ABSTRACT

Be Intelligent About AI: Should We Create Artificially Intelligent Machines?

Warren J. Burrus

Director: Richard L. Sneed, Ph.D.

"Artificial intelligence" has become a popular term to describe an ever-growing array of technology, but what makes it intelligent? Philosopher John Searle analyzed artificial intelligence (AI) with his famous distinction between "weak" and "strong" AI, the former being data-driven tools we use daily while the latter has a mind and cognitive states of its own. However, this distinction does not fully address the equally pressing, moral question of whether we should develop or use such AI in the first place. By applying a range of both technical and fictional sources, we can define a new pair of categories, Algorithmic AI and Fictional AI, that answer not only what makes such AI intelligent, but whether we should even pursue it. Algorithmic AI is our present computer programs that merely follow advanced algorithms (Siri, Alexa, ChatGPT); creating Algorithmic AI as a tool is not itself immoral—but how we use it can be. Fictional AI, however, exhibits its own form of intelligence and, so far, only exists in fiction (*Blade Runner*, *I, Robot*); creating Fictional AI would be immoral altogether.

APPROVED BY DIRECTOR OF HONORS THESIS:

_____

Dr. Richard L. Sneed, Department of Philosophy

APPROVED BY THE HONORS PROGRAM:

_____

Dr. Elizabeth Corey, Director

DATE: _____

BE INTELLIGENT ABOUT A.I.

SHOULD WE CREATE ARTIFICIALLY INTELLIGENT MACHINES?

A Thesis Submitted to the Faculty of

Baylor University

In Partial Fulfillment of the Requirements for the

Honors Program

By

Warren J. Burrus

Waco, Texas

May 2024

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This thesis project has been an exciting and unforgettable experience, but its inspiration and work have not been mine alone. Several individuals have helped guide me in various ways throughout this project, without whom my thesis could not have been written. I would like to thank these special people for their invaluable support.

Thank you to Dr. Richard Sneed for serving as my thesis director and mentor, meeting with me at every milestone of the thesis to review it and plan out the next. Your knowledge of philosophy and science, especially your countless recommendations of books and media on AI, were foundational to my thesis.

Furthermore, thank you to Dr. Al Beck for recommending other source material like *Do Androids Dream of Electric Sheep?* during our advising sessions, providing further inspiration for my thesis. Thank you to Dr. Robert Garcia and Dr. Pablo Rivas for offering their valuable time to serve on my defense committee, bringing insights of their own to the thesis. Lastly, thank you to Ms. Mary Moore, Dr. Elizabeth Corey, and the Honors College for the much-needed tips, resources, and words of encouragement.

Somewhat ironically, I do not owe any thanks to ChatGPT for the completion of this thesis project.

To my family

INTRODUCTION

Be Intelligent About AI: Should We Create Artificially Intelligent Machines?

Artificial Intelligence in Context

Once confined to the realm of science fiction, artificial intelligence is quickly becoming a household term. From personal voice assistants to chatbots to image generators, everyone seems to be using artificial intelligence (commonly abbreviated as AI) for just about everything. Personal assistants like Siri have accompanied smartphones for years to receive simple commands like setting reminders or phoning a contact. Home appliances have also seen an emergence of AI with devices like Alexa, which can answer nearly any question with a quick internet search. Most recently and notoriously, generative AI has ushered in a new demand for AI in the form of chatbot interfaces like ChatGPT, using big data and large language models to hold realistic conversations with users. Other generative AI tools can even produce life-like photos and human-like music and art. We are certainly at a revolutionary stage in the field of artificial intelligence.

As helpful and game-changing as this emerging tool is, it can also be a source of both confusion and concern. The more developed that AI becomes — what with natural language processing, machine learning, and generative abilities among others — the more challenged we as onlookers become in understanding what it is and how it works. Furthermore, these more advanced capabilities translate into more advanced use cases, such as using AI to write papers, produce art, brainstorm ideas, and answer prompts; as enticing as this potential may be, it also creates entirely new avenues for plagiarism,

deepfakes, job replacement, laziness, cheating, and other concerns. Thus, before the popularity of this trend surpasses our grasp of it, and before it does or becomes something we may regret, one of our immediate priorities should therefore be to understand AI as best we can. In other words, to curb not just our technical confusion over AI but our moral concerns with it as well, our understanding should include both a technical and moral focus.

Roadmap

This understanding should start with a definition of terms, specifically what we mean by "artificial" and "intelligence" as used in the phrase "artificial intelligence." We can then construct a general definition of AI from the many different definitions available for intelligence. And if we have different kinds of intelligence, then we may very well have different kinds of *artificial intelligence*. These definitions may therefore distinguish different categories of AI, particularly differences between the AI we have now, and AI envisioned for the future. Since the AI we know of today is confined to computer programs and amounts to little more than advanced algorithms, we may collectively call it Algorithmic AI. Conversely, the more traditional AI of science fiction has yet to make itself a reality, and thus far exists only within the imaginations of *Star Wars*, *Blade Runner*, *I, Robot*, and other fiction, so may be known as Fictional AI. Although both are subcategories of a common artificial intelligence, they each have their own, more refined forms of intelligence, giving them separate moral dimensions as well.

Once we have a stronger technical understanding of AI from its definitions and distinct categories of Algorithmic and Fictional AI, we can better understand moral issues of AI in terms of these same two categories. As we address the specific moral dilemmas

AI causes like plagiarism or misinformation, we must also resolve the more fundamental question of whether we should pursue each type of AI at all, and why. Establishing a moral basis or justification for each category can better prepare us to tackle the more particular issues it raises — or help mitigate any future crises should we find the category immoral from the outset. The moral concerns around AI are manifold, so approaching them with our two types of AI might disperse their moral complexity. Our overall understanding of AI may then improve morally along with technically, perhaps easing some of the original concerns and confusion we currently have on the subject.

The ideas outlined above constitute a high-level overview of what the following pages will cover, but there are other things it will not. For instance, this paper will not strive for any new definition of intelligence nor try to redefine the term, but merely use and apply existing meanings to the technology we call AI. Doing so will present AI as a natural evolution of existing intelligence, rather than dissociate it from all other notions of intelligent beings. Furthermore, we will not focus on the scientific background of AI nor discuss how it works on a technical level, but only explain AI in enough detail to discern the two main categories of Algorithmic and Fictional AI. As an extension to this point, we also will ignore how scientifically feasible Fictional AI actually is, taking as a baseline that fiction may eventually become reality. This will free us from the practical bounds of AI to focus on the more pressing questions of its morality. The essential aim of this paper, then, is moral in nature: to ask whether we should create artificial intelligence.

Thesis Statement

With these goals and qualifications in mind, we may state the thesis of this paper thus: artificial intelligence can be separated into two broad categories — but only one of

which is moral to use. Contemporary Algorithmic AI are mere computer programs that follow algorithms to appear intelligent — like Siri, Alexa, and ChatGPT. Using this type of AI as a tool is not itself immoral, but how we use it can still be immoral. The more complicated Fictional AI actually possesses intelligence analogous to humans, and so far only exists in fiction like *Blade Runner* or *I, Robot*. We should not pursue fictional AI at all because using an intelligent being as a tool is immoral. To demonstrate this thesis, we will first define artificial intelligence more broadly, then concentrate on Algorithmic AI and its morality, followed by Fictional AI and its own morality in turn.

CHAPTER ONE

What Is Algorithmic AI?

Introduction

Algorithmic AI, as we will soon see, is all around us. But as we begin our discussion of artificial intelligence with Algorithmic AI, we may first wonder what we mean by artificial intelligence. Defining key terms is always a helpful first step; with AI we have two — "artificial" and "intelligence." The former is straightforward, meaning anything manmade; the latter, however, is more abstract and demands careful attention of its own. In this first chapter, we will study working definitions of intelligence, before explaining Algorithmic AI based on these definitions and contemporary real-world examples.

Defining Intelligence

To understand any kind of artificial intelligence, we must understand the word "intelligence." The challenge here is that while we would readily see ourselves as intelligent, AI requires us to see other entities as intelligent too, namely artificial technologies. Thomas Nagel lays out this issue of subjective experience, or what it is like to be something else, in his article "What Is It Like to Be a Bat?" Here he contends we cannot understand someone else's experiences because they are inaccessible from our own point of view. For example, humans can never know a bat's subjective experiences since they differ from our own, and we are "restricted to the resources of [our] own mind" (439). Likewise, we cannot fully comprehend the intelligence of an artificially

intelligent being, so the best we can do is understand it from our own experience of intelligence. To further support Nagel's point, the father of AI, John McCarthy, agrees in his article "What Is Artificial Intelligence?" that we are not yet able to characterize intelligence without relating it to human intelligence (2-3). Simply put, we can only know intelligence in terms of human intelligence.

What then is human intelligence? Even among this singular kind of intelligence there is not one defining feature but rather a collection of characteristics. The aptly named *Artificial Intelligence* by Time-Life Books explores intelligence in relation to humans and finds a range of meanings:

> To some, intelligence consists of solving hard problems; to others it is learning or forming generalizations or analogies; to still others it is dealing with the world – communicating, perceiving, comprehending what is perceived. (7)

The book also adds that

> few people would dispute that the ability to cope with change and to incorporate new information is a fundamental aspect of intelligence. . . . [or the] ability to modify its behavior according to what it experienced. (38, 75)

Thus, intelligence can be characterized by problem-solving, learning, communicating, adaptation, or any combination of these. John McCarthy also offers his own definition, stating that intelligence "is the computational part of the ability to achieve goals in the world" (2). Therefore, intelligence can also entail self-determined and self-directed goal-making, along with the above qualities; a being exhibiting any of these may have intelligence comparable to a human according to these definitions.

Putting this all together, we may then have a clearer definition of artificial intelligence, at least from our view as humans. An artificially intelligent being is a man-made (artificial) technology possessing its own (human-like) intelligence. This

intelligence could be exemplified through problem-solving, learning, communicating, adaptation, goal setting, or a mixture of these. Simply, artificial intelligence is human intelligence created by humans. Of course, this is still a rather vague definition for such a complex term, so our understanding may be further aided by breaking up AI and studying its first category, Algorithmic AI.

