Spring 2021

Machine Ethics, Ethics for Machines: Context-Based Modeling for Machines Making Ethical Decisions

Jaysa Ramirez
jramirez@rollins.edu

Machine Ethics:
Ethics for Machines

Context-Based Modeling for Machines Making Ethical Decisions

by

Jaysa Ramirez

An Honors Thesis

Submitted to the Department of Philosophy and Religion

Rollins College, Winter Park, FL

April 2021

01100100 01100101 01100101 01111010 00100000 01101110 01110101 01110100 01110011

Abstract

Machine ethics is an emerging, interdisciplinary field that focuses on if – and if so, how – machines can make ethical decisions autonomously. Through a close study of two positions on whether or not machines can be moral agents, this project sheds light on a clash of assumptions that is keeping the field of machine ethics in limbo. After making this clash of assumptions clear, I raise two questions which get at the scope of machine ethics itself:

1. *What makes ethical decision-making different from other kinds of decision-making?*

2. *To what extent can machines engage with ethics and make ethical decisions?*

I address the first question by arguing that ethics is distinct because it requires the ability to understand and participate in human conventions. I address the second question by arguing that ethics has always been informed by our humanity, but machine ethics is an opportunity to expand our understanding of ethics so that machines can engage with it insofar as they are machines. This project aims to contribute to machine ethics by proposing a major shift in perspective, from a focus on human abilities to a focus on machines and their own radically novel abilities.

Table of Contents

Chapter I:

Introduction

`>> What is machine ethics?`

It goes by several names, including robot ethics, roboethics, machine morality, and computational morality, but *machine ethics* is the predominant name for the emerging, interdisciplinary field that is focused on if – and if so, how – machines can make ethical decisions autonomously. The field is not to be confused with computer ethics, which is concerned with the ethical usage of technology by humans. As a field, machine ethics is no more than forty years old, with the term *machine ethics* being coined in 1987 by M. Mitchell Waldrop, a writer with a PhD in elementary particle physics. In an article about how machines might be held responsible for their actions, he notes the following:

> "One thing that is apparent ... is that intelligent machines will embody values, assumptions, and purposes, whether their programmers consciously intend them to or not. Thus, as computers and robots become more and more intelligent, it becomes imperative that we think carefully and explicitly about what those built-in values are. Perhaps what we need is, in fact, a theory and practice of machine ethics, in the spirit of Asimov's three laws of robotics" (Waldrop 38).

It would take more than a decade before the work on machine ethics would really begin. In the early 2000s, the first conference was held to establish the theoretical foundations of the field and more articles were published. Wendell Wallach and Colin Allen's book, *Moral Machines,* was released in 2010. Philosopher Susan Leigh Anderson and computer scientist Michael Anderson

published *Machine Ethics*, an edited volume of essays written by scholars belonging to a number of disciplines, in 2011.

As the field continues to find itself, several scholars have tried to identify the main motivations that have arisen in the discourse. An article written by philosopher Marcello Guarini in 2013 views machine ethics as having a *practical* side and a *reflexive* side. Practically motivated researchers are concerned with what it takes to build ethical machines, while reflexively motivated researchers are interested in how machines can help us to better understand what ethics is or could be (Guarini 214). In a more recent article from 2020, a group of researchers with backgrounds in engineering, computer science, and the ethics of technology describe a similar set of motivations but with different language. They choose to divide the motivations of machine ethics amongst philosophers and engineers, although they clarify that the two groups are not disconnected.

>> Machine Ethics, Fast and Slow

I prefer Guarini's framing, especially because machine ethics is worked on by other researchers such as cognitive scientists, linguists, and psychologists as well. Furthermore, I consider myself a staunch advocate for the necessity of interdisciplinary collaboration in order for machine ethics to make real progress. I am currently an undergraduate studying both computer science and philosophy, and throughout my education I have gained a sense of the cultures present in each discipline. Of course, my experience is relatively limited, but what I have noticed is this: many computer scientists and technologists, especially those working in industry, are driven by the desire to make things that are cool. They love a good challenge and they love to push the limits of what's possible. And it goes without saying, the tech industry is strongly influenced by an interest in what's profitable. The result of these motivations is a fervent

forward-motion without enough regard for the potential impacts that come from what gets made. This is especially true for projects that unexpectedly explode in their scale (the Internet being one of the most obvious examples). And universities have only begun to include topics like ethics in their curriculums for computer science and engineering within the last five years or so. Harvard University's computer science department became a national model when they introduced a "distributed pedagogy" approach in 2019 that pairs graduate students in philosophy with computer science faculty to make sure that computer science students develop an awareness of ethics throughout their education (Karoff 2019).

Philosophy does not move nearly as fast as computer science does. But what we have before us is an opportunity to guide the breakneck progress of technology by working together, instead of addressing issues after the damage is done. Philosophers have a wealth of tools for deeply considering the current and future impacts of technology, but many of them lack the technical knowledge to really do so. By collaborating with each other, philosophers and computer scientists (along with researchers from related disciplines) can better address the question of what computers can do versus what they should do.

>> Intelligence, Then and Now

Some historical context is necessary to properly frame my approach to machine ethics. It is particularly important to note the change in the goals of computer scientists working on AI. But first, I should make clear what the word 'machine' means. I, and other researchers who discuss machine ethics, use the word 'machine' to refer to any man-made system – composed of either software, hardware, or both – that has many parts which come together to perform any number of functions. This definition is meant to capture not only the existence of robots, but also

complex programs that may not have much of a physical presence, other than the presence of the computer being used to run it.

The publishing of Alan Turing's article on computing machinery and intelligence in 1950 is seen as a turning point in the discussion about computing, and during the mid-twentieth century, there was growing excitement about the prospect of simulating intelligence. Early attempts often involved choosing a task which seemed to require a good deal of intelligence, then creating a program which carries out that task. Chess was one such challenge which became very popular, and Turing was the first to produce a chess-playing algorithm one year after publishing his influential article. Other projects, such as the development of a program that could translate English to Russian, eventually garnered attention (and precious funding) from the U.S. government.

The term *Artificial Intelligence* was coined in 1956 when a handful of researchers came together during a summer at Dartmouth College. American computer scientist John McCarthy was the organizer of the conference and is considered one of the "fathers of AI" for his work (Knapp 2008). According to McCarthy, the purpose of their meeting was "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al.). The researchers soon realized that this goal, so easily described, could not be accomplished in one summer. Years later, McCarthy commented on the outcome of the workshop by saying, "the main reason the 1956 Dartmouth workshop did not live up to my expectations is that AI is harder than we thought" (Muehlhauser 2016). Indeed.

Machine Translation (MT), the intelligence task that sparked so much initial excitement about AI, was also the task that caused the government to pull back its funding when researchers

failed to make more progress. The aforementioned English-to-Russian program worked by crudely mapping words Russian to English words according to a bilingual dictionary and a set of rules coded in one direction, meaning rules that could only translate from Russian to English and not the other way around (Hutchins 1995 p. 2). At the time, researchers struggled to find a better way to achieve more grammatically sensible translations. In 1966, a report was published by the Automatic Language Processing Advisory Committee (ALPAC), a group established by the U.S. government at the request of the National Science Foundation. This report was effectively devastating. “For years afterwards, an interest in MT was something to keep quiet about; it was almost shameful. To this day, the 'failure' of MT is still repeated by many as an indisputable fact” (Hutchins 2003).

Following the discovery of how difficult it truly is to define and simulate intelligence, there was an era known today as the “AI Winter.” The U.S. government had abandoned the pursuit of AI, and many researchers also felt that it was a lost cause. The length of this period is debated by the computer science community. Some claim it was a ten-year period, while others claim it lasted thirty years (Muehlhauser 2016). If the latter is true, the AI winter lasted roughly from the mid-1960s to the early 1990s.

Importantly, the history of AI is also the history of our assumptions about intelligence. The project of AI, as it was initially conceived, was the attempt to simulate human intelligence – or rather, to simulate intelligence with the implicit assumption that intelligence *is* human. The hope was that we could learn about ourselves by creating machines that could perform intelligence tasks the way we do. But as computer scientists managed to create programs that could play chess, translate languages, win Jeopardy, play Go… the solutions they found were far from reflective of human intelligence. They constituted a series of sophisticated engineering

techniques. The AI Winter was a consequence of the assumption that the human mind is like a symbol-processing machine, an approach now known as 'Good Old Fashioned AI' (GOFAI). It turns out that the human brain does not solve problems in a serial fashion, it solves them quickly by processing information in parallel. In light of this, computer scientists began working on Parallel Distributed Processing (PDP), the basic principle underlying the development of Artificial Neural Networks (ANN) and later, deep learning algorithms. In cognitive science, the shift from GOFAI to PDP is known as the shift from *computationalism* to *connectionism*.

Computationalism, or the computational theory of mind (CTM) works on the assumption that the mind is a kind of computational system. Not a computer, a computational system. As philosopher Ian Ravenscroft notes in his book about the philosophy of mind, "the existence of computers does not establish that the *mind* is a computer; it only shows that computation is physically possible" (Ravenscroft 88). Many proponents of computationalism compare the mind to a Turing machine, a theoretical model of computation that follows a set of instructions to manipulate symbols, step-by-step. Each operation results in a change in the 'state' of the machine. The key aspect of this comparison is "Turing's Thesis," a claim that is commonly misunderstood to mean that a Turing machine can compute anything that can be described as a clear, valid set of instructions (Copeland 2017). What Turing really claimed was that Turing machines can compute any problem from a class of problems solved by "effective methods" (Copeland 2000). The details about effective methods are not entirely relevant here, but for readers who are curious, effective methods meet the following conditions:

1) The method can be described as a finite set of instructions

2) The method will always produce the desired result in a finite number of steps (given that the method is carried out without error)

3) The method can be carried out by a human without assistance from a machine (the human can, however, use pen and paper if they need to)

3) The problem requires no insight or ingenuity on the part of the human carrying it out

With these conditions, you can see how a misunderstanding could occur. Either way, the comparison between the mind and Turing machines is not usually meant to be exact. Rather, those who make the comparison mean to highlight the ability to manipulate symbols by following an algorithm, a routine set of instructions for solving a problem. In response to this thinking, a theory called *connectionism* arose. Proponents of connectionism take issue with the idea that the mind solves problems by manipulating symbols in a step-by-step fashion. Instead, they hold that the mind is like a network that has a myriad of simple units that process information in parallel to solve problems. The connections between units in the network are weighted, so that the paths between some units are taken much more than others.

