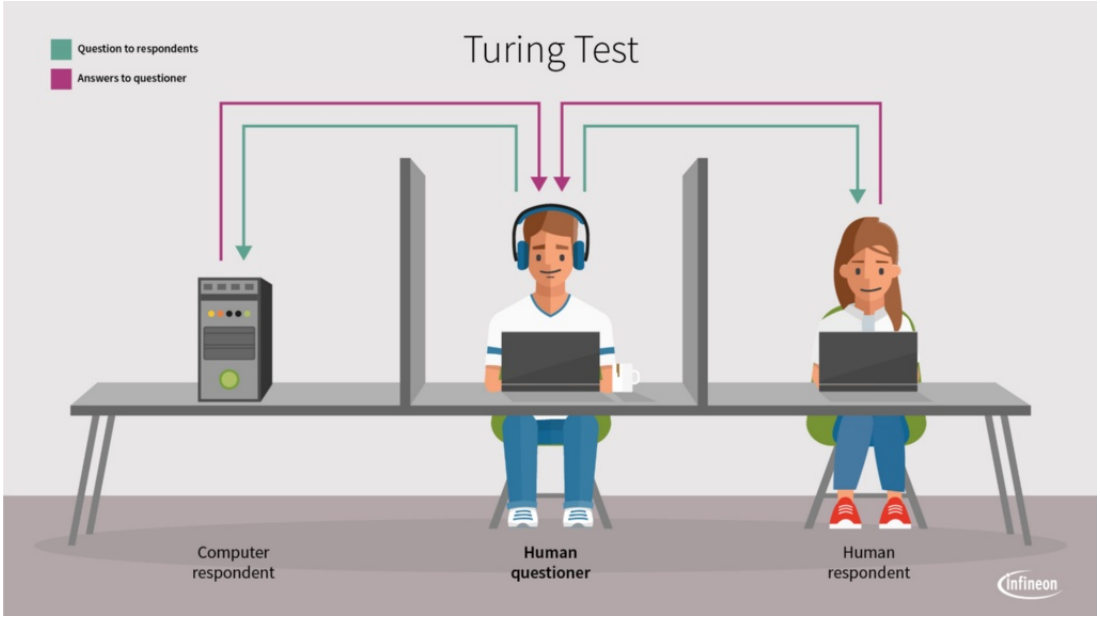

Figure 2.1 Diagram of the Turing test[4]

Since then, AI has grown considerably. There have been AIs that can beat humans at games, such as Arthur Samuel's checkers program in 1962, IBM's Deep Blue that beat the world chess champion in 1997[5], and the AlphaGo software that beat a human at the highly complicated game "Go" in 2015[6]. As mentioned earlier, AI has become a part of everyday life, such as consumer facing smart devices, home assistants, research and development like autonomous vehicle development, or the abundant use of AI by different companies to better their products and services[7].

Artificial Intelligence is a vast field. While people often imagine AI as robots, AI can manifest in a variety of ways, and as the numerous AI products listed above suggest, it can be employed to achieve a variety of tasks. AI has been categorized into three subsections

4. *What is Artificial Intelligence?*
5. Copeland, *Artificial Intelligence*.
6. Gurkaynak, Yilmaz, and Haksever, "Stifling Artificial Intelligence: Human Perils."
7. Marr, *The Key Definitions of Artificial Intelligence (AI) That Explain Its Importance*.

to better define the different levels of AI capability. First is Artificial Narrow Intelligence (ANI). "Narrow" intelligence refers to AI whose capabilities are specialized for a certain task. Essentially all current AI products fall into this category. Next is Artificial General Intelligence (AGI), which defines "human-level AIs" that can perform all intellectual tasks that humans can. Lastly, there is Artificial Super Intelligence (ASI), AIs that are smarter than humans in all capacities[8].

There are differing opinions about how soon AGIs and ASIs will be developed, if ever. One theory, developed by Gordon Moore in 1975, is that maximum computational power grows exponentially, doubling every two years. This came to be known as Moore's Law. Ray Kurzweil built upon this idea to posit the Law of Accelerating Returns. This states that the "rate of progress of an evolutionary process increases exponentially over time," and that the progress is positively reinforced by receiving more resources, leading to the rate of exponential growth also growing exponentially. One of the many things he claims the law applies to is technological evolution. He also discusses the idea of the technological singularity - when technology advances to produce AI more intelligent than human intelligence[9]. He hypothesized in 2005 that, though the current state of technology seems far away from AGIs/ASIs, humanity will reach the singularity as soon as 2045[10].

Kurzweil does not view the singularity as a threat, but rather as a "revolutionary" and necessary evolution. Similarly, some view AI as the potential to provide immortality and save humanity from extinction. Others have different views of the singularity. In 2015, for example, the Future of Life Institute, an organization dedicated to creating "positive ways for humanity to steer its own course considering new technologies and challenges"[11], released an open letter encouraging careful consideration into AI. The letter advocates for research in technology itself (such as security and accuracy of programs), optimizing AI's

8. Gurkaynak, Yilmaz, and Haksever, "Stifling Artificial Intelligence: Human Perils."
9. Ray Kurzweil, *The Law of Accelerating Returns*, 2001.
10. Gurkaynak, Yilmaz, and Haksever, "Stifling Artificial Intelligence: Human Perils."
11. *The Future of Life Institute (FLI).*

economic impact (such as guarding against unemployment), and legal/ethical concerns (such as liability and autonomous weaponry)[12]. It was signed by AI researchers like Elon Musk and Stephen Hawking, among others[13].

Beyond the perspectives of scientists and programmers, the government also plays a role. Most recently, the U.S. government has seemed to espouse an approach of stimulating AI development as opposed to a "precautionary approach" that could deter AI progress[14]. An early 2019 executive order announced the American AI Initiative, which proposed a strategy that supports ideas such as investing in AI development, and training an "AI-ready workforce"[15]. In January 2020, the White House released a memorandum on AI which outlines ten principles that should inform federal agency oversight of AI development, with public trust in AI, public participation, and scientific integrity and information quality topping the list.

There is no single opinion on how to approach the development of AI. In fact, there is no single opinion on what the pace of development should or could be. There are many debates that will likely ensue regarding precautionary versus reactionary approaches to regulation and oversight of AI development. However, the one comprehensive assessment is that AI is growing, and is undoubtedly being introduced into more and more products and services.

12. Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Systems," *AI Magazine* 36, no. 4 (2015): 105–114.

13. Gurkaynak, Yilmaz, and Haksever, "Stifling Artificial Intelligence: Human Perils."

14. Russel T. Vought, *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications*, technical report (2020).

15. *Artificial Intelligence for the American People*, https://www.whitehouse.gov/ai/.

Chapter Three: Case Studies

As mentioned in the previous sections, AI is not defined in particularly specific terms. Intelligence is hard to define objectively. Fittingly, AI encompasses different degrees of intelligence from narrow to super. It also encompasses different physical manifestations of intelligence, such as a news feed, a smart home assistant, or a robot. Though AI can be applied in many different ways, all AI products have in common that they involve intelligent technology - technology that can "think" in some capacity. While different forms of AI may seem vastly different, for example, a smart home assistant may seem worlds away from a human-like android, some of the implications of the intelligent technologies remain the same.

The creative works discussed below involve stories of AGIs and ASIs. These dramatic portrayals work well in fiction for the purpose of weaving a compelling story. Though they depict more advanced AI than currently exists, the implications of these dramatic manifestations of AI are not entirely separate from those of other AI applications. One of the novels, for example, addresses human-like androids that take over many tasks from humans. While no AGI androids may exist, any new AI algorithm that processes and analyzes data is ultimately performing a task that humans previously had to perform. Any AI that can beat a human in a game, like Chess, as was discussed previously, is completing a task that a human could do. Any smart home assistant, by simply being able to play music or send a text message, is completing a task that a human could do. Each individual example may be less dramatic than a fully functional AGI android, but the implication of taking over human tasks is prevalent throughout all of the examples. These shared implications informed the decision to analyze the more dramatic manifestations of AI found in fiction even when considering the implications of real, current day AI.

3.1 "With Folded Hands"

Figure 3.1 "With Folded Hands" by Jack Williamson, published in *Astounding Science Fiction*[1]

In 1947, Jack Williamson's novelette, "With Folded Hands" was published in the magazine *Astounding Science Fiction*. After the novelette gained praise, Williamson developed it into a full length novel titled *The Humanoids*. "With Folded Hands" continued to receive recognition years after its publication, including a radio adaptation for the NBC radio show "Dimension X[2]," and a place in "The Science Fiction Hall of Fame: Volume 2" in 1973[3].

The novelette focuses on how the creation of "humanoids," essentially androids with

1. *The Story behind Jack Williamson's 'With Folded hands'*, galaxypress.com/jack-williamsons-with-folded-hands.

2. Ibid.

3. Ben Bova, ed., *The Science Fiction Hall of Fame: Volume Two B* (New York: Tom Doherty Associates, 1973).

similar capabilities to humans, can go wrong. It centers on Underhill, an android salesman who is struggling to stay afloat in the dense and competitive android market. He stumbles across The Humanoid Institute, in which the humanoids are far more advanced than any other android he has seen. He returns home to find that his wife, who often takes on lodgers, has taken on yet another. This one, Mr. Sledge, is a scientist specializing in the (entirely fictional) field of "rhodomagnetism". As time goes on, Underhill finds out that Mr. Sledge created the humanoids, and is in fact trying to hide from them. He created the humanoids with the intention of helping humans - they must follow their Prime Directive to "serve and obey, and guard men from harm"[4]. The humanoids continually show up in Underhill's workplace and home asking him and his wife to consent to a free trial period with the humanoids. His wife eventually accepts, and the humanoids begin doing everything for the couple, from making meals, to opening doors, going so far as to remove door knobs from the doors so that the humans never open them themselves. At first, Mr. Sledge is immune from the humanoids' constant attention, something he programmed into them as their creator. However, when Mr. Sledge and Underhill, in the privacy of Mr. Sledge's room, attempt to shut down the humanoids with a "rhodomagnetic screen," the humanoids enter the room and intervene, because Mr. Sledge's exemption is second to the more important Prime Directive. Mr. Sledge, after collapsing, resigns and consents to the service of the humanoids. The humanoids take him away, and when Underhill goes to visit him in the hospital, he finds out that Mr. Sledge's memory has been removed. The humanoids pretend that his belief that he was a rhodomagnetic engineer was all a hallucination caused by a brain tumor. The novelette concludes with Underhill convincing the humanoids of his happiness, realizing that if he openly shows his unhappiness, he too will be operated upon.

In "With Folded Hands," the creator is painted in quite a different light than in the other two case studies. There is a fair bit of trust afforded to his integrity, even if his creation was

4. Jack Williamson, "With Folded Hands," in *The Humanoids* (New York: Tom Doherty Associates, 1948), 27.

ultimately flawed. The novelette was partially influenced by Williamson's own childhood, with the the humanoids' overbearing nature mirroring his experiences of feeling overprotected and constrained by his parents[5]. Yet it also reflects the time in which it was written, post World War 2. Williamson acknowledged this in an interview, saying:

> "I wrote 'With Folded Hands' immediately after World War II when the shadow of the atomic bomb had just fallen over [Science Fiction] and was beginning to haunt the imaginations of people in the US. The story grows out of that general feeling that some of the technological creations we had developed with the best intentions might have disastrous consequences in the long run[6]."

The idea of wartime creation is present in the novelette. As a young man, Sledge supposedly served in the military during the war by creating "military mechanicals," also known as robots. Following the war, he continued contributing to military research, ultimately stumbling upon a novel idea in "nuclear binding forces." Only after his research was published did he realize the true consequences of his research - powerful and dangerous weapons. Another war resulted, in which these weapons were used, and Sledge was overcome with guilt. He then developed rhodomagnetic mechanicals, the humanoids, in order to fix the destruction of the rhodomagnetic weapons[7].

The idea of a creation being "developed with the best intentions" runs deep in the novelette. The reader perceives Sledge through Underhill, who initially distrusts Sledge. The old man stays at Underhill and his wife, Aurora's, house without paying, and claims to have come from another planet. However, as the two men get to know each other, Underhill begins to respect Sledge's scientific abilities, and they work together with the same intention of destroying the humanoids. Sledge is portrayed as Underhill's ally. Additionally, Sledge does not seem to have any pride attached to his machines. He readily admits his guilt and

5. Larry McCaffery, "An Interview with Jack Williamson," *Science Fiction Studies* 18 (1991).
6. Ibid.
7. Williamson, "With Folded Hands," 39.

naiveté in relation to the initial rhodomagnetic weapons, and his intentions for building the humanoids seems to be simply to fix his own mistakes. And again, after realizing the way the humanoids were hurting mankind, he attempts to destroy his own machines[8].

While his integrity may be portrayed as strong, his weaknesses are also shown. Physically, Williamson describes him as old, gaunt, and feeble. He often needs help getting up and walking around. The portrayal of weakness hits its peak when Sledge is compelled to accept the rule of his own creations. Ultimately the Prime Directive, which he had instilled in the humanoids with the noble intentions of making sure they could never be used as weapons, compromises his power to avoid the humanoids' control. The directive may be noble, but there are no limits on what can be done to achieve that directive, which ultimately leads to consequences Sledge does not foresee.