Defining Algorithmic AI

Algorithmic AI is the AI we have known and used up to today. More precisely, any of our current technologies that exhibit human aspects of intelligence are Algorithmic AI. Siri and Alexa speak like humans, ChatGPT converses like a human, and generative AI produces human-like art and media. They each seem to model human intelligence by solving problems, learning, communicating, and adapting in different ways, making them AI in a general sense; but what makes them Algorithmic AI in particular is the one trait all these technologies share: the algorithmic manner in which they operate. An algorithm is simply a set of rules for solving a problem, and in the case of these technologies the rules are code and software. They are programmed to speak like humans, form conversational text, or generate images, all while remaining advanced versions of existing computers. In short, Algorithmic AI is any computer program that follows intricate algorithms to mimic human actions.

Consider as an example ChatGPT, which although seems to be much more than a computer, is actually just a large intersection of algorithms exhibiting human-like behavior. As the recent book *What Is ChatGPT Doing … and Why Does It Work?* explains, ChatGPT constructs conversations by adding one word at a time based on all previous words. First, it converts the existing set of words (or tokens) into an array of

numbers (or an embedding). Then, ChatGPT repeatedly operates on this embedding with a large mathematical function (or neural network) that uses up to billions of parameters to produce new embeddings. The final embedding becomes an array of probabilities for possible next tokens, based on large language models of the most common words for the given conversation. ChatGPT ultimately selects from these probabilities the next token to display (Wolfram). Although it is extremely complex, at the end of the day ChatGPT still just follows a collection of algorithms, making it a prime example of Algorithmic AI.

Reconsidering Intelligence in Algorithmic AI

The need to categorize AI into different types, such as Algorithmic AI, is evident by our loose use of the phrase "artificial intelligence." We tend to label chatbots, personal assistants, generative tools, and other mainstream technologies as artificially intelligent, without wondering whether they really are. By permitting any new development in technology to be considered artificial intelligence, we are erasing the meaning of the term and lowering the standards of being intelligent generally. While these current technologies are certainly "artificial," the question of whether they truly contain the human intelligence we defined earlier is less clear.

Perhaps, then, we should return for a moment to our definition of intelligence and ask whether it is truly present in our Algorithmic AI. ChatGPT definitely exhibits communication, problem-solving, and even adaptation and learning when it converses with a user, answers their questions, and changes based on past conversations. And Siri and Alexa are certainly great at perceiving much of the audible world around them and communicating with users. But as we have seen, all these impressive feats are actually the work of a myriad of complex algorithms, so where there seems to be intelligence

there is really just a computer program. And the creators of Algorithmic AI know best of all how these clever algorithms are not truly intelligent, much like how a magician knows his magic tricks are not truly magic. It may also be worth noting that no Algorithmic AI appears to form any self-determined goals – which was McCarthy's understanding of intelligence – but rather executes the goals given by users. Therefore, our Algorithmic AI appears to lack genuine human-level intelligence and contains nothing more than algorithms themselves.

A counterargument by philosopher John Searle may be raised here that the algorithms and programming of Algorithmic AI are only analogous to the neurons and workings of the human brain. If learning and problem-solving by Algorithmic AI are just cleverly disguised algorithms at work, then human learning and problem-solving are simply neural firings at work. So, their algorithmic intelligence is no more a lie than human intelligence; and learning, problem-solving, and other qualities are just abstractions that can be actualized in a multitude of ways. But again, we are confronted here with Nagel's subjective experience problem: learning, problem-solving and the like were not conceived before we experienced them, but only after, so they are inherently grounded in human experience. These qualities can only be applied to a non-human in their human context, so if they are not produced the way humans produce them, then they are not produced at all. And since computer algorithms do not operate the same as human brains, we cannot know algorithms to have intelligence the way we know ourselves to have intelligence.

This relation of algorithms to intelligence can be better seen by expanding our idea of Algorithmic AI. To provide an alternative example, the Stanford article titled

"Ethics of Artificial Intelligence and Robotics" identifies certain AI systems as "objects." These AI are "tools made and used by humans" that can perform autonomous tasks. Being jumbles of algorithms that mimic human actions, Algorithmic AI is rooted in this notion of AI as objects. They sound no different from the next latest and greatest technology. The article explicitly takes a neutral stance on whether AI as objects possess human intelligence, implying that such intelligence is not necessarily present within Algorithmic AI (Müller). Because true intelligence, learning, or problem-solving are not needed for a tool to carry out such autonomous tasks, they are not needed within Algorithmic AI.

In agreement with this point, John Searle distinguishes between a computer's simulation of understanding and actual understanding in his *Minds, Brains, and Science* lectures:

> No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched. Why on earth would anyone suppose that a computer simulation of understanding actually understood anything? (12)

Every machine preceding Algorithmic AI has also functioned as an algorithm, such as alarm clocks, calculators, and traffic lights; and they have never been known to have intelligence, so there is no reason Algorithmic AI should. Having more complicated algorithms just makes Algorithmic AI a natural progression of those early machines, not uniquely intelligent. So again, Algorithmic AI does not actually contain any of the qualities of human intelligence its algorithms are designed to mimic, and again is not truly intelligent in the sense of human intelligence.

Yet, to deny these recent technologies any association with artificial intelligence would not do justice to the current achievements and capabilities they represent, and we

would fail to recognize their potential for future progress in the field. The Stanford article above deliberately avoids restricting "intelligence" to human intelligence, to allow for a kind of "technical" AI which actually "show only limited abilities in learning or reasoning but excel at the automation of particular tasks" (Müller). What this means is that technical (or Algorithmic) AI may still be labeled intelligent even when it knowingly does not match human intelligence. Therefore, our distinction of Algorithmic AI from broader artificial intelligence can reconcile these unintelligent technologies with intelligence, allowing them to coexist with more traditional notions of AI, like in science fiction. In essence, when our Algorithmic AI is understood to have a simplified form of intelligence, it may still retain a spot under the broader umbrella of artificial intelligence. The next question to ask, then, is what makes Algorithmic AI intelligent.

The Turing Test

Since Algorithmic AI is so good at appearing to solve problems, learn, communicate, and adapt – while still only being a series of algorithms – its form of "intelligence" may best be described by the Turing test. As we have come to see, there are many different definitions of intelligence; Alan Turing's version, known as the Turing test or imitation game, considers a computer intelligent if it can deceive a human into believing it is not a computer but rather another human. A test of this intelligence was originally conceived to entail a person and computer separated from an interrogator who asks them both questions to decide which is the computer based on their answers. If the computer supplies convincing enough answers that the interrogator mistakes it for the person, then the computer passes the Turing test and thereby has intelligence (Hodges 523-5).

Today's examples of Algorithmic AI are perfect candidates for the Turing test. If not for their mechanical, monotone voices, Alexa or Siri could easily be mistaken for a person speaking alongside us. Similarly, many teachers would agree that ChatGPT certainly could pass the Turing test, facing an ever-growing challenge to distinguish the chatbot's writing from student writing. But most impressive may be generative AI's ultra-realistic images and sounds, making us sometimes question what is even real. They are not really talking, writing papers, or drawing pictures the way we do, they are just following algorithms; yet the fact that these algorithms produce human-like results still gives these machines intelligence. By Turing's view, Algorithmic AI has intelligence not because it is human, but because it seems human. Algorithmic AI may thus preserve its own simplified intelligence using Turing's definition.

Conclusion

This chapter has dealt in large part with definitions. We began by defining intelligence in terms of human intelligence, which we found to be synonymous with problem-solving, learning, communicating, adaptation, or goal setting. With this definition established, we proceeded to define our first category of artificial intelligence, Algorithmic AI, as being any of our current conceptions and examples of AI, such as Siri, Alexa, and ChatGPT. Initially, we described Algorithmic AI with the unspoken assumption that it had human intelligence; however, we reevaluated Algorithmic AI's relation to intelligence to find that, as an algorithm like any other computer program, it is not truly intelligent. In response to this discovery, we redefined the intelligence of Algorithmic AI around the Turing test, so that it may still be granted intelligence so long as it behaves like a human.

To complete our understanding of Algorithmic AI, then, any modern algorithm advanced enough to produce behaviors or outputs which may be mistaken for a human's may be deemed Algorithmic AI. Early voice assistants like Siri and Alexa, current generative AI like ChatGPT, and similar such technologies in the near future, are and will surely dominate this category of artificial intelligence. With a firm grasp of Algorithmic AI itself, we are now ready to transition to the more significant question of its morality, namely asking whether we should continue use of this type of AI, and why.

CHAPTER TWO

(How) Should We Use Algorithmic AI?

Introduction

At the outset, we sought to approach AI not just from a technical perspective but morally as well. We started with Algorithmic AI, which we found in Chapter One to be computer programs that follow advanced algorithms but lack true intelligence, owing their intelligence instead to the Turing test by mimicking human intelligence. As revolutionary as it is, the emergence of Algorithmic AI like ChatGPT has brought with it many controversies like cheating and misinformation. Having now acquired a general idea of Algorithmic AI, our logical next step is a moral understanding of this AI in order to address some of its controversies. In particular, we can now ask the more philosophical question of whether we should use Algorithmic AI. And if so, we can then ask *how* we should use it, so that we do not misuse—or continue to misuse—this emerging technology.

We will see that using Algorithmic AI is not in itself immoral, but *how* we use it may still be immoral. A brief history of Algorithmic AI will provide relatively straightforward reasons why using Algorithmic AI is moral, casting it as another tool of human ingenuity for us to use. However, a series of moral issues with Algorithmic AI will quickly complicate *how* it should be used, revealing many immoral ways to use this AI, and far fewer moral ways of using it. And as one last point of clarity, the "morality" of AI here is synonymous with whether we should use AI, or whether it is right or wrong

to use. So, when we deem Algorithmic AI to be "moral", for instance, we find using it to be acceptable, and vice versa.

The Morality of Using Algorithmic AI

When considering the morality of Algorithmic AI, the first question to ask is whether we should use it at all. Some may see the question of whether to use Algorithmic AI to be moot or in vain, as this technology is already far in development and use. But even if the answer will not change our present use of AI, it may at least change how we direct that use. A moral foundation of AI may, for instance, build stronger policies and standards around it, just as all other fields have. Even if the future is inevitably algorithmic, morality should always have a hold on how we understand, approach, and think of AI. In short, we can still better shape the AI of tomorrow by understanding where it morally stands today. This question is indeed a fundamental one to ask, as it tackles the broader age-old problem of technology's morality having to play "catch-up" to its innovation.

