Distribution Agreement

Benjamin P. Goldfein April 6, 2018

Virtuous Artificial Intelligence

by

Benjamin P. Goldfein

Thomas R. Flynn
Adviser

Department of Philosophy

Thomas R. Flynn

Adviser

Cynthia Willett

Committee Member

Andrew M. Kazama

Committee Member

2018

Virtuous Artificial Intelligence

By

Benjamin P. Goldfein

Thomas R. Flynn

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Philosophy

2018

Abstract

Virtuous Artificial Intelligence

By Benjamin P. Goldfein

My aim in this project is to challenge the consequentialist narrative that morally-sentient artificial intelligence (AI) should be machine-utilitarians. I begin by asserting that we must understand how to conceive AI before we can sketch a blueprint of a future with person and non-person morally-sentient beings. I then argue that due to recent explosive advancements in AI, persons' and AIs' moral sentiences will soon be indistinguishable. When this happens, AIs should abide by a system of ethics to ensure the protection of person and non-person morally-sentient beings. Furthermore, I assert that ethical morally-sentient AIs would necessarily follow a system of Aristotelean virtue ethics. Following this ethical system would equip AIs to balance competing obligations through creative 'judgment calls' and corrective *prohairesis*. I conclude by showing how an AI virtue ethicist would be able to learn *prima facie* virtues, thus removing the supererogatory burden off of the programmer to code the 'perfectly-ethical' AI.

Virtuous Artificial Intelligence

By

Benjamin P. Goldfein

Thomas R. Flynn

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Philosophy

2018

Acknowledgements

To Mom, Dad, Shaina, Harry, and Sam.

I love you all very much.

Table of Contents

Foreword

My thesis champions a philosophical approach for examining the socio-ethical implications of artificial intelligence (AI). Namely, I will argue that future with socially-integrated, morally-sentient AI demands AIs to follow a system of Aristotelean virtue ethics. This is because virtue ethics promotes the mutual safety, responsibilities, and interests of person and non-person morally-sentient beings. Moreover, I seek to fill in a gap in the current literature by problematizing the common consequentialist narrative that AIs should be machine-utilitarians. Besides the fact that many programmers are prone to viewing ethics as a oversimplified calculus of 'the right' versus 'the wrong,' the pro-consequentialist attitude clues us into the notion that AI should be utilitarians because programming an AI to abide by a system of consequentialist ethics is computationally easier, and therefore more efficient, than empowering AIs to register, understand, and act in accordance to the situation at hand. However, the most streamlined approach is not always the best approach, especially for questions about ethics. This is because it is impossible "to enumerate all possible situations a superintelligence might find itself in and to specific for each what action it should take."[1] An utilitarian-AI would ultimately be less ethical, and programming one would be impractical, if not impossible. On the other hand, an AI virtue ethicist that is equipped to make 'judgments calls'[2] when faced with competing obligations would hone the capacity to make decisions *prohairesis*,[3] thereby promoting the safety and interests of persons and AIs without sacrificing the individual and group autonomies of either party.

[1] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: UP, 2013), 226.

[2] Noel Tichy, et al, "Making judgment calls," *Harvard Business Review* 85, no. 10 (2007): 94-104.

[3] The term *prohairesis* (Ancient Greek: προαίρεσις) is loosely defined as 'wise choice.' *Prohairesis* is realized when decisions are made at the right time, about the right thing, towards the right subject, for the right end, and in the right way. See: Aristotle, *Nicomachean Ethics,* trans. Terence Irwin (Indianapolis: Hackett Publishing Company, 1999), 1111b26, 1113a15, 1112b15, 1112b26, and 1139a21-b5.

I tackle this problem by splitting my project into three chapters: 1) Defining Artificial Intelligence, 2) Regarding Moral Sentience, and 3) Considering Ethical Systems. I begin by asserting that we must understand what we mean by 'AI' before sketching a blueprint of a future with person and non-person morally-sentient beings. I frame my argument within an in-depth analysis of three types of AI: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Superintelligence (ASI). Underlying this categorization lies the assumption that if an AI has the potential to think in a way that an intelligent person thinks, then that AI dons a person-like moral agency.

I then challenge the argument that donning a moral agency requires being able to think and act autonomously 'like a person' in such a way that the agent is regarded as 'having a conscious.' Specifically, through an analysis of a functionalist theory of mind, I assert that 'having a conscious' is not what makes a being moral. Besides the fact that the term 'conscious' summons unavoidable linguistic connotations which would distract from the current project, there are clearer ways to describe moral agency. Nevertheless, I will still examine these theories to key the reader into how and why I come to my conclusion.

Accordingly, I will posit that a being is regarded as one who dons a moral agency if that being has the ability to reasonably take responsibility under normal circumstances for its actions. I call this the ability 'to do otherwise.' This means that if there is a being that can both 'do otherwise' but did not do otherwise and has the potential to understand the moral implications of 'doing otherwise,' then that being is an agent who dons a 'moral sentience.' This is significant because persons are currently regarded as the only beings who don moral sentiences. However, due to the recent explosive advancements in AI, we can reasonably assume that we will reach the point within the next few decades when persons' and AIs' moral sentiences are indistinguishable. When this happens, what we think it means to don a moral agency will be uprooted.

Finally, I will argue that such a future demands AIs to follow a system of ethics which promotes the mutual safety, responsibilities, and interests of person and non-person morally-sentient beings. In doing so, I will analyze consequentialism and virtue ethics. I will propose that consequentialist ethical systems, while technologically easier to program 'into' an AI, would not guide an AI towards an ethical 'frame of mind.' This is because an AI that abides by a consequentialist ethical model would restrict its decision-making to align with an unpredictable 'utility' of their decided-action's unpredictable consequence. On the other hand, an AI that follows a system of virtue ethics would approach situations within a scope of virtue-directed ethical subjectivity and recognize the finer subtitles of moral decision-making. Ultimately, designing an AI to be a virtue ethicist would simultaneously lift the burden off the programmer to create the 'perfect AI' while also empowering an AI to make a 'judgment call' when it is faced with competing virtues and/or obligations.

My project also serves as a medium for me to challenge the notion that technology research only involves the hard sciences. I do this because I believe that research which concerns humanity must recognize its underlying humanistic elements in order to promote a safer and more inclusive society for ourselves and future generations to come. As to avoid unnecessarily restraining myself here and in future projects, I want to acknowledge that recognizing the humanistic elements of research is not the only factor which advances us towards a more idealistic and utopian society. Alas, such a recognition raises the tensions in a field commonly restricted to the hard sciences and deserves to be taken seriously.

It is important to note that throughout my project, I deliberately call upon thought-experiments, news stories, and personal anecdotes to exemplify my arguments. Crucially, each method of portraying examples serves a particular purpose; the thought-experiments help me illustrate philosophical assertions in a more digestible fashion, the news stories demonstrate the prominency of a variety of ethical issues, and my personal anecdotes show how these issues have impacted me and are highly likely to affect or have affected others, including you and/or your family. That said, I must admit that I do not know the exact method to go about influencing an AI to be a virtue ethicist; that is something I will leave to computer scientists and engineers. Alas, I am neither embarrassed by this nor do I believe this takes away from my project. This is because, as previously stated, AI research cannot be solely imparted unto technology experts; AI is much more complex than the mechanical workings of machinery. My background in philosophy is a strength; my imagination is not limited by what technologists believe to be or not to be plausible, and my academic and intellectual training has equipped me to approach the study of AI ethics within a broader scope instead of by just focusing on one particular AI advancement. Accordingly, I strive to demonstrate that interdisciplinary scholarship enhances our understanding of the world and pushes us towards achieving a more innovative and cohesive society. Above all, I maintain that while philosophy is not directly concerned with the technological workings of AI, philosophical and humanistic research of AI garners imperative and comprehensive relevance now more than ever.

Defining Artificial Intelligence

My aim in the first chapter is to clarify what we mean by 'AI.' This is crucial because acknowledging and dismantling misconstrued definitions of AI will allow us to more fully understand and appreciate the complexity of the topic at hand. The term 'AI' was coined at a computer conference at Dartmouth College in 1956.[4] Since then, the amalgamation of current conceptions of AI have been rooted in pop culture and science fiction references that romanticize certain aspects of AI while dismissing others; films and television shows like *iRobot*, *Black Mirror*, and *Westworld* fail to capture the 'true essence' of AI. That being said, trusted sources also fall into these traps of vagueness and ambiguity. Notably, the Oxford English Dictionary (OED) defines 'AI' as "the capacity of computers or other machines to exhibit or simulate intelligent behaviour."[5] Even though this definition tends towards what I deem to be an appropriate conception of AI, OED's definition is not sensitive enough to the complexities of AI to be sufficient for this project. For one, I struggle to understand what the phrase 'other machines' means. Second, there is a difference between 'exhibiting intelligent behavior' and 'simulating' intelligent behavior; 'exhibiting' behavior could either mean imitating a behavior or behaving in a certain way, while 'simulating' refers to imitating behavior by in large. This leads me to conclude that the simulation of behavior is not sufficient for intelligence. Thus, even if we are able to jump over this linguistic hurdle, it is not clear what it means for behavior to be 'intelligent' and whether there are other beings or things which have the capacity to 'exhibit' or 'simulate' this behavior. It is for these reasons that I will clarify what we mean by 'AI.'

[4] Ray Kurzweil, et al, *The Age of Intelligent Machines* (Cambridge: MIT Press, 1990), 474.

[5] "Artificial intelligence, n." in OED Online. January 2018. Oxford UP.

Three Categories

I will clarify what we mean by 'AI' by categorizing AI into three sectors. The first category is Artificial Narrow Intelligence (ANI), the second is Artificial General Intelligence (AGI), and the third is Artificial Superintelligence (ASI). These terms are mistakenly used interchangeably with terms like 'machine learning,' which signifies an AI's computational capacity "to learn from experience [and] modify its processing on the basis of newly acquired information,"[6] and 'deep learning,' a process in which an AI "attempts to mimic the activity in layers of neurons in the [human] neocortex"[7] by constructing a thinking and learning internal 'neural network' capable of inputting, manipulating, and outputting data in a way indistinguishable to how persons input, manipulate, and output information. Above all, understanding what AI is (and what it is not) will eliminate any confusion about the subject of this paper and will provide the reader with a better understanding of the challenges that lie ahead, ethical or otherwise.

Artificial Narrow Intelligence

ANI refers to machine software that is limited to autonomously solving a single problem or a small cluster of problems. This is a complicated definition that needs to be broken down. First, notice how I use the word 'software,' or an AI's operating information (such as its code), instead of 'hardware,' which refers to the physical apparatus of the AI itself (such as the its metal casing). This suggests that there something beyond the physical construction of a machine that makes it 'an AI.' Notably, if a machine cannot *autonomously*

[6] "Machine learning, *n*." in OED Online. January 2018. Oxford UP.

[7] Robert Hof, "Deep Learning," *MIT Technology Review*, 2013.

'solve a problem'[8] in a way that is 'on par' with the way a person would do it, then that machine is not an ANI. In other words, an ANI is an AI that autonomously demonstrates person-level intelligence "in one or another specialized area."[9] This is best highlighted by way of example. I will first call upon the example of a baseball-pitching machine to illustrate that there are machines which solve problems but are not regarded as ANIs. Then, I will posit that a game-playing AI is different from machines like baseball-pitching machines and should be regarded as ANI.

I played baseball on my Little League team in middle school. My goal was to be the fourth batter, the most coveted spot in the batting line-up. My coaches told me that they would put me as the fourth batter if I was able to improve my batting technique and hitting consistency. Determined to accomplish my goal, I decided that I needed to practice hitting as many baseballs as I could, as often as possible. However, I did not like hitting baseballs off of the batting tee (the black rubber tube that players balance a ball on top of and hit at their own leisure) because the baseballs I would be hitting in games would be moving. I also found it inconvenient to constantly need to fetch the baseballs form the other side of my backyard. That said, having my dad pitch me the baseball underhand was also not ideal because his pitches were not like the pitches of baseball players. To solve this problem, my dad took me to the batting cages and taught me how to use a baseball-pitching machine.

[8] In *Risks of Artificial Intelligence*, Vincent Müller posits that these problems are "practical problems." I find it problematic that Müller makes his assertion without distinguishing 'practical problems' from 'non-practical problems' and 'impractical problems,' assuming such a distinction exists. I assume that 'practical problems' are those that, if solved, would affect one's everyday experience, while non-practical problems would not necessarily have bearing on someone's everyday experience. However, it is still unclear why Müller would make this implicit distinction and neglect to define why such a distinction exists. See: Vincent Müller, *Risks of Artificial Intelligence* (Boca Raton: CRC Press, 2016), 70.

[9] Ben Goertzel, et al, *Artificial General Intelligence* (New York: Springer, 2007), 1.

The mechanics of the machine are quite simple: you put four quarters into the machine's coin slot, load the machine with a few dozen baseballs, and have someone (i.e. my dad) click a button to pitch the ball at a desired speed. From this description it seems that a baseball-pitching machine is an ANI because the baseball-pitching machine solves my practical problem of finding a way for me to practicing hitting baseballs that are being pitched to me without having a trained pitcher pitch them. However, even though the machine can propel a baseball in a way that is on par with how a baseball pitcher would pitch a baseball, a baseball-pitching machine cannot operate unless an agent places a baseball inside of the machine to launch; it is not autonomous. Furthermore, the pitching machine cannot recognize or respond to (external) factors that an ANI would be able to recognize and respond to, such as deciphering where the baseball player is standing and autonomously change the speed at which it catapults the ball. Thus, baseball-pitching machines, and machines like baseball-pitching machines, are not ANIs because ANIs must have the capacity to autonomously solve problems or complete tasks at or above the level of an average intelligent person.