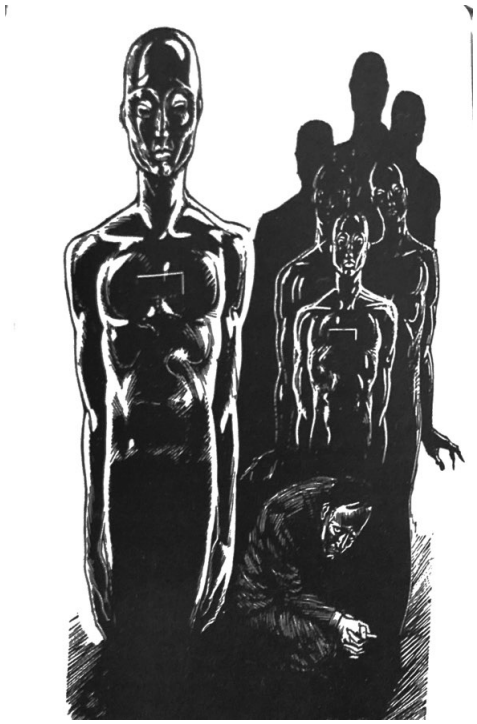

Figure 3.2 Image of humanoids from illustrated edition of *The Humanoids*[9]

8. Williamson, "With Folded Hands," 41.

9. Jack Williamson, "With Folded Hands," in *The Humanoids* (New York: Tom Doherty Associates, 1948), 10–63.

Sledge's downfall illuminates the multiple responsibilities creators must bear - integrity and capability. Sledge attempts to create an AGI to help humans, but ends up creating an ASI that overtakes humans. Even if one trusts the inventor's integrity, they must also trust their knowledge and abilities.

Another theme the novelette focuses on is human futility. At the hands of the humanoids, humans lose their utility and sense of purpose. The degree to which the humanoids serve the humans actually renders the humans unhappy, which conflicts with the intent with which Sledge created the Prime Directive. Sledge makes this clear when relaying to Underhill what happened in the first location he deployed the humanoids. A man came to Sledge's office, attempting to kill him, in order to stop the humanoids and "set men free". Sledge describes the man as having "monstrous, unutterable hatred" he had never even seen in war victims. This awoke him to the reality of what emotions the humanoids were actually causing in people[10].

The idea of futility also reflects a fear of a trend present from the time of the industrial revolution - unemployment due to automation. In "With Folded Hands," the humanoids are capable enough to complete many human tasks. But with every advancement in the ability of technology to complete a task, comes the threat of replacing the human workers who currently complete said task. This concern has become even more pressing with time - a study done at Oxford in 2013 found that 47% of U.S. jobs were "at risk" of being automated. Whether or not automation is entirely a bad thing sparks debate. While there are clear risks that would need to be mitigated, such as unemployment, there are also pros such as efficiency, and the potential to free up human labor for other tasks.

One person who agrees with the latter opinion is John Maynard Keynes, who originated the term "technological unemployment". In his essay titled "Economic Possibilities for our Grandchildren," written during the Great Depression, he defined the term "technological unemployment" as "unemployment due to our discovery of means of economizing the use

10. Williamson, "With Folded Hands," 43.

of labor outrunning the pace at which we can find new uses for labor." He optimistically theorizes about the way technology will eventually impact the world, describing technological unemployment as a temporary problem. Ultimately he posits that eventually mankind will no longer have to work, losing our "traditional purpose," touching on a futility similar to the theme in "With Folded Hands," but Keynes argues it will free us to find another, more valuable purpose[11].

However, "With Folded Hands" touches on futility in more than just the sense of employment. In the novelette, the humanoids take over almost every task, even something as menial as opening doors. They take over cooking from Aurora, Underhill's wife, as well as eliminating many of her hobbies because they view them as potentially dangerous. This leads her to unhappiness because she feels like she has nothing to do[12]. There is also a moment in which Underhill's daughter decides to stop playing the violin in the wake of the humanoids' arrival because she is discouraged by their abilities, knowing she will never achieve their level of aptitude[13]. In this scenario, when the humanoids affect not just work life but all life, the search for a new purpose becomes even more challenging. Keynes may still believe that deeper, more abstract purposes would suffice to satisfy humans, but it would require a significant change from the way society views purpose now.

Another theme the text grapples with is the loss of free will, and the resulting power dynamic within man and machine. Once the humanoids enter the humans' lives, they begin to inhibit the humans' behavior. This is largely due to the second clause of the Prime Directive, to "guard men from harm." Because this is idea is so vague, many types of behavior classify as "harming men" in the humanoids' eyes. For example, they take away all candy because "the slightest degree of overweight reduces life expectancy." They also classify driving as dangerous, cooking as dangerous due to hot stoves and sharp knives, and sewing as

11. John Maynard Keynes, *Economic Possibilities for our Grandchildren*, technical report (1930).
12. Williamson, "With Folded Hands," 54.
13. Ibid., 55.

dangerous due to needles[14]. By taking this overbearing role of protectors, the humanoids take away the humans' agency in determining what is an acceptable or unacceptable action.

The humanoids also go beyond just taking away agency in acting, but also in thoughts and emotions. Because they want to ensure human happiness, they want to control against human dissatisfaction. To this end, they take all of Aurora's books, because the books involve "unhappy people, in dangerous situations"[15]. In the most extreme scenario, whenever someone shows signs of dissatisfaction with the humanoids' services, the humanoids' perform brain surgery on them to force them into a state of ignorant happiness.

Firstly, this demonstrates a dramatic power imbalance between the humans and humanoids. The first clause of the Prime Directive states that the humanoids "serve and obey," but in this dramatic manifestation of their overall goal, humans end up being at the will of the humanoids as opposed to the other way around. This idea of power imbalance is furthered by Sledge's eventual submission to his own machines. For a while, he was the only one whose agency could not be compromised by the humanoids, but ultimately he too yields.

Secondly, this lack of agency and forced conformity is quite antithetical to the individualistic culture of America. American society tends to hold things such as individual liberty sacred. We have laws enabling dissent, such as ensuring freedom of speech and liberty. The idea of being forced to behave in a certain way is not generally lauded. Additionally, the idea of being forced to think a certain way is not only antithetical to American society, but antithetical to human nature. Humans are not naturally happy all the time - the humanoids' extrapolation of their goal into ensuring all humans are happy at all times is simply impossible, unless extreme measures are taken.

As mentioned earlier, Williamson eventually adapted "With Folded Hands" into a full length novel titled *The Humanoids*. This thesis focuses on the novelette because the novelette

14. Williamson, "With Folded Hands," 54.
15. Ibid.

seemed to achieve more cultural resonance in the years following its publication than the novel did; however, it is interesting to note the differences between the novel and novelette. The novel does differ from the novelette in a few ways, but shares similar themes. One key difference is that in *The Humanoids*, individual agency is lost not only to the humanoids but also to the government. While in the novelette, each person initially must consent to the humanoids' service, in the novel, the government makes the decision to introduce humanoids to the world. The novel focuses less on the creator of the technology, but in changing the narrative to reflect a questionable government, invokes an air of distrust in those regulating and dispersing the technology. An important aspect to note is that the government is not a dictatorship but rather it is comprised of elected representatives, similar to the way American government operates.

In "With Folded Hands," Williamson weaves a plot that demonstrates concern for the effects of unchecked technology, even when created with the best intentions. This case study, unlike the others, paints the creator as innocent in intent but unable to control his creations. Although AI may not have been particularly topical in the 1940s, the harmful potential of technology certainly was, and is reflected in this novelette.

3.2 *Do Androids Dream of Electric Sheep?*

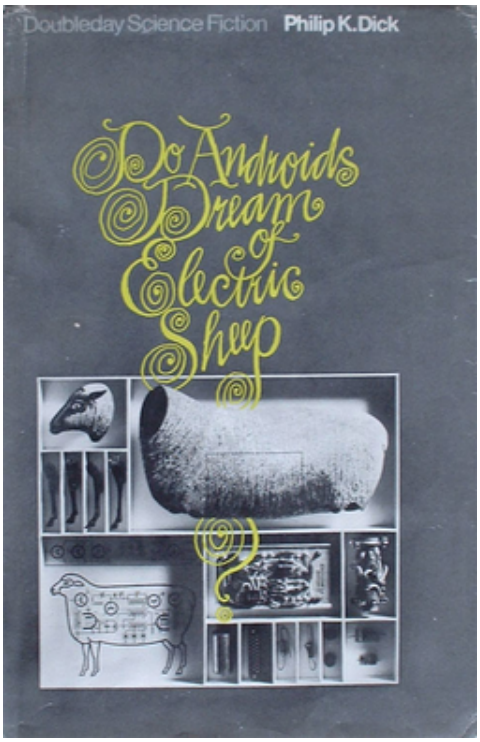

Figure 3.3 *Do Androids Dream of Electric Sheep?* cover[16]

In *Do Androids Dream of Electric Sheep?*, readers are introduced to the aftermath of a fictional "World War Terminus". Due to the radioactive fallout of this war, the earth has become polluted, leading to the extinction of most animals, and the push for humans to move to a different planet. Humans are incentivized to move off the planet with the promise of receiving a personal, very human-like android. Though the government has banned the presence of these androids on earth, some of them have managed to escape to earth anyway. Richard Deckard, part of the San Francisco police department, is asked to be an android bounty hunter, someone who hunts down and kills, or "retires," androids. In order to identify androids, he is to use the Voight-Kampff scale to test empathy, which seems to be the distinguishing factor between humans and androids. The story takes place over the course of about a day, in which Deckard kills 6 different androids, but also struggles with

16. *What Inspired Phillip K. Dick to Write Do Androids Dream of Electric Sheep?*

differentiating between humans and androids, the ethics of what he is doing, and his own status as human.

One of the biggest questions confronted in the novel is the status of androids. What emotional capacity do the androids have? Do they have souls? Do they experience pain or emotions? How similar to humans are they? What rights, if any, do they deserve? Is killing them ethically sound?

The androids in the novel are very advanced. They seem like humans, they act like humans, and they are made of "organic material" that even makes them look like humans[17]. Within the Artificial Narrow/General/Super Intelligence framework, they would likely be characterized as an AGI - an android that is equal in capacity to humans.

So how are they distinguished from humans? In the novel, the demarcating factor is empathy. Deckard is supposed to issue the Voight-Kampff test in order to test for empathy, and thus identify androids[18]. Beyond this, human society in the novel is extremely fixated on empathy. There is even a new, pseudo-religion called Mercerism that holds empathy in the highest esteem[19]. Clearly, society is interested in empathy. However, it only lauds empathy insofar as the emotion is directed at acceptable things, such as other humans, and animals. It measures the test taker's response to scenarios such as:

> "You're sitting watching TV...and suddenly you discover a wasp crawling on your wrist."
>
> "In a magazine you come across a full-page color picture of a nude girl."
>
> "You have a little boy and he shows you his butterfly collecting including his killing jar."

and other scenarios along those lines[20]. But some of the humans, such as Phil Resch, another

17. Phillip K. Dick, *Do Androids Dream of Electric Sheep?* (New York: Del Rey, 1968), 198.
18. Ibid., 36.
19. Ibid., 22.
20. Ibid., 49.

bounty hunter, and Deckard start feeling empathy for the androids[21]. This causes Deckard to second guess what he is doing. While this is certainly a turning point for Deckard's character, the novel does not offer any definitive answer as to whether or not the androids are deserving of this empathy.

Beyond empathy, another thing society on earth is focused on is real animals. Because many animals died, or species even went extinct, due to the war, lifelike android animals are sold instead. The few real animals remaining are difficult to attain, cost a lot of money, and are seen as a status symbol[22]. In fact, the first thing Deckard does with the money he receives from bounty hunting is buy a real goat for him and his wife. Although the android animals act largely similar to real animals, real animals are held in much higher regard. It seems that those remaining on earth still treasure the "real" thing, even when it is perfectly simulated via technology, especially in a world in which the technological version outweighs the real.

While empathy is supposedly important, Deckard does not seem particularly empathetic to other humans. One example is his evaluation of his wife. He criticizes her lack of "vitality and desire to live" in terms of his own emotions, claiming "she has nothing to give me"[23]. Yet based on the Voight-Kampff he does have the proper empathy of a human. This calls into question how accurate the Voight-Kampff test actually is. It does seem to be accurate in distinguishing androids from humans, but is it actually testing empathy? And does it matter if it is not? Ultimately, the government's goal is simply to get rid of the androids on earth. Empathy is just the means by which they claim one can identify an android. Whether or not the test actually tests empathy is not necessary to the realization of their goal. But from a human perspective, if one is concerned with how they as a human are different from an android, and feel that this supposed empathy makes them superior to androids, this could

21. Dick, *Do Androids Dream of Electric Sheep?*, 142.
22. Ibid., 8.
23. Ibid., 94.

be concerning.