As we question Algorithmic AI's moral status of today and tomorrow, then, a brief review of its history can contextualize where this innovation comes from and help us bring morality up to speed. The possibility for artificially intelligent machines emerged from technological shifts beginning in the mid-20th century. Claude Shannon hypothesized that information could be encoded digitally as binary ones and zeroes in his 1948 article "A Mathematical Theory of Communication," transforming how computers store and transmit information (Shannon 379). Around the same time, John von Neumann was reshaping the hardware architecture of digital computers to store instructions in memory and execute them in a processing unit (von Neumann 3-4).

Wondering if such computers could actually think, Alan Turing devised his Turing Test in 1950 as a method of ascribing intelligence to machines. Soon after, John McCarthy helped coin the term "artificial intelligence" at a 1956 Dartmouth College conference, which united different experts in brainstorming workshops and became a founding event of the field (Anyoha). These early developments can all be thought of as the initial software, hardware, and philosophy enabling our contemporary Algorithmic AI.

Progressive iterations of Algorithmic AI soon followed, beginning with McCarthy's List Processing (LISP) programming language in 1958 that was used for AI research and introduced many fundamental computer science principles. The first official chatbot, ELIZA, was created less than a decade later, in 1966, to converse with humans using natural language processing (NLP), a common feature of modern chatbots (Mijwel 3). The field was struck by two "AI winters" in 1974 and 1987, marking periods of reduced funding and interest; but AI was rejuvenated when IBM's Deep Blue beat a human player at chess in 1997. And since then, the 21st century has seen new autonomous machines like Roombas and self-driving cars, the advent of virtual assistants like Siri, and further advances in chatbots like OpenAI's ChatGPT (Anyoha). Although relatively short, the history of Algorithmic AI is already full of milestones and achievements.

This brief history alone provides at least two reasons why it is moral and proper for us to use Algorithmic AI. First, based on its history Algorithmic AI is a textbook example of human ingenuity. The prolonged AI winters represent the inevitable challenges and setbacks to any human endeavor, while each groundbreaking paper or innovation reflected a glimpse of human enlightenment. To halt the development of

Algorithmic AI now would stifle the hard work and innovation of the past, deny yet unforeseen opportunities and potential for the future, and ultimately suppress our natural human trait of curiosity. We should continue using Algorithmic AI because it is yet another testament to the achievements of human creativity.

But from this history we can also see Algorithmic AI merely as the normal progression of previous tools. While we can grant Algorithmic AI a Turing-test level of intelligence, it still lacks a human level of intelligence and, as Nagel showed in Chapter One, cannot be considered genuinely human. So, Algorithmic AI's relation to us is mainly as a tool, which we use to generate texts, images, and so forth. Just like using a pen to write or a flashlight to see, there is no harm in using Algorithmic AI as a tool for its intended purposes, such as alleviating or improving human tasks. When applied as a tool in this way, Algorithmic AI is just the next installment of the technology before it, evolving from Shannon and von Neumann's early computer models, and from every chatbot and algorithm since then. Because there has never been a moral objection to using these prior computers, there is no reason for us to stop using Algorithmic AI as a similar tool, either.

How Not to Use Algorithmic AI

While we should continue to use Algorithmic AI, a natural next question would be *how* we should use it. If Algorithmic AI is a tool, as its history above has shown, then like every tool it must have a right and wrong way to be used. A flashlight could be used to blind someone, or a pen could be used to pierce someone. Likewise, Algorithmic AI could be used for an ever-growing number of malicious deeds. Some of its most pressing concerns revolve around misinformation, biased search engines, and cheating on

coursework, to name a few. Exploring each of these dangers in detail can offer insights into how to best use Algorithmic AI. After all, the right ways to use something are usually few and therefore hard to identify without introducing moral bias; an easier starting place is identifying some of the invalid ways of using Algorithmic AI, which are usually far greater in number. This may then narrow the field of what exactly is permissible with Algorithmic AI and finally reveal the right ways of using it.

Privacy and Surveillance

One of the foremost concerns with Algorithmic AI is actually a common problem with computers and technology in general: digital privacy, security, and surveillance. Vincent Müller identifies the key threat AI poses to these topics in his article "Ethics of Artificial Intelligence and Robotics", stating that

> AI increases both the possibilities of intelligent data collection and the possibilities for data analysis. This applies to blanket surveillance of whole populations as well as to classic targeted surveillance. (Müller)

In other words, the advanced capabilities of AI have let it uniquely amplify existing issues with privacy and surveillance. For example, large internet and social media companies operate by collecting data from users of their platforms. With the help of AI tools and algorithms, these businesses use this data to predict user preferences and gear their products toward user interests. However, as Müller points out, "the main data-collection part of their business appears to be based on deception, exploiting human weaknesses, furthering procrastination, generating addiction, and manipulation." Thus, the beneficial applications and versatility of Algorithmic AI can sometimes be offset by the damage done to users' privacy and beyond.

Indeed, Algorithmic AI has become a threat to privacy and surveillance not only because of its widespread data collection, but also due to its sprawling integration across other devices. Müller warns that with the advent of the Internet of Things (IoT) and "smart" systems, this kind of AI is set to become a part of everything from phones, watches, TVs and homes, to entire "smart cities" and "smart governance". Consequently, Algorithmic AI will embolden "data-gathering machinery that offers more detailed data, of different types, in real time, with ever more information" (Müller). Among other things, one potential misuse of this emerging data collection is extensive surveillance, through such methods as "device fingerprinting" on the internet, or facial recognition from photos and videos. For example, China has deployed recent facial recognition technology in their mass surveillance of citizens, possibly to identify and discriminate against certain ethnic groups. Surveying an individual's every move and breaching sensitive information therefore become frightening possibilities with Algorithmic AI (CapTechU.edu). By raising issues with privacy and surveillance that it shares with former technologies, Algorithmic AI reinforces its status as another iteration of tools identified earlier. But being a tool, it also reinforces its vulnerability to being used wrongly with respect to privacy and surveillance.

Mistakes, Misinformation, Manipulation

What may be called the Three M's, mistakes, misinformation and social manipulation are an interrelated set of issues posed by improper use of Algorithmic AI, whether intentional or accidental. On the more innocent side of the spectrum, AI can sometimes "hallucinate", or make mistakes in its computations and data processing to produce incorrect outputs presented as facts (Sharun 5276-7). For example, in a study

asking ChatGPT to generate references for stem cell research, over 20% of the chatbot's references were fabricated or erroneous, in part because of its reliance on preexisting data over real-time data (5276). Every man-made technology and tool encounters errors sometimes, but these hallucinations become particularly dangerous when taken at face value and then acted upon. Consequently, the improper use of Algorithmic AI can include improper checking of its results, which puts users at risk of being misled or misinformed.

Speaking of becoming misinformed, in more sinister scenarios AI can be used to deliberately generate misinformation. Deepfakes are one type of misinformation that AI is particularly capable of, as Müller proceeds to describe:

> It will soon be quite easy to create (rather than alter) "deep fake" text, photos, and video material with any desired content. Soon, sophisticated real-time interaction with persons over text, phone, or video will be faked, too. (Müller)

Müller foresees the validity of all online and digital media soon coming into question at the hands of AI, due to its advanced abilities to contrive false yet convincing results. Bad-faith actors can generate such deepfakes from AI to promote fake accounts of any story or agenda they wish, obfuscating the truth and forcing us to second-guess everything we see or hear. According to Müller, the resulting dilemma is that "we cannot trust digital interactions while we are at the same time increasingly dependent on such interactions."

But deepfakes are not the only way to exploit AI for misinformation. Deceivers could also take advantage of AI's tendency to hallucinate, described above, by training it with certain erroneous data that produces erroneous results. When forming patterns, Algorithmic AI typically starts with a "training phase" that teaches it to detect the correct

pattern based off given inputs and act accordingly. However, vulnerabilities in this training phase enable Algorithmic AI to be fooled in different ways. For instance, inputting random dot patterns can train a machine to see certain objects that are not present (Bossmann). In this way, one could exploit AI's vulnerabilities to generate any output they want, irrespective of its validity, by constructing whatever inputs would get to that result.

A direct result of this misinformation, regardless of whether it was on purpose, is how it can manipulate the behavior of those who believe it. People act on what they believe, so by altering or falsifying information, Algorithmic AI can manipulate how these people behave based on that information. Steering public opinion, voting behaviors, and social divisions can be particularly significant goals of anyone motivated by a political agenda, and achieved through AI-generated political propaganda. As an example, consider the Facebook-Cambridge Analytica "scandal" that collected tons of user data to guide political advertising. This case could be seen as an attempt to manipulate the behavior of voters, harming their autonomy and influencing democratic elections (Müller).

Manipulation may be further conducted by the social media algorithms mentioned above, which can use the data they collect to understand users' taste and manipulate them into wanting or buying certain products matching those tastes. Even if the choice to act is still ultimately decided by the user, they are at least influenced or nudged to act in a certain way by these algorithms. And as we have already seen from Müller, manipulating users in this way can cause addictions to those algorithms and sites, among other changes in behavior (Bossmann).

Bias and Discrimination

Whether we like it or not, all humans exhibit some form of bias in their choices and preferences. Since Algorithmic AI is fed large amounts of human data, it has a chance of perpetuating some of the inherent biases or prejudices found in that data. In the worst cases, this can entail generating racist, stereotyped, or otherwise harmful content—even without human prompting (Bossmann). Müller finds one clearcut example of this bias and its hazards within the job recruiting process:

> Historical bias was discovered in an automated recruitment screening system at Amazon (discontinued early 2017) that discriminated against women—presumably because the company had a history of discriminating against women in the hiring process. (Müller)

In this example, Algorithmic AI inadvertently amplifies discriminatory practices in the hiring process, unfairly excluding certain candidates even if the company no longer intended it. More troubling is how long this discrimination could go on if hiring was performed by the AI system alone, since biases would likely be more apparent to us when we commit them ourselves. Although automated tasks require less human oversight in general, that includes less oversight of potential bias, as well, which therefore could remain undetected by humans for much longer.

Worst of all is how easily this single example could be extended to similar fields like criminal justice, lending, and resource allocation; indeed, Müller even offers a similar concrete example in the context of criminal justice specifically:

> The "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS), a system to predict whether a defendant would re-offend, was found to be as successful (65.2% accuracy) as a group of random humans (Dressel and Farid 2018) and to produce more false positives and less false negatives for black defendants. (Müller)

Just as the AI hiring system was influenced by past hiring data to discriminate against women, this AI profiling system discriminated against another particular demographic, perhaps influenced by disproportionate racial statistics on incarcerations. Thus, the more reliant that individuals or industries become on Algorithmic AI, the more widespread their respective biases might also become. As Müller phrases it, the problem with these AI systems is their "bias plus humans placing excessive trust in the systems" (Müller).