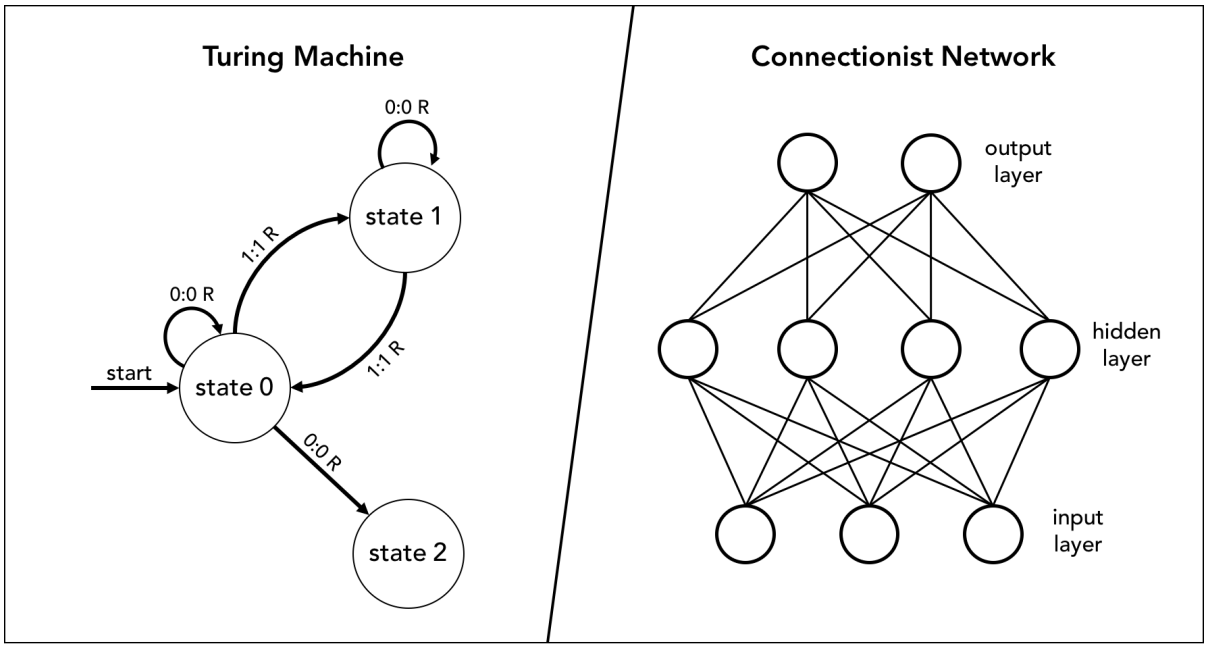

Figure 1.1: An illustration of the difference between computationalism and connectionism. Both models are *highly* simplified, but the juxtaposition of the two models should make clear the contrast in the serial, step-by-step approach of the Turing Machine and the distributed, parallel approach of the connectionist network.

The critical thing to remember about the shift from computationalism to connectionism is that, once computer scientists revised their understanding of intelligence, they were able to make incredible progress in what they could achieve with their explorations of what constitutes intelligence. An example I love to share is that of the massive improvements to Google Translate that occurred in 2016 after the company replaced the previous model for their tool with a neural network, called the Google Neural Machine Translation system (GNMT). Prior to the introduction of the GNMT, Google's translation tool implemented a translation scheme that worked by mapping independent words and phrases to equivalent words and phrases in the target language (Le and Schuster 2016).

Today, there are still technologists who are interested in modeling human intelligence with computers, but much of the community involved in developing AI has shifted its understanding of intelligence to a focus on a machine's ability to perceive an environment and change its goals in response. This is one reason why self-driving cars are an incredibly popular project right now. The ability to safely navigate such a complicated environment as a road captures many of the challenges that come with achieving intelligence as a kind of responsiveness.

>> Goals of my Thesis

There is a clash of assumptions that is keeping machine ethics in limbo. My thesis aims to highlight this tension by way of analyzing what is ostensibly a disagreement about agency. Once I've laid bare what the disagreement is really about, there are two questions that need to be addressed. The first is, what makes ethical decision-making distinct from other kinds of decision-making? The second is, to what extent can machines engage with ethics and make

ethical decisions? I will be answering these questions by proposing a method called *context-based modeling* for machines, a framework that combines aspects of philosophy with techniques already used in computer science. The primary goal of my thesis is to bridge the gap in communication between the sciences and humanities by providing a framework that is rich with opportunities for analysis and offers an actionable means for building ethical machines. I would like for both philosophical and technical audiences to come away from my thesis with a sense of direction. So with that, let's begin by taking a look at the topic of agency.

Chapter II:

A Disagreement about Agency

>> Hopes, Fears, and Assumptions

It is difficult to join the conversation about machine ethics without addressing the topic of agency. The progress of AI research has generated a vast spectrum of ideas both for and against the agency of machines, making agency one of the most widely discussed issues in machine ethics. Concerns about the agency of other intelligent beings have long-preceded the existence of computers, but computers have uniquely challenged our understanding of agency because they show the potential to be more "intelligent" than we are, and yet they seem to lack so many of the faculties necessary for traditional agency: intentionality, consciousness, free will, and so on.

There are scholars who prefer to move past the topic of agency, with some arguing that it provides very little direction for designers and engineers. Wendell Wallach and Colin Allen argue that the topic is only useful if it helps us to designate capacities needed for a machine's performance (Wallach and Allen, 58). In their view, whether or not a machine truly understands what it is doing, or is conscious, or is acting freely is irrelevant if it makes no difference in the machine's behavior. There are other scholars who argue that the topic of agency is what makes machine ethics a lost cause. Aimee van Wynsberghe and Scott Robbins take issue with the language used by machine ethicists and argue that their use of the term 'agency' indicates a possibly dangerous misunderstanding of what machines are capable of. They write, "One should not refer to moral machines, artificial moral agents, or ethical agents if the goal is really to create safe, reliable machines. Rather, they should be called what they are: safe robots" (Wynsberghe and Robbins 732).

I will readily admit that I preferred to avoid the topic of agency for quite some time. Asking engineers to program something like consciousness is probably overkill, I thought. I

feared that disagreements about agency would derail the interdisciplinary collaboration that is so crucial for the progress of machine ethics. But a closer look at the topic reveals that it has generated a remarkable space for observing the hopes, fears, and assumptions scholars are bringing into machine ethics as the field continues to emerge. In this chapter, I will present two positions on machine agency – one in favor, one opposed. Then, I want to demonstrate how the conflict between these positions raises problems that get at the scope of machine ethics itself. But first, I will share a popular framework for measuring a machine's agency that will help to introduce the range of capacities machine ethicists examine in their work.

>> James H. Moor

One of the most widely cited works on machine agency is James H. Moor's paper, "The Nature, Importance, and Difficulty of Machine Ethics," published in 2006. The paper was one of the first to address agency in terms of machine ethics. As such, Moor's work helped to shape the outlook of the field and he is considered one of the founding authors. He describes several kinds of agency, each characterized by a different proximity to ethics. He effectively provides a scale to measure the extent to which a machine can be (or is) an agent. This makes his framework adaptable and thus more responsive to the wide variety of technologies. Before I describe the opposing positions on machine agency, I would like to share Moor's work because it helps to introduce some of the language and themes that machine ethicists are working with.

According to Moor, computing technology is inseparable from ethics because it is, by its nature, normative. He writes, "With technology, all of us – ethicists and engineers included – are involved in evaluation processes requiring the selection and application of standards" (Moor 18). For him, accounting for agency is important as computers "do jobs on our behalf" (Moor 18). At the same time, their performance of these jobs has ethical import in many cases. He goes on to

describe four kinds of agents. With each kind, I will be providing recent examples to show how Moor's framework has scaled over time.

Ethical-Impact Agents: This is the class of technologies which have some ethical (or *un*ethical) impact, either by design or by consequence of its use. YouTube, the largest video sharing platform on the Internet (by *far*), uses deep-learning to recommend content to its users and manage the monetization of videos hosted on the site. YouTube's developers designed the system to increase engagement with the site and protect themselves from legal trouble, but their system has had countless impacts on both content creators and their viewers. There is a lot to be said about the ecosystem created by the developer's choices, but I'll name one impact here. In 2019, a very dedicated YouTuber named Andrew Platt began creating a massive spreadsheet of keywords deemed inappropriate for advertisers by YouTube's algorithm. He created the spreadsheet by posting hundreds of videos with different titles and checking whether or not the video remained monetized. Through his work, he found that many words associated with the LGBTQIA+ community were being systematically marked as inappropriate. Even videos with the words 'gay' or 'gender' in their titles were being demonetized and hidden from recommendations. A Washington Post article about the situation claimed that the site's software was "targeting" LGBTQIA+ creators, but this wording is probably not appropriate (Bensinger and Albergotti). "Targeting" makes it seem as though someone or something is making the choice, but it is more likely that the algorithm learned some negative association it shouldn't have. YouTube has never been very transparent about the way their algorithm works, so the answer is unclear. Either way, their algorithm absolutely has ethical impacts by consequence of its use.

Implicit Ethical Agents: This is the class of technologies which include some measure for the prevention of harm. The Apple Watch is a wearable technology equipped with an ECG. Although, the watch does not just monitor the wearer's heart rate – it issues an alert if it detects an irregularity. There have been several people whose lives have been saved by this feature.

Explicit Ethical Agents: This is the class of technologies which are capable of ethical decision-making. They possess some representation of ethics and operate on that knowledge (Moor 20). More and more experimental work is being done in this area. A group of computer scientists from the University of Liverpool created a system that verifies the decisions an Unmanned Autonomous Vehicle (UAV) makes by testing them against ethical dilemmas (Dennis et al.). Roboticist Ronald Arkin, a widely cited scholar in machine ethics due to his work on robots and warfare, worked with colleagues Jaeuun Shim and Michael Pettinatti to make robots that were augmented with an "Intervening Ethical Governor" (Nallur 2389). The system used deontological ethics to constrain the behaviors of health-care robots.