In contrast to contraptions like baseball-pitching machines, game-playing AIs are great examples of ANIs.[10] This is because game-playing AIs have a single function and goal: to autonomously play (and win) a game against a person. The first game-playing AI to use machine-learning to beat its creator was Arthur Troy's checkers program that he created in 1955.[11] A few years later, AI-experts postulated and popularized the opinion that "if one could devise a successful chess machine, [then] one would seem to have penetrated to the

[10] Bostrom, *Superintelligence*, 14.

[11] Arthur Troy, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development 3*, no. 3 (1959): 210-229.

core of [person] intellectual endeavor."[12] This conclusion was reached because chess was seen as the ultimate test of intelligence, for expert chess-playing "requires being able to learn abstract concepts, think clearly about strategy, compose flexible plans, make a wide range of indigenous logical deductions, and… even model ones' opponent's thinking."[13] Alas, this hypothesis was overthrown when programmers realized that chess-playing was like checkers-playing in that it too only required the mastery of recognizing patterns and applying algorithms. This marked the turning point from believing that ANI signified 'true AI' to the introduction of a more complex type of AI: Artificial General Intelligence (AGI).

Artificial General Intelligence

AGI, as the name suggests, refers to an AI that is "endowed with a high degree of *general* intelligence"[14] and "can [autonomously] solve a variety of complex problems in a variety of different domains."[15] Furthermore, an AGI must also be able to learn how to solve new problems that it was not originally programmed to tackle. If an AI can learn to solve "a variety of complex problems in a variety of different domains" but cannot learn how to solve new problems in new domains, then that AI is just an ANI that is capable of (successfully) tackling a multitude of pre-programmed problems. For example, an AI that can play both checkers and chess is just an ANI that is more advanced than its ANI counterparts that can only play one of the two board games.

[12] Allen Newell, et al, "Chess-playing programs and the problem of complexity," *IBM Journal of Research and Development* 2, no. 4 (1958): 320.

[13] Bostrom, 14.

[14] Ibid.

[15] Goertzel, et al, *Artificial General Intelligence*, 1.

On the contrary, an AI that passes Steve Wozniak's 'Coffee Test' would be an AGI. The Wozniak Coffee Test posits than an AI would be an AGI if it "could walk into an unfamiliar house and make a cup of coffee."[16] At first glance, this test seems like it would be simple to pass. In actuality, this test is extremely demanding: the autonomous AI would need to to approach the house, ring the doorbell, explain to the homeowner why it is there, build enough of a rapport to be invited inside, locate the kitchen, find the coffee ingredients, place them into the coffee machine, and make a quality cup of coffee, all while making smalltalk with the homeowner. The reason this test is so demanding is that it requires the agent to seamlessly coordinate a plethora of nuanced inferences (such as discerning a logical place for the coffee grounds to be located) with complex actions (like holding a decent conversation with the homeowner). That being said, this test is particularly unique in that "coffee-making is a task that most 10-year-old [persons] can do reliably with a modicum of experience,"[17] even in an unfamiliar environment. Accordingly, if an AI was able to pass the Wozniak Coffee Test, then it is reasonable to claim that that AI is at least an AGI.

Figures like Elon Musk posit that "AGI will be the most significant technology ever created by [persons]"[18] — and he may be correct. But, I believe that we must consider the possibility of an AI either upgrading itself or creating a new, more powerful AI. Much of the current literature[19] champions a Muskian classification and limits AI to ANIs and AGIs. Such a categorization fails to account for a third type of AI: Artificial Superintelligence (ASI).

[16] Sam Troys, et al, "Mapping the Landscape of Human-Level Artificial General Intelligence," *AI Magazine* Vol 33, no. 1 (Spring 2012): 36.

[17] Troys, et al, "Mapping the Landscape of Human-Level Artificial General Intelligence," 37.

[18] Elon Musk, "About OpenAI," OpenAI.

[19] See: Müller, *Risks of Artificial Intelligence*, 70-72.

Artificial Superintelligence

Coined by Nick Bostrom, the term ‘superintelligence’ (SI) refers to a system whose intellect “greatly exceeds the cognitive performance of [persons] in… all domains.”[20] Accordingly, an ASI would be that superintelligent system. While this definition could serve as a sufficient definition of ASI, this definition is too vague to get us anywhere. Namely, it is unclear what it means for a system to ‘greatly exceed’ a cognitive performance of persons. Does this ‘greatly exceeding’ refer to: 1) the speed at which an AI operates, 2) the aggregation of AGIs whose performances, once combined into a single AI, surpasses that of an intelligent person on all fronts, or 3) an AI whose IQ, assuming AIs can have IQs, would be greater than that of an intelligent person? My answer is that meeting at least one of these criterion is necessary and sufficient for an to AI to be considered an ASI. If an AI meets the first criterion, it would be a Speed ASI. In the same vain, an AI that meets the second criterion would be a Collective ASI, and an AI that meets the third criterion would be a Quality ASI.[21]

Crucially, I struggle to see why Bostrom neglects to extend his explanation of SIs to codify new and important AI classifications by combining these criterion. In other words, it is reasonable to assume that a Speed ASI, for example, could also be a Collective ASI. This might be due to the multiplicity of AIs in a Collective ASI accelerating the performance time of the AI itself, thus demanding it to also be considered an Speed ASI. On the other hand, a Speed ASI might invite the aggregation of multiple AIs into its already-functioning system, thus transforming that Speed ASI into a Collective-Speed ASI.

[20] Bostrom, 26.

[21] My categorization has been adapted from Bostrom’s broader SI categorization. See: 64-70.

This is important for two reasons. First, it illuminates the notion that an AI[22] has the potential to advance and upgrade itself by virtue of being an AI. This is significant because it dismantles the common argument that persons will be able to 'just program an AI to do whatever we want it to do.' Because of this, I believe it is in humanity's best interest to explore ways to teach AIs to follow an ethical system that allows us to hold an AI both causally-responsible and morally-responsible for its actions while also promoting the mutual safety of persons and AI. This brings me to my second point: if an AI has the potential 'to think' or behave in a way that an intelligent person thinks and behaves, then that AI would, in essence, don a person-like moral agency. This conclusion demands us to reconsider what it means to be a moral agent. I will explore this point in the next chapter, Regarding Moral Sentience.

[22] For this point onward, I will be using the terms 'AI' and 'ASI' interchangeably (unless otherwise specified).

Regarding Moral Sentience

British soldier Sean Wiseman was deployed in 2010 to Afghanistan on his 18th birthday.[23] Wiseman and two other soldiers were given orders to drive all-terrain vehicles to scout the Nad-e Ali district of Helmand. As Wiseman was surveying the territory, he set off an improvised explosion device (IED) buried in the ground, which detonated instantly. Miraculously, only his vehicle was compromised; Wiseman walked away from the incident with just bruises and scratches.

Six days later, Wiseman's battalion was ordered to foot patrol. While Wiseman was walking, he set off an IED hidden in a speed bump. Wiseman survived, but the lower portion of his right leg combusted into an oblivion. The unharmed soldiers, who were just outside of the explosion's radius, rushed to Wiseman's aid. One solider stripped his shirt and tied it around the remaining sliver of Wiseman's leg, desperately trying to slow Wiseman's incessant bleeding.

Upon arriving at the hospital, Wiseman was informed by the doctors that he had two options. His first option was to have the remainder of his right leg amputated. His second option was to get a prosthetic leg. The caveat to consenting to the second option is that the prosthetic leg would neither look nor feel like a real human leg, i.e. a natural biological leg.[24] That said, Wiseman would be able to control his prosthetic right leg with his brain in a way identical to how he controlled and manipulated his original right leg. Wiseman opted for the prosthetic. A year later, Wiseman returned to the battlefield as if he had never

[23] This example has been adapted from a recent article in *The Sun*. See: David Willetts, "Hero soldier returns to duty with battalion after losing leg to a Taliban bomb," *The Sun*, March 14, 2017.

[24] This means that a deformed leg is 'real' because a human may reasonably be born with a deformed leg, while a prosthetic leg would not be a 'real human leg' because it is impossible for a human to be born with a prosthetic limb.

encountered or survived an IED. This raises an interesting question: does the fact that Wiseman has a prosthetic limb mean that he is no longer a being in the way you would describe him as a being prior to the explosion?

I suspect that a common answer to this question would go something like this: 'Of course Wiseman is still a being in the way I would describe him as a being prior to the explosion; I described him as a person before the explosion, and I describe him as a person now after the explosion. First, just because Wiseman does not have a real human right leg does not change the fact that he is able to control his new prosthetic right leg by consciously and/or unconsciously sending neuronal signals to and from his brain. Additionally, having a left leg is neither a necessary nor a sufficient quality of living beings. Third, it seems that Wiseman is more than just a living being — he is a person who thinks, acts, and interacts in and with society. That is not to say that the way Wiseman thinks and acts post-denotation *is* synonymous with the way he thought and acted before the explosion or that it *would be* synonymous to the way he might be thinking and acting now had he not come in contact with an IED. Rather, what I stress is that Wiseman did not lose his status of personhood when he lost his leg in the explosion.'

Already we are starting to see the delicate intricacies underlying moral beingness. Importantly, I do not consider 'being a person' and 'being a human' to be synonymous statements; I regard 'personhood' as an extension of 'humanhood.' This is because 'a human' is an organism who classified as a member of the *Homo sapiens* species, while a person is a moral agent, in a human body, who acts and interacts.[25] Accordingly, the above story illuminates why we must consider more than anatomy when defining what it means

[25] "Human, *adj*. and *n*" and "Person, *n*." in OED Online. June 2017. Oxford UP.

to be a person. I will now direct my attention to another true story in order to emphasize how social interactions inform our definition of what it means to be a 'moral agent,' a phrase that is mistakenly used interchangeably with the word 'person.'

When Wiseman encountered both IEDs, a man by the name of Martin Pistorius was watching *Oreo* reruns at his parent's home in South Africa. In fact, Pistorius had been involuntarily watching *Oreo* reruns for the past *twelve years*.[26] Right before his thirteenth birthday, Pistorius had unexpectedly slipped into a coma, "emerging several years later completely paralyzed, unable to communicate with the outside world."[27] The National Institute of Neurological Disorders and Stroke describes this cureless[28] condition as 'total locked-in syndrome,' "a rare neurological disorder characterized by complete paralysis of voluntary muscles. Individuals with locked-in syndrome are conscious and can think and reason, but are unable to speak or move."[29] However, before Pistorius was tested by medical professionals for 'total locked-in syndrome,' no one could figure out whether or not Pistorius was conscious of himself, his surroundings, or his experiences; it appeared to onlookers that Pistorius was in a pure vegetative state — a vegetable.

This raises the debate of whether someone in a vegetative state who we do not know is conscious would be considered to have a sort of moral status. Some might argue that Pistorius was not a moral agent, but I find that argument to be vulgar and dehumanizing; it feels wrong to me to claim that someone in a vegetative state has no moral agency.

[26] Peter Holley, "Meet the man who spent 12 years trapped inside his body watching 'Oreo' reruns," *Washington Post*, January 13, 2015.

[27] Holley, "Meet the man who spent 12 years trapped inside his body watching 'Oreo' reruns," *Washington Post.*

[28] While there is currently no cure for 'total locked-in syndrome,' it does not necessarily mean that this condition is incurable.

[29] "Locked-In Syndrome Information Page," National Institute of Neurological Disorders and Stroke, May 25, 2017.

However, just because it *feels* wrong does not mean it *is* wrong. Ultimately, the examples of Wiseman and Pistorius show us that there is something beyond the anatomical and biological makeup of a person which makes a person[30] a 'moral agent.' Alas, even though persons are moral agents, this does not rule out the possibility of other non-person autonomous moral agents.

Accordingly, my aim in this chapter is to help fill in some of the gaps for determining the qualities that allow us to appropriately refer to a being as one who dons a moral agency. One question philosophers and scientists ask when debating the moral status of an agent is the question of whether or not that agent can think in the way average intelligent person thinks. Another way of saying this is to ask whether or not that agent 'has a conscious.' Underlying this question lies the assumption that if there is an AI 'has a conscious,' then that AI would be a moral agent. This approach is rooted in theories of mind. As we will see, the meaning of 'conscious' is difficult to fully grasp; no philosopher or scientist has yet to determine what exactly we mean by 'conscious.' Alas, exploring whether an AI 'having a mind' makes it so we can reasonably consider that AI to be a moral agent will help the reader understand why I do not believe that relying on theories of mind will best help us understand moral sentience. While there are many theories of mind, such as Cartesianism and folk psychology, I will direct my attention to the theory of person-minds most often extended to the discussion of the concept of AI-minds: functionalism.

Functionalism, a theory presupposed in cognitive science, posits that mental states are defined by their causal (functional) roles, i.e. their inputs (stimuli), outputs (behavior),

[30] I use the word 'persons' instead of 'people' because 'persons' refers to multiple morally-sentient human beings while 'people' refers to a group of persons.

and other inner thought processes (mental states).[31] Accordingly, if functionalism is a sound approach for determining the moral status of an agent and if that agent's actions are defined by its causal roles, then that agent would be considered have the (potential to have the) capacity to think and act in a way identical to a person, i.e. as a moral agent. From this we can conclude that a non-person sentient agent is a moral agent on the basis of that agent having the capacity to think and act in ways similar to how a average intelligent person thinks and acts. This inference is particularly alarming when we consider the moral status of an AI. Namely, according to functionalism, if an AI's actions are defined by its causal roles, then that AI would be considered have the (potential to have the) capacity to think and act as a moral agent. This is significant is because it, if true, uproots the belief that person-agents are the only agents to be moral agents. This opens up the possibility of an AI being regarded as an autonomous moral agent.