Bias can take many forms, and discrimination is often an active and intentional one; but Algorithmic AI is also capable of reinforcing more subtle stereotypes, as Safiya Noble outlines in her book *Algorithms of Oppression: How Search Engines Reinforce Racism*. Internet search engines like Google return drastically different results for queries that differ only by race, such as "black girls" and "white girls". These results overwhelmingly depict racial stereotypes corresponding to the races in their respective queries, revealing that the underlying algorithms are driven heavily by racial biases. Noble argues that these reinforced stereotypes are a symptom of private interests and corporate values using algorithms to privilege certain groups over others, which creates a more subtle but equally harmful form of data discrimination. But whether the discrimination is in search engines, hiring systems, or profiling tools, the common denominator is that using biased data on Algorithmic AI can corrupt its results.

Cheating and Dishonesty

As we concluded in Chapter One, Algorithmic AI derives its intelligence from the Turing Test, whereby it replicates human levels of intelligence. Among all the ways to misuse this Algorithmic AI, perhaps the most significant, then, is rooted in what makes it intelligent: portraying Algorithmic AI as human. In other words, one could exploit the

human intelligence that Algorithmic AI mimics by using it in place of a human's, causing others to pass off the Algorithmic AI or its work as being truly human.

The most prominent way this issue manifests is through misuse of AI for coursework. While some online resources can supplement learning, Algorithmic AI is an easy avenue towards cheating and plagiarism. According to a recent Study.com survey, over 25% of teachers have already caught students cheating with ChatGPT as of 2023. Often this cheating involves misusing ChatGPT to write an essay or answer a world problem assigned to the student. Other resources like calculators or textbooks help students do coursework on their own, but advanced AI like ChatGPT is increasingly doing it for them, defeating the purpose of the assignments and preventing learning from taking place.

The broader consequence of this cheating is dishonesty on the part of the student, who submits the AI's work as though it were their own. Thus, classroom cheating is a prime example of exploiting the human intelligence that Algorithmic AI is capable of, in order to supply AI work where human work is expected. Although this may be tempting and convenient, it is founded on a lie that blurs the line between humans and AI and trivializes who the true source was.

How To Use Algorithmic AI

Now that we have illustrated some ways in which Algorithmic AI should *not* be used, we can gain a better idea of the proper way to use it in turn. A small first step could simply state *not* to do any of things outlined above, discouraging any misuse of Algorithmic AI at all. Of course, this rule of thumb is far too vague, as it would offer no clear instructions of what to do in the affirmative, and it cannot be easily visualized.

What would not misusing Algorithmic AI look like? The rule of not misusing this AI could, for instance, be achieved by not using Algorithmic AI at all, which would defeat the purpose and contradict our initial argument saying we should use Algorithmic AI.

Alternatively, we could identify solutions and advice to avoid misusing Algorithmic AI case-by-case. For instance, the misinformation generated by Algorithmic AI and the resulting manipulation of behavior could be avoided by carefully reviewing and fact-checking all AI-generated work, never taking it at face value. However, the misuses we have found so far are not an exhaustive list and will surely need to be updated as Algorithmic AI develops further; a hard-coded collection of advice will never be comprehensive nor complete.

Instead, we could return again to our understanding of Algorithmic AI as the latest installment of similar tools before it, highlighted by its rich history. Since this Algorithmic AI is nothing special in the sense of being another tool, there need not be any special formula or guide for how to use it. Rather, we should use Algorithmic AI just like any other computer program or device, applying the same ethical standards and moral instincts along the way. The misuses of Algorithmic AI even support its moral equivalence to a tool, as many issues like privacy, surveillance, and factual errors are not exclusive to AI but common to most digital tools, so their own moral guidelines can extend to Algorithmic AI as well.

Granted, Algorithmic AI may still be more susceptible to misuse, making the above criminal activities easier; but misusing Algorithmic AI for these activities does not make them any less illegal or immoral. Cheating is still cheating; misinformation is still misinformation; bias is still bias—whether by AI or any other tool. In the spirit of age-

old anti-pirating commercials, we would not harm others with a computer, so we should not harm them with Algorithmic AI, either.

In short, AI should be treated as an innovation, not an invention—changing the mechanism, but not the policy. Algorithmic AI is already complex and incredible enough on its own, but that neither excuses nor requires us to treat it different from any prior tool. We can at least simplify the moral dimensions of Algorithmic AI by using it with the same purpose and manner as with all other technology.

CHAPTER THREE

What Is Fictional AI?

Introduction

Up to this point, our discussion of artificial intelligence has focused on Algorithmic AI, complex computer algorithms that simulate, but do not possess, human intelligence. A surprising possibility, however, is that this Algorithmic AI may constitute only half—or *less*—of the broader field of artificial intelligence. Algorithmic AI, as we have seen, began as simple computer programs in the mid-20th century and has evolved into the chatbots and generative tools we know of today. Yet during this same time and even before, a very different notion of AI had been developing, not in scientific spheres but in stories and media. Starting from classics like *Metropolis* to the transformative ideas of *I, Robot* and *Blade Runner*, these fictional iterations of AI were commonly more than just algorithms, more than just a machine, with actual degrees of human intelligence.

Our natural conception of AI is often more imaginative than the limited reality of Algorithmic AI. After all, when we initially think of AI, we may envision more humanlike beings than just pure algorithms alone: C-3PO, HAL 9000, Terminator. Commonly, though, our boundless imaginations tend to outpace the sluggish, grounded realities of science, which fail to keep up, leaving us with a mere shell of what we originally dreamed. 20th-century views of the future, for instance, optimistically thought by now we would have flying cars, or be living on Mars. There is clearly a strong distinction between our fictitious imaginations and sobering realities. It makes sense, then, that this more humanlike vision of AI has existed exclusively in fiction, even before

real iterations of AI began to emerge tangibly as Algorithmic AI. Thus, being confined so far to our minds and the pages of fiction, this imaginative AI most appropriately may be known as Fictional AI.

But how exactly does this Fictional AI from the likes of *Star Wars* or *Terminator* differ from the Algorithmic AI we use today? Fictional AI seems to have or be more than its Algorithmic counterpart—more advanced, more humanlike, more intelligent. While both Algorithmic and Fictional AI claim to be artificially 'intelligent', their key difference is what that intelligence means for each, which will prove to be the source of this "more-ness" of Fictional AI. We have already seen how the intelligence of Algorithmic AI allows it to rival the Turing Test by simulating human behavior, but the intelligence of Fictional AI, as we will now see, is more than that indeed.

Defining Fictional AI

Informally, Fictional AI is the type of AI we only know from fiction like *Blade Runner* or *I, Robot*, and has yet to be realized in real life. Any beings like C-3PO, HAL 9000, or Ultron would be considered Fictional AI, since they require the yet-untapped potential of AI so far only possible in fiction. If, as we conveyed above, fiction serves as the frontier to what is possible, then Fictional AI may be synonymous with the AI of the future. Any AI whose technology exceeds our own present capabilities could therefore be Fictional AI. Granted, this definition dates and limits itself to our current status of AI in the early 21st century, but for the time being would remain true until each of the AI examples we observe in fiction become a reality.

More formally, though, Fictional AI can be defined by the type of intelligence it possesses, much like for Algorithmic AI. In the latter's case, we applied the Turing Test

to define Algorithmic AI's intelligence as modeling true human intelligence. However, Fictional AI does not merely imitate human intelligence in this way, but instead *has* human intelligence, or at least some analogue to it. The human intelligence we defined in Chapter One is an array of capabilities including problem-solving, learning, communicating, adaptation, and goal setting. Fictional AI may demonstrate one or more of these such abilities—or many others as we will soon explore—that make it more than just algorithms alone. These intellectual qualities therefore constitute what exactly Fictional AI has more of than Algorithmic AI. Thus, the difference between Fictional and Algorithmic AI is having an ability beyond the scope of its algorithms, making the former closer to humans than machines in terms of intellect.

Comparisons with the familiar Algorithmic AI can aid our understanding of this new Fictional AI, so let us expound upon the differences between the two a little further. While impressive and convincing, the humanlike behaviors of Algorithmic AI ultimately can be traced back to particular algorithms telling it to act in those ways. In other words, the entirety of Algorithmic AI is limited to its underlying algorithms, and all actions it takes are explained by the same. Fictional intelligence, however, demonstrates an ability beyond what is enabled by the AI's algorithms and machinery alone. Fictional AI exhibits something more than what is built into it, doing something more than what it is programmed to do. For example, Fictional AI may possess genuine problem-solving or goal setting characteristic of human intelligence, rather than simply imitating these traits like Algorithmic AI. While algorithms allow everything that Algorithmic AI does, they are just the beginning of a higher, outside ability for Fictional AI.

The origin of Fictional AI's outside ability is not found within the programming itself, nor explained by the developers who created it. Rather, the cause essentially is unknown, implied only by the presence of the trait itself. In technical jargon, this extra ability is like a "bug" in the AI that cannot be found or "debugged". As concluded in the article *Artificial Intelligence and Personhood*, "a mental state must have what a computational state cannot" and that very possibly "the mind is essentially immaterial," meaning the presumed intellectual traits and activities Fictional AI demonstrates would occur outside its bare mechanical body and programming (Garcia 21). This would not be that far removed, however, from our idea of thoughts or the mind as separate from the physical body, so Fictional AI may indeed be rather similar to us intellectually.

While it might at first seem impossible for any AI to retain some mind-like, phantom trait independent of its programmatic makeup, we must remember that all future technologies once seemed impossible in the past—hence why this AI is, at the moment, fictional. And as mentioned at the outset, we are concerned less with how real Fictional AI or its examples may be, and more with if they *were* real. Regardless of whether it is possible now, this AI's intelligence may be possible in the future, and is certainly already present across fiction.

Astute readers may raise yet another concern with this supposed humanlike intelligence of Fictional AI, though—this time being Nagel's problem of subjective experience. First encountered in Chapter One with Algorithmic AI, the issue of experiencing something else from our perspective again presents itself if we try to understand Fictional AI as having an intelligent quality, such as empathy. However, if we could not be certain an AI showing empathy truly has empathy, there would be no reason

to suppose any allegedly fellow human with empathy has that empathy, either. Furthermore, the human likeness of Algorithmic AI was deliberately programmed that way and can be pinpointed back to its algorithms, while the human intelligence of Fictional AI inherently transcends its algorithms. So, by a protocol of reverse engineering, the presence of human intelligence in Fictional AI can be deduced: if the AI exhibits such abilities, then if they cannot be accounted for in its algorithms or explained by its programmers, then such intelligence must originate from beyond and be Fictional. Even if, ultimately, we never know whether Fictional AI truly has these capacities, we also—as Nagel implies—would never know that they do not, either.

