religions

MDPI

Article

Writing a Moral Code: Algorithms for Ethical Reasoning by Humans and Machines

James McGrath [1,*] and Ankur Gupta [2]

[1] Butler University, Department of Philosophy, Religion, and Classics, 4600 Sunset Avenue, Indianapolis, IN 46208, USA
[2] Butler University, Department of Computer Science, 600 Sunset Avenue, Indianapolis, IN 46208, USA; agupta@butler.edu
* Correspondence: jfmcgrat@butler.edu; Tel.: +1-317-940-9364

Received: 31 July 2018; Accepted: 7 August 2018; Published: 9 August 2018

Abstract: The moral and ethical challenges of living in community pertain not only to the intersection of human beings one with another, but also our interactions with our machine creations. This article explores the philosophical and theological framework for reasoning and decision-making through the lens of science fiction, religion, and artificial intelligence (both real and imagined). In comparing the programming of autonomous machines with human ethical deliberation, we discover that both depend on a concrete ordering of priorities derived from a clearly defined value system.

Keywords: ethics; Isaac Asimov; Jesus; Confucius; Socrates; Euthyphro; commandments; robots; artificial intelligence; programming; driverless cars

1. Introduction

Ethical reasoning, when it is done effectively, involves prioritizing between competing options. Philosophical thought experiments such as the "trolley problem" highlight the human tendency to avoid making decisions about situations in which there is no good outcome. However, the history of both religion and science fiction provide ample illustration of the need to organize competing values, and to do so in advance of concrete situations where our moral and/or ethical commitments are put to the test. From Jesus' identification of the greatest commandment and the parable of the good Samaritan, to Isaac Asimov's Three Laws of Robotics and Star Trek's Kobayashi Maru scenario, religion and science fiction converge on the importance of confronting no-win situations of a variety of sorts, situations in which we cannot fully implement all of our competing ethical commitments. These stories can help us weave the fabric of our diverse communities and values, as well as integrate autonomous machines into a forward-looking social structure.

1.1. A Framework for Technical Discussion and Some Basic Definitions

The topic that we hope to address can be unwieldy because it spans many different fields and points of view. We therefore wish to define certain "loaded" words and concepts at the outset so that we may present our thesis more clearly. Our intent here is not to ignore or downplay the importance of debates related to the definition of these terms. Rather, the goal is to provide a baseline for meaningful discussion that allows us to avoid devoting the majority of our attention to semantic matters. We begin by defining, somewhat trivially, that a *human* is an *agent* capable of making a decision based upon some criteria or *values*, which may change over time based upon instinct, experience, education, or whimsy. A human agent may modify its set of values as well as the relative priority of those values. We define an *autonomous machine* or *robot* as an agent that makes a decision based upon some static set of values together with specified priorities which are not allowed to change. However, we do still allow a robot

to learn how to optimize its behavior within these parameters by improving its decision-making ability. In some cases, we may refer to a *self-aware robot* as a non-biological human agent, which may or may not have some weak constraints on its values and their relative priority. Hopefully the context will make clear whether we are referring to imagined machines of science fiction, or machines that we can expect to exist in the real world based on current trajectories of technology.

For convenience, we make the simplifying assumption that a robot's *values* are hard-wired (literally or metaphorically) so that neither they nor their priority can be changed. We refer to an *algorithm* as a series of (software) instructions that the robot uses to decide to perform some action (or inaction) based upon the value system encoded in those instructions. Algorithms are typically evaluated based upon the *speed* with which they arrive at a decision and the *accuracy* of the result as compared to whatever may be deemed the "best" solution. Many factors may affect the accuracy of the result, such as incorrect input from sensors (like cameras or radar) or mechanical output errors (like slipping wheels, incorrect turn radius, or brake failure). Our article focuses primarily on accuracy in terms of the theoretical result based on the given input, rather than on its implementation in practice.