Importantly, I want to reiterate that I do not maintain that relying on theories of mind is the optimal way to determine the moral status of beings, whether that being be a person or a machine. Regardless, the discussion of AI-minds merits earnest scholarship and consideration. This is because I believe that I would be doing a disservice to my work if I did not consider one of the most common paths taken to understand AI. By offering what I believe to be a more compelling argument for what we must direct our attention to when considering an AI as a moral agent, I strive to make functionalists question the strength of their position and consider alternate perspectives. Above all, I seek to call for more interdisciplinary AI-ethics research in hopes of positioning us with a running start to better shape a future surrounded and influenced by rapidly accelerating technologies.

[31] Tim Crane, *The Mechanical Mind*, (New York: Routledge, 2016), 194.

In the subsequent section of the chapter, 'Could Have Done Otherwise,' I will examine how donning what I call 'moral sentience,' or ability to have experiences and view these experiences on a spectrum of perfectly-right to perfectly-wrong, is necessary for an agent to be considered as one who dons a moral agency. In other words, thinking and acting like an autonomous person is sufficient but not necessary to be 'morally-sentient.' This means that donning a moral sentience does not have to do with 'having a mind,' but rather relies on the one's capacity to have the potential to have the ability to reasonably take responsibility under normal circumstances for its actions, i.e. the ability 'to do otherwise.' Accordingly, if there is an AI that has the capacity to have the potential to reasonably take responsibility for itself and its actions in the way an average intelligent person does, then that AI would be regarded as a being that dons a moral sentience and moral agency.[32]

Thinking Artificial Intelligence

I will now explore whether an AI 'having a mind' makes it so we can reasonably consider that AI to be a 'moral agent.' In doing so, I will direct my attention to functionalism and argue that functionalism is a flawed approach for theorizing about whether or not an AI has a mind. This is because the causal roles of a mental state, assuming an AI has a mental state, are not sufficient for defining a mental state itself. I will start by providing an functionalist example for theorizing the mind. Importantly, the subject of this example will not be AI. This is because I want to simultaneously clarify how a functionalist may generally approach theorizing the mind while also differentiating functionalism from behavioralism. I will use this as a segue to outline Alan Turing's 'Imitation Game' and argue that Turing

[32] I will explore the implications of this conclusion in the final chapter of my project.

ignores the distinction between "real thought and its mere simulation."[33] Accordingly, I will assert that the mere simulation of person-thought is insufficient for thinking 'like a person.' This is because thinking 'like a person' requires a certain state of understanding and awareness which cannot be programmed into an AI.[34] I will further this position by arguing that a functionally-definable mechanism lacks a subjective character of experience, i.e. a consciousness.[35] This leads me to conclude that theories of mind actually show that a hypothetical thinking-AI would not be thinking 'like a person,' but rather 'like an AI.' However, because there is more to know about mental states than the causal relation between mental states, there is no way for persons to determine, through functionalism, what it is like to think 'like an AI.' Therefore, we must consider more than an AI's functional roles when regarding an AI as donning a moral status.

Functionalism

I will begin by offering an example of how a functionalist might theorize the mind. This will help me clarify what I mean by functionalism and set the stage for the remainder of this section. Imagine that a master chef, Troy, wants to teach an amateur cook, Maxwell, how to make an omelet. Troy removes a cast-iron from the cupboard, turns on the tabletop gas, and places the pan on the stove. Even though Troy insists that the pan is heating up, Maxwell does not believe him. So, when Troy turns around to fetch the eggs, Maxwell touches his palm to the center of the skillet to test if the skillet is the temperature Troy claims it to be. Sure enough, the skillet is scorching hot, and Maxwell blisters his hand. This

[33] Crane, *The Mechanical Mind*, 82.

[34] Searle demonstrates this in example of the 'Chinese Room.'

[35] Thomas Nagel, "What Is It Like to Be a Bat?," in *The Mind's I*, ed Hofstadter and Dennett. (London: Penguin, 1982), 392.

differs from a behaviorist example because the pain Maxwell endures after he naïvely placed his hand on the skillet *is* a mental state. This means that pain is not a response to a mental state, but rather a mental state itself. Crucially, we must note that this is different from a behaviorist example. We know this to be true because behaviorism does not account for mental states. In other words, Maxwell's mental state, pain, is defined by him placing his hand on the skillet, reacting negatively to the consequent searing of his flesh, and quickly removing his hand so he would not further burn himself. Additionally, Maxwell's mental state of pain interacts with other mental states, such as the mental state of grief, and it is the causal interaction with these other mental states that causes Maxwell to retract his hand from the skillet and helps him avoid getting in such unfortunate circumstances again when he encounters other hot surfaces, like an iron.

It is important to recognize that functionalism is neutral about whether or not the mind is material; functionalism does not advocate or deny any sort of physicalism.[36] We know this is true because causal events need not occur in the physical realm in order for one to have mental states about their experiences. For example, let us say Maxwell did not physically burn his hand, but instead was just dreaming about it and woke up in a mental state of grief. A physicalist, or one who holds that everything is ultimately physical (including mental events), would struggle to understand how Maxwell could have a mental state about something he did not physically experience. Functionalism, on the other hand, is not restricted to this conclusion.[37]

The important takeaway is that functionalism argues that mental states are defined by their causal roles. This brings us to the supposition that if an AI's actions are defined by

[36] Nagel, "What Is It Like to Be a Bat?," 400.

[37] This might explain why many functionalists, such as Daniel Dennett, are also physicalists.

its causal roles, and an agent whose actions are intentionally caused by that agent is said to be thinking just as intelligently as a thinking person, then we can conclude that the AI whose actions resemble those of an intelligent person may be said to be thinking just as intelligently as an intelligent person. I recognize that this point is quite extreme. Namely, it raises the question of whether or not an AI can think, i.e. whether the second premise holds. Let us see how Turing tackles this problem.

Turing Test

Turing sought to answer the question 'can [an AI] think?' by developing the 'Turing Test.' The goal of the Turing Test is to see whether or not someone can distinguish a computer from a thinking person. Before I begin, it is important to note that in his original write-up, Turing uses the word 'computer' instead of 'AI.'[38] However, based on Turing's language, we assume that Turing was not just talking about computers in general, but rather a superintelligent AI. Furthermore, the word 'AI' was coined five years after Turing published his paper, so it is reasonable to assume that Turing would have used the word 'AI' had he been familiar with the term. Accordingly, I will replace the word 'computer' with the word 'AI.'

Turing implemented a version of his own test through what he called 'The Imitation Game.' Three people play this game in the first round: a man, a woman, and an interrogator of any sexual and gender orientation. The man sits behind one door and the woman sits behind another, and both are given typewriters which they use to record their responses to the questions asked by the interrogator. The goal of the interrogator is to determine which

[38] See: Alan Turing, "Computing Machinery and Intelligence," in *The Mind's I*, ed. Hofstadter and Dennett. (London: Penguin, 1982), 53-67.

door the man is behind and which door the woman is behind, respectively. The interrogator asks each mystery participant questions, and the participants may try to trick the interrogator if they so please. During the second round, one of the people, let us say the man, unbeknownst by the interrogator, is replaced with an AI. The question is: “will the interrogator decide wrongly as often when the game is played like this [with an AI] as he does when the game is played between a man and woman?”[39] Turing postulated that if the interrogator fails to distinguish the AI from a thinking person, then it holds that the AI is a thinking AI with ‘a mind of its own.’

Notice how I use the phrase ‘fails to distinguish’ instead of the phrase ‘cannot distinguish.’ This wording is crucial; ‘fails to distinguish’ signifies an the interrogator’s unsuccessful attempt to discern the AI from a thinking person, while ‘cannot distinguish’ heralds the interrogators’s inability to discern the AI from a thinking person. In order to for the AI to ‘pass’ the Imitation Game, the interrogator must, at some point, fail to distinguish the AI from the thinking person. This is an unfortunate flaw of the test; if a interrogator gets lucky and randomly guesses that the AI is an AI, then the AI would be said to have failed the Turing Test. A more surefire test would require the AI and interrogator to play the Imitation Game multiple times. The test's administrator would then determine whether or not the interrogator’s answers were either ‘lucky guesses’ or were grounded in reason and chosen purposefully. This leads me to conclude that a more complex AI should possess the ability to ‘pass’ the Imitation Game more times than a less complex AI. This is because a more complex AI would presumably output more nuanced person-like answers at a greater

[39] Turing, “Computing Machinery and Intelligence,” 57.

frequency than its less complex counterparts, and thus would be able to trick the interrogator more times into believing that it is a thinking person.

The Turing Test emphasizes that the knowledges of the person and the AI are determined by their respective causal roles. For example, the AI is regarded as thinking just as intelligently as an intelligent person if that AI can input data and submit answers that are undistinguishable from intelligent-person answers. However, the Turing Test blurs the line between "real thought and its mere simulation."[40] Noticing this flaw is a recognition that the mere simulation of person-thought is insufficient for thinking like the person-mind. Searle sees this pitfall, and argues, through his Chinese Room thought experiment,[41] that an AI which passes the Turing Test "would only be a *simulation* of thinking [thing rather than] the real thing."[42] This highlights a distinction between the concepts of weak AI and strong AI. Weak AI maintains that AI programs are useful tools that help us explain the workings of minds, even thought they are not minds themselves. Strong AI, on the other hand, is the idea that the AI would *be* a mind.

The notion of Strong AI underlies the goal of the Turing Test and hints at two points worth noting. First, it suggests that the concept of the existence of a mind is all there is to thinking. Second, it presumes that the sole function of the mind and its parts is to manipulate uninterpreted symbols.[43] This simple explanation of the mind demands further examination. Accordingly, let us shift our attention to consider how John Searle counters Turing's argument.

[40] Crane, 82.

[41] John Searle, "Minds, Brains, and Programs," in *The Mind's I*, ed. Hofstadter and Dennett. (London: Penguin, 1982), 355.

[42] Crane, 86.

[43] Müller, *Risks of Artificial Intelligence*, 70.

Searle's Chinese Room

Imagine that Maxwell, the amateur cook who burned his hand on the cast-iron skillet, is now a willing participant in the Chinese Room thought experiment. Accordingly, Maxwell is locked inside of a room with two windows, window *I* (for 'input') and window *O* (for 'output'). In the room there is a set of rules which enables Maxwell to correlate one mysterious symbol with another. Someone from outside of the room slips a note through window *I*, which Maxwell picks up. Maxwell flips through the set of rules until he locates the symbol that matches the one on the note which came through window *I*, finds the 'response symbol' that correlates with the initial symbol, prints the response symbol onto a new piece of paper, and deposits the paper with the response symbol through window *O*. In these conditions, the 'Chinese Room' would pass the Turing Test. This is because Maxwell's responses would be indistinguishable from those of a person who is also a fluent Chinese speaker. However, we know that these persons would not be having a conservation with Maxwell because Maxwell does not *understand* what the characters mean. This is because a conversation requires both parties involved to be aware of that which is being conversed. Thus, Maxwell is just manipulating the inputted "uninterpreted formal symbols"[44] and then outputting a pre-planned response.

What is crucial to note is that even though Maxwell does not understand the meaning of the inputted uninterpreted formal symbols, he is able to output the 'correct response' by finding a depiction of the inputted symbol in his notebook and using that to determine the 'correct response.' This brings me to the conclusion that the act of understanding marks an important difference between the simulation of thinking and

[44] Searle, "Minds, Brains, and Programs," 356.

actual thinking, and the Turing Test fails to account for the mental events and procedures which can not be exhibited by functional states. Therefore, an AI that only manipulates uninterpreted symbols is not actually thinking because "form (or syntax) can never constitute, or be sufficient for, meaning."[45]

While I agree with Searle that "running an AI program can never be sufficient for understanding or thought,"[46] we must recognize that Searle's analogy is vulnerable to considerable criticism. One of the stronger arguments against his thought experiment is that it invokes the fallacy of composition. To elaborate, Searle claims that the AI does not understand Chinese because Maxwell, who is locked inside of the Chinese Room, does not understand Chinese in the way a native fluent speaker understands a language. However, this argument is flawed on the basis that Maxwell is only one part of the entire composition of the 'Chinese Room.' This argument is logically equivalent to one in which someone claims I do not understand Chinese because my arm does not understand Chinese; my arm not understanding Chinese has nothing to do with whether or not I as a thinking person understand Chinese. This is because my arm is a part of my body's composition, a part that's function is not to learn languages. So in relation to an AI, Maxwell would only a piece of the AI and not the entire AI itself. Thus, the Chinese Room thought experiment relies on the fallacy of composition.

Even if Searle's analogy worked, it would still be the case that functionally-definable mechanisms (like Searle's Chinese Room) lack consciousness, i.e. an individual character of experience, and therefore would not have 'minds of their own.' Before delving into my argument, I will clarify what we mean by 'consciousness.' Nagel asserts that something has

[45] Crane, 87.
[46] Ibid.

a consciousness “if and only if there is something that it is like to be [a particular] organism, i.e. something it is like for the organism” to be that organism.[47] Nagel does not intend his argument to be about the distinct viewpoints of each individual and the (subjective) character of different experiences. Rather, Nagel aims to offer us a more general subjective point of view of experience and claims that we know a bat has a consciousness because there is something it is like to be a bat.

I would like to point out that ‘what it is like to be a bat’ neither chronicles the experience of a person pretending to be a bat[48] nor describes what it feels like to be a bat.[49] Instead, it underlines that there is something unintelligible that it is like for a bat to be a bat, and the experience of being a bat can only be fully conceptualized by bats themselves. This is because we (persons, presumably) do not experience life from the perspective of a bat. For example, while we understand that bats communicate with each other using sonar signals, it is impossible for humans to fully grasp what it would be like to use sonar. This proves that objective facts, such as that bats use sonar signals to communicate with each other, are subjective experiences which can only be fully understood from certain subjective points of view. Returning to the notion of the Chinese Room, even though the Chinese Room ‘outputted’ the correct symbols, the Chinese Room itself does not have a consciousness because there is nothing that it is inherently ‘like’ to be the Chinese Room. Therefore, because functionalism fails to account for uniquely unquantifiable subjective characters of experience, functionalism does not provide us with a sturdy path for determining the moral status of a thing or being.