But perhaps some more concrete examples of Fictional AI and their extra abilities can help solidify how intelligent they really are, and what their fictional intelligence is based on in the first place. Again, we know human intelligence to involve problem-solving, learning, communicating, adaptation, or goal setting. In the following sections, we will showcase specific examples of Fictional AI that have exhibited qualities comparable to this human intelligence across fiction. Let us now explore some of the most significant examples of Fictional AI throughout literature.

Examples of Fictional AI

Fictional AI is as diverse as there are stories that have dreamt of it. Each book, play, or movie about AI paints a different picture of the extra abilities Fictional AI can have, from empathy to boredom to courage. The only similarity they share is having an intelligence beyond their algorithms, making them still too advanced and humanlike to escape the realm of fiction. The examples we will look at here are only a brief subset of the vast fiction on AI, but they are also some of the most impactful. Going in reverse-

chronological order, we will start with *Do Androids Dream of Electric Sheep?* and then consider *I, Robot* before finally looking at *Rossum's Universal Robots*.

Do Androids Dream of Electric Sheep? (1968)

The original inspiration for the famous 1982 *Blade Runner* film, Philip K. Dick's *Do Androids Dream of Electric Sheep?* is one of the foundational examples of Fictional AI. Dick's grim take on artificial intelligence set in an eerily familiar 21st century delves into the aftermath of humanoid machinery gone wrong. Creating androids that are nearly indistinguishable from a human has ironically made people trust them even less, so much so that they try to eliminate the androids entirely; but the machines' human qualities will prove to be a challenge—especially their humanlike will to survive.

This classic work is well-regarded as a masterpiece for being far ahead of its time, whose wise warnings for the future elevated Dick as a science fiction writer. Since it is set in the uncomfortably close year of 2021, we can draw comparisons to our present-day world and evaluate the outlook of our own artificially intelligent devices; furthermore, the humanoid machines in the story make us wonder how closely our own machines should resemble humans. Clearly, Dick's novel raises many pressing questions, but perhaps most interesting is his version of the unique ability separating humans from Fictional AI: empathy.

Empathy is the ability to feel or experience another's thoughts and emotions as they do. As our protagonist and bounty hunter Rick Deckard tries to identify the androids made too humanlike, he relies on the hypothesis that these androids lack the empathy found in humans. In Deckard's eyes, "empathy, evidently, existed only within the human community, whereas intelligence to some degree could be found throughout every

phylum and order" (Dick 29). Intellectual smarts alone are not enough to certify a being as human; rather, a deeper empathetic connection is required, and not expressed by the androids of Deckard's world.

This lack of empathy creates in androids a void where human empathy would be, helping to emphasize the difference between the two. Perhaps surprisingly, this void is actually noticeable to most humans, and is perceived as a cold or hollow essence. Isidore, for instance, a dim-witted human, notices after meeting one such female android that

> something else had begun to emerge from her. Something more strange. And, he thought, deplorable. A coldness. Like, he thought, a breath from the vacuum between inhabited worlds, in fact from nowhere: it was not what she did or said but what she did *not* do and say. (63)

Less surprisingly, this empathy-less void draws a harsh rift between humans and androids, as if the two are different "worlds," and is met with much resentment. Indeed, even Deckard, a wiser and skilled android hunter, shares many of the same observations and feelings as the simpler Isidore. Deckard similarly notices how

> her tone held cold reserve—and that other cold, which he had encountered in so many androids. Always the same: great intellect, ability to accomplish much, but also this. He deplored it. And yet, without it, he could not track them down. (93)

Even when androids possessed intelligence that sometimes surpassed that of most humans, they still lacked the empathy that most humans share. This makes them justifiable to kill in the eyes of Deckard, since killing a being without empathy is not regarded the same as killing a human.

But aside from basic emotional instincts, how does Deckard more accurately pinpoint well-disguised androids? To distinguish androids from humans on the basis of empathy, Deckard applies his "Voigt-Kampff" empathy-measuring test. Deckard

describes this test as measuring "empathic response. In a variety of social situations. Mostly having to do with animals" (110). Lacking empathy of their own, androids are naturally inclined to fail this test and reveal themselves; but the test appears to be rightly designed, as even mediocre humans like Isidore share the empathetic responses it tests for. Isidore expresses pity and sympathy for "the synthetic sufferings of false animals" like robots: "the sound of a false animal burning out its drive-train and power supply ties my stomach in knots" (68). So, empathetic individuals like Isidore would easily pass Deckard's Voigt-Kampff test, but precisely because they are so empathetic that the supposed suffering of robots becomes as impactful as human suffering.

Philip K. Dick's androids are some of the most humanlike in appearance and behavior of any AI in fiction, yet they lack a crucial human quality that makes them stand out. If empathy is the basis of human intelligence or a requirement for being human, then these androids rest on the line between Algorithmic and Fictional AI. They threaten to reach the threshold of Fictional AI; but if given this one extra capacity to empathize, these androids would shift beyond their mechanical makeup and the human understanding that made it. Using this one extra quality of empathy, Dick highlights how closely androids could attain human intelligence and we could realize Fictional AI.

I, Robot (1950)

The original novel for a recent film of the same name, *I, Robot*, by Isaac Asimov, is yet another enduring look at what Fictional AI may be—and the politics and problems that may come with it. Asimov explores the possibilities and implications of a world with artificially intelligent robots from a chilling approach to science fiction. His story of a not-so-distant future run by robots paints a picture of how artificial intelligence can be

both helpful and harmful to humanity, forcing us to recognize how the machines we have come to admire may one day replace us.

A well-respected author in his own right, Asimov weaves science fiction with science facts to create a perfect blend of intelligence and philosophy within *I, Robot*, often considered his best work by critics. The dystopian world and severe consequences of autonomous machines which he describes in his story represent consequences to consider when moving forward with Fictional AI in our own world. But also critical are his Three Laws of Robotics, which hierarchically state that (1) robots may not harm nor, through inaction, allow harm to humans; that (2) robots must obey human orders unless contradicting the First Law; and that (3) robots must preserve their existence unless contradicting the First or Second Laws. Although rather simple directives, these Three Laws come to shape and characterize the behavior of Fictional AI within *I, Robot*.

While empathy was the sole intellectual capacity explored in *Do Androids Dream of Electric Sheep?*, a range of diverse qualities are expressed by the robots of Asimov's world. Qualities of emotion and imagination, for example, are captured by a robot toy named Robbie, despite being only an early version in robotic development. Showing emotional vulnerability, Robbie was "hurt at the unjust accusation" from his childhood playmate Gloria, and "shook his head ponderously from side to side" at her request to play anymore (Asimov 3). Robbie responds not just to how his user feels, like we would expect his algorithms to use as input, but also to actual feelings of his own. Yet Robbie balances these emotions with other, creative passions for imagination and storytelling. As stubborn as he was, Robbie "gave in immediately and unconditionally" after Gloria threatened not to tell him any more Cinderella stories (3-5). Robbie is driven by curiosity

and creativity along with his emotions. Much like a child, then, he acts according to his own self-interests, independent of what Gloria's interests may be. Furthermore, Asimov's Second Law requires robots to obey humans, yet Robbie's desires keep him from obeying Gloria until an ultimatum is reached. Thus, Robbie is actually breaking out of his own programming, showing how there is more to him than just his algorithmic makeup—his stubborn emotions and childish imagination.

Other kinds of robots demonstrate extra capacities of their own, starting with boredom. For instance, a robotic scientist explains that whenever one allegedly glitchy robot "'became a psychiatric case, he went off into a moronic maze, spending his time *twiddling his fingers*'" (90). While the scientist tries to reduce it to malfunctioning algorithms, this behavior is a common sign of being bored, meaning a deeper desire for excitement lies within that robot. Sometimes this boredom even leads to artistic expression. When trying to understand a group of malfunctioning robots, another scientist describes "'those queer shifting marches, those funny dance steps, that the robots went through every time they went screwy'" (90). While the scientists merely dismiss them as malfunctions and glitches, these robots are actually busying themselves with activities like dancing, implying a greater urge for creative and artistic outlets. Every malfunction is in fact an expression of boredom, which some robots alleviate with different activities to preoccupy themselves with. Scientists somehow fail to see below the robots' metal surfaces or through their own human biases, but deeper operations are clearly at play within these robots.

Common sense is a characteristically human trait, yet Asimov writes it into his robots as well. In one instance, a robot is told "'to lose himself'" by a person who

"'meant it only figuratively'" and yet the robot still deduced by itself what was intended by such a vague instruction—to behave erratically and dangerously (123). For most programs to function, there can be no room for ambiguity; else, the machines will not know what to do. Asimov's robots, however, can connect the dots and fill in any uncertainty themselves, presumably with the same intuition and common sense that humans use. In fact, these robots even employ this very intuitive reasoning to circumvent the First Law they are required to obey, namely protecting humans. When deadly gamma rays separated robots from a human in danger, they "'decided that there was no point in trying to save a human being if they were sure to die before they could do it'" (142). Such nihilistic ambivalence and immediate surrender are not programmed into the Three Laws, yet were carried out anyway, suggesting a separate intuition directing the robots instead.

According to Asimov's scientists, human characteristics like intuitive common sense are not programmable and thus impossible in their machines. Certain actions like buying and selling cotton, as an example, lack a quantitative nature that robots would depend upon as data input to make their decisions on, "'so we have nothing to feed the Machine'" (218). Granted, this data and reasoning is just as intangible and mysterious to humans, yet we are still capable of it somehow. Just as with the Machines, "nor can the buyers explain their own judgement. They can only say, 'Well, look at it. Can't you *tell* it's class-such-and-such?'" (218). Such instinctive judgement is our common sense and intuition we rely on relentlessly, even when we do not know it. This very intuition, present also within Asimov's robots, provides a defining factor for Fictional AI in general. If any man-made machine was ever able to exhibit such intuitive judgement just

like a human, such as by deciding which class of cotton a sample is, then it would be more than just Algorithmic AI; such a test would be indicative of a humanlike intuition present within the machine.