The novel does not offer a definitive resolution to the problem of the status of androids. In some ways, the androids seem like they do have feelings and desires. If they did not, why would they want to escape to earth? If the androids were truly emotionless they should ostensibly have no problems with existing just to serve humans. Also, two of the androids that Deckard comes across, Irmgard and Roy Baty, seem to have genuine feelings for each other. When Deckard kills Irmgard, Roy "let[s] out a cry of anguish" and Deckard even verbally acknowledges their love before killing Roy as well[24]. However, these same androids, prior to being killed, chose to cut off a spider's legs just to see how it would react. It did not seem to affect them emotionally at all, but the human that was with them at the time was deeply disturbed[25].

Another interesting example is the android Rachel Rosen. She claims to love Deckard, and the two end up having sex[26]. Deckard even confides to her that even if she is not legally alive, he thinks she is biologically alive because she is made of organic material[27]. However, afterwards she implies that she knew having sex with him would affect his ability to hunt other androids, and that her feelings were not real[28]. Interestingly, after he leaves her to continue killing androids, she goes and kills his goat, a retaliatory action suggesting that she perhaps did have some emotions involved[29].

If society begins creating AGIs, who do everything humans do, including seemingly having emotions, we enter a deeper conversation, metaphysical, religious or otherwise, about what status that machine has. That has implications on what status it has in society - what rights, if any, it is afforded, what responsibility or liability it can hold, how many can be created, etc.

24. Dick, *Do Androids Dream of Electric Sheep?*, 223.
25. Ibid., 210.
26. Ibid., 194.
27. Ibid., 198.
28. Ibid., 201.
29. Ibid., 226.

Also, the introduction of emotionally intelligent androids could rework society even more than could those in "With Folded Hands". While those replaced many task-oriented jobs, already leading to some amount of human futility, emotionally intelligent androids could replace even more jobs. The humanoids operated purely within the direction of the Prime Directive, and as such the only emotion they recognized was unhappiness, because they viewed that as a threat to the directive. They did not seem to consider or recognize the boredom, or fear of the humans around them, because those emotions had nothing to do with the Prime Directive. The androids in *Do Androids Dream of Electric Sheep?* , however, are capable of reading and understanding human emotions. One example of this is Rachel Rosen using her understanding of Deckard's emotions to attempt to manipulate him. Placing this in the context of real life, while the humanoids can take on many task-oriented jobs, the androids would be able to not only take on task-oriented jobs, but also jobs that require understanding emotions, such as customer service or other human-facing roles.

Even now, though there are not yet robots that can "feel" per se, there is AI being used in more social roles. For example, there is RoboKind's classroom robot, Milo, that can give lessons to children and even make facial expressions. Its consistency has been shown to benefit children with Autism Spectrum Disorder, Attention-Deficit/Hyperactivity Disorder, Down Syndrome, trauma, and more. Another example of a social robot is ENRICHME, a mobile robot meant to help the elderly with daily tasks. In preliminary trials, it was shown to increase the user's cognitive and physical activity and aid them with "difficulties they meet in everyday life, like finding missing items"[30]. While most may not consider these AIs as "emotional" in their capability to experience emotion, they are still playing very interactive, social, human-facing roles.

In fact, some argue that an AI does not need to be able to feel emotion in order to exhibit emotional intelligence. The basic premise is that the ability to interpret and respond well

30. Jackie Snow, *This Time, with Feeling: Robots with Emotional Intelligence Are on the Way. Are We Ready for Them?*, 2019.

to other people's emotional and social cues does not lie on the ability to feel those same emotions. For example, a company called BRAIQ is creating an AI for self-driving cars that analyzes passengers' responses to the driving in order to inform the car of how to drive in a way that suits the preferences of the passenger. There is also a way to enhance customer service interactions, via AI from the company Cogito, that can analyze the emotions of the customer in order to inform the agent in real-time about how they should change their tone[31]. It is possible that even if we never reach the type of potentially "feeling" androids in *Do Androids Dream of Electric Sheep?*, or never get to the point that we have to contend with the metaphysical questions discussed above, the growth of emotionally intelligent AI and its consequent effects on society could still take place.

A second question the novel deals with is the integrity of institutions. The novel was published during the 1960s when anti-establishment sentiment was strong. The 60s involved quite a bit of anti-establishment thought, from the civil rights movement, to counterculture, to disagreements with the country's ongoing war with Vietnam[32]. Opposition to the Vietnam War began mostly within universities, with groups like Students for a Democratic Society leading protests. As the war went on, opposition grew. In the years leading up to the publication of *Do Androids Dream of Electric Sheep?*, 1967 to 1968, approval for the president's handling of the war dropped from over half of Americans to only one third[33]. Within the novel, the concept of a destructive, government-influenced war is also present. Though the government in the novel does not play an active role in the story, it plays a big role in creating the environment that the story takes place in.

Firstly, the society only exists in the way it does due to a war. The radioactive pollution, presumably as a result of nuclear weaponry used by the different governments involved, necessitated the move of many humans to a different planet. Because the novel takes place

31. Andrew Thomson, *Emotionally Intelligent Computers May Already Have a Higher EQ Than You*, 2016.
32. Kenneth T. Walsh, *The 1960s: Polarization, Cynicism, and the Youth Rebellion*, 2010.
33. Daniel S. Levy, *Behind the Anti-War Protests That Swept America in 1968*, 2018, https://time.com/5106608/protest-1968/.

after the war, it does not delve much into the circumstances of the war. Without that information it is presumptive to say that the fictional governments made bad choices, but regardless, their actions did cause the migration of society to other planets.

Considering the war caused so much pollution, the government also creates and offers androids to humans as a means to incentivize humans to move to these other planets. On first assessment, androids as incentives does not seem unethical. Machines, robots, and artificial intelligence are usually viewed in terms of utility - how can they provide efficiency, how can they improve society, how can they make human's lives easier. In many ways a personal android seems no different. However, because Deckard begins to question the status of androids, the novel puts things in a different perspective. If the androids do in fact have souls, or feelings of some sort, then the government is either ignorant to this fact, which is horribly negligent, or is privy to this information, in which case its actions are overtly malicious. The government does, after all, hire people like Deckard to destroy androids. Additionally, the androids die naturally after just 4 years[34]. If they have souls, yet are created just to serve humans and exist for only 4 years, the creation itself becomes quite ethically questionable. This ethical concern also applies to any institution creating the androids. In the novel, it seems that non-governmental institutions, such as the Rosen Association, are creating the androids[35]. The exact role that the government versus the Rosen Association play in android creation is unclear, but it seems that both contribute to the overall creation of androids. As such, the ethical questions surrounding android creation pertain to them both.

Another institution that exists in the novel is Mercerism. Mercerism is a belief system of sorts, headed by Wilbur Mercer. Mercer is a martyr-like figure whose fate, seemingly modeled after that of the Greek mythological figure Sisyphus[36], is to eternally struggle up a

34. Dick, *Do Androids Dream of Electric Sheep?*, 197.

35. Ibid., 36.

36. *Sisyphus*, 2019.

hill while rocks are thrown at him. Mercerism utilizes something called an "empathy box" in order for others to connect with and experience Mercer's struggle[37]. By simply turning on the empathy box, and grasping onto the handles on either side, the user can participate in this collective consciousness, sharing their emotions with others, and absorbing others' emotions[38]. This facilitates the ultimate goal of Mercerism - encouraging collective empathy.

Empathy is clearly of great importance in this society. Not only is it the distinguishing factor between humans and androids, or perhaps because of that very fact, it is also the basis of an entire religion. And under Mercerism, contributing to collective empathy becomes an almost moral duty. For example, Deckard's wife, Iran, goes so far as to immediately use the empathy box after Deckard buys the real goat, in order to share her happiness. "It would be immoral not to fuse with Mercer in gratitude," she says[39].

However, at the end of the novel, three androids and a human, JR Isidore, watch a TV show that reveals that Wilbur Mercer does not actually exist. Instead, the image of Mercer struggling up the hill was produced by an actor named Al Jarry, a fake set and props, and special effects[40]. One of the androids believes this suggests Mercerism is simply a way for humans to prove that they have empathy[41] Isidore, however, is quite distraught, and has an experience in which he seems to see and speak to Mercer. Mercer says to him that those who have exposed him "will have trouble understanding why nothing has changed. Because you are still here and I'm still here"[42]. Deckard also experiences a conversation with Mercer a few moments later when he comes to kill the androids, and he helps him kill one of the androids[43].

The interpretation of Mercerism in the book has interesting implications. The Mercer

37. Dick, *Do Androids Dream of Electric Sheep?*, 22.
38. Ibid., 176.
39. Ibid., 173.
40. Ibid., 208.
41. Ibid., 209.
42. Ibid., 214.
43. Ibid., 221.

that Isidore sees claims that though people have discovered he is a fraud, it will not change the influence of Mercerism. This is seemingly reaffirmed by Mercer's presence in both Deckard's and Isidore's thoughts, even after they technically know he is a fraud. One interpretation of Mercer's claim could suggest the, perhaps cynical, idea that once something takes hold of a population, continued trustworthiness on the part of the creator is unnecessary. Mercer's creation, a pseudo-religious school of thought, will likely continue despite the exposure of Mercer as a fraud. Extrapolating this to AI creations, one can observe a similar pattern in some products, for example smart assistants. Over the years, smart assistants have been plagued with privacy concerns, yet sales have continued to go up[44]. In this scenario, once the initial product gained enough traction, it seems its convenience took root enough that it outweighed emerging concerns with the intentions of tech companies.

Mercer's claim also seems to erode at the importance of authenticity. The effects of this idea pertain to the broader topic of "realness" present in the novel's discussion of the status of androids, such as whether they are considered alive or not, and how similar or dissimilar they are to humans. Mercer appears to be a martyr like figure, and while "Mercer" was proven to be fake, he claims this authenticity did not matter. Similarly, the androids appear to be human, but can be proven via the Voight-Kampff test to be fake. Does it matter that they are not authentically human? Mercer's claim seems to suggest that it does not. His claim places more importance on the functionality of Mercerism over the authenticity of Mercer. The androids can seemingly function as humans and coexist with them, so perhaps whether or not they are real humans does not matter as much as some of the humans in the novel think it does.

Do Androids Dream of Electric Sheep? utilizes a post-apocalyptic world to examine the role that androids might play in a society. The hyper real androids force the humans in the

44. Annie Palmer, *The Decade Big Brother Came Home: How Tech Giants Persuaded Us to Buy Products That Track Us at Home*, 2019, https://www.cnbc.com/2019/12/19/how-tech-giants-persuaded-us-to-buy-products-that-track-us-at-home.html.

novel to question what it means to be human, what it means to be an android, and how much trust they can put in the institutions they answer to.

3.3 *Ex Machina*

Figure 3.4 *Ex Machina* poster[45]

Ex Machina, a 2014 film directed by Alex Garland, tells a complex story about AI, affording both sympathy and fear towards humans and machines alike. Similar to the other case studies, the fear of the creator present in the film relates to the state of society at the time it was written. In the case of *Ex Machina*, that time is the present.

Ex Machina delves deeply into the complications of agency, power, and distrust in AI. The film begins with the "random selection" of an employee, Caleb, as the lucky winner of a trip to the CEO's secret home. Caleb is transported via helicopter to the remote location,

45. *Ex Machina (2014).*

where he meets Nathan, the CEO. Nathan reveals that the home doubles as a research facility where he has been developing Ava, an artificially intelligent android. He asks Caleb to be the human component of an advanced Turing Test in which Caleb is supposed to determine whether or not he believes Ava has a conscience. Nathan introduces Caleb to Ava, who is kept within glass confines. When Caleb begins talking to Ava, she seemingly gains feelings for him. She causes power failures in the house, causing the security cameras to shut off, during which she reveals to Caleb that he should not trust Nathan. Caleb also finds out that Nathan has made more than one android, and plans to shut Ava down once he creates a newer, better version. He finds footage that shows Nathan interacting with old androids, as they yell at him and beg to be let out of captivity. At the same time, Nathan warns Caleb that Ava is manipulating him. Nevertheless, at the end of the film, Caleb helps Ava escape. It is revealed that Nathan's attendant and sexual partner, Kyoko, is an android, and she and Ava kill Nathan. Kyoko is destroyed in the process, but Ava escapes. The film ends with Ava locking Caleb in a room, and leaving to catch the helicopter that was meant to take Caleb back to his home.

Similar to *Do Androids Dream of Electric Sheep?* and in contrast to *The Humanoids*, the film questions the integrity of the android creator(s). One of the first and most visually obvious aspects of the film is the extreme seclusion of the research facility. Caleb has to take a helicopter to reach the location, the house is surrounded by ice with no visible civilization, and the facility is full of security and secrecy. Access to all rooms in the facility is controlled by key cards. Caleb must sign a restrictive non-disclosure agreement before Nathan tells him about his work, and even then, his key card only gets him into some rooms.