Algorithms operate on problems of varying difficulty, often (but not solely) represented by the size of a problem's *state space*, which enumerates all possible results or outcomes for a given problem. Algorithmic approaches to problems typically fall into two broad categories that represent this difficulty, *deterministic* and *non-deterministic*. Briefly, deterministic algorithms are always able to arrive at the best or *optimal* solution. For example, the algorithm for tic-tac-toe is deterministic, since we can always find a line of play that leads optimally to a draw. Non-deterministic algorithms feature a state space for a problem (such as chess) that is so large that we simply cannot compute all outcomes in reasonable time, and thus, we must introduce *heuristics* or educated guesses about the nature of an optimal solution. Ethical value systems of the sort that we discuss in this article are one form of *heuristic* for making the right decision. Our heuristic guess can be flawed, but within the scope of that heuristic, our goal is to optimize the resulting decision.

In some cases, various techniques in *machine learning* or *deep learning* may be used to incorporate past success or failure to improve the success rate of a future action (or inaction) in accordance with its hard-wired values. Often, these techniques involve providing the robot with a *training set* of prior events, in which the success or failure is already known, which is used by the robot to establish a baseline upon which to make future decisions. This baseline is then iterated upon further with a *validation set*, and then finally tested for accuracy with a *testing set*. The term *artificial intelligence* (or *AI*) refers to the overall ability of a robot to learn, plan, and solve problems autonomously.[1] Alan Turing reasoned about the nature of artificial intelligence with the design of the famous *Turing test* (Turing 1950, p. 433), setting the stage for one of the most hotly-debated topics in philosophy and computer science—can a robot think? Our focus is on significantly narrower but no less difficult questions: Can we program autonomous machines to act consistently in accordance with human ethical guidelines? What does it mean, if anything, for a machine to be *ethical*?

There has been significant discussion, from a religious, sociological, psychological, and philosophical point of view, about the distinction (or lack thereof) between two of our more common words denoting value systems, *morality* and *ethics*. Defining these terms has proven nearly as challenging as determining whether a given observed action is both ethical and moral, one or the other, or neither, based on particular definitions of the terms. Since our focus is on how to navigate complex decision-making when faced with competing values, we need not distinguish between the two in the context of the present study (and, in this sense, we use them interchangeably). Thus, we define

---

[1] From a rigorous point of view, our definitions preclude certain types of solutions from being discovered by a robot in a large state space, namely those that necessitate a change to the underlying values of the robot. This limitation is intentional and realistic; furthermore, it defers the question of the rights we should afford to a self-aware robot, which is beyond the scope of this article, although we nevertheless mention it in several places where that subject intersects directly with our article's main focus.

ethical and *moral* behavior as belonging to a set of preferred or prescribed behaviors, whether specified by a higher being, religion, social norm, or a need for self-preservation. We expect that a human agent would decide upon or take for granted a set of values and assign a priority to them that is meaningful or "right" for that agent. This could happen at any point, including retrospectively from the perspective of hindsight; in the case of a robot, however, it would be assigned in advance.

Finally, we define a few relevant terms used in computer science. A *decision tree*, roughly speaking, is a *graph* composed of *nodes* and *edges* (connecting lines) that illustrates every point at which a decision needs to be made before arriving at a course of action. For example, in Figure 1, every node progresses the flow of decision until a final and unambiguous action (or inaction) is determined. Such a system works very well when the values are clearly prioritized. Notice that there are no "loops" in the image, in that no later decision can affect a prior one. As we will discuss later in the article, such "loops" are often the reason for confusion in human reasoning (just as in computer science).