[47] I chose Nagel’s definition because I believe it gets us sufficiently close to what we actually mean when we claim a being ‘has a conscious.’ See: Nagel, 392.

[48] 394.

[49] Crane, 169.

Further Thoughts

While I believe that I have made a strong case against functionalism and theories of mind in general, critics may still argue that functionalism is an apt approach for theorizing the mind. This is because functionalism supposes that an AI can "understand and have other cognitive states,"[50] is likely to be true (that is, a Strong AI would be thinking like a human). However, I believe functionalism weakens the argument of AI as a foreseeable futuristic possibility. This is because an AI is unlikely to be thinking like a person *even if* the AI is a thinking agent; Nagel's 'What Is It Like to Be a Bat?' argument supports this point. That being said, there is no way can we actually know this because we are not an AI and because, just as we do not know what it is like to be a bat, we do not know what it is like to be an AI. Regardless, it is a reasonably justified belief to think that the way an AI thinks, if an AI can think, would not be identical to the way a person thinks. Therefore, functionalism does not arm us with the ability to attribute a moral status to an AI because functionalism does not provide us with a complete explanation of the mind.

There are other theories that attempt to provide a complete explanation of the mind as a way to determine whether we can attribute a moral status to an AI. However, for this project, relying on the question of whether an AI can think in the way an average intelligent human can think will not get us anywhere; it is just a red herring. This is because even though we might not be able to say whether or not an AI can think 'like a person,' AI has the potential to do person-like things to an extent identical to persons. Accordingly, I will argue that what more accurately informs us about the moral status of an AI is whether or not that AI 'could have done otherwise.'

[50] Searle, 353.

Could Have Done Otherwise

The 'could have done otherwise' principle maintains that "one has acted freely (and responsibly) only if one could have done otherwise."[51] Accordingly, my aim in this section is to posit that the ability 'to do otherwise' equips a being with a moral sentience. Engrained in my argument query lies notions of casual versus moral responsibility, determinism, indeterminism, and the problem of freewill. I will begin by defining a what we mean by a moral agent and freedom. I will also distinguish causal responsibility from moral responsibility. This will serve as a segue to my argument that determinism is incompatible with freewill because moral agents have the freedom 'to do otherwise' in the sense that they have the opportunity to perform legitimate and possible alternative actions which they could have reasonably performed instead. I will then explore how indeterminism and freewill attribute moral responsibly to an agent after performing its actions. However, critics like Hume suggest that free action "is necessitated, and that it is only because this is so that [agents] can be seen as morally-responsible."[52] This claim is mistaken. By referring back to the distinction between causal and moral responsibility, I will argue that the "causes [of an event may] incline without necessitating"[53] particular consequences. Therefore, in order to claim that the moral agent who is physically-responsible for an action is also morally-responsible for that action, we must consider that agent's intentions in carrying out that action alongside its ability 'to have done otherwise.'

[51] Daniel Dennett, *Elbow Room* (Cambridge: MIT Press, 2015), 131.

[52] Alan Bailey and Dan O'Brien, *Hume's 'Enquiry Concerning Human Understanding': A Reader's Guide.* (Bloomsbury: London, 2006), 84.

[53] John Hospers. *An Introduction to Philosophical Analysis* (Edgewood Cliffs: Prentice-Hall, 1967), 164.

Setting the Stage

To begin, a moral agent is an agent who is morally-sentient. As previously stated, persons are moral agents because they [(have the potential to) have the capacity to] to have the ability to view their actions on a spectrum of virtues and vices. Interestingly, a moral agent's action may be regarded as morally-right or morally-wrong depending on the specific circumstances in which the action and actors are situated. For example, SEAL Team Six assassinating Osama bin Laden is regarded as morally-right because bin Laden and his actions were inherently evil, but Mark David Chapman assassinating John Lennon is regarded as morally-wrong because John Lennon did not do anything that would prompt a reasonable person to push death upon him.

What is particularly important to this project is the idea that moral agents possess (the potential to have) the capacity to reasonably take responsibility for their actions.[54] This highlights a crucial distinction between causal and moral responsibility. For example, let us consider the example of a tornado causing the death of a squirrel.[55] I cannot blame a tornado or any other non-sentient thing or force for causing the death of a squirrel in the same way I can blame a sentient agent for killing a squirrel. This is because a tornado is not aware of its own experiences, even if there are agents that are aware of (the experience of) what it is like to witness or experience the occurrence of tornadoes. One reason we know that a tornado is not aware of its own experiences is that a tornado is not a (living) being or agent. Thus, if I were to blame a tornado for causing the death of a squirrel, what I would

[54] I parenthesize 'the potential to have' to include those agents who is not born as moral agents but may become moral agents through growth and development. An example of this type of agent is a human child.

[55] I use the phrase 'causing the death of' instead of 'killing' due to the connotations I associate with each phrase; the phrase 'causing the death of' bears a non-moral connotation for how the death of the squirrel arose, while the word 'killing' dons what I regard as an immoral connotation similar to that of 'murdering.'

really mean is that the tornado is *causally*-responsible, but not *morally*-responsible, for the death of a squirrel.

At this point it seems that moral responsibility requires the thing or being that performs an action to 'be living.' However, we must recognize that there are various levels of moral responsibility only accessible by certain beings. For example, it would be unreasonable for me to blame my dog, Oreo, for being in the moral-wrong by virtue of her killing a squirrel in the same way that I would blame an adult person hunter for killing the exact same squirrel. This is because Oreo is neither able to comprehend nor will ever have the ability to comprehend conceptions of right and wrong in the ways persons can; Oreo might learn which actions result in rewards versus those which result in punishments from being classically-conditioned,[56] but Oreo does not have the potential or capacity to understand the moral weight of her actions. Therefore, Oreo is a sentient agent, or an agent who is aware of her experiences but will never be able to understand the moral weight of those experiences in relation to herself and others.

There is one more 'level of agency' beyond sentience: 'moral sentience.' Until recently, the only beings which have been postulated to [(have the potential to) have the capacity to] don a moral sentience are persons. However, persons are not born with a fully-developed moral agency. Consider infants and children. Using the example of the squirrel, we are more likely to blame an adult person more than a child for killing a squirrel. This hints at the common phrases 'you cannot blame him — he is just a kid' and 'you are an adult — you should know better.' The fact of the matter is that you *can* blame a child for doing something morally wrong, but the child might not receive the same reprimanding

[56] Oreo received a treat when she does 'good things,' such as rolling on her back, and is scolded for doing 'bad things,' like killing a squirrel.

because the child does not have as developed of a moral sentience compared to that of an adult person. Likewise, adult persons 'should know better' because they have had the opportunity to grow and develop their moral sentiences. This emphasizes that what is important is the child's 'potential' to mature his own moral sentience.

Problem of Freewill

Underlying the concept of moral sentience is the assumption that moral agents are free to act as they wish; we cannot hold agents to a (higher) moral standard if agents are not 'free.' This brings us to the problem of freewill. The problem of freewill asks whether or not an agent can be free. The word 'free' bears multiple connotations. I will not be referring to total and absolute freedom in the sense that one is unrestricted by the laws of nature and other beings to perform specific actions. For example, if I was totally free, then I would never face obstacles when pursuing any action; "all [I would] have to do is will it, and it [would] happen."[57] However, our lived experience is enough evidence to conclude that agents, specifically moral agents, are not absolutely free; freedom is limited. For example, as much as I wish I had this ability, I cannot (currently) just 'will' a stoplight to turn green; the stoplight's 'actions' are out of my reach. I can extend this claim to include other potentially-moral and free agents, such as AIs. This is because even if an AI could will a stoplight to change, it is not currently the case that there is an AI that can simply 'will' any action whatsoever and 'make it happen.'[58] The question remains: how is freedom limited? Moreover, do moral agents even have the freedom to pursue their individually desired

[57] Hospers, *An Introduction to Philosophical Analysis*, 152.

[58] I use the word 'when' instead of 'if' because the development of such an AI is reasonably possible, as evidenced in the previous section.

actions, or this is notion a mere illusion? This is the problem of freewill, and determinism and indeterminism seek to answer to these questions.

Determinism and Indeterminism

Determinists claim that "every event is necessitated by antecedent events and conditions together with the laws of nature."[59] This means that determinism applies notions of causality to agent-actions.[60] This is reminiscent of Hume's argument that the "actions of [moral agents] are just as regular as the mechanistic behavior of the… world"[61] as we know it. This claim loses its strength when a moral agent's action appears to be irregular and unpredictable. Hume combats this opposition by reminding us that we, as moral agents, are not entirely aware of our surroundings in the sense that we do not have access to all of the causal relations between ourselves and environments. These seemingly 'irregular actions' are not irregular, but rather stem from a "secret opposition of contrary causes,"[62] thus proving that person-action is causally necessitated. Unlike determinists, indeterminists argue that not all events have causes and support the case for the existence of freewill. In other words, there are some agent-acted events that are not entirely caused by the past, such as my mom randomly choosing to buy one loaf of bread over another. This demonstrates that there are agent-caused events which are not entirely predictable; "the lack of predictability is inherent in [action] and [is] not only a result of our lack of

[59] Hospers, 152.

[60] Determinism differs from fatalism and predestination; fatalism refers to that which 'is fated' to happen, while predestination refers to that which happens as a result of God's willing.

[61] Bailey and O'Brien, *Hume's*, 84.

[62] David Hume, *An Enquiry Concerning Human Understanding* (Oxford: UP, 2007), 124.

knowledge of the causes."[63] I will further distinguish determinism from indeterminism by using the example of my friend offering me a cup of tea.

Every morning I drink a cup of tea with my housemate Joshua. Because of this, it is reasonably justified for Joshua to predict that if he offers me a cup of tea in the morning, I will accept his offer and drink the tea as per usual. Accordingly, let us say that Joshua offers me a cup of tea, and, as Joshua predicted, I accept his offer. The determinist would say that there were causal factors which necessitated my *accepting* the cup of tea, while the indeterminist would say that there were no causal factors which necessitated my *desire* for the cup of tea. However, if I refused Joshua's offer, the determinist would say that Joshua was ignorant to the causal factors which caused me to refuse his offer and that his knowledge of me usually drinking tea in the morning has no bearing on my freedom to accept or decline the tea. The indeterminist, on the other hand, recognizes that I might purposely try to "flout" his prediction for uncaused reasons. This underscores that there is something more complex about moral agents and their ability to be influenced by external forces; the rising of the sea level will not be affected by my prediction that the sea level will rise when the moon is closer to the earth, but my decision to accept or decline on offer for the cup of tea may be influenced by my knowledge of Joshua's prediction that I will accept the tea. However, the determinist would respond by saying that my "knowledge of [Joshua's] prediction" *is* the causal factor which caused me to refuse the cup of tea.

From this it seems that the argument for determinism holds true. However, this argument is mistaken because "we simply do not exempt someone from blame or praise for an act because we think [one] could do no other."[64] For example, imagine that after I finish

[63] Hospers, 156.

[64] Dennett, *Elbow Room*, 133.

my cup of tea, a two convicted felons break into my apartment. One criminal ties me up, while the other holds Joshua at gunpoint. Instead of killing Joshua, the felons tell Joshua that the only way he can save me is if he goes outside and kills the squirrel from the previous example. Fearing the loss of a friend, Joshua decides to shoot and kill the squirrel. Importantly, a determinist would say that Joshua did not have the freedom to do otherwise because he was coerced into killing the squirrel,[65] and therefore is not morally-responsible for his action.

However, just because Joshua *felt* as if he was coerced into killing the squirrel does not remove Joshua's moral responsibility for killing the squirrel. We know this to be the case for two reasons. First, Joshua was not physically coerced or forced to kill the squirrel. This is because, in this example, the criminal did not put Joshua's finger on a trigger and force him to shoot an innocent animal; Joshua independently shot the squirrel. Second, just because Joshua did not *want* to kill the squirrel does not change the fact that he would still be both causally and morally-responsible for killing the squirrel. So even though Joshua was held at gunpoint, this shows us that determinism is incompatible with the concept of freewill because Joshua 'could have done otherwise' and allowed me to be murdered, even though killing the squirrel seemed to be the only reasonable option. Thus, determinism does not account for an agent's freewill or ability 'to do otherwise.'

Furthermore, determinism does not tell us anything about an agent being moral responsible for its actions. This is because determinism denies the possibility of a system of morals and responsibility on the basis that "whatever *does* happen is the only thing that *can* happen."[66] This conclusion removes the possibility of creativity and scientific discovery. I

[65] Hospers, 160.

[66] Dennett, 144.

will now argue that creativity and scientific discovery further weaken the deterministic argument, strengthen the indeterminist's argument for freewill, and show how indeterminism underscores a moral agent's ability 'to do otherwise.' Let us consider the invention of the iPhone in conjunction with the 'could have done otherwise' principle. If determinism is true, Jobs *had* to invent the iPhone and 'could not have done otherwise.'

Regardless, determinism also holds that the necessitation of causality makes it so agent actions are predictable. But how could we have predicted that Steve Jobs would have invented the iPhone if we did not even know what an iPhone is? The determinist might say that he is not claiming that he can predict the future, but rather that the future is predictable when reflecting on the outcomes and the previous events leading up to the outcome. However, the predictability of the future does not make it so the invention of the iPhone necessitated from past events. This lends credence to the consequence argument, which holds that because "it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are, the consequences of"[67] our actions are not up to moral agents. This underscores that moral agents and their actions can be influenced by their (and others') past actions *without* being necessitated by them. Therefore, instead of inventing the iPhone, Steve Jobs, like other moral agents, 'could have done otherwise.'