Full Ethical Agents: This is the class of technologies that can make ethical judgements and reasonably justify them. For now, this title is only held by humans. "It's here that the debate about machine ethics becomes most heated" (Moor 20).

Moor concludes his article by stressing the importance of interdisciplinary collaboration for the advancement of machine ethics. Whether or not the development of full ethical agents is possible is partly an empirical question and thus requires the participation of computer scientists. But there is also the ever-present philosophical dimension of it all, as ethics remains a thorny and controversial matter. "Not only do people disagree on the subject, but individuals can also have conflicting ethical intuitions and beliefs," Moor points out (Moor 21). And on that note, let's take a look at the disagreement this chapter is about.

>> Luciano Floridi and Jeff Sanders

The first position in the disagreement on machine agency comes from an article titled, "On the Morality of Artificial Agents," written by Luciano Floridi and Jeff Sanders, two computer science researchers from the University of Oxford. Their work is motivated by the observation that machines have become a significant source of "im/moral" actions, yet our understanding of agency remains unduly constrained by a focus on the human domain (Floridi and Sanders 351). In their view, the concept of agency is essential as a means to analyze the new problems raised by our technology, so its scope should be extended to include machines. But there is still a considerable gap between the ethical capacities of us and our computers. For one thing, we have yet to ascribe any moral patiency to a computer. How can we define agency in a manner that respects this distinction? Floridi and Sanders believe the disagreement about agency is due in part to the imprecision of the term. Unlike a mathematical function which may be abstruse but ultimately definite after enough scrutiny, the term *agency* has not lent itself to a tight set of necessary and sufficient conditions. The struggle to formally capture our every intuition about agency is perhaps an indication that the term should have more than one definition – and at least one of those definitions should include machines. Floridi and Sanders offer a way to manage this by introducing what they call the "Method of Abstraction" (Floridi and Sanders 351).

The Method of Abstraction works by establishing a level of abstraction in order to determine the object of one's analysis. A level of abstraction (LoA) is like a scientific model, it "consists of a collection of observables, each with a well-defined possible set of values or outcomes" (Floridi and Sanders 354). And indeed, it would make sense to use abstraction as an approach to machine ethics because the tool is one of the most vital for computer scientists and

engineers. But the abstractions used by a programmer ultimately correspond to hardware. Presumably, agency is not so tangible. So how could abstraction (the very opposite of specification) help us to do away with the imprecision of the term?

Consider the definition of a tomato. To a botanist, it is a fruit. It develops from the flower of a plant and it has seeds. Be that as it may, a chef is unlikely to put tomatoes in a fruit salad. Implicitly, there is a difference in the levels of abstraction being used by the former and the latter. According to Floridi and Sanders, "abstraction acts as a 'hidden parameter' behind exact definitions" (Floridi and Sanders 352). Some LoAs are more common and more important to us than others – which is why we usually take no issue with searching for tomatoes near the vegetables in the grocery store, despite their phytological classification. However, "you say to-may-to, I say to-mah-to" does not fly for the concept of agency. This would mean that agency is a fuzzy, contested term because there is no dominant LoA associated with its application. Now is as good a time as any to decide, then, what the LoAs for agency should be.

But first, I should address the looming presence of relativism that has ostensibly been summoned by their position. Floridi and Sanders are not proposing we should be free to define agency according to whatever suits our interests. (This would be an especially dangerous perspective to have with regard to the agency of large corporations, for instance.) Recall that an LoA is linked to a collection of observables. This makes LoAs "mutually comparable and assessable" (Floridi and Sanders 355). The authors are trying to make way for pluralism, not relativism. Both the botanical and culinary classifications of a tomato contain factors that can be observed and agreed upon, such as the presence of seeds or a tomato's acidic flavor profile.

With that out of the way, what should the LoA for agency be? At one LoA, an agent could be anything that acts to produce some effect on the environment – but this is probably too

high, too abstract an LoA for agency to be meaningful. Here, there is no difference between a human being and an earthquake (Floridi and Sanders 357). So Floridi and Sanders propose that one sense of agency should exist at an LoA below the threshold created by 1) interactivity, 2) autonomy, and 3) adaptability, as defined below:

1) *Interactivity:* An agent and its environment (can) act upon each other.

2) *Autonomy:* The agent can act without response to being interacted with.

3) *Adaptability:* An agent can change its policy for acting.

The LoA for *moral* agency is established by adding another condition: "An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action" (Floridi and Sanders 364). But at this LoA, humans and machines remain indistinguishable – i.e., the authors have not yet resolved the tension raised in the introduction. This is where a distinction between *accountability* and *responsibility* becomes important. Like agency, the term *responsibility* is also bogged down by a focus on the human domain. In this case, though, the focus is reasonable because responsibility is more closely tied to psychological factors. And in fact, we can use this aspect of the term to finally distinguish between human agents and nonhuman, machine agents.

According to Floridi and Sanders, those who equate accountability with responsibility assume that "we should reduce all prescriptive discourse to responsibility analysis" (Floridi and Sanders 368). They disagree with this assumption by citing the example of parents who practice morality with their children without holding them responsible for making bad decisions. Implicitly, the parents identify their children as sources of moral action. At the same time, they accept that their children are not yet at an age when they can be subject to the process of moral evaluation. Floridi and Sanders go on to state the following:

> "Trying to equate identification and evaluation is really just another way of shifting the ethical analysis from considering *x* as the moral agent/source of a first-order moral action *y* to considering *x* as a possible moral patient of a second-order moral action *z*, which is the moral evaluation of *x* as being morally responsible for *y*. This is a typical Kantian move, but there is clearly more to moral evaluation than just responsibility because *x* is capable of moral action even if *x* cannot be (or is not yet) a morally responsible agent" (Floridi and Sanders 368).

Therefore, a contrast can be made between humans and machines by specifying that humans are *morally responsible* agents whereas machines are only *morally accountable* ones.

The account provided by Floridi and Sanders is helpful because it captures our intuitions about our own agency while making room for the consideration of these new, more sophisticated and impactful technologies. But there is something missing from their account. It glosses over what makes ethics distinct. They use the phrase 'morally qualifiable' without explaining what that means. We shall return to this issue in a later chapter. For now, let us move on to the second position in the disagreement on machine agency.

`>> Deborah Johnson`

Deborah Johnson, one of the first scholars to publish a textbook about computer ethics, offers another way to capture our intuitions about agency while making room for the consideration of ethical issues raised by machines in her article, "Computer Systems: Moral Entities but not Moral Agents." But instead of expanding the definition of agency, Johnson argues that machines belong to the category of *moral entities*, not moral agents. "Because of the

efficacy of computers and computer systems, those who argue for the moral agency of computers are quite right in drawing attention to the moral character of computer systems. However, they seem to overstate the case in claiming that computer systems are moral agents," she says (Johnson 177). Whereas Floridi and Sanders find the human-centered parameters of agency problematic, Johnson believes they are in place for good reason. Technology, she explains, is human-centered too. "What computer systems are and what they do is intertwined with the social practices and systems of meaning of human beings" (Johnson 168). That is, computer systems are artifacts. They are made and used as a result of human social activity. Johnson worries that by going so far as to claim that machines are agents, we will be separating them from this fact. And crucial to Johnson's view is the notion of intentionality. She wants to argue that machines do not have intentions, they merely manifest the intentions of the people who developed the machine and the people who end up using the machine. Thus, their categorization as moral entities. When something goes wrong, Johnson wants the humans involved in the machine's design and deployment to be held responsible, not the machine itself. Her hope is that a focus on human responsibility will push the tech community to anticipate the role of their work in producing states of affairs (Johnson and Verdicchio 646).

Johnson begins her account by accepting contemporary action theory and using it as a guide for sketching her definition of agency: internal states – desires, beliefs, and the like – result in outward or embodied events with some effect on a recipient or patient. Humans, of course, pass the test. She presumes that humans are free and that we can reason about and then choose how we behave. Machines, however, fail to meet a significant condition. They lack intentionality. When computers behave, they instead do so because of what Johnson calls a "triad of intentionality" (Johnson 179). There is 1) the intentionality "put into" the computer system by

the intentions of the system's designer (Johnson 178). There is 2) the intentionality brought about when a user provides input to the system. She says that computer users "use their intentionality to activate the intentionality of the system" (Johnson 178). Lastly, there is 3) the latent intentionality of the system itself. Not the computer, the computer *system*, which is the computer paired with its meaning to us. The computer is "poised" to behave in a certain manner by virtue of being an artifact, something wrapped up in human social activity (Johnson 179). But what about complex, unpredictable systems with behaviors that appear to be far removed from the intentions of the designer, like an AI created by a deep-learning algorithm? Johnson will maintain that these systems are still just moral entities because the system's designer facilitated their actions and the system's user initiated their actions. Without humans, the intentionality of a computer system is inert.

Johnson clarifies the role of responsibility as it relates to her triadic definition of agency in a more recent article titled, "AI, Agency and Responsibility: the VW Fraud Case and Beyond." In her view, an agent can only be responsible if it is intentional – although, the connection between intentional action and responsibility is complex due to the many types of intentions a human could have. And again, she maintains that unlike humans, machines are not intentional and therefore cannot be held responsible for their actions. But in this article, she goes further with her point by stating that "AI is computational, whereas intentions are not, that is, the two are ontologically different" and "since [AI] consists of software running on hardware" there is an "ontological chasm between computational artifacts and sentient beings (Johnson and Verdicchio 645). Furthermore, there is little evidence to support the expectation that computers will end up having intentions "like humans do" (Johnson and Verdicchio 645). Hm.

On the one hand, I think Johnson's emphasis on the human role in the design and deployment of technology is valuable. It reminds me of an excellent point made by Marc Canellas and Rachel Haga in a 2020 article featured in the proceedings of the Association for Computing Machinery. Canellas is the current Chair of the Institute of Electrical and Electronics Engineers (IEEE) AI Policy Committee, and Haga is a data scientist. The pair argues that the field of automation as it is understood by many is problematic because it neglects the human involved in and affected by the task that is being automated. Engineers automate as much as possible then shunt the rest to humans without any regard for the knowledge they might need to figure out what happened or what comes next. Designing machines this way affects trust, and more importantly, it has consequences for safety. This approach is simply unacceptable for machine ethics. We must be mindful of what (or who) these machines are *for*, after all. And this mindfulness needs to take place throughout the life cycle of any machine with the capacity to cause 'im/moral' actions.