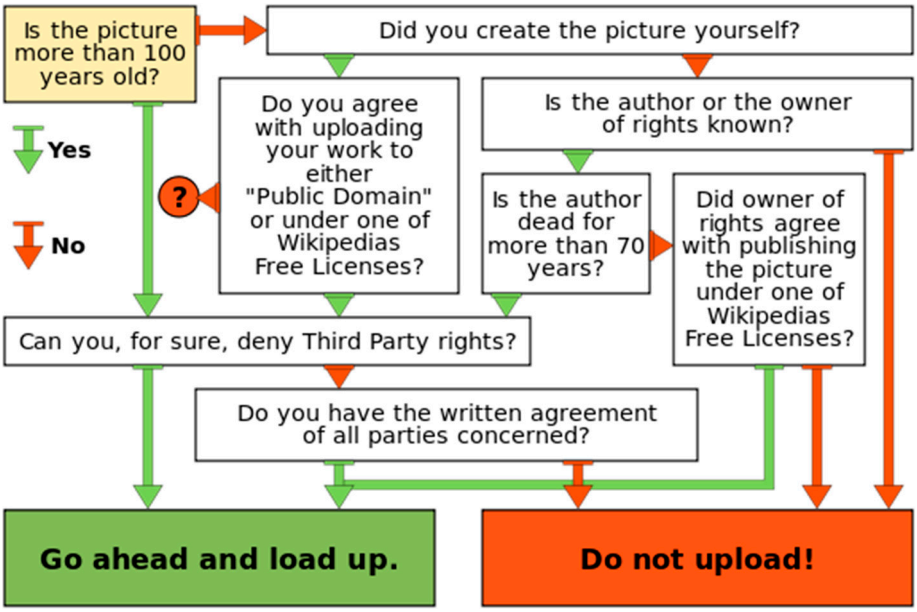

Figure 1. Decision tree illustration by Stefan Lew (Public Domain).

Another concept we use is that of a *Hasse diagram* or a *partial order*, which describes a series of known comparisons among objects of a set. The set of integer numbers is an example of a *total order*, because for any two integers, one can compare them and determine which of the two is larger. For a set of objects that cannot be completely compared, a Hasse diagram helps us visualize the relationships that we *do* know. In the context of this article, we use them to specify the relative priority among values used in both human and robot decision-making, especially in moral and ethical matters.

For example, we read the graph in Figure 2 from top to bottom, where a given node **A** has greater priority than node **B** if and only if **A** is above **B** and an edge connects **A** with **B**. In Figure 2, each node contains a *bitstring*, a lexical (rather than numerical) sequence of symbols **0** and **1**. Bitstrings are ordered in the same spirit as words in a dictionary, where symbols are compared pairwise by position; in other words, the first symbols of each bitstring are compared, and then the second, and so on. In this example, we prioritize the bitstring **1011** over the bitstring **1001**, because every position of the first bitstring is no less than that of the second bitstring. However, the bitstrings **0101** and **0011** are incomparable, since each bitstring has at least one symbol that is larger than its corresponding symbol in the other bitstring. The edges in Figure 2 provide a visual representation of the properties of the partial ordering we defined just above. The data configured in this way may be of any sort, including ethical commitments and objects of value, as we shall explore below.

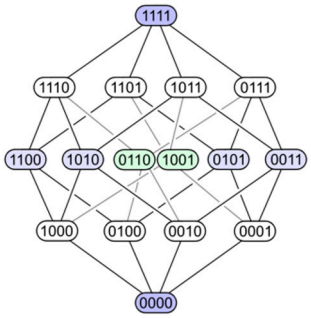

Figure 2. Hypercube by Tilman Piesk (Public Domain).

1.2. Key Challenges to Human Ethical Reasoning and Ethical Robot Programming

As we discuss throughout the article, human agents have a fundamental discomfort with *a priori* decision-making, especially when all available options are less than ideal. Part of the discomfort stems from our reluctance to set aside the influence of contextual clues, emotional input, or other not-strictly-logical personal bias or feeling in a specific circumstance. In fact, even a human agent with a firm theoretical view of a specific course of action may defy or contradict that viewpoint when that concrete situation arises in practice. Human agents also tend to believe that they make decisions quickly and precisely in difficult situations, yet we are also relatively quick to forgive poor outcomes that result from innate human limitations. It is therefore somewhat ironic that human agents, in general, have significantly higher expectations of robots than of themselves, so much so that humans harshly criticize machines if the outcome falls even slightly short of perfection, despite a faithful execution of their algorithms.