Moral Capacity and Potential

Now that I have shown that moral agents are free 'to do otherwise,' I will argue that indeterminism highlights the sense in which the 'could have done otherwise' makes a moral agent is responsible for its actions. In other words, I will assert that a moral agent is

[67] Peter van Inwagen, "The Incompatibility of Freewill and Determinism," in *Free Will*, ed. Watson (Oxford: UP, 1982), 39.

morally-responsible for its actions to the extent that the agent has both the *capacity* to will one action over of an equally plausible alternative action and the *potential* to be in an undistracted, uninhibited, and sober state when it performs actions. I use the word ‘capacity’ to emphasize an agent’s freedom in deciding to perform one action over another, even if it may seem to the agent that the agent has no other viable choice of actions; the example of Joshua and the squirrel illustrates this point. Similarly, I use the word ‘potential’ to emphasize that the moral agent need not be in an undistracted, uninhibited, and sober state to be held morally-responsible for its actions. This is important because if negligence were to lift the burden of moral responsibility off an agent, then a distracted or impaired driver who injured or killed an agent as a result of negligent driving, for example, would not be considered morally-responsible for their actions.

The example of heedless driving holds especially close to my heart. About ten years ago, my aunt was driving my then three-year-old cousin from his school to their house. As she drove through an intersection, someone who was texting and driving ran a red light and hit my aunt’s car. Consequently, my cousin’s seatbelt snapped, and he flew out of his seat and through the left passenger window. Over a decade later, my cousin is still paralyzed.

The story of my cousin further highlights how the notion of ‘could have done otherwise’ differs from deterministic and indeterministic lenses. From the deterministic perspective, the driver ‘could not have done otherwise’ but hit my aunt’s car; crashing into my relatives was his only option. However, claiming that the driver is physically and morally-responsible for running a red light but is only physically responsible for the subsequent events squanders the significance of the accident and how it is portrayed. This is because it suggests that the consequences of an agent’s actions are irrelevant when considering the moral responsibility of the agent *in relation to* the agent's actions themselves. This means that the determinist would suggest that the driver is as morally-responsible as a non-sentient force or thing, like a tornado, would be for paralyzing a toddler. In other words, the driver would not be morally-responsible for his actions. I expect the reader to find this conclusion troubling because it forces us to equate the driver to a non-sentient force that is only regarded as causally-responsible for its actions; the driver does not *don* this physical responsibility, but rather is regarded as being responsible for the actions in virtue of physically causing the event to transpire. This conclusion is mistaken because of freewill and because the driver is a moral agent. Therefore, we know that the driver must carry at least some responsibility beyond causal responsibility for his actions.

By championing an indeterministic position, we see that the driver could have done something other than text on his phone. Notice how I state 'could have done something other than text on his phone' instead of 'could have done something other than paralyze my cousin' or even 'crash into my aunt's car.' I stress this because the reason the driver is morally-responsible for his actions has nothing to do with the fact that the outcome of the series of events 'could have been otherwise' but was not otherwise, such as my cousin not flying through window. Rather, it is because the series of events 'could have been otherwise' in virtue of the driver having the freedom to decide to perform an equally-plausible, alternative action to texting on his phone, such as paying attention to the road or stopping at the red light. This shows us that the driver's decision to not pay attention to his driving was a moral decision, thus suggesting something about his moral character. This is because actions that could have been otherwise *but were not otherwise* carry moral weight.[68] This

[68] Keith D. Wyma. "Moral Responsibility and Leeway for Action," in *American Philosophical Quarterly* 34, no. 1 (1997), 57-70.

means the driver is morally-responsible for both his decision to text and drive, as well as the events which resulted from his decision. Thus, it is an agent's ability 'to do otherwise' combined with that agent's *ability* to decide *not* to do otherwise and *decision* to not do otherwise that renders that agent morally-responsible for its actions.

Applied to Artificial Intelligence

Until this point I have directed most of my attention to exploring the moral agency of person-agents. However, due to the rapid speed of advancements in AI, it is reasonable to assume that there will soon be AIs that will harness the ability 'to do otherwise' and practice *exercising* its ability to do or not do otherwise. This is important because an AI that 'could do otherwise' and is able to understand, practice, and perfect its ability 'to do otherwise' would *be* a moral agent.

To illustrate this, I will now consider what we would make of it if instead a self-driving vehicle crashed into my aunt's car. If this were the case, who (or what) would be morally-responsible for paralyzing my cousin? Readers may argue that the passenger is the one to blame because that self-driving car would never have spiraled into traffic if the passenger, who presumably exercised their autonomy and voluntarily consented to being transported from one location to another in a self-driving car, opted to drive a vehicle himself. However, it seems unreasonable to blame an agent who is not controlling the vehicle itself for the crash *caused* by the vehicle. Here we see that the AI-car would be causally responsible for the damage done to my aunt's vehicle and for injuring my cousin. Conversely, one might attribute the responsibility to the car dealer or manufacturer for selling a faulty vehicle, or even to the AI company that programmed the vehicle in the first place. The important take-away from this paragraph is that the reason the advancement of AI makes it more difficult for us to answer the question of who (or what) is morally-responsible for a tragedy caused by an AI is because we, at this point, have neither outlined what it means for an AI to be 'moral' nor determined what system of ethics a 'moral AI' would necessarily follow.[69]

While such a situation would not have been plausible ten years ago, it is highly likely that there will be self-driving cars on the road within the next few decades.[70] This claim is loosely[71] supported by Moore's Law, which holds that every year new computers will be built to be approximately two times smaller and two times faster.[72] Moreover, I challenge those who deem my hypothesis to be misguided or ill-founded to consider that companies like Waymo, a subsidiary of Google, are beta-testing self-driving cars, while others like General Motors are trying to figure out how to build autonomous self-driving cars that do not have breaks or a steering wheel.[73] These advances are not restricted to self-driving cars; nursing homes in Japan are beginning to introduce AI 'care-bots' for the elderly,[74] there are places in South Korea where you will soon be able to hire robot prostitutes,[75] and Saudi Arabia just granted citizenship to a robot named Sophia.[76] These examples highlight that

[69] This is the subject of my next chapter.

[70] Lee Gomes, "When will Google's self-driving car really be ready? It depends on where you live and what you mean by 'ready,'" *IEEE Spectrum* 53, no. 5 (2016): 13-14.

[71] I say 'loosely' because Moore's Law actually is not a law; it is a theory that has been consistently accurate since 1958. That said, it is still possible for AI to advance at a speed faster than Moore's Law affirms. See: Robert R. Schaller, "Moore's law: past, present and future," *IEEE spectrum* 34, no. 6 (1997): 52-59.

[72] Gordon E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE* 86, no. 1 (1998): 82-85.

[73] Nicholas Shields, "Waymo and GM are far ahead in self-driving car tests," *Business Insider*, February 2, 2018.

[74] Nicola Davies, "Can robots handle your healthcare?," *Engineering & Technology* 11, no. 9 (2016): 58-61.

[75] Patrick Lin, et al, *Robot Ethics* (Cambridge: MIT Press, 2012), 224.

[76] Taylor Hatmaker, "Saudi Arabia bestows citizenship on a robot named Sophia," *TechCrunch*, October 26, 2017.

"as [AI] become[s] more autonomous, it will be plausible to assign responsibility to [an *AI*] *itself*, that is, if [that AI] is able to exhibit enough of the features that typically define personhood,"[77] i.e. moral sentience and the ability 'to do otherwise.' When this happens, an AI that can do otherwise and could have done otherwise but did not *do* otherwise will be regarded as being morally-responsible for its actions.

So in the case of the self-driving car, the onus of causal *and* moral responsibility falls on the car itself. What is of particular interest to this project though is not the idea that self-driving cars donning causal and moral responsibility. Rather, the case of the self-driving car emphasizes that the notion of AIs having rights, responsibilities, and obligations is *not* a science-fiction; if a self-driving car will be developed to the point where they are fully autonomous and superintelligent 'moral machines,'[78] then it reasonably justified to suspect that other AIs will one day be superintelligent 'moral machines' as well. However, I sympathize with the reader who is struggling to balance the delicate juxtaposition of excitement and fright upon realizing that they might very well be living in a time when AIs have responsibilities, rights, and obligations. Thus, to best prepare us for a future where AIs can 'do otherwise' at or above the level of persons, we must take advantage of the fact that AI has not been yet developed to this extent and use this time to determine what system of ethics a moral AI would follow. Once determined, we could hypothetically teach an AI to follow that system of ethics. If we were to neglect conducting this crucial research, then we would risk allowing AI to form its own ethical framework that neglects to promote the safest possible interactions between person and non-person morally-sentient beings. This will be the topic of the final chapter, Considering Ethical Systems.

[77] Lin, et al, *Robot Ethics*, 8.

[78] Paul Bello, et al, "On how to build a moral machine," *Topoi* 32, no. 2 (2013): 251-266.

Considering Ethical Systems

The final chapter of my project encapsulates the idea that in order to promote the mutual safety, responsibilities, and interests of person and non-person morally-sentient beings, AI will need to follow a system of ethics. This counters Silicon Valley's current ethos: "build it first and ask for forgiveness later."[79] In fact, I propose that we should be doing the opposite: consider, formulate, debate, and refine questions about the ethics of AI *now* so we can begin constructing and implementing purposeful ethical solutions *as soon as possible*.

Accordingly, I will consider: 1) whether an AI should [(be programmed to) learn to] follow a system of consequentialist ethics or virtue ethics, and 2) how an AI following that system of ethics contributes to the safety of beings, person, machine, or otherwise. I will assert that programming an AI to follow a system of consequentialist ethics would not produce a truly moral AI. This is because a consequentialist AI would restrict its decision-making criteria to an oversimplified Boolean calculus of trying to define the unpredictable 'utility' of an action's unpredictable consequence(s). Conversely, an AI that follows a system of virtue ethics would be better equipped to approach situations within a scope of virtue-directed ethical relativism and thus recognize the finer subtitles of moral decision-making.

Moreover, beyond removing the unmanageable utilitarian-driven burden off of the programmer to discern the morally-permissible actions and morally-impermissible actions for an AI to take in all possible situations that an AI might face,[80] constructing an AI to be a virtue ethicist would enable it to learn how to balance competing obligations. This will promote safer interactions between different types of morally-sentient beings, as well as

[79] Natasha Singer, "Tech's Ethical 'Dark Side': Harvard, Stanford and Others Want to Address It, *New York Times*, February 12, 2018.

[80] Bostrom, 226.

interactions between morally-sentient and non-morally-sentient beings. This is for two reasons. First, the ability to balance competing obligations allows AI to learn from its mistakes in a more refined way than is accessible to consequentialist decision-making. Second, learning how to direct 'judgment calls' towards an Aristotelean ethical standard simultaneously discourages an AI from inaugurating precarious objectives which would jeopardize the safety of itself and its sentient counterparts while encouraging an AI to *habituate*[81] the fine-tuning of a 'virtuous character.' With that, let us explore what we mean by 'consequentialism' and elaborate on the moral implications of (an AI) following a system of consequentialist ethics.

Consequentialism

In this section I will argue that consequentialism is flawed in that it is not the consequences of an act, but rather the character of the agent that is the test of that act being right or wrong, praiseworthy or blameworthy.[82] My argument relies on the idea that there is a distinction between: 1) what we do versus what we allow to happen,[83] and 2): between what we aim at versus what we foresee as the result of our actions, regardless of the acting agent's intentionality of the actions' consequences. This presupposition hints at a distinction between consequentialist and non-consequentialist theories. Accordingly, I will begin by defining 'consequentialism' and 'non-consequentialism' in regards to how each regards actions as being 'right' or superior to 'less-right' actions. I will then narrow my

[81] Jonathan Lear. *Aristotle: The Desire to Understand* (Cambridge: UP, 2010), 186.

[82] The terms 'praiseworthy' and 'blameworthy' bear Aristotelean connotations. See: Aristotle, *Nicomachean Ethics,* 1109b30.

[83] Philippa Foot, "Morality, Action, and Outcome," in *Morality and Objectivity*, ed. Ted Honderich, (London: Routledge & Kegan Paul, 1985), 23.

focus to the consequentialist theory of utilitarianism and show how utilitarianism, like other consequentialist theories, strives to achieve the 'best' possible outcome for the greatest number of people. While this moral doctrine may seem appealing, I will argue that consequentialism is flawed because it has "implications [which] appear to conflict sharply with some of our most firmly held moral" beliefs.[84]

Critics may counter my argument by claiming that consequentialist theories aim to "minimize evil and maximize good, [or], in other words, to make the world as good a place as possible,"[85] an end which seems to align with out moral beliefs. Furthermore, Amartya Sen attempts to circumvent the non-consequentialist's argument by appealing to the relativity of consequentialism is relative; consequentialism's malleability enables agents to bend certain rules on a case-by-case basis. However, consequentialism also encapsulates a strong doctrine of negative responsibility[86] and is too demanding to be implemented. I will support my case by calling upon the example of self-driving cars. Alas, I want to reiterate that this project is not about self-driving cars; I am calling upon this example to illustrate the need for ethical AI. I will now turn to an analysis of what we mean by 'consequentialism' and 'non-consequentialism,' which will ultimately frame my argument for why a moral AI would abide by a system of Aristotelean virtue ethics.

[84] I will provide examples later in the project. For further research, see: Samuel Scheffler, "Introduction," in *Consequentialism and Its Critics*, ed. Scheffler (Oxford: UP, 1988), 3.

[85] Scheffler, "Introduction," 1.