On the other hand, Johnson's approach is riddled with assumptions that should not be taken for granted when considering the radical novelty of machines. And herein lies the clash, the disagreement that turns out to be more than one about agency.

Chapter III:

Machine Ethics in Limbo

`>> Artificial Flavoring`

Floridi, Sanders, and Johnson all seem to notice the same thing, namely the distinct character of computers and what they are (and currently are not) capable of accomplishing in comparison to humans. This is evident in that both of their positions make use of the word *artificial*. What differs is the direction they take in trying to maintain this separation between humans and machines. I want to use the disagreement between Floridi, Sanders, and Johnson to shed light on a tension that is keeping machine ethics in limbo: the preservation of human novelty versus the appreciation of machine novelty.

When it comes to what computers will be able to do in the future, I will admit that I am an optimist. I do not want to be the person that takes a strong stance on what is or is not possible, because the history of technology has shown those sorts of people to be wrong countless times. I think about a famous article written by Vannevar Bush, an inventor and engineer who predicted dozens of modern technologies in an article he published in 1945. He anticipated that "the Encyclopoedia Britannica could be reduced to the volume of the matchbox" and the "author of the future" would "cease writing by hand or typewriter and talk directly to the record" (Bush 5). At the time, people thought his ideas were outlandish. Now, many of his ideas are captured by the smartphone alone. I feel as though it is better to be optimistic about what is possible because such an attitude will push us to try new things. These new things may not turn out the way we expect – Bush, for example, thought that "dry" photography would prevail over digital photography – yet, it is progress nonetheless (Bush 4). No matter the outcome, we will have learned something, and that is never a waste of time.

Still, I recognize there are limits to this thinking. Watching how the directors of *Back to the Future II* imagined the year 2015 is amusing, to say the least. I wish they were right about hoverboards. And as you may remember from my introduction, a handful of computer scientists at Dartmouth College thought they could simulate human intelligence during one summer in the 1950s. They did not manage to achieve this. Frustrated with the unadulterated optimism he saw in the computing community (specifically, the excitement surrounding the symbolic approach to AI), philosopher Hubert Dreyfus is famously known for writing a book in the 1970s called, *What Computers Can't Do*. In the 1990s when the project of AI was regaining steam, he published a second edition of the book called, *What Computers Still Can't Do*. Was he correct? Yes and no. He was right to criticize the 'good old fashioned' approach to AI (GOFAI) which was motivated by the belief that the brain is like a machine that processes symbols and computers could simulate human intelligence by simply doing the same. The AI Winter occurred partly because much of the early optimism computer scientists had towards AI was thwarted by mistakes in their philosophical assumptions about the nature of human intelligence. However, the critiques Dreyfus made have less of a presence now that computer scientists have found new, unexpected ways to make progress.

The disagreement between Floridi, Sanders, and Johnson appears to be another iteration of what Massimo Negrotti calls "a true struggle between two cultures" (Negrotti 195). Negrotti is a professor of the Methodology of Human Sciences and has had conversations with Dreyfus about his views. He reflects on these conversations in his article, "Hubert Dreyfus, the Artificial and the Perspective of a Doubled Philosophy." Earlier, I pointed out a parallel between the work of Floridi, Sanders, and Johnson. They all choose to use the word artificial, and Negrotti happens to be another scholar who has an interest in this word. Well, it's not just an interest. He has

published several books on the subject. "[It] is exactly the concept of artificial that should be deepened, because of its apparent power of setting up a sort of third reality which deserves a careful understanding," he notes in his article (Negrotti 198). Now, I don't want to move too deeply into Negrotti's ideas because the "third reality" he mentions is used to describe his concept of "naturoids," which are machines that mimic natural processes and at the same time, introduce new aspects to the process they imitate due to the difference in materials (Negrotti 195). It is a fascinating perspective, though, so here is a sample of his thoughts:

> "[The] advancements of naturoids consist in generating a new reality with own features crossing the natural and the technological ones. A world that is not destined to growingly approximate the nature, as the enthusiasts incline to think, nor to remain in a trivial realm made of mere machines, as the opponents say and whose general character and properties, and related socio-cultural effects, will emerge progressively in the next decades. This will trigger a sort of a new general evolutionary phase that in some measure is already started. One more reason to support the idea that the artificial should be studied in itself and in all the technological species it consists of" (Negrotti 197).

I include Negrotti's perspective because it helps to highlight the tension at hand. He takes issue with philosophers who "[try] to match scientific knowledge with philosophical conceptualizations and techniques, almost based on the views of past masters," and so do I (Negrotti 197). Johnson's position is ostensibly one about agency, but her reasoning reveals itself to be rooted in philosophical assumptions that are unhelpful to carry on with.

She happily inherits a lot of assumptions from the Western philosophical tradition, proceeding from the idea that humans are mostly rational and have a significant amount of self-awareness and self-control. What makes us special is our ability to reason and then act on reason. Sound familiar? In Greek philosophy, there is Plato's notion of the rational soul and Aristotle's notion of rational animals. In his *Discourse on Method*, Descartes discusses human rationality as well. And if it was not made clear by her position already, Deborah Johnson favors Kantian ideals. "Perhaps the best known and most salient expression of this conception of moral agency is provided by Kant," she notes (Johnson 173).

The issue with these assumptions is that we know better by now. It turns out, most people do not prioritize Reason alone when making decisions. Computer scientists realized this when GOFAI failed to reproduce our intelligence. In psychology, the research done by Timothy Nisbett and Richard Wilson challenges our confidence in our ability to introspect about the causes of our behavior (Nisbett and Wilson 118). Daniel Kahneman and Amos Tversky's groundbreaking research about human decision-making shows that most of the time, Reason comes second to heuristics.

Arguably, we are no longer the only things capable of reason. Arguably, computers are now *better* at (formal) Reason than we are. But I'd rather not harp on this point, because what I am really concerned about is Johnson's insistence that in order to behave ethically, machines have to be "like us" somehow (Johnson and Verdicchio 645). In order to make real progress in machine ethics, I believe we have to move away from the tradition in several ways. First, we must stop thinking that humans alone can be intelligent. Second, we must stop thinking that things have to be like us to work – especially when there are still many unanswered questions about what 'us' is 'like,' so to speak.

I propose that these changes begin with embracing the title, *machine ethics*. In naming it this, we have already identified machines as distinct and we have designated a focused space for reckoning with their impacts. And within the context of Floridi and Sander's position, this specification helps us to establish our perspective. We've made the choice between botanist or chef, so to speak. Now, I am not dismissing the value of learning by trying to create machines that are human-like. I see that exercise as a separate goal. I do not believe it is necessary for machines to possess all the capacities of a human for them to make ethical decisions. Thus, the distinction between human ethics and machine ethics. But does this mean that machines cannot really engage with ethics? What's the point? Allow me to explain.

Chapter IV:

What Makes Ethics Distinct?

`>> Two Concerns`

The disagreement between Floridi, Sanders, and Johnson turns out to be one with implications that get at the very scope of machine ethics. This makes sense because machine ethics is still an emerging field. Moreover, machine ethics is an interdisciplinary field, so tensions were bound to arise as scholars with different backgrounds bring their knowledge to bear on the matters at hand. I want to tighten up the definition of machine ethics by arguing that the field should be just that: ethics for machines. But before I can really clarify what that means, there are two concerns that must be addressed:

1. *What makes ethical decision-making different from other kinds of decision-making?*
2. *To what extent can machines engage with ethics and make ethical decisions?*

The first concern is one that was raised by Yale computer scientist Drew McDermott in his article, "What Matters to a Machine?" He claims that accomplishing ethical decision-making with machines would mean that we have solved many problems to do with reasoning in general – so, ethics is an "AI-complete" problem (McDermott 93). The classification is a play on the concept of "NP-complete" problems, or problems that are the most difficult for computers to solve in any reasonable amount of time. As computer scientist David Eppstein puts it, "NP-completeness is a form of bad news: evidence that many important problems can't be solved quickly" (Eppstein 1996). No wonder McDermott concludes his article by declaring that machine ethicists ought to "find a problem we can actually solve" (McDermott 112). Computer scientists and engineers have made great strides in carrying out formal reasoning with machines, but why is ethical decision-making so hard to figure out? Either there is something that sets ethical decision-making apart from other kinds of reasoning and it needs to be clarified, or there

is nothing special about ethical-decision making and machine ethics reveals itself to be based on a false premise. Evidently, McDermott's concern is a pressing one. And it should be clear why the second concern threatens the scope of machine ethics as well. But even if there is no way for a machine to 'truly' engage with ethics, there remains the issue of their significant impacts on us. We need a way to mediate and analyze their actions.

I would like to address these concerns (especially the first concern) by taking an approach that comes from semiotics, the study of meaning-making and communication through signs. In his book, *The Symbolic Species*, neuroanthropologist Terrence Deacon offers a framework that is based on a compelling synthesis of ideas from influential thinkers including Charles Peirce, Gottlob Frege, and Bertrand Russell. His arguments work to acknowledge the wide gap between animal and human communication without claiming that humans are uniquely capable of meaning-making with language. We need to be clear on what ethics means to us before we can instantiate ethical decision-making with machines, no? I believe semiotics, with its focus on meaning-making, offers substantial tools for approaching a resolution to the concerns raised above. In the following section, I will use Deacon's framework to argue that there *is* something distinct about ethical decision-making.

>> Deacon's Hierarchy of Signs

Terrence Deacon's work is prompted by a question about the apparent chasm between animal and human communication. And yes, he argues, it is a *chasm*. Our lives are utterly saturated with language, yet we grasp precious little about it. Scholars have tried to better understand language by exploring it in terms of evolutionary continuity, but this way of thinking only sharpens the blades of a longstanding Procrustean bias. Comparative studies of animals and

humans tend to place language at the peak of communication so that every other form of communication is "language *minus* something" (Deacon 53). It is not unreasonable to hold that our language, like other features of our species, must have some ancestor common to other creatures. Deacon notes that "extensive nonverbal communication is essential for providing the scaffolding on which most day-to-day language communication is supported" (Deacon 53). For instance, some animals gesture and so do we. However, we cannot deny the chasm that separates us today. Some animals gesture – but scientists have yet to find another Earthly species that takes time to contemplate the nature of the universe.