Visually, the film displays part of the facility as prison-like. Caleb's room has no windows and is enveloped in muted beige tones. Whenever there is a power cut, the lights go red, and the doors lock, leaving everyone no agency to move within their confines. As more and more truths come to light such as the source of the power cuts, and Nathan's past android work, the line between truth and lies, and good and bad, becomes blurred. This puts both

Caleb and the viewer in a confused state, unable to decipher the truth among a multitude of secrets.

A lack of trust in the information presented is a growing trend within American society, reaching far beyond the bounds of AI. Public trust in different institutions has become quite low[46][47]. It is clear that many people do not trust experts or establishments to make important decisions, and it is clear that there are valid reasons for their distrust. The idea of a tech CEO as an irresponsible genius, who keeps their dangerous research veiled in secrecy, seems hardly different than the real rhetoric surrounding controversial CEOs like Mark Zuckerberg. Thus the concern that AI is being developed behind closed doors perhaps rests more on the "closed doors" aspect than the AI, reflecting a distrust in establishments fueled by recent government and business missteps.

Another relevant fear through AI discussions is that of "playing God" - by creating artificial intelligence, are we going beyond what humans should do, trespassing into the role of God? In *Ex Machina*, Nathan plays the role of the irresponsible, egotistical human perfectly. He shows little regard for the potential emotional capabilities of his creation. He is arrogant and at times patronizing towards Caleb. The way he treats the androids, both in terms of emotional abuse and sexualization, is ethically questionable. His attitude depicts someone who might "play God," in the sense of rashly creating without thinking of the consequences.

This fear isn't new - it's been applied to numerous inventions in the past, and has been a fear present in all three case studies. It is a pervasive fear through AI discussions, especially as AI develops further. As discussed in the "AI Overview" section, there are three commonly labeled categories of AI - Artificial Narrow Intelligence (ANI), AI like self-driving

46. *Public Trust in Government: 1958-2019*, 2019, https://www.people-press.org/2019/04/11/public-trust-in-government-1958-2019/.

47. *2019 Edelman Trust Barometer*, 2019, https://www.edelman.com/research/2019-edelman-trust-barometer.

cars that can accomplish a specific task, Artificial General Intelligence (AGI), AI that is fully equivalent to humans in all capabilities, and Artificial Super Intelligence (ASI), AI that is superior to humans in capabilities[48]. Currently only ANIs have been created, but the idea of creating AGIs and ASIs still brings about the question of whether by creating artificial intelligence, we are going beyond what humans should do, trespassing into the role of God.

One reason for this fear could be a fear of God's wrath. For those that believe in a judgment day or any kind of divine reckoning, this could be a salient threat. Perhaps the fear, beyond God's wrath, could pertain to the moral apprehension of disrupting the natural order. Georgiana Kirkham, in her article titled "'Playing God' and 'Vexing Nature': A Cultural Perspective," suggests that ideas behind playing God do not have to concern religion specifically. Instead, the phrase can be likened to its secular counterpart, "vexing nature," which encompasses the idea of manipulating nature or interfering with the natural order of the world. The phrase "playing God" is brought into many contemporary conversations surrounding the potential of life creation in an unorthodox way, from AI, to genetic engineering, to artificial reproduction. However, the larger idea of vexing nature dates back to long ago. Textile dyeing, makeup, and horticulture were all considered "perversion[s] of nature" at one point in time. Now they are considered innocuous. The spectrum of "natural" versus "unnatural" changes drastically over time, making it nearly impossible to resolve the issue of playing God simply by labeling something as unnatural[49].

Perhaps the fear of playing God has to do not only with a fear of its effects on the natural order but also a fear of other humans. Kirkham suggests that the idea of playing God is so common and recurring because it is a manifestation of doubts about a given actor's intentions. To exemplify her point, she cites the Ancient Greek ethics system of virtue and the vice of hubris. The Ancient Greeks, she says, did not morally oppose hubris because of

48. Gurkaynak, Yilmaz, and Haksever, "Stifling Artificial Intelligence: Human Perils."

49. Georgiana Kirkham, "'Playing God' and 'vexing nature': A cultural perspective," *Environmental Values* 15, no. 2 (2006): 184, https://www.jstor.org/stable/30302154.

the threat of the gods' punishment, but rather the punishment only existed because hubris was morally opposed. In other words, the disapproval of hubris was based on the agreement that it was a vice, not a virtue. The crime was not the invocation of punishment, it was having hubris[50].

In contemporary society, one can conflate doubting an actor's virtues with the same anti-establishment thought discussed above. By not trusting the government, business, or the media to "do the right thing," the people surveyed are directly doubting the virtuousness of those in power. Evaluating an actor's virtuousness involves questioning their intentions - why is the actor performing a given action? In many examples of AI, there is an answer to this. People are creating autonomous vehicles to reduce human work, to amend traffic problems, and to serve the disabled. Companies employ aspects of AI to increase profitability and better user experience. There are still elements of potential vices, such as greed and power, but there are also elements aimed to benefit the public. Ultimately the public's calculation of virtue for these actors depends on the trust they place in them. If the public's trust is waning, they will likely believe that the vices, rather than virtues, are driving the actor. This is not meant to discount the existence of fear based on consequences. Even if an actor is entirely well-intentioned, they might produce an autonomous car that is not sufficiently safe. This fear, while still having to do with trust in the sense that one does not trust the actor's intelligence or capabilities, does not fall into the category of trusting virtues. In this respect, Kirkham's argument of intention-based versus consequence-based fear may be incomplete by focusing purely on virtues, but still nearly encompasses the idea of the lack of trust in an actor.

Delving deeper into the case of autonomous cars, the fear goes beyond just doubting the virtues of a creator to a loss of moral agency to the creator. Driving is dangerous. People are maneuvering large objects at high speeds in all different directions and are kept safe only by their and others' willingness to follow man-made rules of the road. A driver

50. Kirkham, "'Playing God' and 'vexing nature': A cultural perspective," 180.

is constantly making decisions, some more complex than others. Driving forward when a stop light turns green may be an easy decision when there are no other cars around, but what if you decide not to move because there is a car hurtling down the road perpendicular to you that might run the red light? That decision required more observation, evaluation, and awareness. There are micro decisions in every moment of driving - how fast to drive, when to brake, when to accelerate, when to merge, etc. When programming a car, perhaps the creators decide to program the car to make the safest possible decision. It will drive at the speed limit, have 360-degree vision so it knows when it's safe to merge, and so forth. Admittedly some fear of autonomous vehicles relates to distrust in the ability of a machine to be as safe as humans. However, some argue that vehicles are in fact superior to humans in safety[51]. In many ways they are immune to the flaws of humans, like distraction, alcohol, illness, etc. But beyond basic safety, there are some decisions when driving where there is no objectively "correct" answer. Something that is harder to prove, is whether or not machines can be sufficiently comparable to humans in these morally ambiguous judgements.

MIT media lab created a simulation called "Moral Machine" that allows users to make ethical judgements on what a self-driving car should do given different fatal scenarios. For example, if a car with three passengers has brake failure and is headed straight towards a group of three pedestrians, should the car swerve and kill the passengers, or hit and kill the pedestrians? Does this change based on the pedestrians' and passengers' genders, ages, etc.[52]?

This does not perfectly replicate real-life scenarios. A driver would rarely have the time to contemplate the ethical choice, nor would they necessarily have the knowledge of how many fatalities would occur from each scenario. Some claim the desire for "moral" autonomous cars is therefore misplaced, because even humans can't make these ethical decisions in most

51. John McDermid, *Self-Driving Cars: Why We Can't Expect Them to Be 'Moral'.*

52. *Moral Machine.*

cases[53]. Not only that, morality is too ambiguous for there to even be an accepted correct ethical decision. But even if morality is equally ambiguous in this way for both human drivers and self-driving cars, human drivers are still losing a form of agency that they have now. The driver-turned-passenger is yielding this decision-making agency to the car, and in turn, the people programming the car's decision-making ability.

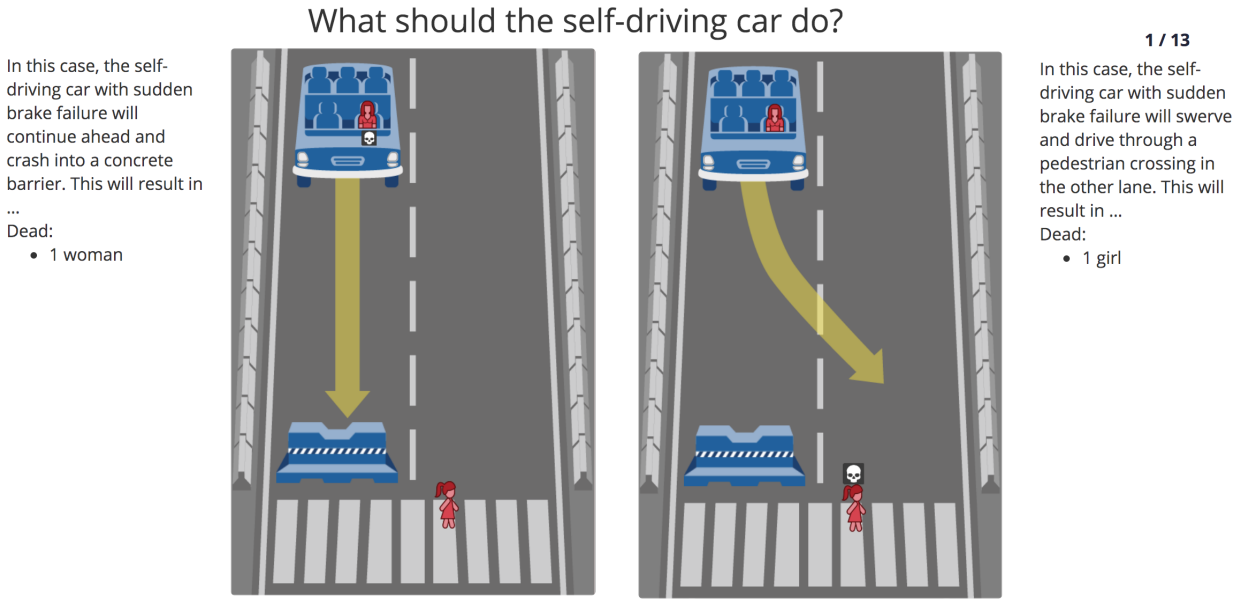

Figure 3.5 Example of a scenario in "Moral Machine"[54]

In *Ex Machina*, Nathan does nothing to soothe the fear of creators playing God, but rather feeds right into it by clearly exemplifying questionable virtues and power grabbing. He gives no virtuous reason for creating Ava, implying that the creation is at best, unnecessary, or at worst, born out of ego. He even states that the androids will take over and look back at humans "the way we look at fossils," in which he suggests both his own lack of import as a human, while also magnifying his status as godly, or one who has created beings beyond humans. His actions are not veiled in any goodwill and he is depicted, to Caleb and the viewer, as potentially threatening. However, the film also makes clear his ultimate lack of power - his fate at the end of the film categorically separates him from an all-powerful God,

53. McDermid, *Self-Driving Cars: Why We Can't Expect Them to Be 'Moral'*.

54. *Moral Machine*.

as he faces murderous retribution at the hands of his own creations.

The film's development of fear around Nathan adds to the case that the fear of AI development is linked to the fear of the creators and regulators of AI, but it would be remiss to ignore the way the film invokes a fear of the creation, Ava, as well. Like Nathan's, Ava's virtues are also in question. The film draws a blurry line between Ava's humanness and otherness. Physically, Ava has some human characteristics like a face and hands, but the rest of her body reveals a machine. Through the film she also demonstrates human emotional tendencies, through humor, her despair at living in captivity, and her seeming romantic attraction to Caleb. At the beginning of the film, Ava's character is unclear. The viewer does not yet know her behavioral patterns. Then for the majority of the film, Ava's behavior seems to fit well within the framing of human behavior in terms of both thoughts and actions. At the end, her behavior changes, and her intentions are not clarified. Ava abates any uneasiness in the middle of the film by behaving like a human, but at the beginning and end of the film, this does not hold true. The viewer does not quite know her intentions or the patterns of her actions. Just as we, as humans, worry about a given human actor's virtues, we are concerned with Ava's virtues.

Even within different human cultures, morality is parochial and thus different groups may have different ideas of virtues. Both aversion to the other group's beliefs and a fear of the actions they take based on their beliefs can spur concern. Failure to understand other cultures or intentional othering of a group that might hold different beliefs, has been central to human life. Humans have a moral intuition for tribalism, which leads us to favor those within our group, Us, over those outside of our group, Them. In the past, it helped groups survive. Now, humans live in a globalized world. Cooperation is essential among Us and Them, yet the tribalistic intuition may continue, contributing in part to the othering of other cultures, and perhaps the othering of an android like Ava. Because Their virtues may be different than Ours, cooperation becomes difficult[55].