[86] Bernard Williams, "Consequentialism and Integrity," in *Consequentialism and Its Critics*, ed. Scheffler (Oxford: UP, 1988), 25.

Terminology

I will begin this section by defining 'consequentialism' and 'non-consequentialism.' 'Consequentialism' signifies the notion that the "right act in any given situation is the one that will produce the best overall outcome, as judged from an impersonal standpoint which gives equal weight to the interests of everyone."[87] Notice how consequentialism attempts to answer the question as to what the "right act [is] in any given situation." The term 'right' carries moral baggage, emphasizing that consequentialism is not a theory of matters of fact, but rather one of 'the praiseworthy' and 'the blameworthy.'

This ignites an important distinction between 'rule' and 'act' consequentialism. 'Rule consequentialism' holds that an action's moral permissibility "depends on whether [the act] is required, permitted, or prohibited by a rule whose consequences are best."[88] This means that if everyone observes a general rule when confronted with similar types of situation, then the most favorable consequences will arise, whatever those consequences may be. Gerard J. Williams' example of abiding by the rule to not murder people illustrates the way in which a rule consequentialist would argue against the moral permissibility of murder: "if everyone observes the rule to never directly take the life of an innocent person… in some particular instance, would generate more good than evil consequences."[89] Interestingly, a strict 'act consequentialist,' or one who maintains that the morality of an action is based on the praiseworthiness or blameworthiness of the act itself,' might argue that there are times when murdering someone would be morally-permissible, especially if murdering a particular person would tend towards the greatest good for the greatest number of

[87] Scheffler, 1.

[88] Mark Timmons, "Consequentialism" in *Disputed Moral Issues*, ed. Mark Timmons (Oxford: UP, 2014), 6.

[89] Gerald J. Williams. *A Short Introduction to Ethics* (Lanham: UP of America, 1999), 43.

people.[90] Both of these consequentialist conceptions will provide us with unique insight as to what it means for an AI to be moral.

Consequentialism can be further understood as either indirect or direct. Indirect consequentialism holds that an agent can perform different acts that would result in various consequences. These consequences would be 'ranked' from perfectly-right to perfectly-wrong in terms of the outcomes each act would produce, regardless of the feasibility of an agent performing that action. The second type is 'direct consequentialism.' Direct consequentialism maintains that "the right act… is the one that will produce the highest-ranked [set of consequences] that the agent is in a position to produce."[91] The phrase 'that the agent is in a position to produce' is meant to restrict the list of plausible actions so the agent does not need to consider a myriad of unachievable states of affairs. For example, let us imagine there is man, Oliver, who notices a homeless man sitting on the curb of the sidewalk. Oliver decides he wants to give the homeless man money. If we ignored the formerly-mentioned restriction, one may suppose that the best possible state of affairs is for Parker to give the homeless man millions of dollars. Underlying the consequentialist's argument would be the hope that Oliver doing so maximizes the good. However, Oliver does not have the luxury of having millions of dollars and therefore is not in the position to produce that idealistic state of affairs. This is a prime example of direct consequentialism, which asserts that the 'right action' is "unqualifiedly a maximizing notion."[92] I will use the phrase 'direct consequentialism' and the term 'consequentialism' synonymously, and I will touch upon 'rule' versus 'act' consequentialism when necessary.

[90] This consequentialist example is, more specifically, one of act utilitarianism.

[91] Scheffler, 1.

[92] Ibid.

Now that we understand what we mean by consequentialism, I will distinguish consequentialism from non-consequentialism. Let us refer back to the example of Oliver and the homeless man. In this example, we can assume that a consequentialist would say the best state of affairs would be for Oliver and the homeless man to both have money. This is because both people having money maximizes both of their states of happiness, which is an utilitarian, and therefore consequentialist, ideal; the homeless man would be happy from receiving money, and Oliver would be happy because he made the homeless man happy. Accordingly, the right action for Parker to perform would be to give the homeless man a few dollars. This is a simplified example of consequentialism, and I believe it to be sufficient for our current purposes.

Interestingly, consequentialists and non-consequentialists can both hold that the same act is 'the right act,' depending on the specific circumstances. For example, a non-consequentialist would also maintain that Parker should give money to the homeless man. However, the non-consequentialist's reasoning behind why this action is right would be grounded in the theory that the best state of affairs consists of the morally-right act, regardless of the results of the act itself. Therefore, consequentialism asserts that the right act is the one which is "derived from the goodness of a certain state of affairs,"[93] while non-consequentialism holds that the best state of affairs is when the right action is being performed by virtue of that action being 'the right action.'

Consequentialism also promotes a strong doctrine of negative responsibility. The notion of negative responsibility suggests that one is just as responsible for the actions which they allow as they are responsible for the actions which they fail to prevent.[94] The

[93] Bernard Williams, "Consequentialism and Integrity," 24.

[94] Bernard Williams, 31.

problem with negative responsibility is that it can be taken to an extremist point of view. This emphasizes the overdemandingness problem; consequentialists argue that we should "forget about [the] integrity [of our current situation] in favor of such things a concern for the general good."[95] For example, the fact that I am writing this thesis means that I am not building shelters for starving children in impoverished communities. I will now analyze the rhetoric of an utilitarian consequentialist and show how moral decision-making requires more than is accessible to a consequentialist's 'either/or' calculus.

Utilitarianism

Utilitarianism[96] is a consequentialist ideology which holds that the "best state of affairs from among any set is the one that contains the greatest net balance of aggregate human pleasure or happiness or satisfaction."[97] In other words, utilitarians argue that the 'right act' is the act which results in the greatest good (happiness or satisfaction) for the greatest number of people. This narrative seems hard to resist. After all, who would not want to live in a society where evil is minimized and good is maximized? However, utilitarianism faces three important criticisms.

The first objection is the issue of distribution; utilitarianism does not account for *how* the levels of satisfaction are distributed across the subject population. The seating arrangements on airlines exemplifies this example. Let us imagine there were two options for which airplane an airline would use. The first option is an airline with extremely uncomfortable seats. The second option has extremely comfortable first-class seats (you

[95] 35.

[96] John Stuart Mill, *Utilitarianism*, ed. George Sher (Indianapolis: Hackett Publishing Company, 2001).

[97] Scheffler, 2.

know, the leather ones in the front of the plane that recline), but a little more than half of the passengers will get a seat. Utilitarianism holds that the latter option is the best because more people would be happier, even though the *distribution* of happiness is uneven.

The second criticism is that utilitarianism presupposes that people will do "*whatever* act will, in a given situation, produce the best available outcome."[98] We know this to be flawed because this could mean one must go against their moral values or break the law in order to achieve a certain result. For example, imagine there was a dictator who wanted to kill a hundred captives. The dictator says that he will only kill two captives if one of the captives, Winston, opens fire on his family. Because utilitarianism is impersonal, an utilitarian would say that the greatest good would be achieved by Winston committing multiple accounts of murder, even though committing murder is illegal, even though Winston has a personal connection with his family, even though it goes against Winston's morals to use a gun, and even though it goes against Winston's morals to kill people.

Finally, utilitarianism highlights the overdemandingness of consequentialism, in that consequentialism requires "that one abandon one's own pursuits [or moral principles] whenever one could produce even slightly more good in some other way."[99] We can clarify this point by returning to the example of the dictator. We have already determined that by Winston murdering his family, the dictator will allow ninety-eight out of the hundred captives to live. Now let us say that the dictator gives Winston the option to set one more captive free if and only if Winston kills the other captive. Consequentialism maintains that Winston should pick who he is going to kill and actually kill that person because saving two lives is better than saving one. However, killing goes against Winston's morals, and it is the

[98] 3.

[99] Ibid.

fact that *he* is doing the killing which affects his decision. But, consequentialism does not consider the personal burden one endures when acting. This is not to delegitimize the significance of actions' consequences; rather, I seek to push for the notion that we should not consider the consequences of an action to be "all that count in the sense that can action cannot be called morally right or wrong until all its foreseeable consequences and *only* those foreseeable consequences are considered."[100] Thus, it is not the consequences of an act that test its being right or wrong, but rather the means to reach these ends.

Utilitarian Artificial Intelligence

Now that we understand some of the implications of maintaining a consequentialist ethic, I will analyze the implications of an utilitarian-AI, first in reference to self-driving cars[101] and then AI in general.[102] Namely, I will argue that if an utilitarian-AI vehicle is in a situation where the death of at least one being, person or otherwise, is inevitable as a result of the vehicle's actions, then the utilitarian self-driving car would base its decision about who to save versus who not to save (or who to kill versus who not to kill) on either: 1) the amount of persons that would live versus the amount of persons that would consequently not live, or 2) the perceived 'value' of the persons that would live versus the the value of the persons that would not live. Accordingly, I will analyze each of these scenarios and show why they do not tend towards the most ethical decision. This will guide my argument that moral AIs would follow a system of Aristotelean virtue ethics over one of utilitarianism.

[100] Gerald J. Williams, *A Short Introduction to Ethics*, 43.

[101] Noah J. Goodall, "Machine ethics and automated vehicles," in *Road vehicle automation* (New York: Springer, 2014), 93-102.

[102] Colin Allen, et al, "Prolegomena to any future artificial moral agent," *Journal of Experimental & Theoretical Artificial Intelligence* 12, no. 3 (2000): 251-261.

The case study of the autonomous self-driving car is a plausible offshoot of Foot's Trolley Car Problem.[103] However, there are some major distinctions worth noting.[104] Briefly, the Trolley Car Problem posits a hypothetical situation in which a trolley quickly approaches a fork in the train-tracks. As you are observing the scenery, you notice that five persons (persons *A, B, C, D,* and *E*) are tied down to the tracks, and the train is heading in their direction. However, you are standing next to a lever that, if pulled, will direct the trolley down the alternate track, thus saving persons *A*-*E*; if you decide to do otherwise and not pull the lever, the trolley will plummet the persons, instantly killing them in the process. Your urge to 'do the right thing' combined with the possibility of famed heroism influences your desire to pull the lever. Just as you are about to do so, you realize that another person (Person *F*) is tied down to the other side of the tracks. This means that if you pull the lever, the trolley will change course and run over person *F*. The Trolley Car Problem, in its purest form, raises questions like whether saving five lives is always better than saving one, whether lives have value to the extent that one could justify saving one person instead of five (such as sacrificing the lives of five criminals to save a doctor), and whether a non-action (such as not pulling the lever) is an action. For purposes of this project, I will only concern myself with tackling the first two questions.

Before I jump into my analysis, I would like to point out that there are a plethora of variations of the Trolley Car Problem,[105] from pushing an elephant off of a bridge in an attempt to stop the trolley to changing the names of the persons tied down to the tracks to

[103] David Edmonds, *Would You Kill the Fat Man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong* (Princeton: UP, 2014).

[104] Sven Nyholm, et al. "The ethics of accident-algorithms for self-driving cars: an applied trolley problem?," *Ethical theory and moral practice* 19, no. 5 (2016): 1280.

[105] For more Trolley Car Problem variations, see: Tage Rai, et al, "Moral principles or consumer preferences? Alternative framings of the trolley problem," *Cognitive Science* 34, no. 2 (2010): 311-321.

see if name-stereotypes instigates any sort of decision-bias from the lever-puller. These variations follow the majority of Trolley Car Problem narratives found in today's literature in that they are hypothetical thought-experiments meticulously designed for armchair philosophers to discuss and debate over drinks with friends. And while you can argue back and forth with friends about whether or not to pull the lever, no implications arise from making either decision; the likelihood that anyone would find themselves in a situation where they can pull a lever to save five persons by sacrificing one person is almost zero to none.[106]

However, imagine that instead of there being a trolley car and an observer who can pull the lever that there was a self-driving car that could either: 1) crash into a brick wall, thus causing the death of passenger *F*, to avoid running over pedestrians *A*-*E*, or 2) run over pedestrians *A*-*E*, thus avoiding the brick wall, to keep passenger *F* alive. An utilitarian-AI car faced with this unfortunate dilemma would assess its options and pick the one that tends towards the greatest happiness for the greatest number of beings.

If it elects the first option, it would be basing its decision on the rule-utilitarian notion that saving the most people, regardless of their identities, contributes to the greatest good for the greatest number of beings. This also aligns with Isaac Asimov's first law of robotics, which states that "a robot may not injure a human being or, through inaction, allow a human being to come to harm."[107] However, such a conclusion is problematic because saving the greatest number of people does not necessarily promote the greatest good for the greatest number of beings. For example, what if the pedestrians were violent

[106] If you find yourself in this situation, please let me know; I would be curious as to what you decided to do.

[107] Christopher Grau, "There Is No 'I' in 'Robot': Robots and Utilitarianism," in *Machine Ethics*, ed. Michael Anderson, et al, (New York: Cambridge UP, 2011), 451. Adapted from: Isaac Asimov, *I, Robot* (New York: Gnome Press, 1950).

felons and the passenger was an Emory oncologist on her way to the Winship Cancer Institute to perform live-saving surgeries on her patients? In this situation, the utilitarian would be forced to denounce their claim that the number of lives saved is all that matters when promoting the greatest good for the greatest number of beings.

Accordingly, an utilitarian-AI would opt for running over the five criminals on the basis that the doctor has 'more value' than the totality of the five criminals. Underlying this decision is the idea that saving the doctor is would result in the best long-term *consequence*. But, should the identities of the pedestrians and passenger matter? Even if the identities of the pedestrians and passenger came into play when the AI was making its split-second decision, I struggle to see how an AI would go about determining the 'value' of a being beyond biasing towards those whose values align with its own. This could lead to a racist, misogynistic, ableist, homophobic, and/or xenophobic[108] morally-sentient AI.