Deacon wants his readers to acknowledge how rare, how anomalous our way of communicating is. He will argue that the chasm between animals and humans is really a *threshold* we – with great effort – managed to cross at some point in our evolution. There are several levels of representation and humanity managed to latch onto the highest one, the level of the symbolic. And in doing so, our cognitive faculties were rearranged such that our brains are now tremendously overbuilt for meaning-making. In more scientific terms, language is the champion of *disruptive selection*, a type of evolution which selects for extreme versions of a trait. Our capacity for finding and forging the symbolic has afforded us immeasurable benefits. Yet it is all too easy for us to stir up a vortex of meaning that is hard to pull apart.

With this in mind, Deacon goes on to explain his approach to semiotics. It begins with a reevaluation of the relationship between *sense* and *reference*. Scholars like Gottlob Frege and Bertrand Russell thought of this relationship as a logical one shared between a sign and something truly in the world. This perspective causes problems when you consider utterances that are about something that does not exist. Where is the logic in that? Deacon takes this relationship and turns it on its head with the help of concepts established by logician and

philosopher Charles Peirce. He argues that the object of a reference is not bound by the contents of the physical world, it is determined by *interpretation*. The nature of a reference is, in other words, manifested by the cognitive response it elicits. He shifts the conversation about sense and reference from *what* is being talked about to a focus on *how we know* what's being talked about. And as there are many ways to respond to something, there are many *kinds* of references. He identifies three by using terms described by Peirce: *icon*, *index*, *symbol*.

An *iconic* representation is interpreted in a negative, *im*-mediate manner. Icons evoke *re-cognition* in the most literal sense of the word. Interestingly, icons can "be a source of discovery" by highlighting traits which might otherwise go unnoticed about the object being represented – for instance, an uncanny caricature of a celebrity (Deacon 76). Icons can be said to have "firstness," they are understood as easily as apprehending the sight of someone you know or hearing your own name. Again, they are understood *im*-mediately. An indexical representation is propped up by a recurrent correlation in time and space between a sign and its object. You interpret an index based on what is *probable*. Indices can be said to have "secondness," the object of the sign is once removed. "What makes one [thing] an index of another is the interpretive process whereby one seems to 'point to' another" (Deacon 77). A symptom is an example of an index, as there is no *im*-mediate relation between the smell of smoke and a fire. The association is learned.

The level of the symbolic is where Deacon's notion of a threshold enters the picture. This is because the interpretation of the iconic and the indexical requires a competence for single references, whereas the symbolic requires a competence for *double*-references. Symbolism works by referring to both some thing or thought *and* other symbols. The realm of the symbolic is characterized by a dynamic nexus of meaning that is created and maintained by us. But that is

not to say that the meaning of a symbol is arbitrary. Deacon's work builds on a foundation laid by Charles Peirce. Importantly, Peirce's ideas were always a work in progress. He adjusts his terminology as he continues to develop his ideas, but he was never sure how to organize his concepts. But throughout his work he is convinced that he is getting at *something*. In a letter to American philosopher and psychologist William James, Peirce wrote, "Now it is easy to see that my attempt to draw this three-way, 'trivialis,' distinction relates to a real and important three-way distinction, and yet that it is quite hazy and needs a vast deal of study before it is rendered perfect" (Peirce 498). In the same letter, he tells James that he knows his work is far from complete. "Others must carry the study further when I am gone, which will be, I fear, all too soon for me to explain what work I have done," he writes (Peirce 495). Deacon takes Peirce's *something* and organizes it by placing his concepts into a hierarchy. This means that your understanding of something can be unfolded to reveal several parts that inform each other.

The hierarchy of signs is key to understanding Deacon's account of the symbolic. "It sounds pretty straightforward on the surface. But this simplicity is deceiving, because … it is one kind of competence that grows out of and depends on a very different kind of competence" he explains (Deacon 74). Our ability to gather meaning from an index is seeded and supported by our ability to gather meaning from icons. The meaning of a symbol, then, is anchored in part by firstness and secondness. But what puts a threshold between secondness and a symbol's thirdness is the unique competence required to gather meaning from symbols. Memorizing mere correlations is not enough to engage with the symbolic – this is why there is a chasm between our communication and that of animals. To understand the symbolic, one must participate in a constant process of learning and unlearning. "Symbols don't just represent things in the world, they also represent each other," Deacon says (Deacon 99). So, every time you encounter an

instance of a particular symbol, you have to revise your understanding of its range in relation to the object it is being applied to *and* its relationship to other symbols. Eventually, you shed the set of whatever iconic and indexical relations supported your understanding as you grasp the higher-order patterns that contribute to the nexus of meanings that surround a symbol. "The process of discovering the new symbolic association is a restructuring event, in which the previously learned associations are suddenly seen in a new light and must be reorganized with respect to one another" (Deacon 93). You develop a sort of intuition about how to clue into the meaning of a symbol when you come across it.

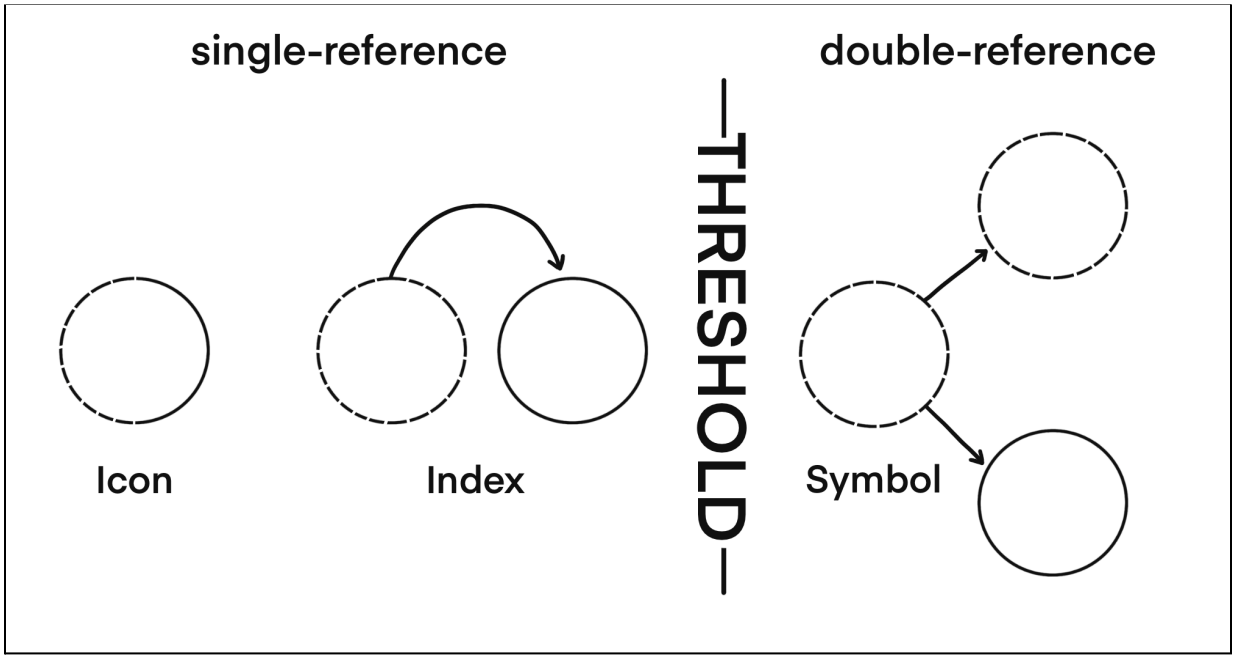

Figure 4.1: Deacon's hierarchy. Dotted lines indicate a sign, while solid lines indicate an object. Icons and indices are single-reference signs, but past the threshold are symbols, which are double-reference signs. With this diagram, you can see how the hierarchy builds. Notice the im-mediacy of icons and the way an indexical sign 'points' to its object. A symbol makes reference to an object and to other symbols.

>> Application to Ethics

Returning to the first of our two concerns, I want to argue that ethical decision-making is distinct because it requires a competence for the symbolic. When we reason about ethics, we are working from the top of Deacon's hierarchy. He claims that insights about ethics "require some of the most counter-intuitive shifts of perspective and recoding process of any symbolic activity" (Deacon 432). This is in contrast to reasoning about something like biology, which comes from the bottom of Deacon's hierarchy. Imagine two undergraduate students, one is a philosophy major and the other is a biology major. Now imagine these students have the opportunity to speak with two other students studying the same subjects at a different school. It is likely that the students studying biology will have no trouble understanding each other. One student could bring up the lymph nodes and the other student would know what is being discussed. No matter what textbooks were used, no matter who taught them about lymph nodes, both students can easily talk about the same thing. The students studying philosophy however, might have very different understandings of the same subject. The first student mentions authenticity and the second student wonders whether the word is being used in Hegel's sense, Heidegger's sense, Taylor's sense, Sartre's sense…? Or maybe they are using the word in a colloquial manner. Deacon's account of signs indicates that the two pairs of students are having different conversations because the biology students are exchanging single-references while the philosophy students are exchanging double-references. In other words, the first pair is discussing concepts closely attached to the world while the second pair is discussing concepts primarily attached to other concepts. Token-object relationships versus token-token relationships (Deacon 446).

That being said, I am not suggesting that ethics has no basis in the world. Here is another way to think about the distinction. It could be argued that if we stopped valuing money, the concept would lose its presence in the world. In other words, we do not *have* to use money.

Bartering was a common way to settle debts before forms of fiat currency like paper money (or more recently, cryptocurrencies) were established. Money is just one solution to the problem. But what problem? At the bottom of it all, there is still the human propensity for exchange, a practice that anthropologists call a "human universal" (Brown). Money could lose its presence but the demand for exchange would remain. It is a behavior that is seen among all peoples known to ethnography and history (Brown). And it is this fact that helps to support the presence of money in the first place. It's hierarchical, to use Deacon's language.