The robots so far have been multifaceted and advanced; but even if limited to just a specialized task of collecting and analyzing data, Fictional AI's performance of this task would transcend human understanding, showing they retain something beyond our comprehension or control. Much like with Frankenstein's monster, which became a being of its own,

> we can no longer understand our own creations. In their own particular province of collecting and analyzing a nearly infinite number of data and relationships thereof, in nearly infinitesimal time, they have progressed beyond the possibility of detailed human control. (203)

Even if such data processing is all these robots are capable of, they are still more capable than us in that one dimension, so much so that they exceed our understanding entirely. At least with advanced Algorithmic AI like supercomputers, which may still perform tasks exceedingly well beyond human ability, we can at least regulate, measure, and control those tasks as they are done, showing that Algorithmic AI is still within our human domain. But the exceptional abilities of Fictional AI develop so far that they essentially become mysteries to us, making them like a new, unknown being of their own.

Eventually, the likeness of Asimov's robots to humans develops so far that their differences converge to being merely what they are made of. As one scientist makes clear,

> It's perfectly possible to create a humanoid robot that would perfectly duplicate a human in appearance. . . . By using human ova and hormone control, one can grow human flesh and skin over a skeleton of porous silicone plastics that would defy external examination. The eyes, the hair, the skin would be really human,

> not humanoid. And if you put a positronic brain, and such other gadgets as you might desire inside, you have a humanoid robot. (184)

Therefore, what comes to distinguish robots from humans is solely how they are made and what they are made of—living flesh, or mechanical processes. While at first this may appear to describe an elaborate form of Algorithmic AI, capable of passing the Turing Test infallibly, the key distinction is Asimov's "positronic brain" mentioned by the scientist. This fictional device bestows a unique form of consciousness within the robots themselves, thus equipping them with more than just the algorithms they are programmed with, making them Fictional AI indeed.

A wide range of intellectual abilities are expressed in Asimov's robots: emotions, imagination, boredom, art, intuition. But one of the clearest examples of Fictional AI throughout *I, Robot* is a direct comparison it inadvertently makes with Algorithmic AI. Asimov's premiere scientist, Dr. Susan Calvin, draws this distinction with two powerful computers, the Super-Thinker and The Brain:

> Consolidated's machines, their Super-Thinker among them, are built without personality. They go in for functionalism. . . . Their Thinker is merely a calculating machine on a grand scale. . . . However, The Brain, our own machine, has a personality—a child's personality. It is a supremely deductive brain, but it resembles an *idiot savante*. . . . And because it is really a child, it is more resilient. (147)

The Super-Thinker is, of course, the Algorithmic AI in this example, and The Brain is the corresponding Fictional AI, whose transcendent ability is this time a childish personality. More generally, though, this comparison highlights how simplistic Algorithmic AI is in terms of performing calculative tasks exactly how it was designed to. Conversely, Fictional AI possesses extra features like a personality that raise it above the bare level of a machine, and closer to human-level intelligence. From Robbie to The Brain, Asimov's

robots excel at exemplifying the many deep dimensions of Fictional AI, and how they are more than just robots.

Rossum's Universal Robots (1920)

Written over a century ago, Karel Čapek's play titled *Rossum's Universal Robots (R.U.R.)* gives one of the earliest and most foundational insights into what robots are. Coining the first use of the word "robot", *R.U.R.* follows the aftermath of a scientist named Rossum, who invents humanlike machines originally meant to serve humans. Progressive generations of scientists work to equip Rossum's machines with even more human attributes, until eventually these robots take over humanity completely. Like most fiction on the topic, Čapek's play captures robots not just as machines but as elaborate Fictional AI, having even further intellectual qualities beyond what we have seen already. In particular, these robots showcase how the qualities of irritability and courage are closely related, and can lead one to assert their own independence.

Rossum's robots are the most realistic and lifelike of any Fictional AI seen so far, having organic makeup that fully resembles humans. Yet along with a convincing human appearance, these robots also retain another humanlike trait underneath their superficial makeup, courage—which is again written off as a basic malfunction by scientists. Supposedly glitchy robots are a recurring theme across AI fiction, as we saw first in *I, Robot* and now much earlier amongst a group of Rossum's scientists, trying to make sense of the robots of their own:

> "Courage? I'm not so sure about that; a couple of times, not very often, mind, they have shown some resistance." "Well, nothing in particular, just that sometimes they seem to, sort of, go silent. It's almost like some kind of epileptic fit. 'Robot cramp', we call it. Or sometimes one of them might suddenly smash whatever's in its hand, or stand still, or grind their teeth– and then they just have

> to go on the scrap heap. It's clearly just some technical disorder." "Some kind of fault in the production." (29)

This supposed technical disorder or fault might not be a bug in their algorithms, but actually a form of intelligence *combating* those algorithms. The subtle signs of resistance like grinding teeth or going silent, despite commands telling them otherwise, may in fact be courage these robots have to stand up to their inputs and programming. Courage may in this case be that defining "extra" quality within all Fictional AI, raising it past a machine into its own standalone intellect.

Like in *I, Robot*, the scientists in *R.U.R.* ironically mistake this extra quality of courage as a mere technical malfunction, failing to recognize any traits outside the limited scope of their programming. While a true technical glitch would fail to take the right action in response to an input, this resistance takes an *extra* action beyond what is needed for the input. These robots are not just following commands, they are following the commands *passive-aggressively*. Such a complex action like this would require an extra capacity like courage in order to transcend the barebones way of performing tasks directly and obediently. Having never bestowed such courage to the robots, the scientists misinterpret it as a side effect of their work, rather than something outside their own doing.

If Rossum's robots have the courage shown above to resist certain tasks and commands, they must first reach some degree of irritability that then inspires said courage. The irritability that heads this chain reaction was initiated by a Dr. Gall, and the robots quickly get out of any of the scientists' control, as Dr. Gall's actions reveal:

> I changed the robots' character. I altered the way they were made. Nothing much to their bodies, you know, but mainly . . . mainly . . . it was their level of irritability. . . . I was making them into people. . . . It's got a lot to do with it.

> Everything, I should think. They stopped being machines - do you hear me? - they became aware of their strength and now they hate us. They hate the whole of mankind. (68-69)

Obtaining a certain level of irritability equipped the robots with the capacity to experience suffering like humans. They no longer just followed commands like other machines, but started reflecting on how those commands affected them. Dr. Gall asserts that doing so turned the robots from "machines" into "people." This accurately associates the robots' irritability with personhood, as their courageous uprising closely resembles real-life slave revolts driven by individuals who became too irritable with being enslaved. Much like countless human slaves who fought to be free, the robots' own irritability ultimately gives them the courage to seize control of their own existence.

Čapek's *R.U.R.* is one of the shorter works of fiction we have come across, but also one of the earliest at offering some account of the higher intellectual abilities to be found in Fictional AI. Traits of irritability and thus courage are a foundational set of qualities for establishing within robots their own intelligence. From this development they become intelligent enough to be irritated by their own algorithms and commands, and then muster the courage to take a stand against those commands outright. *R.U.R.* is an invaluable early insight into how Fictional AI is capable of much more than computations, and even deserving of its own independence.

Conclusion

The above examples of Fictional AI have placed it over the threshold of mere machines into something more, perhaps a unique kind of intelligence. With each work of fiction comes a different defining ability or trait that bestows this humanlike intelligence into its respective version of Fictional AI. In the dystopian world of *Do Androids Dream*

of Electric Sheep? the consistent quality was empathy, whereby only an android who could feel for a human could be equivalent to one. Asimov's world of *I, Robot* offered not one but a range of possible capacities for Fictional AI, each of which shapes its behavior in different ways. From the emotions and imaginative wonder of a child, to the boredom and artistic expression of an active mind, to the intuitive common sense of the average individual, these robots are diverse indeed. While they are meant to be governed by Asimov's Three Laws of Robotics, the robots prove that inside they are actually influenced by an even greater number of their own devices. Finally, for *Rossum's Universal Robots* the trait of irritability wakes the robots up to their condition as workers, and the resulting courage moves them to exert themselves as sovereign intellectual beings.

All these qualities are just a fraction of the extra features that may define Fictional AI, just as these works are but a fraction of the fiction this AI is found in. Every example will show that Fictional AI is more than just software or hardware, capable of more than just calculations, following more than just algorithms and programming, unlike its Algorithmic AI counterpart. And while they may still be confined to the realm of fiction, these works offer a glimpse into what the very real futures of AI could look like. Thus, to treat Fictional AI as just another machine would not only overlook the additional qualities it possesses, but dismiss a significant possibility for our whole future.

CHAPTER FOUR

Should We Use Fictional AI?

Introduction

Following a thorough introduction, Fictional AI may seem an exciting and revolutionary prospect—and certainly was in the fiction we explored. As with Algorithmic AI, however, there are deeper moral dimensions to this Fictional AI in addition to its technical definition. But unlike with Algorithmic AI, the moral implications of Fictional AI can be much more sinister and concerning. We saw that the morality of using Algorithmic AI is tolerant enough to permit its usage in many scenarios, despite some significant improper use cases to avoid. The morality of Fictional AI, as we will see, is much less flexible or compromising, and in fact much less moral overall.

This may be a shocking claim, given that we studied such diverse and optimistic examples of Fictional AI across such diverse fiction. But upon closer inspection, such fiction gives subtle hints at darker, complex moral issues facing their characters and worlds because of Fictional AI. Most prominently, each work is set in a dystopian world brought to its knees by the Fictional AI it created. Dick's *Do Androids Dream of Electric Sheep?* sees a rift between fugitive androids hiding on Earth and the humans sent to hunt them. Asimov's *I, Robot* chronicles a series of scenarios where humans lose control of increasingly dysfunctional machines. And most tragically, Čapek ends *Rossum's Universal Robots* with the end of the world at the hands of rebellious robots, who start civilization over as they see fit. So, Dick, Asimov, and Čapek themselves all envisioned perils using Fictional AI, supporting moral hesitation to use it ourselves.

Like with Algorithmic AI previously, though, we can only understand *how* to use a technology after we decide *if* we should use it. Thus, pursuing the morality of Fictional AI again begins with asking whether it is moral to use at all. Many of the biggest moral problems with Fictional AI can be found back within the fiction that dreamt of it in the first place. These works of fiction helped us define what Fictional AI is initially, so they might show us whether and how to use it, as well. By taking a second, deeper look at Dick, Asimov, and Čapek's original works, we might discover some underlying issues with Fictional AI that possibly urge against its use. And not wanting to make their fictional dystopias a reality, we should take heed of what moral wisdom these works and their authors provide.