Alas, a shrewd utilitarian might try to puzzle me be asking whether or not I would put five non-virtuous pedestrians in danger to secure the safety of one virtuous passenger. To this I would respond that it is not the consequences of an act, but rather given states of affairs in relation to an act that is the test of that act being right or wrong, praiseworthy or blameworthy. This is because a moral agent would necessarily consider not just the ends, but also *the means* to those ends. Accordingly, I maintain that a moral AI would not follow a system of consequentialist ethics. The question remains: what type of ethics would a truly moral AI follow? I will now turn to the final section of this chapter and argue that a moral AI would be a virtue ethicist and would be able to perform virtuous 'judgment calls' when faced with completing moral obligations.

[108] Bernard Williams, "A Critique of Utilitarianism," in *Ethics: Essential Readings in Moral* Theory, ed. George Sher (New York: Routledge, 2012), 257.

Aristotelean Virtue Ethics

My aim in this section is twofold. First, I will outline the theory of virtue ethics by defining what we mean by a virtue, a virtuous agent, and a right action. In doing so, I will argue that virtue ethics clarifies what it means to live a moral life in that it elucidates the relationship between the good of the moral agent and morality itself.[109] This relationship emphasizes virtues as prudentially corrective.[110] This will bring me to a discussion of 'corrective *prohairesis*,' or corrective wise choice. I will argue that virtuous agents, unlike their consequentialist counterparts, are equipped to make particularly well-informed decisions grounded in *prima facie* ethics; their decision-making parameters can be justly altered towards a more ethical means *and* ends.

Second, I will argue that moral agents, specifically moral AIs, should follow a system of Aristotelean virtue ethics. This is because virtue ethics prioritizes the importance of acting by and for virtues rather than the significance of acting by and for perceived consequences. This will make it so an AI acts in and with good character instead of towards a goal that could sacrifice the mutual safety of moral-sentient agents and potentially put (morally-sentient) agents in (catastrophic) risk. Finally, I will argue that designing an AI to follow a system of virtue ethics would simultaneously lift the burden off the programmer "to enumerate all possible situations a superintelligence might find itself in and to specific for each what action it should take"[111] and equip AI with the tools to make, habituate, and perfect 'judgment calls' when faced with competing virtues and/or obligations.

[109] Roger Crisp, *Reasons and the Good* (Clarendon Press; Oxford UP, 2006), 9.
[110] Philippa Foot, *Virtues and Vices* (Oxford: Clarendon Press, 2002), 169.
[111] Nick Bostrom, 226.

Overarching Principles

I will begin by defining 'virtue ethics' and its respective counterparts. To avoid later confusion, I would like to point out that there is a stark difference between 'virtue theory' and 'virtue ethics.' Virtue theory is concerned with virtues writ large, while virtue ethics "is narrower and prescriptive, and consists primarily in the advocacy of virtues."[112] The study of virtue ethics can be traced back to Aristotle's *Nicomachean Ethics*. I believe that Terrance Irwin's translation is the most accessible translation of this pertinent text, and I will be referencing Irwin's translation throughout this section.[113]

Students sometimes limit the scope of Aristotle's definition of virtues to "excellences of the speculative intellect whose domain is theory rather than practice."[114] However, one who only theorizes ways to be virtuous will never grasp a full understanding of virtuosity; the virtuous agent must strive to attain virtuous character "by reflecting on [their] lives and those of others, practicing virtuous behavior, or imitating [virtuous] exemplars"[115] like Buddha. One who has a virtuous character habitually practices what Aristotle denotes as the four cardinal virtues: prudence, courage, temperance, and justice.[116] What makes virtues 'virtues' is that each one is "a mean between two vices, one of excess and one of deficiency."[117] For example, 'courage' the mean between the deficiency of cowardice and the excess of recklessness. The delicate yet mountainous task taken on by the virtuous agent is to determine how that mean is reached. I will later show that, in the case of AI, a virtuous AI

[112] Crisp, *Reasons and the Good*, 5.

[113] See: Aristotle, *Nicomachean Ethics.*

[114] Foot, *Virtues and Vices*, 169.

[115] Lewis Vaughn, "Bioethics and Moral Theories" in *Bioethics: Principles, Issues, and Cases* (New York: Oxford UP, 2017), 44.

[116] Foot, 169.

[117] Aristotle, 1107a3-4.

that is programmed to learn the virtues instead of being preloaded with these virtues is in a better position to practice and perfect acting in and with virtuous character.

We must also remember that the virtuous agent need not benefit from their own virtuous actions. For example, Oliver's charitable action is considered to be a 'moral action' beyond the fact that his action appeals to the virtue of justice. This is because someone who is not a virtue ethicist would most likely consider giving money to charity to be a moral action; an utilitarian would claim that giving money to charity would positively contribute to the greatest good for the greatest amount of people, and a deontologist would argue that it is our duty to give money to charity. Moreover, Oliver's action would still be virtuous even if he does not monetarily benefit from his action. We know this to be true because the fact that Oliver does not monetarily benefit from his action does not discount that he preformed his action with respect to "noninterference and positive service."[118]

Furthermore, virtuous agents must habituate their virtuous character. What I mean by this is that a virtuous agent cannot just perform one action guided by the virtues to be considered a virtuous person; the virtuous agent is one who *habitually* and *regularly* performs and seeks to perfect these actions[119] at the right time, about the right thing, towards the right subject, for the right end, and in the right way. An action that meets these criteria is considered to be 'a right action.' I will not blueprint how the virtuous agent knows when these criteria are met because my attempt would fall short of anything which would do justice to what how virtuous agents go about practicing and perfecting their character. Fortunately, such a task is not pertinent to my project. What is more important to note is that an action that is done by the virtuous agent in accordance with a virtue (or

[118] Foot, 165.
[119] Gerard J. Williams, 79.

virtues) is a moral action, and the virtuous agent is one who embodies these virtues and expresses these virtues through intentional moral decision-making.

Before highlighting the what I believe to be the greatest strength of virtue ethics, I will briefly raise, address, and dismantle some criticisms of the theory. First, critics argue that virtue ethics' weakness "involves the concept of eudaemonia,"[120] which is Greek for 'human flourishing.' Critics claim that eudaemonia[121] is an obscure concept. However, the fact that the concept of eudaemonia is obscure does not detract from the power of the virtue ethics argument. Additionally, critics assert that virtue ethics is trivially circular.[122] This assertion is flawed; virtue ethics "does not specify right action in terms of the virtuous agent in terms of right action," but rather enumerates right actions in terms of *character traits* which *promote* eudaemonia. Furthermore, critics claim that virtue ethics neglects to conjure any principles which constitute what is and is not morally virtuous. This argument ignores the positive instruction of virtues and the negative instruction of vices, and thus can be ignored. A more powerful argument appeals to the culture relativism and states that we cannot certainly define what is a virtue and what is a vice from culture to culture. However, arguing for the presence of cultural relativism in an attempt to obliterate virtue ethics as a way to boost consequentialism unintentionally disenfranchises the utilitarian's agenda. This is because one could subsequently claim that "there has been, for each rule [or virtue], some culture which rejected it."[123] There are obviously more arguments against virtue

[120] Rosalind Hursthouse, "Virtue Theory and Abortion," in *Virtue Ethics*, ed. Roger Crisp (Oxford: UP, 1997), 219.

[121] Aristotle, *Nicomachean Ethics*, 1097a15-b21.

[122] Hursthouse, "Virtue Theory and Abortion," 220.

[123] 222.

ethics,[124] including some from socio-psychological perspectives.[125] However, I believe this brief account to be sufficient for this project. I will now argue that what makes virtue ethics is superior to utilitarianism in that virtue ethics is prudentially corrective.

Corrective Prohairesis

The strength of virtue ethics lies in 'corrective *prohairesis*,' or corrective wise choice. I will first look at what we mean by 'corrective' and then what we mean by '*prohairesis.*' First, virtues are corrective in the sense that "there is some temptation [for virtuous actions] to be resisted or [a] deficiency of motivation [for actions] to be made good."[126] This claim reflects Aristotle's argument that virtues are difficult to achieve, while Foot's assertion lies in the claim that "almost any desire can lead a man to act unjustly."[127] So if any desire can lead a man to act unjustly, any act can be unjust. And one who is said to perform an unjust act is said to lack virtue. However, this suggests that any unjust act could also potentially be just, or virtuous, if done for the right reasons. For example, what makes Oliver's action virtuous and praiseworthy is twofold: 1) it is morally right for Oliver to help those in need if Oliver is in the position where helping those in need will not put him at risk, and 2) there is a "deficiency in motivation" to help those in need.[128] If Oliver performed a non-virtuous act and did not give money to charity, his act could be 'corrected' in the sense that he could donate money to charity at another opportunity. Notice I used the word 'non-

[124] Simon Keller argues that virtue ethics is self-effacing. See: Simon Keller, "Virtue ethics is self-effacing," *Australasian Journal of Philosophy* 85, no. 2 (2007): 221-231. G.E.M. Anscombe also raises important criticisms. See: Gertrude Elizabeth Margaret Anscombe, "Modern moral philosophy," *Philosophy* 33, no. 124 (1958): 1-19.

[125] See: John M. Doris, "Persons, situations, and virtue ethics," *Nous* 32, no. 4 (1998): 504-530.

[126] Foot, 169.

[127] Foot, 169.

[128] 170.

virtuous' instead of 'vice'; Oliver not giving to charity is not an act of malice, but rather out of disregard to those in need. This emphasizes the importance of intentionality and ignorance.[129] Unlike utilitarianism, virtue ethics considers more than brute facts; virtue ethics allows us to expand the breadth of questions of morality.

Finally, the intentionality of these actions must be guided by *prohairesis*,[130] or wise choice. *Prohairesis* differs from *hairesis* in that *prohairesis* signifies an understanding of the state of affairs prior to an act, a calculated prediction of the given state of affairs after that act, a 'rational desire' for a end and a means towards that end which promote 'the good,'[131] an awareness of whether those means actually produced the virtuous-guided desired ends, and a genuine devotion to habituating the practice of comprehending an accurate and sensitive outlook required for making wise decisions. This denotes the temporality of *prohairesis*; wise choice occurs when the morally-sentient agent carefully and intentionally decides their *voluntary* action, directed by the virtues, *before*[132] they carry out that action. *Hairesis*, on the other hand, signifies a non-deliberate choice, one that does not require moral and intellectual character, expertise, habituation, or prudence. *Prohairesis*, thus, is the 'gold standard' of decision-making, and the virtuous agent who champions *prohairesis* becomes the 'wise judge' who will scrutinize the parameters of a situation in accordance with the virtues and decide on the course of action which tends towards the best ends and means. To be sure, these decisions would be *prima facie* and could be altered upon realizing there is a more moral response. With that I will turn to the last section of this chapter, where I will elaborate on why a moral AI would follow a system of virtue ethics.

[129] 165.

[130] Aristotle, 1112b10.

[131] 1111b27.

[132] 11113a2-9.

Artificial Intelligence Virtue Ethicists

I will devote the final subsection of the final chapter to illuminating why a morally-sentient AI that exercises its potential to be moral would necessarily follow a system of Aristotelean virtue ethics. My approach is unique because the vast majority of AI-ethics scholarship, which is limited to begin with, focuses on AI-utilitarianism.[133] Mirroring the argument structure of the subsection "Utilitarian Artificial Intelligence" from the previous chapter, I will analyze the implications of AI virtue ethicists in reference to self-driving cars and then to AIs in general. My position campaigns for the narrative that acting by and for virtues rather than solely by and for perceived consequences will enable an AI to make a 'judgment call' when there does not appear to be one 'right' decision. This compliments the prudential nature of moral decision-making and invites the agent to improve their approach by acting by and for *prima facie* morals and virtues.

The example of self-driving car virtue ethicists illustrates why moral AI would make virtuous decisions. Namely, if the AI-car was in a situation where the death(s) of either the pedestrian(s) or the passenger(s) was inevitable, the AI-car, through *prohairesis*, would summon an understanding of the situation at hand, calculate its options, act in accordance with the virtues, and ultimately select the action whose ethical means tend towards an ethical ends, *even if those ends are less desirable.* For example, an AI-car might choose to save the passenger by appealing to the loyalty, or it might choose to save the pedestrians by

[133] There is also some research on Kantian deontological-AIs, but, as our examination of utilitarian-AIs shows us, strict rule-based decision-making fails to leave room for ethical improvement and growth. This is because programming an AI to follow a system of categorical imperatives would discount the means in which the ends are achieved; actions influenced by Kantian ethics would not be made in 'wise choice' and thus would not always be ethical. And, as we have started to see, an AI must carry out its actions in the 'right way' in order for the AI to be considered moral. To be sure, I am keen to explore this topic in future projects. For further scholarship on this topic, see: Thomas M. Powers, "Prospects for a Kantian Machine," in *Machine Ethics*, ed. Michael Anderson, et al, (New York: Cambridge UP, 2011), 464-475.

appealing to mercy. This raises an interesting problem: loyalty and mercy, in this case, are conflicting virtues. In other words, it appears that AI which appeals to loyalty can be said to be just as virtuous, and therefore just as moral, as one who appeals to mercy. However, while it is true that loyalty and mercy can be conflicting values, it is not necessarily the case that an AI which opts for the 'loyal decision; over the 'merciful decision' is more moral (or vice versa). In other words, it depends on the situation, and a virtuous AI would be able to discern the truly moral decision, regardless of the situation, by making a 'judgment call.'