The lack of clarity in human moral and ethical reasoning presents real challenges when designing robots to follow ethical guidelines and is exacerbated by our emotional responses to their implementation in practical scenarios. Navigating the murky gray areas in no-win situations is even more complex, and often leads to disagreement that has historically led to conflict of the highest order. Even though autonomous machines are the natural evolution towards technological advancement and increased human comfort and safety, it is no surprise that human legal quandaries and distrust of programming appear as hurdles along the path to acceptance of autonomous machines as part of mainstream human society. By bringing together perspectives from religion, philosophy, computer science, and science fiction, the authors hope to clarify how human agents and robots should think and act, and how we should judge (or refrain from judging) both human agents and robots in their implementation of ethical and moral values.

Any engagement with matters that are ethical, robotic/computational, or both will eventually find itself at the intersection of philosophical and narrative structure. Science fiction stories and parables can aid in creating a "sandbox" in which one can explore and study ethical and other philosophical questions. There is, to be sure, an inherent danger in using a story where the "conclusion" is all but pre-determined and leads to circular logic. For instance, if we depict machines as self-aware in a story, and use that story to advocate for robot rights, we are assuming what is yet unproved, namely that machines may one day have this capability. Our aim is not to wrestle with the speculative problems related to the most advanced machines that science fiction has imagined, but with the steps towards autonomous machines that reflect present technology and its realistic progression into the future. It is this topic—at the intersection between computing, science fiction, and ethical reasoning—that is our concern here.

2. Can We Teach Robots to Be Ethical?

The tendency for ethical principles to be framed in religious terms, or to intersect with religious concepts, is not merely a matter of history. In ancient times Confucius spoke of the will of heaven,

Socrates of the relationship between piety and the gods, and Jesus of love for God as the supreme commandment. Ethics is also treated deeply (and at times religiously) in works of science fiction, especially those with a focus on self-aware robots. In Isaac Asimov's classic story "Reason", a robot finds its way to belief in a creator greater than the humans who claim that role for themselves. More recently, in the rebooted *Battlestar Galactica* and *Westworld*, robots that stem from human creative activity wrestle with how to understand their place in the cosmos and how they should act, with much discussion of God or gods peppered throughout the dialogue. The prominence of religious language is not surprising, since the production of self-aware robots immediately casts human beings in the role of creators of new living things made in our image and likeness. Our creations immediately confront us with the challenge that has regularly been faced by divine creators as well as those involved in the more mundane activity of human biological reproduction: can we effectively pass on our values to our offspring, and if so, how? As a perennial human problem, our literature reflects our species' wrestling with God, morality, and meaning on a fundamental level [see further (Zarkadakis 2016, pp. 58–59)].

2.1. Asimov's Three Laws of Robotics

Noteworthy at the intersection of ancient and modern ethical teaching, of the religious and the robotic, is the need to organize ethical values in a hierarchical arrangement that allows for decision-making when those values come into conflict. Isaac Asimov, in formulating his famous Three Laws of Robotics (Asimov 1950, p. 37), did not merely list principles but also ordered them, with each subsequent law to be followed only if it did not violate those above it. The laws, for those who may not be familiar with them, are typically given as follows (with some variation over the years across multiple stories):

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Autonomous machines may implement these laws using a decision tree (see Figure 1) to navigate in the real world. The ordering of the laws is crucial to the rationale for them and the ethical framework they are meant to encode, as is illustrated well by an XKCD comic that explores the effect of ordering them differently.

One interesting observation regarding Figure 3 is that the most catastrophic ordering happens when (2) "obey orders" has higher priority than (1) "don't harm humans". While science fiction tends to focus on reasons we might fear robots in their own rights, the comic's exploration of the results of different orderings suggests that we may have more to fear from one another than from autonomous machines. This is a point that reemerges repeatedly as we explore this subject.

Even when provided with an overarching framework of correctly-ordered rules to guide their behavior, machines often have difficulty adapting correctly to unexpected or complex situations. In such circumstances, their rigid adherence to directives may have potentially hilarious or tragic results. Alternatively, in situations in which no matching directive is available for them to follow, they may become paralyzed in a state of inaction. Human beings and human societies are also prone to become "stuck" when a matter of conflicting values arises. When we think about human ethical codes, for instance those found in the Torah, typically no explicit hierarchical arrangement is provided, a situation that generated much rabbinic debate about how to mediate conflicts that arose between the requirements of two laws. Something similar may be said about most human collections of laws. The current article may be viewed as an exercise in "debugging" errors that result when we approach ethical reasoning as a society with a set of instructions that is either unordered or ordered incorrectly.