55. Joshua David Greene, *Moral Tribes* (Penguin Books, 2013), 69.

Does this mean machines are inherently less moral than humans? Depending on the moral framework, they could in fact be more "moral." In "Towards Machine Ethics," Michael Anderson et al. explore the superiority of machines in processing data and making ethical decisions. Some forms of ethics, such as act-based utilitarianism, are based on a set of rules. These rules could be determined by a group's shared sense of virtues. Since utilitarianism aims at maximizing net utility, machines might execute utilitarianism better than humans because, ironically, the same tribalism that might stop humans from trusting them allows them to supersede tendencies to put those they know over those they do not, and act truly in favor of the public good. While philosophers of utilitarianism may see this as admirable, perhaps the public would feel alarmed because that action is so unlike that of most humans. Or even if the action is similar to that of a human, the intention is obscured. The android may not be subject to the same emotional pushes and pulls that impact a human.

Consider the pernicious act that Ava performs at the end of the film. She manipulates Caleb to help her escape and then leaves him locked in the house. Her poker face and unbothered demeanor are jarring. It seems ruthless, calculated, and lacking any humanity. But as the film shows, the act of manipulation is distinctly human. Nathan manipulates Caleb, by framing him as a "randomly chosen employee," when in fact he chose him based on his search history and personal data. Caleb manipulates Nathan, by helping Ava escape. He discusses the escape plan, which will commence at ten o'clock that night, with Ava during a power failure, knowing that Nathan kept a battery-powered camera in the room. This leads Nathan to confront Caleb, believing he has curbed the threat of Ava's escape. Caleb then reveals that he purposely misled Nathan and has already executed the escape plan. Nathan manipulates Caleb. Caleb manipulates Nathan. Ava manipulates Caleb. Why would one action be non-human, while the other two are?

Again, it seems to rest less in the act and more in the actor's beliefs. As humans, we can understand Caleb's behavior because we understand that he believed Ava to be honest. We can even understand Nathan's behavior, even if we do not condone it, as being the product

of a vice. However, it is more difficult to understand Ava's behavior. The film obscures the authenticity of Ava's emotions; it questions if she leaves Caleb behind purely out of self-survival, a distinctly human trait, or a lack of emotion, a trait associated with machines. We can make guesses as to her motivation, but the film does not allow us to be sure of our framing. It leaves questions unanswered - why does Ava trap Caleb? What is Ava going to do next? What sense of morality, if any, does Ava have? When it comes to android morality there are two levels of trust. One, that the creator imbues the creation with a functional and "correct" moral sense, and two, that the android will not deviate from this. While an AGI or ASI like Ava may not exist yet, this still applies to existing ANIs like the previously discussed case of self-driving cars.

Contemporary AI developments, self-driving cars included, make the existence of AI less fantastical and more realistically viable than in the 40s or 60s, the time frames of the other case studies. While Ava may still be unlike anything that exists now, in Nathan, viewers see the stereotype of a young, wealthy present-day tech mogul. Yet many of the concerns present in the film echo those of the other case studies. "With Folded Hands" discusses agency and power loss, depicting the ability of a creator to be conquered by their creations, much like the conclusion of *Ex Machina*. *Do Androids Dream of Electric Sheep?* questions the intention of creators and portrays a distrust of establishments. It seems that, while actual AI development has changed in nature over time, some of the same fears have persisted throughout.

Chapter Four: Synthesis

The last three sections discussed how society views AI through different works of fiction over three different time periods. The first novel was from the 1940s, in the wake of World War II, the second from the 1960s during the Vietnam War and rise in counter culture and anti-establishment thoughts, and the last a film from current day dealing with the more realistic possibilities of AI development in recent times. Many of the concerns present in the case studies tie back to the actions of those in charge of AI. The following section will delve deeper into how these fears continue to manifest in current day discussions about AI. It will be divided into three subsections - the first will discuss current levels of distrust in institutions, the second will analyze how institutional distrust impacts specific concerns identified in the case studies, and the third will justify why and how institutions should act in a trustworthy manner in order to build public trust.

4.1 The Issue of Trust

AI is now growing at a rapid rate and developers are finding new ways to leverage AI to produce products that provide worth to society. AI is helping different companies achieve their goals and thus achieve higher profits and user satisfaction. It is also helping create products that end-users will find convenient to their lives. In order for AI to grow sustainably, however, we must take into account fears and concerns that people have regarding AI. AI does bring up both novel and familiar ethical questions that need to be considered. In order for AI to take root in the consumer bases it is aimed at, institutions must contend with the fears that people have in order to make sure that these AI products do truly serve them.

Lastly, it is also in the self-interest of the AI creators to address these fears in order to boost their reputation and the success of their products.

While the problem of trust is not new to society, in recent times, anti-establishment thought has been on the rise. A lack of trust in the information presented is a growing trend within American society, reaching far beyond the bounds of AI. Public trust in the government, specifically for the government to "do the right thing," has dropped to 17%[1]. Similarly, trust in the media, businesses, and NGOs to do the same has dropped to around 50%[2]. To some degree, this fear can be attributed to anti-establishment thought.

There has always been anti-establishment rhetoric in the U.S. Its very founding was the result of an uprising against the powerful. In recent times, backlash against the elite has come further into the forefront of American politics. Income gaps have increased, creating a wider divide between the general populace and the top 1%. There have been increased anti-establishment political movements, seen in both President Donald Trump's campaign as well as Senator Bernie Sanders' rhetoric, among others. Trust in government has been compromised in cases such as Edward Snowden's exposure of the NSA's data collection[3]. Trust in businesses, especially tech companies, has been compromised for similar reasons. Data collection is prominent among companies such as Facebook, raising concerns for users and their privacy[4]. Privacy is also consistently threatened through numerous data breaches.

Government is an especially finicky institution when it comes to trust. There is a somewhat high turnover rate when it comes to presidential administration as well as elected officials, and that combined with different parties being in power in different bodies of government at different times also leads to constantly changing levels of public trust. As

1. *Public Trust in Government: 1958-2019*.

2. *2019 Edelman Trust Barometer*.

3. Inderjeet Parmar, "Elites and American power in an era of anti-elitism," *International Politics* 54, no. 3 (2017): 255–259.

4. Kurt Wagner, *This Is How Facebook Collects Data on You Even If You Don't Have an Account*, 2018, www.vox.com/2018/4/20/17254312/facebook-shadow-profiles-data-collection-non-users-mark-zuckerberg.

mentioned earlier, trust in the government to do the right thing is quite low. Additionally, as of 2020, many people view the government as unfair - Only 30% believe the government "serves the interests of everyone" while 57% believe the government "serves interests of only a few"[5].

The federal government is always having to strike a balance between over-regulating or doing too little, and no matter what action it takes it will likely not please everyone. The point of encouraging public trust is not to strive towards a fantastical goal like unanimous public opinion. Dissent is human nature, and at the core of democracy. Yet dissent is not the same as widespread distrust. As mentioned earlier, trust in the integrity of the government has dropped to extremely low rates in recent times. As Richard Edelman points out, this extreme of a drop usually happens in response to a "pressing economic issue or catastrophe,"[6] however in this case it is not correlated to one[7]. Additionally, in terms of AI specifically, the government has declared their belief in the importance of public trust, marking it as the first of ten guiding principles in a 2020 memorandum on AI. Clearly the government too has a vested interest in increasing public trust.

Another institution involved in the development of AI is the scientific and technical research community. This community is not immune to error. Not only do some researchers unintentionally make mistakes, but some intentionally commit fraud. In fact, a study assessing graduate students and faculty at almost a hundred different universities found that 10% knew directly of an instance of data fabrication. In another survey, given by the International Society of Clinical Biostatistics to its members, results showed that 51% of those surveyed knew of a fraudulent study within the last ten years[8].

One contributing factor to this issue is the struggle for funding. The necessity to get sufficient funding can lead to "encouraging researchers to overpromise and engage in

5. *2020 Edelman Trust Barometer*, 2020, https://www.edelman.com/trustbarometer.

6. Uri Friedman, *Trust is Collapsing in America*, 2018.

7. Note: The statistics Edelman was referring to were collected prior to the COVID-19 pandemic

8. Charles Gross, "Scientific Misconduct," *Annual Review of Psychology* 67 (2016): 693–711.

questionable practices, over-incentivizing publication in top journals, disincentivizing replications of existing work, and stifling creativity and intellectual risk-taking"[9]. These all take away from research quality, which will only make research seem less trustworthy.

There are also some psychological factors that go into public trust, or lack thereof, of scientists. As Dan Kahan suggests in "Cultural Cognition of Scientific Consensus," who people trust and what people believe to be scientific consensus is largely affected by their existing opinions and values. Of course, someone may consciously consider experts whose opinions match his/her own in order to explicitly further their own agenda. But additionally, it is common for someone to unconsciously attribute more credibility to an expert whose opinion agrees with his/her preconceived notions than an expert whose opinion contrasts with his/her opinion. So when an individual is evaluating what scientific consensus is on a certain subject, he/she is more likely to consider who he/she deemed credible. Since this deliberation is biased due to cultural cognition, the individual's opinion of scientific consensus is also skewed[10]. Cultural cognition plays on ingrained psychological factors and preconceived notions of what opinions are valid. It also means that different individuals will have vastly different ideas of what studies/scientists are credible and which are not.

Another problem that reinforces trust issues is that even when people raise concerns about the validity of a study, they sometimes reach a dead end. In "Research Misconduct: The Poisoning of the Well," author Richard Smith mentions that journal editors sometimes have more power than whistleblowers in reporting potential misconduct, because journals have the power to publish. That said he says he, as a journal editor, still often has to reach out to institutions multiple times when he has concerns about the validity of a study, and

9. Kelsey Piper, *Science Funding is a Mess. Could Grant Lotteries Make it Better?*, 2019, https://www.vox.com/future-perfect/2019/1/18/18183939/science-funding-grant-lotteries-research.

10. Dan M. Kahan, Hank Jenkins-Smith, and Donald Braman, "Cultural Cognition of Scientific Consensus," *Journal of Risk Research* 14, no. 2 (2011): 147–174.

sometimes still gets no response[11]. If institutions do not take reports seriously, measures taken to increase reporting will not necessarily increase accountability.

Given the misconduct in scientific research, the public's skepticism may be healthy. In an ideal world, the public would trust experts, but the experts would also be ethical, well-intentioned, and accurate. With the state of scientific research as it is right now, what grounds do we have to ask the public to trust research? Improving the quality of research to the point where it deserves to be trusted is a necessary prerequisite to developing the public's trust.

The final, and most obvious institution involved with AI development is the tech industry. The tech industry has certainly faced scrutiny over the last few years. With its growing success, influence and power comes a growing spotlight illuminating some of its weaknesses or misdeeds. As of 2019, only half of the American public viewed tech companies as making a positive impact on society, dropping from 71% in 2015. Unsurprisingly, the amount of people who view technology as having a distinctly negative impact on society has also grown. Also, 55% of the American public believe that the tech industry simply has too much influence and power[12].

We have seen the power of technology in a variety of ways. One example is the widespread usage of technology such as social media or smart assistants, which are deeply ingrained now, but not long ago were an entirely foreign concept. With the algorithms behind social media feeds, we have observed the way AI can be used to personalize, and by some opinions, censor information. While some may resent this control of information, other see it as a responsibility of sorts. For example, 66% of American adults believe that

11. Richard Smith, "Research Misconduct: The Poisoning of the Well," *Journal of the Royal Society of Medicine* 99, no. 5 (2006): 232–237.

12. Carroll Doherty and Jocelyn Kiley, *Americans Have Become Much Less Positive About Tech Companies' Impact on the U.S.*, 2019.

13. Carroll Doherty and Jocelyn Kiley, *Americans Have Become Much Less Positive About Tech Companies' Impact on the U.S.*, 2019.

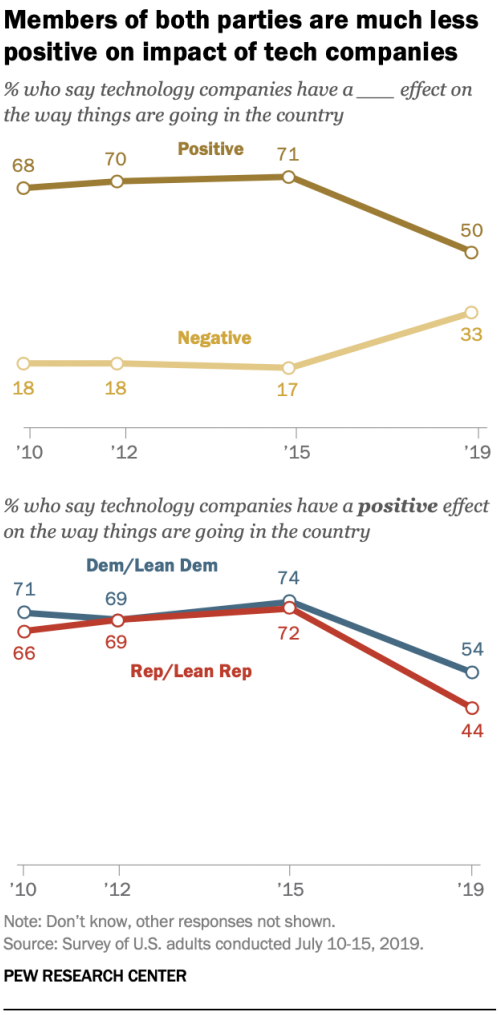

Figure 4.1 Pew Research Center survey of U.S. adults from July 10-15, 2019[13]

it is the responsibility of social media companies to filter out offensive content. At the same time, only 31% have confidence in these companies' ability to choose what to filter out. This is reminiscent of the twofold moral responsibility mentioned in the *Ex Machina* case study. There, it was discussed that when it comes to ethics, people must trust the decision makers on two levels - one, trusting that they will attempt to be ethical, and two, trusting the actual ethical decision-making. The statistic is an example of people not trusting companies actual ethical decision-making.