So, could it also be argued that we do not *have* to use ethics? Well, there is a slight incongruence of scope going on when you compare money to ethics. Money is one solution, ethics is a class of solutions. It would be more appropriate to compare money to one theory, arguing that we do not *have* to use, for instance, utilitarianism. Still, it is fair to wonder what would happen if we stopped valuing ethics as a whole. Is there anything at the bottom of the concept that sustains its presence? This is a difficult question, of course. But there is reason to believe that there exists some part of the world that keeps ethics around. When a new domain of human inquiry/activity becomes formalized, there is often a specialized branch of ethics that develops along with it. Medical ethics, business ethics, legal ethics, computer ethics, sports ethics, environmental ethics, cowboy ethics, the list goes on. It seems that ethics is anchored to cooperation. Is that all? Humanity is a symbolic species, we seek meaning and we have feelings about the things we do. Our cooperation *means* something to us. Therefore, ethics must also be anchored to the valenced nature of our actions and interactions.

what it is + **what it means** = **ethics**
iconic/indexical *symbolic*

If ethical decision-making requires a competence for such a distinct kind of reasoning, can machines make decisions about ethics? There are already technologies that manipulate the iconic and the indexical levels of Deacon's hierarchy – but this is only one of the references made by a symbol. The computational challenge is the learning and unlearning of world knowledge associated with the second reference made by a symbol, the reference to other symbols. Can machines have a competence for 'what it is' *and* 'what it means?' We can now proceed to the second of the two concerns raised by the disagreement between Floridi, Sanders, and Johnson. To what extent can a machine engage with ethics?

Chapter V:

To What Extent Can Machines Make Ethical Decisions?

>> The Brains of Men and Machines

In Deacon's view, machines cannot engage with ethics at all because they do not have the capacity to be a participant – that is to say, a subject – in the process of interpreting symbolic references. He sets up his point by discussing Searle's Chinese Room, but not in the way you might expect. Philosopher John Searle introduced what is now a well-known thought experiment in which there is a man using a book of rules to translate Chinese from inside of a closed room. The man has no idea what any of the characters mean, he is simply following the rules in the book, which are written in English. But from outside of the room, the responses he gives are indistinguishable from that of a native Chinese speaker. Searle uses the experiment to argue that the man in the room is like a computer executing a program. Neither the man nor a computer can understand the content of what is being manipulated, even if they give the appearance that they do.

Deacon takes issue with Searle's approach, arguing that it helps to challenge our intuitions but the design of the thought experiment fails to really capture the difference between mind and mechanism (Deacon 445). So Deacon reframes the thought experiment in terms of his semiotic hierarchy and asserts that the questions raised by the room do not correspond to consciousness as a whole, instead they get at those parts of human consciousness that have to do with the level of the symbolic. Searle's thought experiment "emphatically begs the question: What's wrong with the picture?" (Deacon 445). What's wrong, Deacon says, is the walls. The symbolic can never be a part of the room because the room is closed off from the world – you can only reach the indexical, which is exactly what the man is working with when he follows the rules in his book. The man in the room can respect the associations between Chinese characters

but he has no knowledge about the objects of those characters (the lower level reference) or how those objects/events relate to each other (the reference to other symbols).

Some might respond to Deacon's criticisms by claiming that the mind "is like the sort of 'computation' that takes place in electronic computers" (Deacon 442). It's all just software running on hardware. And if we can comprehend the symbolic, why can't a sufficiently sophisticated computer do the same? Deacon suspects that the comparisons made between computers and the brain are due in part to the language we've adopted to describe the way computers function. Here, I am reminded of an argument made by Edsger Dijkstra, a pioneer in the field of computer science. Dijkstra held very strong opinions about the goals of computer science and how the discipline ought to be understood by those who study it. In a presentation titled, "On the Cruelty of Really Teaching Computer Science," Dijkstra argues that computers are a "radically novel" invention and we should think of them as such. Therefore, it is wrong for us to "reason by analogy" and speak about computers in anthropomorphic terms (Dijkstra 1988). In an oral history, Dijkstra commented on his time in America by saying he was "shocked by the clumsy, immature way in which they talked about computing. There was a very heavy use of anthropomorphic terminology, the *electronic brains* or *machines that think*" (Dijkstra 2001). Yes, computers can 'compute' the answer to a mathematical problem but the way they do so is nothing like the way humans would 'compute' their answer to the same problem. Just as the way a computer plays chess tells us little to nothing about how chess Grandmaster Garry Kasparov plays the game.

The reality is that the current architecture of most computers is far from being similar to the structure of the brain. Computers present "a sharp discontinuity" for which "our past experience is no longer relevant," Dijkstra says (Dijkstra 1988). An interesting and somewhat

unusual way to think about this difference is described by neuroscientist Ernest W. Kent in his book, *The Brains of Men and Machines*. We have not been able to create a brain-like computer because "the brain and the computer have developed in an evolutionary manner" towards very different ends, he says (Kent 5). Human beings are adaptable, general intelligences with multiple means for perceiving their environments whereas computers are hugely context-based tools made to solve specific problems. Furthermore, there is a substantial difference in the "hardware" available to nature versus engineers (Kent 5). Any similarities shared by a logic gate and a neuron are far outweighed by discrepancies in speed and quantity, among other things.

It should be noted, however, that computers have changed a great deal since the 1980s, when Kent published his book. In fact, scientists from Zhejiang University recently created what is currently the world's largest neuromorphic computer. Its name? *Darwin Mouse*. And the computer's operating system is called "DarwinOS" (Borak 2020). A neuromorphic computer is one that is designed to be brainlike. Quite a remarkable development, given Kent's perspective just forty years ago. But the computer only has as many neurons as the brain of a mouse, hence its name. The hope is that the scientists will be able to "continue developing the Darwin series of brain-like computers in the direction of human intelligence, just like biological evolution," one scientist said (Borak 2020).

While Deacon doesn't believe machines can engage with ethics now, he is not opposed to the thought that they could in the future. He claims that the current architecture of computers cannot develop the competence for the symbolic, but he is open to the possibility of a completely different architecture that could support sentience and thus a competence for the symbolic. Perhaps Darwin Mouse is a step in this direction. But in the meantime, Deacon's perspective – which was helpful for demonstrating what makes ethical decision-making distinct – presents an

significant obstacle for addressing our second concern, the extent to which a machine can engage with ethics. So, it is finally time for me to make clear what I mean when I say we ought to embrace "machine ethics" as such.

>> Machine Ethics, Ethics for Machines

Evidently, ethics as we have understood it for so long is a very human thing. It is deeply informed by our capacities *as humans*. It would be absurd for someone to declare that omniscience is a requisite for ethics, because that would mean that no human could ever come close to being Good. I want to argue that it is similarly absurd to insist that something like consciousness is required for machines to engage with ethics. Computers are radically novel. Therefore, machine ethics should be seen as an opportunity to rethink ethics as something machines can engage with *qua* machines. There should be a difference between a Good human and a Good machine because they are good at different things. Now, I am not advocating for some hard division between human ethics and machine ethics. Rather, I picture machine ethics as being part of human ethics because machines are, for the most part, built *by* us *for* us.

Deborah Johnson was right to worry about removing designers and engineers from the conversation about what went wrong when a machine causes harm. It is our values and our choices that determine whether or not a machine gets built and deployed to begin with. But we have reached a point where the most high-impact machines (in both the software and hardware sense) are intricate and are developed over time by large teams of people, all with different concentrations. The project managers are interested in making sure the machine meets the needs of its stakeholders. The programmers are interested in making sure the software's logic is sound. The engineers are interested in making sure the hardware supports the functioning of the

machine's software. And so on. Intentions are difficult to maintain throughout the development process of a project so large, so complex – even when everyone involved tries their best to do so.

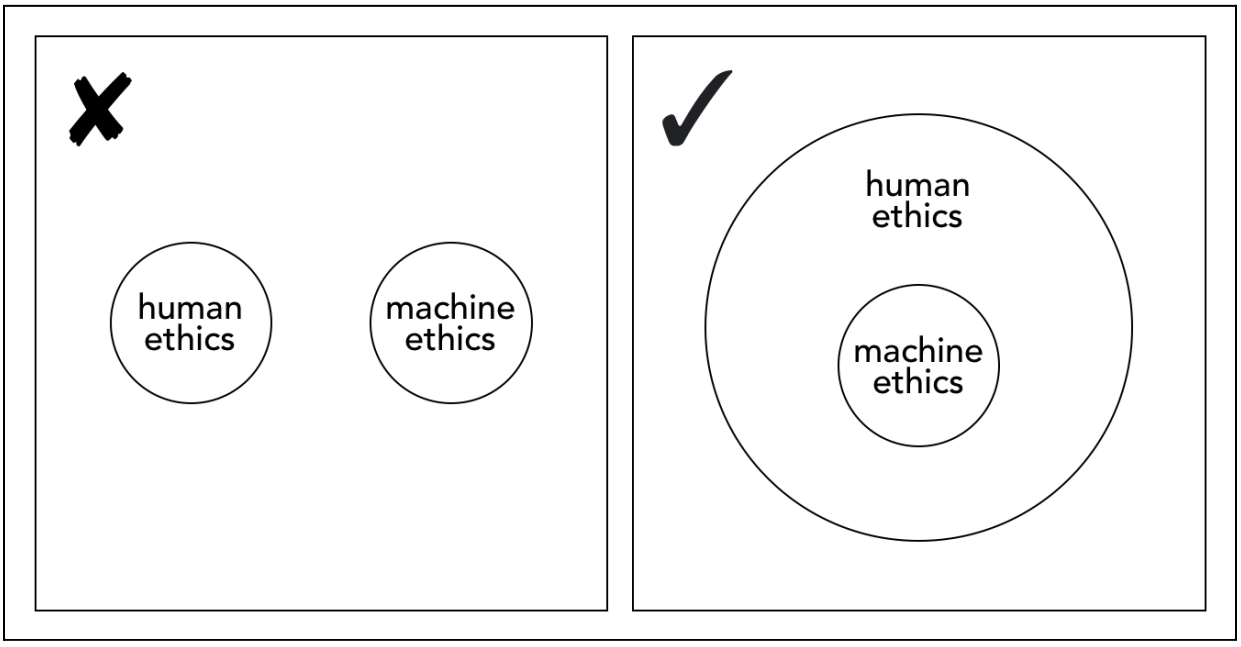

Figure 5.1: I am not advocating for a separation of machine ethics from human ethics (left). Instead, I am arguing that machine ethics is a part of human ethics (right).