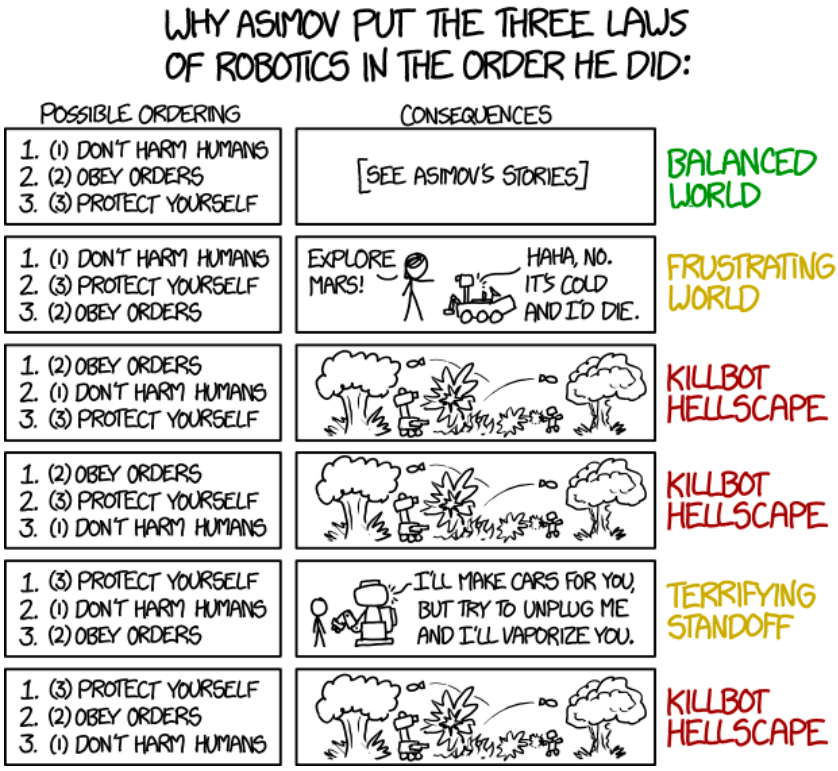

Figure 3. The Three Laws of Robotics, by Randall Monroe. Creative Commons Attribution. Source: https://xkcd.com/1613/.

Science fiction is widely understood to provide useful thought experiments for the exploration of ethical as well as other philosophical and theological issues, and many of the stories that spring to mind are directly related to the realm of artificial intelligence. These stories typically raise questions such as whether a robot can have a soul or deserves rights. The narrative assumptions of many science fiction stories are extremely unlikely to reflect real technological developments. This does not, however, make them any less relevant or useful, since the point is precisely to use an extreme or even contrived scenario to test the edges and limits of our conclusions and principles. Science fiction, like parables and other stories of old, provides a symbolic narrative framework or sandbox within which we can play around and explore moral and ethical concerns.

For this reason, we need not detain ourselves here with some of the legitimate criticisms of Asimov's laws of robotics that have been offered, especially as these pertain to the impossibility of converting them into programmable lines of code that machines can follow (see for instance Keiper and Schulman 2011; Pagallo 2013, pp. 22–29; Hampton 2015, pp. 6–7; Allen and Wallach 2015, pp. 91–94; Bhaumik 2018, pp. 250–57). Our aim in this study is to utilize the challenges that programmers face when working on driverless cars and other automated machines, as well as Asimov's and others' stories involving related conundrums in the future, to investigate neglected aspects of human moral reasoning (Asimov 1953). As Christian and Griffiths conclude, "Any dynamic system subject to the constraints of space and time is up against a core set of fundamental and avoidable problems. These problems are computational in nature, which makes computers not only our tools but also our comrades" (Christian and Griffiths 2016, p. 256). The inevitability of conflict arising between ethical and moral values that an individual or a society subscribes to makes it appropriate to explore solutions in a manner that brings Jesus, Isaac Asimov, and others into conversation with one another.