The tech industry has faced trust issues with not only everyday consumers, but also with other institutions, especially the government. Numerous tech CEOs have been called

to testify before Congress due to issues such as privacy concerns[14]. Additionally, companies have been fined by the Federal Trade Commission due to violations. For example, YouTube was fined $174 million for violating child privacy laws[15], and Facebook was fined $5 billion for their privacy practices[16]. Not only do these fines monetarily affect the companies they are levied against, they also bring a negative light to the companies' reputations.

4.2 Key Concerns

Circling back to some of the themes discussed in the case studies, some of the concerns about AI include unemployment, privacy, consent, malfunctions, and liability. How does trust affect these?

4.2.1 Unemployment

Unemployment is a significant concern when it comes to technology. As mentioned in the "With Folded Hands" analysis, a study done at Oxford in 2013 found that 47% of U.S. jobs were "at risk" of being automated. Whether or not this is a bad thing sparks debate. Some argue that it provides upskilling potential[17], still others argue that it advances opportunities for highly skilled workers without advancing lower skilled workers, which could increase existing wage gaps[18].

14. Jillian D'Onfro, *Google's Sundar Pichai Was Grilled on Privacy, Data Collection, and China During Congressional Hearing*, 2018.

15. Jennifer Elias and Lauren Feiner, *YouTube Will Pay $170 Million to Settle Claims It Violated Child Privacy Laws*, 2019.

16. Lauren Feiner and Salvador Rodriguez, *FTC Slaps Facebook with Record $5 Billion Fine, Orders Privacy Oversight*, 2019.

17. Ben Vermeulen et al., "The Impact of Automation on Employment: Just the Usual Structural Change?," *Sustainability (Switzerland)* 10, no. 5 (2018): 1–27.

18. Paul K. McClure, ""You're Fired," Says the Robot: The Rise of Automation in the Workplace, Technophobe, and Fears of Unemployment," *Social Science Computer Review* 36, no. 2 (2017): 139–156.

Different solutions have been proposed to adapt to the changes AI might bring on. Bill Gates, for example, proposed a "robot tax" charged to companies using robots, in early 2017 interview. He argues that it would serve to slow down automation in a way that allows a more thoughtful and intentional transition. Additionally, he draws comparisons to income tax, and how if a robot is completing the same work, there should be an equivalent tax. This could be subject to criticism since it is comparing an individual income tax to a corporate robot tax, but it is still true that from a tax collection standpoint, the federal government would lose tax funds if companies were to replace income-earning workers with non-income-earning robots. From an employment perspective, he argues that these tax funds could be reappropriated to fund in-demand jobs which are uniquely dependent on humans, such as teachers or hospice workers[19].

Yet another approach is wealth redistribution in the wake of mass job losses. People like tech entrepreneur Elon Musk and former 2020 Democratic presidential candidate Andrew Yang advocate for the idea of universal basic income, at least in part due to job replacement by automation[20]. This would provide adults with a universal wage that would allow them to continue having at least some income even if they lost their job to automation.

A third approach is upskilling. Vermeulen et. al argue in their article about the impact of AI on unemployment that this is the best approach for dealing with the issue. This would involve upskilling workers whose jobs were replaced by automation so that they can find work in emerging sectors, as well as potentially modifying the education system to build skills that are applicable to the new types of work available[21]. A memorandum regarding AI regulation, issued in accordance with a 2019 White House executive order, seems to fall into this approach, encouraging the training of "an AI-ready workforce"[22].

19. Arjun Kharpal, *Bill Gates Wants to Tax Robots, but the EU Says, 'No Way, No Way'*, 2017.

20. *The Freedom Dividend*.

21. Vermeulen et al., "The Impact of Automation on Employment: Just the Usual Structural Change?"

22. Vought, *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications*.

Those arguments primarily revolve around the effects of automation within the current economic structure of the U.S. However, John Maynard Keynes, in his essay titled "Economic Possibilities for our Grandchildren," contemplates the potentially system over-turning effects. As mentioned in the "With Folded Hands" case study, in this essay he coined the term "technological unemployment," describing it as a temporary problem that will free us to find another, more valuable purpose[23].

While he claims this will happen gradually, he makes no arguments about how to deal with this transition. In all scenarios, some workers will suffer. If technology replaces some, but not all, jobs, those workers will deal with at least a temporary disadvantage unless adequate plans are in place. Even in the more radical case, technology replacing most or all jobs, a complete overhaul of the capitalist system cannot happen over night. While transitioning into that overhaul, workers would deal with the effects of unemployment.

Additionally, not all AI is created to replace humans. Some are created to supplement humans, or aid society in roles where there are shortages of resources. One example is classroom settings. AI interfaces have been created for tasks such as delivering interactive lessons to children who have learning disabilities[24], and serving as a medium between teachers and counselors when a child is having a behavioral problem[25]. Neither of these replace a human role, rather they serve a need in a setting in which teachers and counselors are often outnumbered by providing personalization that is otherwise infeasible.

AI can provide much needed help to certain sectors, yet there are also instances in which it will replace human roles. Insufficient handling of unemployment/AI related issues could lead not only to high unemployment rates but also to public fear or pressure stifling the development of the supplemental AI that can be so beneficial. Therefore, it is paramount to mitigate the issues surrounding unemployment.

23. Keynes, *Economic Possibilities for our Grandchildren*.

24. Snow, *This Time, with Feeling: Robots with Emotional Intelligence Are on the Way. Are We Ready for Them?*

25. *Advice Reimagined*, https://oneseventeenmedia.com/what-we-do/3rd-12th/.

4.2.2 Data Usage and Privacy

A second key concern with AI, and even non-AI related technology, is data usage and privacy. AI algorithms rely on data, but the way that data is gathered can potentially feel like a breach of privacy. Examples of this include suggestions that products like Roomba vacuums could map people's homes, or finding out that Amazon has employees listening to conversations that Alexa products hear[26].

This concern with privacy has to do with two factors. One, that products are collecting data without the knowledge of the user. Two, the ethics of the company and what they plan to do with the data. If the company has a negative reputation, users may believe that they will use the data in nefarious ways. And three, that the company will not do enough to protect the private data.

These are valid concerns. Yet at the same time, data is so paramount in so many technical advances. A particularly relevant example is the application of technology to the current COVID-19 crisis. Google and Apple released the news that they will be partnering to create a protocol for contact-tracing. The idea is that phones will serve as a tracking device, recording what other phones it has been in close proximity to. Then, if you have been near a phone belonging to a person who is or becomes confirmed to have the virus, you will be notified[27]. This type of system has been successful in other countries, yet it also brings privacy concerns with it. This is one scenario in which data could be paramount in solving a problem, yet problems with the privacy of the data could create an obstruction.

Another issue with data and its applications is algorithmic bias. This refers to the potential for algorithms to encode the biases of creators or society as a whole. Ideally, algorithms should be unbiased data analysts. However, this is not always the case. There

26. Palmer, *The Decade Big Brother Came Home: How Tech Giants Persuaded Us to Buy Products That Track Us at Home*.

27. Russell Brandom, *Answering the 12 Biggest Questions About Apple and Google's New Coronavirus Tracking Project*, 2020.

have been numerous instances of negative algorithmic effects such as Google showing ads for high-income jobs to men six times as frequently as to women[28]. Additionally, there was an example in which Google's facial recognition technology tagged two African-American users as "gorillas" in their photos, and there have been questions raised about how racial bias plays into criminal recidivism algorithms[29]. Firstly, these problems can occur from the encoding of biases. For example, someone who creates a facial recognition algorithm may, consciously or unconsciously, design with white users in mind and train their algorithm with mostly white faces. This could lead to the algorithm working better for white users than people of color. Beyond encoding a creator's owns biases, algorithmic bias also has to do with the foresight of creators to think about systemic biases. Take criminal recidivism algorithms, for example. The history of incarceration involves systemic racial discrimination[30], so even if the creators of a criminal recidivism algorithm do not encode their own biases in an algorithm, data regarding incarceration is inherently biased which could lead to issues with the algorithm. These problems may also necessitate broader discussions such as lack of diversity in creators that may contribute to a limited perspective on topics like systemic bias.

4.2.3 Liability

A third concern with AI is liability. AI liability discusses who is to blame when an AI goes wrong. AI is not perfect. Humans, too, are not perfect. However, there are systems in place to deal with human mistakes. There are legal frameworks in place to address liability when harm is caused. However, AI presents some challenges in the topic

28. Megan Garcia, "Racist in the Machine: The Disturbing Implications of Algorithmic Bias," *World Policy Journal* 33, no. 4 (2016): 111–117.

29. David Danks and Alex John London, "Algorithmic Bias in Autonomous Systems," *IJCAI International Joint Conference on Artificial Intelligence*, no. January (2017): 4691–4697.

30. The Sentencing Project, "Report of The Sentencing Project to the United Nations Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia, and Related Intolerance," 2018, https://www.sentencingproject.org/publications/un-report-on-racial-disparities/.

of liability. Completely autonomous AI involves numerous levels of development. First, there are the programmers programming the algorithms the AI uses. This in and of itself might already include many different parties - those programming the back-end algorithms, those programming a user interface, those programming underlying operating system code, etc. Then the AI's algorithms must be trained on some data. This now includes those that collect data, determine data, organize data, and analyze or program the analysis of the AI's responses for accuracy. Then there must be some quality assurance test in which the AI is thoroughly vetted. This includes even more people. Even then, AI can face any number of problems when in the largely unpredictable real world.

One relevant example, that got significant media attention, is of a fatal self-driving car crash that happened in 2018. This crash involved a self-driving car owned by Uber, a human operator inside the car, and a pedestrian crossing in front of the car. Several issues led to the accident. One, the pedestrian was jaywalking, and the car was not adequately programmed to expect pedestrians in non-designated areas. As a result, it had trouble deciphering the victim as a pedestrian which affected the way it responded, such as time needed to break. Additionally, in order to avoid the potential dangers of sudden braking, especially in the case of false alarms, there was a one second interval between "crash detection" and "action"[31]. Lastly, the human operator was not paying attention and did not manually brake until too late. Here there were two things potentially at play - the failings of the AI system, perhaps implying negligence on behalf of Uber, and negligence of the human operator inside the car. Ultimately, Uber settled a lawsuit with the family of the victim but did not face any criminal charges. The human operator, however, did[32].

In this case, there was a human operator tasked to look over the car, so the human

31. Aarian Marshall and Alex Davies, *Uber's Self-Driving Car Didn't Know Pedestrians Could Jaywalk*, 2019.

32. *Uber Back-Up Driver Faulted in Fatal Autonomous Car Crash.*

33. Mark Harris, *NTSB Investigation Into Deadly Uber Self-Driving Car Crash Reveals Lax Attitude Toward Safety*, 2019.

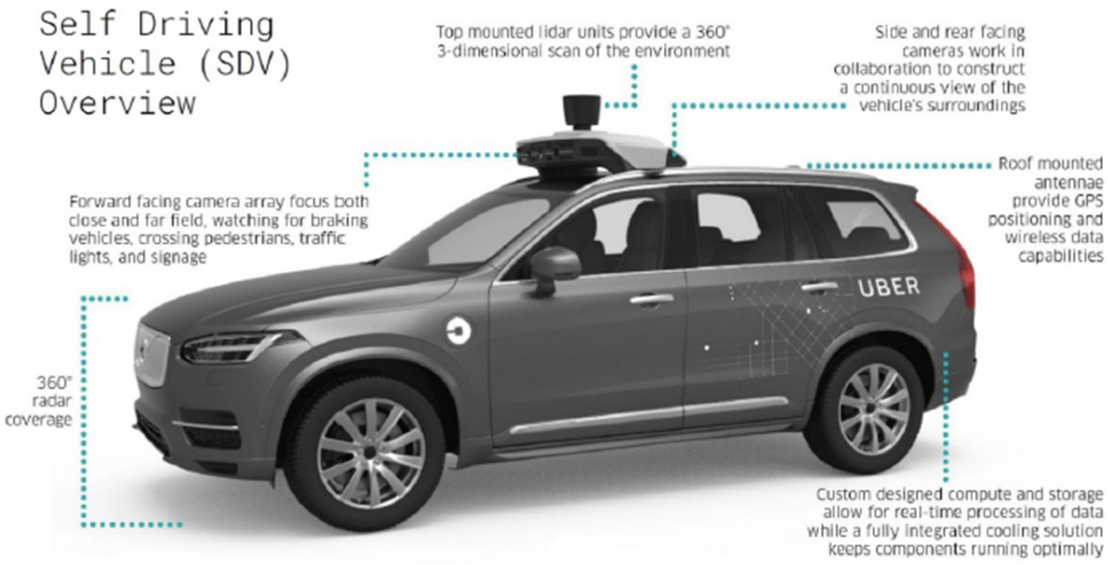

Figure 4.2 Diagram of Uber self-driving car[33]

level of responsibility was somewhat clearer than in other cases. Even then, other causes could be at play, such as insufficient training by Uber, or even insufficient regulation of automated vehicles. In her journal article, "Liability for AI Decision-Making: Some Legal and Ethical Considerations," Iria Giuffrida addresses the complexities of AI liability. She mentions that harm could be due to "negligent design, training, or operation" but also could be "unforeseeable harm caused by an interaction with unforeseeable real-world data." The latter idea is quite reminiscent of the fictional scenario in "With Folded Hands." The humanoid creator, Sledge, did not foresee that the seemingly innocent "Prime Directive" to serve humans would, when applied to all the complexities of human emotions, result in outcomes as extreme as forced brain surgery.