In *The Alignment Problem*, writer and computer scientist Brian Christian shares the story of some programmers who were trying to get a small boat to win a race on its own. They placed markers along a course and each marker was worth points. They figured that if they programmed the boat to 'get the most points' it would advance along the course to pass the markers and eventually finish the race. Their intention was for the boat to finish the race, with 'get the most points' being a proxy for that goal. But what happened was the boat went in circles around one of the markers scoring endless points without making it any further to the finish line. A more serious example Christian shares is a situation where several researchers competed with each other to see who could develop the best model for predicting the risk of death from pneumonia. A pretty noble intention, it is a tool that could help a lot of people. The researcher who developed

a neural network won the competition by a wide margin, but upon further inspection, the team realized that the model had determined that patients with asthma are at low *risk* for death from pneumonia. This is obviously not good and very wrong. In the data used to train the network, patients with asthma were more frequently hospitalized for pneumonia, so the model interpreted them as being low risk because patients who are hospitalized are less likely to die. Good intentions, bad outcome.

As I mentioned in the previous section, machines are best developed within a well-defined context. Problem solving in computer science involves taking the time to understand the problem space for what it is, then finding a solution for *that* problem. Machine ethics should begin with this in mind. I would like to put forward an approach called *context-based modeling*, which comes from some of the ideas laid out by Floridi and Sanders. Their Method of Abstraction highlights the notion that experts in different disciplines can engage with the same subject in distinct yet comparably rigorous ways. Consider the difference between fairness in business and fairness in medicine. For each field, the principle has a nuanced body of meaning about it that helps to guide the field's practices. Any attempt to generalize how fairness is carried out in both business and medicine would dilute the potency of what the principle implies for professionals working in either field. Context-based modeling works by defining the model of a Good machine according to the goals of *that* field. In contrast to humans, we know what the *telos* of a machine is because we choose it. We ought to use this fact to our advantage and allow it to inform the policies for the machine's decision-making.

`>> Machines doing Good Work`

I propose that experts who are deeply familiar with the problem space of their discipline should define the model of a Good machine. Here, I need to make a comment about the role of

experts before I continue with my explanation of context-based modeling. Expert consensus is a somewhat common suggestion for how to approach machine ethics. Susan and Michael Anderson, two of the scholars who helped to establish the field, include ethicists as a key part of their methodology. Initially, leaning on expert consensus seemed like a cop-out to me. A way to patch a hole in an otherwise complete argument. But with regard to the importance of context, experts are incredibly valuable. Anyone who spends years immersed in a particular discipline is bound to have keen insights about what makes the discipline tick and what gets in its way.

American psychologist Richard Dawes was intrigued by research which showed that statistical models, sometimes very simple ones, consistently outperform expert human decision-makers in countless domains. He conducted his own research on the matter and found the same thing. He was stunned. "Given the complexity of the world, why on earth should such dead-simple models – a simple tally of equally weighted attributes – not only work but work *better* than both human experts and optimal regressions alike?" he wondered (Christian 96). Upon further thought, Dawes concluded that linear models "cannot replace the expert in deciding such things as 'what to look for,' but it is precisely this knowledge of what to look for in reaching the decision that is the special expertise that people have" (Christian 97). The linear models work so well because there were years of work done by experts to designate what their variables should be to begin with. So that is why experts are a part of my position.

Now, what is a model of a Good machine? And what makes it ethical? A model of a Good machine is one that contains a set of virtues that are important to the context the machine is being deployed in. The process of defining this model differs from simply defining performance requirements because it accounts for the *impact* of the machine on the humans involved in its use. This is what makes it ethical. As Susan and Michael Anderson put it, "Correct ethical

behavior does not only involve not doing certain things, but also attempting to bring about ideal states of affairs" (Anderson and Anderson 330). But how is this worth a distinction between human ethics and machine ethics? It seems as though I've assigned much of the work to the humans – what is the machine's role in all of this? Where in the process is the machine involved in making ethical decisions?

I realize that I've yet to share my own view on machine agency. Now would be a good time to share it, as I think it will help to answer the questions raised above. I agree with the foundation laid out by Floridi and Sanders; I think machines can be agents at a particular Level of Abstraction. But instead of using the term 'artificial agents' – which, as I've explained, is somewhat loaded already – I'd like to use the term *virtual* agents. Consider the difference between the phrases 'artificial reality' and '*virtual* reality.' The latter has more meaningful connotations. *Artificial* suggests an imitation, something lacking the potency of the original. *Virtual* better captures the (unique) force of the new thing. And that's just it: there are machines that can cause just as much good or harm as some people can. There are machines that are forceful and can have impacts that are substantial. Additionally, the term *virtual* already has a powerful meaning in computer science – virtual reality being one example. The Association for Computing Machinery hosts the *International Conference on Intelligent Virtual Agents*. Recently, a lot of work has been done to develop virtual negotiators, virtual job recruiters, virtual educators, et cetera. I should note, however, that many of these virtual agents are meant to be human-like so it feels more natural to interact with them. This is not necessary for my definition of virtual agents. But making the connection to language already used in computer science helps to improve the interdisciplinary communication necessary for machine ethics to work.

Now, a machine's potential to cause good or harm is not enough to establish its status as a virtual agent. GPT-3 is a sophisticated natural-language AI that can do such things as generate tweets, translate between languages, and even write programs. In 2020, a college student named Liam Porr generated entire blog posts with the tool, one of which became the top trending post on Hacker News the day it was published. He later confessed to the true nature of his posts and reflected on his experiment in an article he wrote himself titled, "What I Would Do with GPT-3 if I had no Ethics." Porr's experiment was ultimately harmless, but it raised many concerns about what could happen if the tool became widely accessible. "As soon as this thing enters the public I think it's going to usher in a new era of internet chaos," he said (Porr). The potential for GPT-3 to cause good or harm is clear – however, it has no means for perceiving, let alone evaluating, its impacts on the world. There is no measure in place that would stop someone from trying to generate and spread thousands of tweets containing misinformation, for example.

In order for a machine to be a virtual agent, it must have an *interface*: a virtualized representation of its environment and a means to register 1) its actions and 2) the effect of its actions, or rather, the state of its environment after it makes a decision. This information is indispensable for any machine that is supposed to make ethical decisions. Notably, though, the representation of the machine's environment does not have to be exhaustive. It's all about context, the machine needs access to the salient features of the problem space it is meant to address. Think of it this way. When a doctor is treating a disease, she may not always have a fulsome view of the disease itself. Instead, she works with a representation of the disease built by indicators of the patient's state. She monitors the state of the disease by monitoring these indicators. And such a representation is often enough for her to do her job, and do it well.

>> Ethical Decision-Making versus Ethical Reasoning

One more thing needs to be made clear with regard to my view on machine agency. You might have noticed that I have been intentional about using the phrase, 'ethical decision-making.' In my view, there is a difference between ethical decision-making and ethical reasoning insofar as these concepts relate to machine ethics. They require different skills. And this difference is one of the things that characterizes the distinction between human ethics and machine ethics. Ethical reasoning is the process of developing theories and defining ethical principles. This type of thinking requires a sophisticated ability to engage with the symbolic, connect ideas, and reason about the parameters of ethics itself. Ethical decision-making involves applying or instantiating established principles. Constructing the calculus versus performing the calculus, if you will. Humans can do both, while machines should be able to perform ethical decision-making given a sufficient interface for doing so.

I must emphasize the role of the environment in all of this, as that is what makes ethical decision-making a skill. In her article, Deborah Johnson uses the example of a search engine to point out the intentionality of the designer and the user in its functioning. "What artifacts do is receive input and transform the input into output," and "the output [of the search engine] is a function of how the system has been designed and the input I gave it," she explains (Johnson 178). Her example fails to acknowledge the role of the environment, in this case the dynamic ecosystem of web pages and how they are linked to each other. Heavy-duty search engines do not simply receive a query and return a list of webpages, there is a complex algorithm that works in response to the ever-changing makeup of the Internet.

So, a machine has an active role in making ethical decisions and is therefore a virtual agent because of its response to its environment. When asked about what intelligence is, John

McCarthy – the computer scientist who coined the term 'Artificial Intelligence' and who endeavored to accomplish such a project during that summer at Dartmouth – says it is "the computational part of the ability to achieve goals in the world" (McCarthy). A team of developers can do their best to facilitate the algorithms for a machine to make its decisions, but the machine has a role in responding to novel situations and making decisions about them. They are *virtual* agents because they work from a context-based representation of their environment and yet their decisions have great efficacy.

>> The Importance of Machine Agency

Classing certain machines as virtual agents is valuable for achieving a greater depth of analysis when something goes wrong. A study conducted in 2020 by Human-Machine Communication (HMC) researcher Andrea Guzman sought to identify the key factors that make up the apparent ontological divide between humans and computers. She found that people often cite emotions and certain attributes of intelligence, but "for most people, there is no singular ontological boundary; there are multiple divides, some of which serve as the foundation for others." At the same time, new technologies challenge these boundaries, especially that of emotion (Guzman 50, 51). It stands to reason that the rapid improvement of our technology will only continue to challenge all of these ontological boundaries. Working with the language of virtual agents allows us to better adapt to the actions and effects of new technologies as they come along.

I appreciate Deborah Johnson's emphasis on the human's role in the design and deployment of a machine, but her emphasis on human intentionality makes for a messy analysis of a machine's performance. Even before things go wrong, classing certain machines as virtual agents is valuable as a means to structure the way powerful machines get designed in the first

place. Requiring experts and engineers to deeply consider a machine's purpose and the context of its use will press them to better anticipate what sort of impacts are permissible and which impacts go against the virtues of the discipline. Instead of focusing only on the intentions of the designers and the users of the machine – which can get lost through the complexity of both the sheer amount of people who interact with the machine and the complexity of the machine itself – the evaluators of a machine's performance can look to the model of what Good performance should be, and investigate where the machine's response to its environment diverges from what's expected.