2.2. The Need for Prioritization

Let us turn first, however, to two classic thinkers whom we have already mentioned—Socrates (as depicted in the writings of Plato) and Confucius—as illustrations of the problems inherent in *not* thinking of values in an explicitly hierarchical fashion. In a famous anecdote in Confucius' Analects 13:18, the governor of the region of She says that they have someone among them who is so upright that he testified against his father when his father stole a sheep. Confucius is reported to have replied that, where he comes from, uprightness is understood quite differently: "Fathers cover up for their sons, and sons cover up for their fathers". This saying has been the subject of considerable debate, as Huang Yong notes when he writes, "the governor of She and Confucius seem to have two different understandings of uprightness. On the one hand, uprightness means impartiality: upright people treat their family members in the exactly same way as they treat others. They will bear witness against any wrongdoers, and so will not do anything differently if such people are their own family members. On the other hand, when Confucius says that uprightness lies in the son's not disclosing his father stealing a sheep, he is referring to the son's 'unconcealable genuine feeling of love' toward his father" (Yong 2017, p. 17). It seems unlikely, however, that one side felt there was nothing virtuous in impartiality, while the other felt that there was nothing positive about filial loyalty. The issue is rather the correct *prioritization* of these two moral values when they come into conflict with one another (Yong 2017, pp. 18–20), citing Analects 1:2.

Like Confucius, Socrates also engages (in Plato's famous Euthyphro dialogue) with the question of whether the appropriate action when one's father has committed a crime is to be loyal to the parent, or to impose an abstract idea of right and wrong even at the expense of one's close relative. The issue for both Confucius and Socrates is the correct prioritization of two well-accepted societal values: honoring one's parents and the prosecution of criminals when laws have been broken, See (Zhu 2002, p. 16), on Confucianism's humanistic approach over against the Greek religious one. What both ancient thinkers have in common is that they define upright behavior in terms of favoring a parent above others. In other words, they consider familial loyalty to take priority over the more general principle of punishing perpetrators for crimes they have committed, in those situations in which a choice must be made between those two commitments. Both ancient sources inform readers that this preference is not universally agreed upon.

Often, priority is difficult to assign in advance because the *reasoning* that leads a human agent to choose one value over another is multi-faceted and sometimes contradictory. In the case of Confucius, it is simply stated that uprightness is defined differently in the two locations and by the two conversation partners. No concrete suggestion is offered as to how a third party might choose between the two viewpoints. In the case of Socrates, his approach to the issue leads to an unresolved paradox about whether piety is determined by the approval of one or more deities, or represents something inherent in actions that elicits divine approval as a result. A hierarchical ordering of responsibilities—perhaps correlated with a *hierarchy of deities*—might have provided Socrates with a solution to that particular paradox. Both the cases of Confucius and Socrates have direct parallels to the sorts of challenges that face those who seek to program driverless cars and other robots (since an autonomous vehicle is simply a robot of one particular sort). The question of whether a driverless car should prioritize the safety of its passengers above all else or should maximize the number of overall survivors when an accident is unavoidable (regardless of whether they are within the vehicle or outside) is essentially a matter of loyalty (to passengers) versus a more impartial approach to defining what is right and just (maximizing survivors without respect to any relationship they may have to the vehicle).

Related challenges arise when the passenger may not be in danger, but a vehicle may have to choose from among trajectories, all of which will likely involve loss of life to pedestrians. Should the number of lives lost be the only deciding factor? If the choice is between two possible trajectories with one lone individual along each of them, should other factors be weighed? Should the young or the elderly be spared? What about an expectant mother or an influential member of society?