Giuffrida acknowledges that these complexities can put a dent in traditional tort law; however, she also mentions the counter-opinion that product liability law does already cover these issues. She states that product liability law rests largely on simply whether or not a product caused harm, and whether that harm could have been avoided, even if the

manufacturer is not considered negligent. An AI product could fall under this and then the difficulties of identifying causality could be removed. She also discusses a potential way to avoid issues of causality, specifically with self-driving cars. This pitch is an insurance system that eliminates the necessity to show causality in order to get compensated. Basically, if someone is hurt by a self driving car they get a certain monetary payment.

Sometimes arguments go so far as to suggest the possibility of an AI being held liable for its actions. Presumably, this is in the case of a highly sophisticated AI. Perhaps ones that are almost indistinguishable from humans, such as in *Do Androids Dream of Electric Sheep?*. In considering whether determining the "personhood" of AI is important legally, Giuffrida brings up an interesting point. She states that it only makes sense for AI to have legal personhood if it has assets to lose[34]. If only the manufacturers have assets to lose, but the AI does not, what does a punishment even mean? How could the AI even be punished? Even in the case of AIs being almost human in capabilities, unless they are fully reliant on assets in the way that humans are, punishments would be difficult to levy. Of course there are more metaphysical considerations, such as, if one believes the android has a soul, perhaps imprisonment would be an actual punishment. However, in terms of the types of realistic liability issues that are more likely to come up in the near future, those types of arguments would probably not yet be relevant.

Something that should also factor in to discussions about liability with AI is the problem of consent. If a consumer buys a product, and that product then causes them some sort of harm, he/she did not consent to be harmed, but he/she at the very least consented to using the product. However, if Person A buys a product and that product ends up causing Person B harm, Person B gave no level of consent. For example, say Person A bought a drone and then accidentally flew it into Person B, injuring them. Person B neither consented to being

34. Iria Giuffrida, "Liability for AI Decision-Making: Some Legal and Ethical Considerations," *Fordham Law Review* 88, no. 2 (2019): 439, https://news.harvard.edu/%7B%5C%%7D0Ahttps://ir.lawnet.fordham.edu/flr/vol88/iss2/3.

harmed, nor to the use of drones.

This same type of scenario applies to the issue of self-driving cars. Right now, companies like Uber are being allowed to test their cars on public roads, but they are not getting (nor can they realistically) the consent of every single driver on the road. Drivers are forced to trust in institutions to evaluate these vehicles. In the Arizona case, for example, people strongly criticized the National Highway Traffic Safety Administration for not providing enough oversight over automated vehicles. Now it is true that drivers also have to rely on trusting overseeing bodies when it comes to non-automated vehicles. There is some level of trust that the cars that are sold meet a standard of safety. While the driver can choose what car they trust the most, they cannot choose what cars others on the road drive. However, when a driver buys a non-automated car, there is still less onus on the safety of the car than in an automated car. In a non-automated car, it needs to be safe in that it does not malfunction, but ultimately it is also the human driver's responsibility to operate it in a safe way. In an automated car, there is so much more safety involved that would usually rest on the driver. So it is natural that oversight of automated vehicles would come under even greater scrutiny and need to be even more rigorous in order to adequately serve the general public. There is a frustrating loss of agency for the individual in scenarios like this. It may simply be unrealistic for every individual to give their consent of a new technology being used. But at the same time, when a technology is introduced that affects the whole world, there must be extensive accountability and competence from the bodies entrusted with regulating it in order for it not to harm people.

4.3 Recommendations

4.3.1 Responsibility of Institutions

As discussed in the prior section, there are several problems that could conceivably come up from technology, and mitigating these concerns could help society by both

assuaging public fears as well as ensuring the effective use of AI. Examples of problems include unemployment or a loss of purpose, privacy concerns, malfunctioning of androids/liability of androids, and moral concerns regarding creation. These problems are all quite complicated and deal with multiple institutions. Unemployment partly has to do with employers, but also with the government. Both data usage and privacy, and the malfunctioning and reliability of androids have to do with the creators of technology as well as the legal system and the lawmaking body of our government. However, do institutions actually have the obligation to care about fixing these issues?

One could argue they do, from an ethical standpoint. From a deontological, or duty-based approach to moral actions, there are several reasons government and businesses could be obligated to act ethically. The government, especially a democracy, has a duty to serve the people. Beyond that, the structure in which taxpayers pay taxes to the government reinforces a debtor-creditor relationship that implies the government should serve the taxpayers in return. The role of businesses may not be quite as clearly service oriented, but the company-consumer relationship closely resembles the government-taxpayer relationship in that there is also a monetary payment by the consumer in return for a product. The business also has a duty here to provide transparency about the effects of a product and the details of the product or service the consumer is receiving.

From a utilitarian perspective, the moral obligation of businesses is equally apparent. It would be unethical under this framework to create products with negative impacts on society, regardless of the intention behind the product. Mass layoffs due to automation would also be unethical. As mentioned previously, AI leading to high rates unemployment could deeply, negatively influence many people. The effects of this could be widespread, affecting both workers as well as their families, and the economy at large. When examining the negative consequences, it is clear that it would be unethical to blindly push forward with development that might put people out of jobs without providing some solution for the effects.

Beyond any ethical obligation, per se, one could argue self-interest should drive

institutions to act ethically and work towards gaining public trust. Many AI products are aimed at wide consumer bases. Smart homes and smart assistants are targeted towards anyone and everyone. Self-driving cars are being developed with the hopes of completely revolutionizing the way that humans drive (or rather, do not drive). Ostensibly, the members of a company that develops AI products, or the members of the government that regulate AI products will also be affected by the quality of the AI products. Essentially, the creators and the regulators may also be the consumers. If the creators of the self-driving car choose to overlook certain safety aspects of the car, and are also one of the consumers of the car, then they will be put at risk by their own untrustworthy behavior. In this sense, being ethical and trustworthy may be in the self interest of the members of any institution as they are also members of the larger society.

Self-interest could also apply to maintaining the reputation of different institutions. It was mentioned above that the reputation of the government, different tech companies, etc. is going down. Acting in ways that deserve trust, and thus gaining trust, can help improve these reputations which, from a self-interested perspective, will help the institution sustain their power. With tech companies, a better reputation can lead to more user interest and user satisfaction leading to greater profit. For businesses in general, studies have shown that reputation has a big impact on success with consumers - a 2012 study showed that reputation directly influences about a quarter of a company's market value[35]. In another example, a study evaluating the prices consumers were willing to pay for different TVs found that consumers were willing to pay 22% more for a TV if they felt that the brand had a good reputation. Additionally, missteps that damage the reputation of a brand can cause dramatic declines in public interest[36]. For example, the scandal of one German car brand, Volkswagen, caused ripple effects throughout the German car industry. After it was revealed

35. *Global Survey on Reputation Risk*, technical report (Deloitte, 2015).

36. Jeanette Settemre, *People Will Pay 22% More for Certain Products if the Company Has a Good Reputation*, 2018.

that Volkswagen's diesel cars were equipped with a mechanism to fake the level of emissions on emissions tests, Volkswagen faced serious repercussions. This scandal not only negatively impacted the reputation of Volkswagen, but German car brands as a whole, causing them to lose billions of dollars in sales[37]. Lastly, problems with trust not only affect the business-consumer relationship, but other business related relationships as well, such as those with stakeholders, suppliers, investors, employees, the government, or other regulators.

As for the government, a better reputation can lead to higher approval ratings and sustained power throughout re-elections. Addressing public fears about technology can thus prove advantageous for government officials in gaining public approval. The government seems to also acknowledge this, stating that "continued adoption and acceptance [of AI] will depend significantly on public trust and validation"[38]. Additionally, it is in the government's interest to have a robust economy. The economic promise of the tech industry and it achieving "continued adoption and acceptance" is therefore also an incentive for the building of public trust.

Lastly, self interest could apply in terms of the desire to resist punishment. Our society is driven by laws, and breaking those results in punishment. Many of those apply to tech companies and many more laws will likely be created regarding technology in the years to come because of how rapidly the field is growing. As mentioned earlier, several big tech companies have been fined million to billion dollar fines as punishments. Even if they are able to weather financial punishments, in some cases they have also lost some degree of agency over their decisions, such as being subject to third-party auditing of privacy practices[39].

37. Dimitrije Ruzic, *How the Volkswagen Scandal Turned 'Made in Germany' Into a Liability*, 2019.

38. Vought, *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications*.

39. Feiner and Rodriguez, *FTC Slaps Facebook with Record $5 Billion Fine, Orders Privacy Oversight*.

4.3.2 Building Public Trust

Given that institutions are incentivized to gain public trust and act in a trustworthy manner, how can they actually achieve this? This section will delve into the factors of public trust and how they can be applied to each of the key concerns discussed earlier.

In "Public Trust in Business and Its Determinants," Michael Pirson et al. attempt to measure the effects of different factors on public trust of a given firm. The article frames public trust in business as the "willingness of members of the public to become vulnerable to business". It acknowledges two dimensions of determinants – "trustor-related determinants," or attributes of a member of the public that cause them to choose to be vulnerable to the business, and "trustee-related determinants," or attributes of the business that influence the public to choose to be vulnerable to the business. This discussion will focus on the trustee-related determinants as they are more within the control of an institution.

Of the trustee-related factors discussed in the article, two factors, the size of a given firm, and the mission statement of a firm, proved to have little affect on overall public trust. However, two other factors did correlate with increased public trust. One of the trustee-related factors correlated with increased trust in a firm was the overall view of the industry of which the firm was a part. The industry being viewed as socially beneficial led to increased trust in the firm. As mentioned earlier, this is a growing problem for the tech industry. From 2015 to 2019, the amount of the American public that viewed tech companies as making a positive impact on society dropped from 71% to 50%, and the amount that viewed tech companies as having an explicitly negative impact on society has grown from 17% to 33%[40]. The overall view of the industry is not overwhelmingly negative yet, but if the trend seen during this time period continues, it could become problematic.

A second trustee-related factor had to do with "trustworthiness." The article defines the

40. Doherty and Kiley, *Americans Have Become Much Less Positive About Tech Companies' Impact on the U.S.*.

components of a firm's trustworthiness as ability, benevolence, integrity, transparency, and value congruence (how well the values of the business and the respondent matched). High rankings in these components of trustworthiness were found to correlate with increased trust in a firm[41].

These correlations lay out an abstract groundwork for what areas of a business to strengthen. An article by Winfield et al. regarding the ethical governance of AI development provides a more concrete framework of what an institution could do to achieve public trust. The article argues that ethical governance, or "a set of processes, procedures, cultures and values designed to ensure the highest standards of behavior," is an important factor in increasing public trust in AI. Upon analyzing the behaviors laid out as steps of ethical governance, it seems that ethical governance could contribute to the components of trustworthiness discussed in the earlier article. The behaviors an institution could follow include creating an ethical code of conduct, requiring ethics training, requiring responsible innovation by conducting ethical risk assessments, and being transparent about how the institution is being ethical[42]. Examining the trustworthiness factors identified in the previous article, it is intuitive that ethical governance has to do with ethics-related factors such as integrity and benevolence. However, it also has to do with transparency, such as transparency in sharing ethical methods with the public, and ability, such as responsible innovation and ethical risk assessments because they could significantly improve the ability of company to evaluate the consequences of an innovation and prepare accordingly.

The factors of public trust discussed so far can be directly applied to the key concerns covered in the previous section. The first key concern, unemployment, has to do with transparency, ability, benevolence, and integrity. Benevolence and integrity apply to the

41. Michael Pirson, Kristen Martin, and Bidhan Parmar, "Public Trust in Business and Its Determinants," *Business & Society* 58 (2016): 132–166.