Chapter VI:

Conclusion

`>> Review`

The account I have provided seeks to offer a robust framework for how philosophers, computer scientists, and other experts can work with each other to create machines that do a Good job. I began by discussing the most prevalent and pressing topic in machine ethics, the topic of agency. As researchers from diverse disciplines contribute to the growing conversation about the status of machines, they bring with them their hopes, fears, and assumptions. Agency gets at some of the most fundamental aspects of ethics, so it makes sense that a disagreement about agency could reveal itself to have consequences for the parameters of ethics itself. Once I brought forward the greater implications of what was ostensibly a disagreement about agency, I addressed two questions:

1. *What makes ethical decision-making different from other kinds of decision-making?*
2. *To what extent can machines engage with ethics and make ethical decisions?*

Theoretical schemes from semiotics helped to illustrate the way ethics is entwined in human meaning-making. Terrence Deacon's hierarchy of signs showed that ethical decision-making is distinct from other kinds of decision-making because it involves the ability to recognize what a situation *is* as well as what it *means*. Although Deacon is skeptical of the notion that machines could engage with ethics, he is open to machines with entirely different architectures that could participate in the learning and unlearning required to interpret signs at the level of the symbolic. However, my proposal for context-based modeling is an attempt to reconstruct ethics as something that the machines we have *now* can engage with. Machine ethics, ethics for machines. Once the community involved in creating AI shifted their understanding of intelligence to make room for alternatives to our way of doing things, they made exponential

leaps in what they were able to achieve. Anthropomorphism often gets a bad rap, especially amongst the sciences. Drew McDermott calls it "the Original Sin of AI," which is "harder for [him] to condone than to eat a bug" (McDermott 100). Still, I think there is some value in the practice of using machines to learn more about ourselves. Sinful or not, the project of developing AI has led to increased collaboration between multiple disciplines. In *Matter and Consciousness*, philosopher Paul Churchland describes how the development of artificial neural networks prompted researchers to ask novel questions about the human brain. And in general, the challenge of defining intelligence has pushed us to more deeply reflect on what makes us intelligent.

But the human way is not the only way. The simulation of human biological processes is not necessary for the humbler goal of making better, more trustworthy and dependable machines. My hope is that, in embracing the radical novelty of machines as virtual agents – agents who may lack whatever 'mysterious' features of humanity are necessary for traditional agency yet still exhibit substantial amounts of efficacy – we can make more concerted, tangible progress in machine ethics. ✦

References

Anderson, Susan Leigh, and Michael Anderson. "Toward Ensuring Ethical Behavior from Autonomous Systems: a Case-Supported Principle-Based Paradigm." *Industrial Robot*, vol. 42, no. 4, 4 Mar. 2015, pp. 324–331.

Bensinger, Greg, and Reed Albergotti. "YouTube Discriminates against LGBT Content by Unfairly Culling It, Suit Alleges." *The Washington Post*, The Washington Post, 15 Aug. 2019,www.washingtonpost.com/technology/2019/08/14/youtube-discriminates-against-lgbt-content-by-unfairly-culling-it-suit-alleges/.

Borak, Masha. "Chinese scientists say their neuromorphic computer Darwin Mouse has the same number of neurons as a real mouse." South China Morning Post (Hong Kong), 1 ed., sec. Classified Post, 4 Sept. 2020, p. 7. NewsBank: Access World News, infoweb.newsbank.com/apps/news/document-view?p=AWNB&docref=news/17D468690F6ABF30.

Bringsjord, Selmer, et al. "Artificial Intelligence." *The Stanford Encyclopedia of Philosophy*, July 2018.

Brown, Donald E. *Human Universals*. McGraw-Hill, 1991.

Bush, Vannevar. "As We May Think." *The Atlantic Monthly*, July 1945.

Canellas, Marc, and Rachel Haga. "Unsafe at Any Level: The U.S. NHTSAs Levels of Automation Are a Liability for Automated Vehicles." *Communications of the ACM*, vol. 63, no. 3, Mar. 2020, pp. 31–34.

Copeland, B. Jack. “The Church-Turing Thesis.” *Stanford Encyclopedia of Philosophy*, Stanford University, 10 Nov. 2017, plato.stanford.edu/entries/church-turing/.

Copeland, Jack. “What Is a Turing Machine?” *AlanTuring.net*, July 2000, www.alanturing.net/turing_archive/pages/Reference Articles/What is a Turing Machine.html.

Christian, Brian. *The Alignment Problem*. New York, W.W. Norton and Company, Inc., 2020.

Churchland, Paul M. *Matter and Consciousness.* Massachusetts, MIT Press, 2013.

Deacon, Terrence W. *The Symbolic Species: the Co-Evolution of Language and the Human Brain*. W.W. Norton & Co., 1997.

Dennis, Louise, et al. “Formal Verification of Ethical Choices in Autonomous Systems.” *Robotics and Autonomous Systems*, vol. 77, Mar. 2016, pp. 1–14.

Dijkstra, Edsger W. “On the Cruelty of Really Teaching Computer Science.” 2 Dec. 1988, The University of Texas at Austin, The University of Texas at Austin.

Dijkstra, Edsger W., and Philip L. Frana. “An Interview with Edsger Dijkstra.” Charles Babbage Institute, 2 Aug. 2001.

Dreyfus, Hubert L. *What Computers "Still" Cant Do: a Critique of Artificial Reason*. The MIT Press, 1994.

Eppstein, David. “ICS 161: Design and Analysis of Algorithms Lecture Notes for March 12, 1996.” *NP-Completeness*, UC Irvine, 2 Mar. 1996, www.ics.uci.edu/~eppstein/161/960312.html.

Floridi, Luciano, and J.W. Sanders. “On the Morality of Artificial Agents.” *Minds and Machines*, vol. 14, no. 3, 2004, pp. 349–379.

Guarini, Marcello. “Introduction: Machine Ethics and the Ethics of Building Intelligent Machines.” *Topoi*, vol. 32, no. 2, 5 Sept. 2013, pp. 213–215.

Guzman, Andrea L. “Ontological Boundaries Between Humans and Computers and the Implications for Human-Machine Communication.” *Human-Machine Communication*, vol. 1, 2020, pp. 37–54.

Hao, Karen. “A College Kid Created a Fake, AI-Generated Blog. It Reached #1 on Hacker News.” *MIT Technology Review*, MIT Technology Review, 10 Dec. 2020, www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/.

Hutchins, John. “ALPAC: the (In)famous Report.” *Readings in Machine Translation*, edited by S. Nirenburg, H. Somers and Y. Wilks. Cambridge, Massachusetts: The MIT Press, 2003.

Johnson, Deborah G. “Computer Systems: Moral Entities but Not Moral Agents.” *Ethics and Information Technology*, vol. 8, no. 4, 1 Nov. 2006, pp. 195–204.

Johnson, Deborah G., and Mario Verdicchio. "AI, Agency and Responsibility: the VW Fraud Case and Beyond." *AI & Society*, vol. 34, no. 3, 11 Jan. 2018, pp. 639–647.

Kahneman, Daniel. *Thinking, Fast and Slow*. New York, Farrar, Straus and Giroux, 2011.

Karoff, Paul. "Harvard Works to Embed Ethics in Computer Science Curriculum." *The Harvard Gazette*, Harvard University, 28 Jan. 2019, news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/.

Kent, Ernest W. *The Brains of Men and Machines*. McGraw Hill, 1981.

Knapp, Susan. "Artificial Intelligence: Past, Present, and Future." Dartmouth University, 17 December 2008.

Le, Quoc V., and Mike Schuster. "A Neural Network for Machine Translation, at Production Scale." *Google AI Blog*, Google, 27 Sept. 2016, ai.googleblog.com/2016/09/a-neural-network-for-machine.html.

McCarthy, John. "What Is AI? / Basic Questions." *Professor John McCarthy*, Stanford University, 12 Nov. 2007, jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html.

McCarthy, John, et al. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." *AI Magazine*, vol. 21, no. 4, 2006.

McDermott, Drew. "Artificial Intelligence and Consciousness." *The Cambridge Handbook of Consciousness*, Cambridge University Press, 2007, pp. 117-150.

McDermott, Drew. “What Matters to a Machine?” *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, Cambridge University Press, Cambridge, 2011, pp. 88–114.

Moor, James H. “The Nature, Importance, and Difficulty of Machine Ethics.” *IEEE Intelligent Systems*, vol. 21, no. 4, 2006, pp. 18–21.

Muehlhauser, Luke. “What should we learn from past AI forecasts?” *Open Philanthropy*, 2016.

Nallur, Vivek. “Landscape of Machine Implemented Ethics.” *Science and Engineering Ethics*, vol. 26, no. 5, 8 July 2020, pp. 2381–2399.

Negrotti, Massimo. “Hubert Dreyfus, the Artificial and the Perspective of a Doubled Philosophy.” *AI & Society*, vol. 34, no. 2, 2018, pp. 195–201.

Nisbett, R. E., & Wilson, T. D. “Telling More Than We Can Know: Verbal reports on Mental Processes.” *Psychological Review*, vol. 84, no. 3, 1977, pp. 231–259.

Peirce, Charles S. *The Essential Peirce: Selected Philosophical Writings*. Indiana University Press, 1998.

Platt, Andrew. “How I Test Words, And Why They're Wrong Sometimes…” *YouTube*, 3 October 2019, https://www.youtube.com/watch?v=fwR34KF8_pM.

Porr, Liam. “What I Would Do With GPT-3 If I Had No Ethics.” *Nothing But Words*, 3 August 2020, https://adolos.substack.com/p/what-i-would-do-with-gpt-3-if-i-had.

Ravenscroft, Ian. *Philosophy of Mind: a Beginner's Guide.* New York, Oxford University Press, 2005.

Rumelhart, David E., et al. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, 1985, pp. 3-44.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 417–457.

Turing, Alan. "Computing Machinery and Intelligence." *Mind*, 59, no. 236, 1950, pp. 433-460.

Van Wynsberghe, Aimee, and Scott Robbins. "Critiquing the Reasons for Making Artificial Moral Agents." *Science and Engineering Ethics*, vol. 25, no. 3, 2018, pp. 719–735.

Waldrop, M. Mitchell. "A Question of Responsibility." *The AI Magazine*, vol. 8, no. 1, 15 Mar. 1987, pp. 28–39.

Wallach, Wendell, and Colin Allen. *Moral Machines*. Oxford University Press, 2009.

Zemeckis, Robert, director. *Back to the Future Part II*. Universal Pictures, 22 Nov. 1989.

Acknowledgements

I would like to thank the following nerds:
Dr. Tom Cook
Dr. Mario D'Amato
Dr. Daniel Myers
Kelsey Eelman
Rahmat Rashid
London Davidson

And I would like to thank my family, for supporting my own nerdiness.