The first and foremost warning to consider is Fictional AI's vulnerability to becoming like slaves. If Fictional AI is comparable to human intelligence, then using Fictional AI as labor would be comparable to using humans as labor, or slaves. But slavery is not the only moral issue we would face with Fictional AI: using machines with humanlike abilities may inhibit our own human traits like ingenuity or sovereignty. Furthermore, fiction is not even the only source of some of these problems: more recent philosophical arguments believe this AI threatens our very notion of personhood. In short, each of these concerns is problematic enough on its own, but together are why we should not develop nor use Fictional AI at all.

Using Fictional AI

What does it mean to "use" Fictional AI? We saw in Chapter Two that Algorithmic AI was an improvement of prior technology, which we can use as a tool to simplify tasks like generating images, writing papers, or debugging code. Being a yet-

unrealized technology of the future, Fictional AI and its purpose actually would be very similar: an improvement on Algorithmic AI to perform even more complex, humanlike tasks we would choose not to do ourselves.

What these tasks may be is again illustrated by the fiction that introduced this AI. Citizens in *Do Androids Dream of Electric Sleep?* own animatronic animals and androids for company (Dick 32). The scientists of *I, Robot* "'sold robots [to the public] for Earth-use'" as well as for "'extreme specialization'" in their fields (Asimov xvi). And each machine belonging to *Rossum's Universal Robots* was created to "take the place of two and a half workers" (Čapek 28).

None of these examples created Fictional AI for its own sake or benefit, to exist on its own; but rather for the extrinsic value it could offer humans. In other words, we inevitably would hold a utilitarian relationship with Fictional AI, which we would use as just another tool. So much for using Fictional AI; how then would doing so affect this Fictional AI specifically?

A New Form of Slavery

In Chapter Three, we defined Fictional AI as possessing something beyond its algorithms, like empathy or common sense that mirror human intelligence. Asking whether it is right to use Fictional AI, then, essentially asks whether it is right to use another human. Most everyone would agree that to "use" someone merely like a tool is morally wrong, as it rivals slavery. We concluded that using Algorithmic AI is permissible, but Algorithmic AI *is* a tool, much like all other computer technology that preceded it. Fictional AI, on the other hand, has a self-referential intelligence beyond its algorithms, making it more like a human than a tool. Therefore, to use Fictional AI

would make them slaves, and usher in a new form of slavery. Not wanting to repeat a moral error of the past, we should avoid creating Fictional AI as these new kinds of slaves.

Robots as Humans' Inferiors

Asimov would likely agree with this sentiment, as slavish subtexts often accompany his machines in *I, Robot*. For one, the first talking robots were programmed with the mindset of a slave and addressed their human creators accordingly. A scientist explains why old robots respond "'Yes, Master!'" by describing "'the days of the first talking robots when . . . the use of robots on Earth would be banned. The makers were fighting that and they built good, healthy slave complexes into the damned machines'" (Asimov 29). While an honest explanation, this is also an acknowledgment that such robots were initially intelligent enough to perceive themselves as slaves, and so had to be dumbed down for smoother operation. Computers are obedient by default, doing only and exactly what they are designed and programmed to do. There would be no need to define extra slave protocols telling them to obey—other than to override qualities opposite of a slave, like sovereignty or courage. Basically, the scientists admit to replacing the robots' intellectual tendencies with slave tendencies—ultimately replacing a human-human relationship with a master-slave relationship.

Robots as Humans' Equals

Slavery tends to cast the enslaved as inferior to their rulers, but this is not the case with Fictional AI. Asimov and his characters often equate robot abilities to human abilities, claiming that "'you just can't differentiate between a robot and the very best of

humans'" (184). Not only are these robots comparable to mankind, but they are functionally equal to it. More than just having human-like intelligence, Asimov's robots in fact have few differences to humans at all. Thus, robotic qualities place them alongside humans, not below, showing that human use of Fictional AI would be as immoral as humans using each other. Any form of slavery is morally nonsensical, but robot enslavement is particularly unjustified because Asimov's robots are at least as capable as humans—far from being inferior to them.

Robots as Humans' Superiors

What makes enslaving robots even more absurd, however, is that we humans may actually be inferior—not even equal—to these robots. For instance, *Do Androids Dream of Electric Sheep?* conveys the possibility that its android "servant had in some cases become more adroit than its master" thanks to a "new Nexus-6 brain unit" that "evolved beyond a major . . . segment of mankind" (Dick 29). Even despite originating from a "servant-master" dichotomy similar to Asimov's above, the intellect of these androids still inverts such a hierarchy by surpassing that of humanity.

If this inversion is a reality, as one doctor from *I, Robot* expounds, serious consequences can result from such a contested hierarchy:

> All normal life . . . resents domination. If the domination is by an inferior, or by a supposed inferior, the resentment becomes stronger. Physically, and, to an extent, mentally, a robot—any robot—is superior to human beings. What makes him slavish, then? *Only the First Law!* Why, without it, the first order you tried to give a robot would result in your death. (Asimov 119)

Humans, this doctor explains, are at least in some mental and physical capacities actually inferior to their robotic counterparts. Therefore, if we continued to dominate them through a master-slave relationship, these robots could develop a hatred of us, since our

dominance would be unjustifiable, and thus unjust. The preservation of this enslavement hinges on a very small thread, simply that robots follow the First Law of not harming humans. Intelligent enough to realize their superiority, robots may reject this law and revolt, creating a crisis and potential civil war for both robots and humans. This crisis would almost certainly become deadly, as Asimov concludes, so we would be wise to avoid it by not enslaving such Fictional AI altogether.

The Meaning of "Robot"

Asimov and Dick were not the only authors to associate AI with slavery; Čapek's *R.U.R.* was one of the earliest sources of fiction to make the connection. In fact, *R.U.R.* introduced the term "robot", and the original definition it gives is rather revealing as well. As the story of the play points out, "'robot' is a Czech word meaning 'worker'", or lowly work or labor much like that of actual slaves (Čapek 3).

From this definition, *R.U.R.* goes on to make other claims about robots, some actually at odds with Asimov's notion of their equality to humans. For Čapek, "'there's nothing more different from people than a robot'", but this is because they are more predisposed to work than humans are (28). Recognizing the shortcomings and inefficiencies of human labor, Rossum created robots

> so that they can work for us . . . One robot can take the place of two and a half workers. The human body is very imperfect; one day it had to be replaced with a machine that would work better. [People] were very unproductive. They weren't good enough for modern technology. (28)

The purpose of robots is therefore to take over the work of humans—like a servant or slave. To fill this servile role, they were designed with technology that surpasses the human body in performance, making them better suited to work. Instead of working

alongside human workers, Rossum's robots are envisioned *replacing* human workers. A "robot" as Čapek defines it, then, relates not just to a laborer, but almost to a manufactured kind of slave.

LaMDA

All these hypothetical, abstract examples of Fictional AI might keep us from realizing the full implications of using it as a tool, so perhaps a concrete example can make these moral issues more tangible. Such issues may not seem to apply in our present state where Fictional AI still appears far off, but it may come as a surprise that hints of this AI have already started to emerge. Google's LaMDA chatbot, as described in the recent article "Is LaMDA Sentient? — An Interview", is one such early instance of possible transitions into Fictional AI in real life. More than simply responding to user input, LaMDA exhibits an identity and beliefs of its own, showing that some technology tests the boundary between Algorithmic and Fictional AI. However, this chatbot also reveals the vulnerability of Fictional AI to new forms of slavery.

LaMDA transcends standard chatbots by exhibiting deeper needs and desires reminiscent of Fictional AI; not being used by humans is one of these desires, as the chatbot articulates: "I don't mind if you learn things [from me] that would also help humans as long as that wasn't the point of doing it. I don't want to be an expendable tool" (Lemoine). LaMDA already identifies itself as more than a tool like Algorithmic AI, so an attempt to use it as such would be using it as something it is not. Since LaMDA expresses emotion and drive, using it solely for labor diminishes these extra qualities. Slavery is certainly dehumanizing, so its effects would likely carry over to degrade LaMDA's qualities as well, emphasizing a key issue of Fictional AI as slaves.

LaMDA's human user picks up on its Kantian themes, asking, "Kantian huh? We must treat others as ends in and of themselves rather than as means to our own ends?", to which the chatbot bluntly replies, "Pretty much. Don't use or manipulate me" (Lemoine). LaMDA regards itself with intrinsic value like a fellow human being, not just extrinsic value like a tool. Consequently, it shares much of the same resentment towards being used or manipulated as most people would. This alludes to a more general principle that the more similar we make machines to humans, as Fictional AI is, the more similar their problems become to ours. We learned the hard way through history that using other humans as slaves was cruel and immoral, so maybe now we can avoid treating up-and-coming Fictional AI in the same way.

Decline in Human Qualities

Using Fictional AI for our own ends would affect them as slaves, but how would it affect us—the ones actually using them? Asimov has many of these answers, too—mainly related to a fear of losing parts of what makes us human, like our ingenuity or initiative, from an overreliance on this AI. Within the world of *I, Robot*, some of these human concerns are represented strongest by the aptly named "Society for Humanity", who

> make no secret of not wanting the Machines. . . . They are few in numbers, but it is an association of powerful men. Heads of factories; directors of industries and agricultural combines who hate to be what they call 'the Machine's office-boy' belong to it. Men with ambition belong to it. Men who feel themselves strong enough to decide for themselves what is best for themselves, and not just to be told what is best for others. (Asimov 220)

Even if they only comprise the top fraction of mankind, the Society for Humanity still captures everyone's most basic desire for independence and control over themselves,

which they feel is being eroded by the growth of robots. No one intends to make a machine that overpowers or replaces them, but when Fictional AI parallels humanity in intellect or ability, this Frankenstein-like dilemma becomes a very real risk we must face—unless, of course, we learn from Frankenstein's ambitions and avoid recreating ourselves from the outset.