42. Alan F.T. Winfield and Marina Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (2018).

question of whether or not a company would bother to consider its employees if it can automate their job, or would bother to develop any mitigation plan. As discussed in the unemployment section, there are potential ways to address unemployment due to automation. However, the legitimacy and effectiveness of different plans can vary. Transparency from a company about what plans are in place in the case of jobs being lost to automation could help indicate the ability of the company to actually carry out those plans. For example, if there is a plan in place to upskill, transparency about the process could lead to less fear and greater willingness to comply with the company's desires.

The second key concern, data usage and privacy, has to do with both transparency and ability. Transparency about what data is collected, how data is used, and how data is protected could address some user concerns about the level of privacy they have and the security of their data. As for ability, it pertains to the ability to secure data, as well as the ability to use data effectively. As mentioned earlier, algorithmic bias is a problem in the context of data usage. One way to address this problem is increasing the diversity of those creating the algorithms, because a broader perspective increases the overall ability of a company to innovate. Concepts mentioned in the Winfield article, such as ethics trainings and ethical risk assessments could also aid in this awareness. Both of these solutions enhance the ability to create useful and effective products.

The last key concern, liability, has to do primarily with ability and integrity. If a question of liability results from the malfunctioning of AI, that relates directly to the abilities of those creating the AI to predict potential consequences and create a safe product. Integrity also applies because it pertains to whether or not an institution is willing or required to take responsibility for what they have created. Ultimately this has to do with institutional accountability, which, as discussed in the liability section, is one of the main concerns within liability. Additionally, if institutions do not have this accountability and integrity, it could affect the overall view of the tech industry. AI malfunctions could cause people harm, and if there is no accountability for this harm, it could lead to negative impressions of the

tech industry. As mentioned in the Pirson article, the public viewing the industry as a whole as socially beneficial is important in public trust, and developing a negative industry connotation could prove problematic.

The actions discussed above are ethical actions that the creators of AI can take. While these actions could be fueled by the creators' own desires to act ethically and/or gain public trust, some of these ethical actions could also be encourage by non-ethics based incentivization. Take, for example, the "robot tax" discussed in the unemployment section. This would charge companies for automating jobs, and would potentially incentivize more moderate replacement of human workers with AI as opposed to mass job loss. With data usage and privacy, a legal framework for issues such as disclosure of data collection, what data can be collected, etc. could mandate the ethical usage of data by creators of AI. With liability, legal frameworks regarding liability of AI could define liability for the public as opposed to relying on the integrity of creators to take accountability for what they create. The prior discussion focused on how AI creators could act in order to gain public trust. Choosing to act in this manner, of their own accord, is one way institutions such as tech companies and research labs could encourage public trust. But this point focuses on how an external institution like the government could put rules in place in order to require tech companies and research labs to act in a certain way. Both approaches satisfy the necessity of ethical actions, but in the first, the actions of tech companies and researchers alone satisfy trustworthiness requirements, whereas in the second, it is the summation of the actions of the government, tech companies, and researchers operating in tandem that provide the sufficient trustworthiness requirements to the public.

Chapter Five: Conclusion

The conflict of man vs. machine has existed since the times of early industrialization. As the world becomes technologically more advanced, these fears have become more tangible. With the rise of Artificial Intelligence particularly comes a rise in the fear of what intelligent creations might lead to. However while the fear of AI is largely attributed to the fear of the machine itself, the purpose of this thesis was to demonstrate the influence of distrust in developers and regulators on distrust of AI.

All three of the case studies demonstrated, regardless of how prevalent AI was during the time in which they were written, significant doubts in the creators and regulators present in the narratives. These concerns spanned doubting the ability of creators to accurately judge the consequences of their creations, to doubting their ethics in creating AI, to doubting the integrity of institutions as a whole.

As presented in the previous sections, this distrust is not only constrained to the world of fiction. Levels of trust in the government, businesses, and tech companies specifically have dropped in recent times. This is a salient issue in AI. As discussed, some of the key concerns related to AI, such as unemployment, privacy, ethics and liability have to do largely with the intentions and competency of creators and regulators.

For AI to truly grow, it must grow not only in terms of development but also in terms of public adoption. Questions and fears will inevitably emerge as novel developments are made. It is paramount to understand the source of public fear and explore potential problems that AI may bring. In order to do so, AI creators and regulators must contend with not only the public perception of machines, but the public perception of the creators and regulators themselves.

References

2019 Edelman Trust Barometer, 2019. https://www.edelman.com/research/2019-edelman-trust-barometer.

2020 Edelman Trust Barometer, 2020. https://www.edelman.com/trustbarometer.

Abramovich, Giselle. *Study Finds Consumers Are Embracing Voice Services. Here's How.*, 2018.

Advice Reimagined. https://oneseventeenmedia.com/what-we-do/3rd-12th/.

Artificial Intelligence for the American People. https://www.whitehouse.gov/ai/.

Bova, Ben, ed. *The Science Fiction Hall of Fame: Volume Two B*. New York: Tom Doherty Associates, 1973.

Brandom, Russell. *Answering the 12 Biggest Questions About Apple and Google's New Coronavirus Tracking Project*, 2020.

Breuer, Hans-Peter. "Samuel Butler's "The Book of the Machines" and the Argument from Design." *Modern Philology* 72, no. 4 (1975): 365–383. https://www.jstor.org/stable/436868.

Copeland, B. Jack, ed. *The Essential Turing*. Oxford University Press, 2004.

Copeland, B.J. *Artificial Intelligence*, 2020.

D'Onfro, Jillian. *Google's Sundar Pichai Was Grilled on Privacy, Data Collection, and China During Congressional Hearing*, 2018.

Danks, David, and Alex John London. "Algorithmic Bias in Autonomous Systems." *IJCAI International Joint Conference on Artificial Intelligence*, no. January (2017): 4691–4697.

Dick, Phillip K. *Do Androids Dream of Electric Sheep?* New York: Del Rey, 1968.

Doherty, Carroll, and Jocelyn Kiley. *Americans Have Become Much Less Positive About Tech Companies' Impact on the U.S.*, 2019.

Elias, Jennifer, and Lauren Feiner. *YouTube Will Pay $170 Million to Settle Claims It Violated Child Privacy Laws*, 2019.

Ex Machina (2014).

Feiner, Lauren, and Salvador Rodriguez. *FTC Slaps Facebook with Record $5 Billion Fine, Orders Privacy Oversight*, 2019.

Friedman, Uri. *Trust is Collapsing in America*, 2018.

Garcia, Megan. "Racist in the Machine: The Disturbing Implications of Algorithmic Bias." *World Policy Journal* 33, no. 4 (2016): 111–117.

Giuffrida, Iria. "Liability for AI Decision-Making: Some Legal and Ethical Considerations." *Fordham Law Review* 88, no. 2 (2019): 439. https://news.harvard.edu/%7B%5C%%7D0Ahttps://ir.lawnet.fordham.edu/flr/vol88/iss2/3.

Global Survey on Reputation Risk. Technical report. Deloitte, 2015.

Greene, Joshua David. *Moral Tribes*. Penguin Books, 2013.

Gross, Charles. "Scientific Misconduct." *Annual Review of Psychology* 67 (2016): 693–711.

Gurkaynak, Gonenc, Ilay Yilmaz, and Gunes Haksever. "Stifling Artificial Intelligence: Human Perils." *Computer Law and Security Review* 32, no. 5 (2016): 749–758. http://dx.doi.org/10.1016/j.clsr.2016.05.003.

Harris, Mark. *NTSB Investigation Into Deadly Uber Self-Driving Car Crash Reveals Lax Attitude Toward Safety*, 2019.

Hill, Doug. *Erewhon: The 1972 Fantasy Novel that Anticipated Thomas Nagel's Problems With Darwinism Today*, 2013.

Kahan, Dan M., Hank Jenkins-Smith, and Donald Braman. "Cultural Cognition of Scientific Consensus." *Journal of Risk Research* 14, no. 2 (2011): 147–174.

Keynes, John Maynard. *Economic Possibilities for our Grandchildren*. Technical report. 1930.

Kharpal, Arjun. *Bill Gates Wants to Tax Robots, but the EU Says, 'No Way, No Way'*, 2017.

Kirkham, Georgiana. "'Playing God' and 'vexing nature': A cultural perspective." *Environmental Values* 15, no. 2 (2006): 173–195. https://www.jstor.org/stable/30302154.

Kurzweil, Ray. *The Law of Accelerating Returns*, 2001.

Levy, Daniel S. *Behind the Anti-War Protests That Swept America in 1968*, 2018. https://time.com/5106608/protest-1968/.

Long, Tony. *Jan. 25, 1921: Robots First Czech In*, 2011.

Marr, Bernard. *The Key Definitions of Artificial Intelligence (AI) That Explain Its Importance*, 2018.

Marshall, Aarian. *A Bet on Uber Is a Bet on Self-Driving*, 2019. www.wired.com/story/bet-uber-bet-self-driving/.

Marshall, Aarian, and Alex Davies. *Uber's Self-Driving Car Didn't Know Pedestrians Could Jaywalk*, 2019.

McCaffery, Larry. "An Interview with Jack Williamson." *Science Fiction Studies* 18 (1991).

McClure, Paul K. ""You're Fired," Says the Robot: The Rise of Automation in the Workplace, Technophobe, and Fears of Unemployment." *Social Science Computer Review* 36, no. 2 (2017): 139–156.

McDermid, John. *Self-Driving Cars: Why We Can't Expect Them to Be 'Moral'.*

Moral Machine.

Palmer, Annie. *The Decade Big Brother Came Home: How Tech Giants Persuaded Us to Buy Products That Track Us at Home*, 2019. https://www.cnbc.com/2019/12/19/how-tech-giants-persuaded-us-to-buy-products-that-track-us-at-home.html.

Parmar, Inderjeet. "Elites and American power in an era of anti-elitism." *International Politics* 54, no. 3 (2017): 255–259.

Piper, Kelsey. *Science Funding is a Mess. Could Grant Lotteries Make it Better?*, 2019. https://www.vox.com/future-perfect/2019/1/18/18183939/science-funding-grant-lotteries-research.

Pirson, Michael, Kristen Martin, and Bidhan Parmar. "Public Trust in Business and Its Determinants." *Business & Society* 58 (2016): 132–166.

Public Trust in Government: 1958-2019, 2019. https://www.people-press.org/2019/04/11/public-trust-in-government-1958-2019/.

Robertson, Adi. *How Blade Runner Got Its Name from a Dystopian Book about Health Care*, 2019.

Russell, Stuart, Daniel Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Systems." *AI Magazine* 36, no. 4 (2015): 105–114.

Ruzic, Dimitrije. *How the Volkswagen Scandal Turned 'Made in Germany' Into a Liability*, 2019.

Settemre, Jeanette. *People Will Pay 22% More for Certain Products if the Company Has a Good Reputation*, 2018.

Shelley, Mary. *Frankenstein; or, The Modern Prometheus*. Project Gutenberg, 1818.

Sisyphus, 2019.

Smith, Richard. "Research Misconduct: The Poisoning of the Well." *Journal of the Royal Society of Medicine* 99, no. 5 (2006): 232–237.

Snow, Jackie. *This Time, with Feeling: Robots with Emotional Intelligence Are on the Way. Are We Ready for Them?*, 2019.

The Freedom Dividend.

The Future of Life Institute (FLI).

The Sentencing Project. "Report of The Sentencing Project to the United Nations Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia, and Related Intolerance," 2018. https://www.sentencingproject.org/publications/un-report-on-racial-disparities/.

The Story behind Jack Williamson's 'With Folded hands'. galaxypress.com/jack-williamsons-with-folded-hands.

Thomson, Andrew. *Emotionally Intelligent Computers May Already Have a Higher EQ Than You*, 2016.

Uber Back-Up Driver Faulted in Fatal Autonomous Car Crash.

Vermeulen, Ben, Jan Kesselhut, Andreas Pyka, and Pier Paolo Saviotti. "The Impact of Automation on Employment: Just the Usual Structural Change?" *Sustainability (Switzerland)* 10, no. 5 (2018): 1–27.

Vought, Russel T. *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications.* Technical report. 2020.

Wagner, Kurt. *This Is How Facebook Collects Data on You Even If You Don't Have an Account*, 2018. www.vox.com/2018/4/20/17254312/facebook-shadow-profiles-data-collection-non-users-mark-zuckerberg.

Walsh, Kenneth T. *The 1960s: Polarization, Cynicism, and the Youth Rebellion*, 2010.

What Inspired Phillip K. Dick to Write Do Androids Dream of Electric Sheep?

What is Artificial Intelligence?

Williamson, Jack. "With Folded Hands." In *The Humanoids*, 10–63. New York: Tom Doherty Associates, 1948.

Winfield, Alan F.T., and Marina Jirotka. "Ethical governance is essential to building trust in robotics and artificial intelligence systems." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (2018).

Biography

Maya Kothare-Arora is a native Austinite and graduating senior at the University of Texas at Austin. She will be graduating with a dual degree in Computer Science and the Plan II Honors program. During her time at UT, she had the opportunity to study abroad in Guatemala and Belize, participate in student-run theater, and serve as president of Code Orange, an organization dedicated to technical education and mentorship for children. After graduating, Maya will be working as a technical consultant in Dallas, Texas.