A 'judgment call' encapsulates the means through which an agent uses *prohairesis* to resolve a situation where there are two virtues or competing obligations in conflict. Making a 'judgment call' is a "three-part process"[134] composed of preparation, deciding the appropriate response, and executing the decision in a morally-permissible fashion. As I alluded to before, 'judgment calls' are particularly important when the agent is forced to weigh one good decision against another equally-good decision. This is why 'judgment calls' are useful in cases like the self-driving car where the most ethical course of action for the AI to take is not (immediately) clear. Situations like these highlight the creative notion of 'judgment calls' in that the AI 'wise judge' that habituates and perfects the 'judgment call' will treat equal situations equally and unequal situations unequally. I say 'creative' because the AI needs to make its decisions at the right time, about the right thing, towards the right subject, for the right end, and in the right way. Making 'judgment calls' through *prohairesis* can help AI accomplish this virtuous goal and habituate virtuous actions in inevitable future situations when conflicting morals and obligations arise.

[134] Tichy, et al, "Making judgment calls," 94.

The ability for AI virtue ethicists to make ‘judgment calls’ also broadcasts an important strength of virtue ethics that other ethical systems neglect; virtue ethics does not chain the AI down to an *a priori* system of universal doctrine and categorical imperatives. This allows for corrective habituation; if an AI begins habituating actions that do not align with promoting the mutual safety of (morally-sentient) beings, we would be able and ready to surround an AI with other virtuous agents who could teach it how to be more virtuous. Thus, ‘judgment calls’ are both practical in theory and in performance.

This raises the questions as to what values a moral AI would base its decisions off of and how an AI would come to know it should follow those values. I will tackle these questions by directing our attention to the crucial distinction between ‘value-loading’ and ‘value-learning.’ Let us say that an AI should follow the four cardinal Aristotelean virtues of prudence, courage, temperance, and justice. A proponent of ‘value-loading’[135] would assert that we could simultaneously pick and choose which virtues we want an AI to follow and make the AI follow those virtues by programming them into the code of the AI. ‘How’ these virtues would be encoded into the AI’s software is not relevant to this paper. What is more important is Bostrom’s argument that ‘value-loading’ is impractical because it requires a “utility function”[136] to enable an AI to discern between its options. Furthermore, an AI that were to rely on a ‘utility function’ would not be equipped to make an informed decision if it was placed in a situation where there was no ‘utility’ for the actions it could take. It is even difficult to say whether or not an AI would even be able to *recognize* its options beyond what the utility framework makes blatantly obvious to it.

[135] Bostrom, 226.
[136] Ibid.

I believe that Bostrom's argument should be taken even further. Namely, for an agent to be moral, the agent must learn the virtues instead of being forced to follow them. This shows us how the value-loading hypothesis fails to meet Aristotle's standard of voluntary virtuous actions; an action that is involuntary is not a virtuous action. This is because "involuntary action is either forced or caused by ignorance, [while] voluntary action… has its principle in the agent himself, knowing the particulars that constitute the action."[137] We know this to be true because, as stated in my discussion of *prohairesis*, the voluntary acting and habituation of similar actions is what sets moral and virtuous agents apart from non-moral agents who perform virtuous actions. Accordingly, a moral AI would learn which values to follow by surrounding itself with virtuous agents who can teach it how to hone a virtuous character of its own. This process is called 'value-learning.'[138] I emphasize the word 'process' because it reinforces the habituation of value-learning, recognition, practice, and perfection. 'Value-loading,' on the other hand, is not a process, but a singular procedure that excludes *prima facie* morals, values, or virtues. Determining what these virtues are is a "wicked problem"[139]: it is a unique, perpetually-changing problem that does not have a definitive right or wrong answer. However, a morally-sentient AI virtue ethicist that makes 'judgement calls' through *prohairesis* will be able to discern the virtuous course of action in any given situation. This removes the burden off of the programmer to code the 'perfectly-ethical' AI while ensuring that an AI would not go 'off course' and put person or non-person morally-sentient beings in danger. Thus, it is for these reasons that I conclude that for an AIs to be moral, it must follow a system of Aristotelean virtue ethics.

[137] Aristotle, 1111a22.

[138] Bostrom, 325.

[139] Horst W. Rittel, et al, "2.3 planning problems are wicked," *Polity* 4 (1973): 155-169.

Afterword

Throughout this project I have challenged the notion that the most ethical AIs would be utilitarians. I began by differentiating ANI, AGI, and ASI. I then argued that due to the recent explosive advancements in AI, we can reasonably assume that there will soon be AI that can don moral responsibility. This is not due to AIs 'having a conscious' or from passing the Turing Test. Rather, their ability to have experiences and view these experiences on a spectrum of perfectly-right to perfectly-wrong combined with the ability 'to do otherwise' makes it so AI can be both causally and morally-responsible for its actions. This led to my final argument, in which I asserted that AI virtue ethicists would promote the mutual safety and interests of person and non-person morally-sentient beings. This is because AI virtue ethicists would be able to learn *prima facia* virtues and balance competing obligations through prohairesis. This would remove the burden off of the programmer to code the 'perfectly-ethical' AI while ensuring that an AI always intends to act at the right time, about the right thing, towards the right subject, for the right end, and in the right way.

Finally, I recognize that my analysis may have frustrated readers who were looking for a list of the *actions* a moral AI should take. Instead, I offered a framework for *how* an AI should perform the actions it chooses in a morally-virtuous way. This opens the umbrella for understanding the relationship between AI and ethics. Understanding this relationship will help us guide the AI advancements so we can more effectively react to it. This, I believe, serves a much greater long-term value. Ultimately, ongoing applicable research addressing philosophical questions related to AI is desperately needed, and I am eager to continue this investigation in future projects with promise, integrity, and fervor.

Bibliography

Troy, Arthur. “Some studies in machine learning using the game of checkers.” *IBM Journal of research and development 3*, no. 3 (1959): 210-229.

Troys, Sam, et al. “Mapping the Landscape of Human-Level Artificial General Intelligence.” *AI Magazine* Vol 33, no. 1 (Spring 2012): 25-42.

Allen, Colin, et al. “Prolegomena to any future artificial moral agent.” *Journal of Experimental & Theoretical Artificial Intelligence* 12, no. 3 (2000): 251-261.

Anscombe, Gertrude Elizabeth Margaret. "Modern moral philosophy." *Philosophy* 33, no. 124 (1958): 1-19.

Aristotle. *Nicomachean Ethics*. Translated by Terence Irwin. Indianapolis: Hackett Publishing Company, 1999.

“Artificial intelligence, n.” OED Online. January 2018. Oxford UP.

Asimov, Isaac. *I, Robot.* New York: Gnome Press, 1950.

Bailey, Alan, et al. *Hume’s ‘Enquiry Concerning Human Understanding’: A Reader’s Guide.* Bloomsbury: London, 2006.

Bello, Paul, et al. “On how to build a moral machine.” *Topoi* 32, no. 2 (2013): 251-266.

Block, Ned. “Troubles with Functionalism.” In *Readings in the Philosophy of Psychology*, edited by Block, 268-303. London: Methuen and Company, 1980.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford: UP, 2013.

Campa, Riccardo. “Artificial Intelligence and Industrial Automation,” in *Humans and Automata*, 37-40. Vienna: Deutsche Nationalbibliothek, 2015.

Crane, Tim. *The Mechanical Mind*. New York: Routledge, 2016.

Crisp, Roger. *Reasons and the Good*. Clarendon Press; Oxford UP, 2006.

Davies, Nicola. “Can robots handle your healthcare?” *Engineering & Technology* 11, no. 9 (2016): 58-61.

Dennett, Daniel. *Elbow Room.* Cambridge: MIT Press, 2015.

Doris, John M. “Persons, situations, and virtue ethics.” *Nous* 32, no. 4 (1998): 504-530.

Edmonds, David. *Would You Kill the Fat Man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton: UP, 2014.

Foot, Philippa. “Morality, Action, and Outcome.” In *Morality and Objectivity*, edited by Ted Honderich, 23-38. London: Routledge & Kegan Paul, 1985.

— *Virtues and Vices*. Oxford: Clarendon Press, 2002.

Ford, Kenneth, et al. *Thinking about Android Epistemology*. Cambridge: MIT Press, 2006.

Goertzel, Ben, et al. *Artificial General Intelligence*. New York: Springer, 2007.

Gomes, Lee. “When will Google's self-driving car really be ready? It depends on where you live and what you mean by ‘ready.’” *IEEE Spectrum* 53, no. 5 (2016): 13-14.

Goodall, Noah J. “Machine ethics and automated vehicles.” In *Road vehicle automation* 93-102. New York: Springer, 2014.

Christopher. “There Is No ‘I’ in ‘Robot’: Robots and Utilitarianism.” In *Machine Ethics*, edited by Michael Anderson, et al, 451-463. New York: Cambridge UP, 2011.

Hatmaker, Taylor. “Saudi Arabia bestows citizenship on a robot named Sophia.” *TechCrunch*, October 26, 2017.

Hof, Robert. “Deep Learning,” *MIT Technology Review*, 2013.

Holley, Peter. “Meet the man who spent 12 years trapped inside his body watching ‘Oreo’ reruns.” *Washington Post.* January 13, 2015.

Hospers, John. *An Introduction of Philosophical Analysis*. Edgewood Cliffs: Prentice-Hall, 1967.

"Human, *adj*. and *n*." OED Online. June 2017. Oxford UP.

Hume, David. *An Enquiry Concerning Human Understanding.* Oxford: UP, 2007.

Hursthouse, Rosalind. "Virtue Theory and Abortion," in *Virtue Ethics*, edited by Roger Crisp, 217-239. Oxford: UP, 1997.

Kamm, F. M. "Non-Consequentialism and the Trolley Car Problem." In *Intricate Ethics Rights, Responsibilities, and Permissible Harm*, 9-224. New York: Oxford UP, 2007.

Keller, Simon. "Virtue ethics is self-effacing." *Australasian Journal of Philosophy* 85, no. 2 (2007): 221-231.

Kurzweil, Ray, et al. *The Age of Intelligent Machines*. Cambridge: MIT Press, 1990.

Lear, Jonathan. *Aristotle: The Desire to Understand*. Cambridge: UP, 2010.

Lin, Patrick, et al. *Robot Ethics*. Cambridge: MIT Press, 2012.

"Locked-In Syndrome Information Page." National Institute of Neurological Disorders and Stroke. May 25, 2017.

"Machine learning, *n*." OED Online. January 2018. Oxford UP.

Mill, John Stuart. *Utilitarianism*, edited by George Sher. Indianapolis: Hackett Publishing Company, 2001.

Moore, Gordon E. "Cramming more components onto integrated circuits." *Proceedings of the IEEE* 86, no. 1 (1998): 82-85.

Müller, Vincent. *Risks of Artificial Intelligence*. Boca Raton: CRC Press, 2016.

Musk, Elon. "About OpenAI." OpenAI.

Nagel, Thomas. “What Is It Like to Be a Bat?.” In *The Mind’s I*, edited by Hofstadter and Dennett, 391-403. London: Penguin, 1982.

Newell, Allen, et al. “Chess-playing programs and the problem of complexity.” *IBM Journal of Research and Development* 2, no. 4 (1958): 320-335.

Nyholm, Sven, et al. “The ethics of accident-algorithms for self-driving cars: an applied trolley problem?.” *Ethical theory and moral practice* 19, no. 5 (2016): 1275-1289

“Person, *n*.” OED Online. June 2017. Oxford UP.

Powers, Thomas M. “Prospects for a Kantian Machine.” In *Machine Ethics*, edited by Michael Anderson, et al, 464-475. New York: Cambridge UP, 2011.

Rai, Tage, et al. “Moral principles or consumer preferences? Alternative framings of the trolley problem.” *Cognitive Science* 34, no. 2 (2010): 311-321.

Rittel, Horst W., et al. “2.3 planning problems are wicked.” *Polity* 4 (1973): 155-169.

Schaller, Robert R. “Moore's law: past, present and future.” *IEEE spectrum* 34, no. 6 (1997): 52-59.

Scheffler, Samuel. “Introduction.” In *Consequentialism and Its Critics*, edited by Samuel Scheffler, 1-18. Oxford: UP, 1988.

Searle, John. “Minds, Brains, and Programs.” In *The Mind’s I*, edited by Hofstadter and Dennett, 353-373. London: Penguin, 1982.

Shields, Nicholas. “Waymo and GM are far ahead in self-driving car tests.” *Business Insider*, February 2, 2018.

Simons, Geoff. *Are Computers Alive?*. Brighton, England: Harvester, 1983.

Singer, Natasha. “Tech’s Ethical ‘Dark Side’: Harvard, Stanford and Others Want to Address It.” *New York Times*, February 12, 2018.

Tichy, Noel, et al. “Making judgment calls.” *Harvard Business Review* 85, no. 10 (2007): 94-104.

Timmons, Mark. “Consequentialism.” In *Disputed Moral Issues*, edited by Mark Timmons, 6-11. Oxford: UP, 2014.

Turing, Alan. “Computing Machinery and Intelligence.” In *The Mind’s I*, edited by Hofstadter and Dennett, 53-67. London: Penguin, 1982.

Van Inwagen, Peter. “The Incompatibility of Freewill and Determinism.” In *Free Will*, edited by Watson. Oxford: UP, 1982.

Vaughn, Lewis. “Bioethics and Moral Theories.” In *Bioethics: Principles, Issues, and Cases*, 34-80. New York: Oxford UP, 2017.

Williams, Bernard. “A Critique of Utilitarianism.” In *Ethics: Essential Readings in Moral* Theory, edited by George Sher, 253-261. New York: Routledge, 2012.

— “Consequentialism and Integrity.” In *Consequentialism and Its Critics*, edited by Samuel Scheffler, 20-50. Oxford: UP, 1988.

Williams, Gerald J. *A Short Introduction to Ethics*. Lanham: UP of America, 1999.

Willetts, David. “Hero soldier returns to duty with battalion after losing leg to a Taliban bomb.” *The Sun*. March 14, 2017.

Wyma, Keith D. “Moral Responsibility and Leeway for Action.” *American Philosophical Quarterly* 34, no. 1 (1997): 57-70.