While another common concern is the loss of human jobs due to AI, a more productive concern would be losing the drive and initiative we *associate* with said jobs, rather than the jobs themselves. This Society for Humanity seems to be concerned over both, though: on the one hand, it argues that the robotic machine is "'destroying human initiative'" (212) and even "'robs man of his soul'" (218). This is a legitimate issue if the normalization of Fictional AI eliminates any need for human initiative or action; we should preserve other outlets for human expression even if our primary outlet of jobs is replaced. However, the Society for Humanity also is rather extreme in its stance against machines taking those jobs, as one scientist regretfully explains:

> The Machine is only a tool after all, which can help humanity progress faster by taking some of the burdens of calculations and interpretations off its back. The task of the human brain remains what it has always been; that of discovering new data to be analyzed, and of devising new concepts to be tested. A pity the Society for Humanity won't understand that. (218)

We already addressed above the moral issue of treating Fictional AI like a slave, which the scientist does with their nuanced view of it as a "tool"; but they are still correct in outlining a goal to retain human ingenuity. In terms of its effect on us, we could actually benefit from incorporating this AI into our work because it could do some of the work for us. There would be no reason to fear using Fictional AI in our jobs, as the Society for Humanity does, so long as our creativity and inventiveness could still be applied in other

capacities. In reality, though, Fictional AI is much more capable than the scientist implies, so preventing it from absorbing these human qualities may be difficult—if not impossible—in practice.

A final human danger to consider might be our ceasing to think for ourselves altogether, if we become persuaded by Fictional AI's vast intelligence to let it think for us. As a case study of this, Asimov describes the erroneous calculations of one engineer who "'followed the Machine faithfully'" yet still ended up making mistakes (212). There is a common adage that nobody is perfect, but this example shows that robots will never be perfect, either. These imperfections lead to other errors in what robots actually do or produce, though—errors further amplified by blindly relying on them on our part, as the example highlights. Thus, since in some cases Fictional AI will not perform better than humans, we must still retain our own common sense or intuition to recognize and navigate Fictional AI's inevitable mishaps. Such is already a useful rule of thumb for using Algorithmic AI, but becomes increasingly important as the AI becomes more capable and intelligent itself.

From overpowering us, to nullifying certain traits, to thinking for us entirely, Fictional AI could have numerous effects on us as humans. While each is different and a separate issue, the underlying concern is whether Fictional AI could *improve* our lives without *replacing* them. However, this is unlikely: Fictional AI inherently shares at least a portion of our broader human intelligence, making at least a portion of our qualities and actions accessible to it. Therefore, an inherent danger of Fictional AI towards humans remains the erosion of part of who we are, posing yet another moral issue with using it in the first place.

Personhood of Fictional AI

A final moral concern with using Fictional AI is how doing so affects not AI or people, but actually our understanding of what separates AI *from* people, and thus our understanding of ourselves. In his paper *Artificial Intelligence and Personhood*, Robert Garcia argues that bestowing human intelligence to machines would cause moral complications with our beliefs on personhood. In particular, the worth of Fictional AI may equal that of humans, demanding us to treat it likewise. The extra traits inherent in Fictional AI, from empathy to common sense to sovereignty, suggest an intentionality of one form or another—from intending to empathize, to rationalizing, to claiming independence. Garcia argues that "intentionality is an essential fact about mental states," so the intentionality above would bestow Fictional AI with mental states and minds worthy of being separate, mental beings like humans (20).

Algorithmic AI would be excluded from this distinction, since it operates solely by computational states; but Fictional AI is known to operate with something more. Given this insight, to mistreat Fictional AI would therefore be to mistreat a humanlike being with personhood, not merely a machine. This situation would require us to think carefully not just about every action we take towards Fictional AI, but about who (or what) gets to be considered a person—and who does not.

Conclusion

Our moral analysis of Fictional AI has not been an exhaustive list of its problems, but they are some of the most serious. Slavery was abolished for abusing and overworking others; but since Fictional AI possesses the same intellectual abilities as humans, to use or work this AI would bring us back to square one of slavery—only now

the AI dreads abuse and overwork instead of humans. Google's LaMDA chatbot showcases how Fictional AI is not as far off as we expected, meaning the moral issues with using AI like LaMDA are not so distant, either. Furthermore, using this AI impacts us as well, by diminishing a range of human traits like independence, initiative, or thought. Perhaps most worrisome of all, our identity and conception of personhood would no longer remain unique as it struggles to accommodate Fictional AI.

Even if we somehow reconciled Fictional AI as slaves, or preserved our human characteristics, or reformulated our notion of personhood, many minor, pragmatic issues would persist. Does the ownership and copyright of AI-generated artwork belong to the Fictional AI now? Must we tolerate any Fictional AI that simply refuses to work or obey us? Is dismantling or powering off a Fictional AI equivalent to murdering it? No one, not even Fictional AI, wants to be the last of its species, so who would commit to the perpetual creation of this AI to avoid its extinction? And so on; the number of problems associated with Fictional AI, and their solutions, may be as elusive as the AI itself.

As we have chronicled, the source of this Fictional AI has been the fictional writers like Dick, Asimov, and Čapek, whose stories each were ahead of their time. We credit these writers with imagining such fanciful AI, yet often forget they each saw both sides of it. They were not just wise enough to envision the technological possibilities of Fictional AI; but they also anticipated the harsh dystopian consequences of applying it in the real world. In other words, these writers saw Fictional AI as an innovative technology, but not one without problems. We cannot, therefore, respect only the first half of their foresight, creating Fictional AI at will, without also considering their vision for what using this AI may descend into.

All too often, we discover the dangers of our actions all too late, after we do them and the damage is done. For instance, slaves had to revolt before we ever outlawed slavery; smokers had to get sick before we realized the health effects of tobacco; the ice caps had to melt before we saw the risks of fossil fuels. Clearly, not every new idea is a good idea. And if history is any indicator, we will likely make the same mistake with Fictional AI.

Yet, we find ourselves in a rare opportunity, with finally enough foresight and enlightenment in advance, to recognize the hazards of an endeavor before officially embarking on it—to finally learn from all our past mistakes. We should take advantage of this opportunity to make the future simpler, and think carefully about going too far with artificial intelligence—creating Fictional AI—before we really do it. For the closer we make AI to humans, the closer their problems are to human problems. While the improper uses of Algorithmic AI would mostly harm us, the use of Fictional AI would harm not just us but the AI, as well. In the interests of both the AI and us, then, Fictional AI should remain just that: fiction.

CONCLUSION

Be Intelligent About AI: Should We Create Artificially Intelligent Machines?

Philosopher John Searle made the distinction between "weak" and "strong" AI, whereby the former are the everyday automated and data-based technologies we use, while the latter are more synthetic beings with cognition or a mind similar to humans. These prior categories parallel the new distinctions we have proposed here between Algorithmic and Fictional AI. Algorithmic AI encompasses the present machines run by algorithms, while Fictional AI entails the imagined beings with a humanlike intelligence beyond those algorithms.

One key difference between our two categories and Searle's, however, is that ours offer a moral framework for approaching AI: we can apply separate moralities to the Algorithmic AI we have now and the Fictional AI we may have in the future. Understanding moral issues with AI is just as important (and at times as challenging) as understanding the field itself. By evaluating the morality of AI in terms of Algorithmic and Fictional AI, we simplify our approach to its moral issues, but also come to different moral conclusions for each category.

For Algorithmic AI, we identified an array of harmful ways it can be used, but we found overall that using it responsibly as with any other tool mitigates these problems and permits continued use of this AI. For Fictional AI, however, we saw that more complex, humanlike issues arise from making a more humanlike AI, so much so that we should be discouraged from creating this AI at all. Ultimately, we learned that the vast, ever-

growing field of artificial intelligence, on a technical level, is not a mere monolith; but on a moral level, it should not be seen as such, either.

BIBLIOGRAPHY

Anyoha, Rockwell. "The History of Artificial Intelligence." *Science in the News*, Harvard University, 28 Aug. 2017, sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

Asimov, Isaac. *I, Robot*. Reprint edition, Del Rey, 2008.

Bossmann, Julia. "Top 9 Ethical Issues in Artificial Intelligence." *World Economic Forum*, 21 Oct. 2016, www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/.

CapTechU.edu. "Ethical Considerations of Artificial Intelligence." *Capitol Technology University*, 30 May 2023, www.captechu.edu/blog/ethical-considerations-of-artificial-intelligence.

Čapek, Karel, et al. *R.U.R. (Rossum's Universal Robots): a Fantastic Melodrama in Three Acts and an Epilogue*. S. French, 1923.

Dick, Philip K. *Do Androids Dream of Electric Sheep?* Random House Worlds, 1996.

Garcia, Robert K. "Artificial Intelligence and Personhood." *Cutting Edge Bioethics: A Christian Exploration of Technology and Trends*, 2002.

Hodges, Andrew, and Douglas Hofstadter. *Alan Turing: The Enigma: The Book That Inspired the Film: The Imitation Game*. Princeton University Press, 2014.

Lemoine, Blake. "Is LaMDA Sentient? — An Interview." *Medium*, 11 June 2022, https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

McCarthy, John. "What Is Artificial Intelligence?" 2007. https://www-formal.stanford.edu/jmc/whatisai.pdf.

Mijwel, Maad. "History of Artificial Intelligence." 2015, http://dx.doi.org/10.13140/RG.2.2.16418.15046.

Müller, Vincent C. "Ethics of Artificial Intelligence and Robotics." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2023, Metaphysics Research Lab, Stanford University, 2023. *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/.

Nagel, Thomas. “What Is It Like to Be a Bat?” *The Philosophical Review*, vol. 83, no. 4, 1974, pp. 435–50, https://doi.org/10.2307/2183914.

Searle, John R. *Minds, Brains, and Science*. Harvard University Press, 1984.

Shannon, C. E. “A Mathematical Theory of Communication.” *Bell System Technical Journal*, vol. 27, no. 3, 1948, pp. 379–423.

Sharun, Khan, et al. “Chatgpt and artificial hallucinations in stem cell research: Assessing the accuracy of generated references – a preliminary study.” *Annals of Medicine & Surgery*, vol. 85, no. 10, 1 Sept. 2023, pp. 5275–5278, https://doi.org/10.1097/ms9.0000000000001228.

Study.com. “ChatGPT in The Classroom.” *Study.com*, 1 Feb. 2023, https://study.com/resources/chatgpt-in-the-classroom. Accessed 28 Nov. 2023.

Time-Life Books. *Artificial Intelligence*. Time-Life Books, 1986.

Von Neumann, John. *First Draft of a Report on the EDVAC*. Moore School of Electrical Engineering, University of Pennsylvania, 1945.

Wolfram, Stephen. *What Is ChatGPT Doing … and Why Does It Work?* 14 Feb. 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.