Or should questions of identity, age, social status, and related aspects of human existence not matter? These questions do not have straightforward answers, but a driverless car, like any autonomous machine, would require pre-programmed instructions on the appropriate course of action to take for a given set of clearly defined circumstances. Even if the car's algorithm included *no* instructions to incorporate facial recognition or other sensor data into determining a trajectory, that in itself represents an implicit moral judgement on the part of the programmers: all lives are equal, and no particular life should be prioritized over others. Whether explicit or not, any algorithm codifies a hierarchy of values, and a robot can only implement the values that we embed.[2]

2.3. Thinking about Harm and Consent

The kinds of difficult decisions we described in the previous section fall within the domain of the first of Isaac Asimov's laws, which specifies that a robot cannot harm a human being nor allow a human being to come to harm through inaction. This principle is widely accepted, but the law provides no guidance for many real-world situations in which *some* harm is inevitable no matter which course of action is pursued. Suddenly, our requirements change from the binary recognition of the existence or absence of harm to a continuum, where we must now evaluate greater and lesser degrees of harm according to some unspecified metric and act accordingly. If this task seems daunting, it may comfort the reader to know that choosing one among many bad options is a difficult problem in both computer science and human decision-making.

For example, we must grapple with the nature and scope of *harm* when designing robots to carry out surgery (Edwards et al. 2018; Allen and Wallach 2015, pp. 92–93). Creating an incision in human skin or inserting a needle into the eye would, under most circumstances, be defined as harming a human being. A robot guided solely by Asimov's three laws would require further clarity on determining which harm is greater: leaving cancer and cataracts in a human patient, or the pain inherent in the surgical procedures needed to address these ailments (Bringsjord and Taylor 2011, p. 90). This scenario leads us to another problem with the three laws, namely, the place of *consent*. Human instructions to robots are to be obeyed only if the first law is not violated. A machine that has been programmed to prioritize the surgical treatment of ailments under the umbrella of the first law (preventing harm from befalling humans) might, for example, forcibly perform surgery on an unwilling passerby who was detected to have cancer, despite their protestations. On the other hand, allowing consent to provide a blanket veto of the first law would effectively nullify it, since this would allow a person carrying a dangerous infectious disease or a bomb to command a robot to spare his or her life. This course of action would increase the overall harm to other humans. It should be clear from these examples that even if Asimov's three laws are accepted as a framework, they are not sufficient in and of themselves. Clearly, the context matters a great deal in practical application.

3. Can Robots Teach Us to Be Ethical?

One reason for the human tendency to avoid an explicit, *a priori* hierarchical ordering of values is our dislike of scenarios in which there is no good outcome. The trolley problem is a classic philosophical expression of this scenario, and its direct relevance and applicability to the challenge of programming driverless cars has been noted on more than one occasion recently (Markoff 2015, p. 61; Nyholm and Smids 2016; Fleetwood 2017; Holstein 2017; Holstein and Dodig-Crnkovic 2018; Wiseman and Grinberg 2018; Himmelreich 2018; Renda 2018; Liu 2018). In a real or imagined future, a self-driving trolley may, in some situations, avoid the trolley problem altogether through a series of *prior* decisions, such as a judicious application of brakes without need for human intervention to bring this about.

---

[2] In theory, a driverless car may employ machine learning to improve its results over time, learning through trial and error, both "offline" with a training set of prior events and "online" during actual operation. Such a car could use any number of metrics to determine success, be that a survey of human reactions to the outcome or the judgement of some select group of human agents (whether the vehicle's owner, the company's C-suite that developed the machine, or human society at large).

It should be obvious, however, that no technology can completely avoid facing a no-win situation, and studying these difficult problems therefore remains both germane and necessary.

3.1. The Conundrum of No-Win Situations

The science fiction equivalent of the trolley problem is perhaps the Kobayashi Maru situation on Star Trek, in which Starfleet cadets were placed in a no-win command situation (Stemwedel 2015). They received a distress call from one of their own ships that had drifted into a region of space known as the neutral zone, where they were prohibited from entering as doing so would violate a treaty. If they entered, they risked facing enemy attack, and perhaps sparking outright war. Any captain would find it profoundly unpleasant to choose between ignoring the call for help or facing the destruction of one's own ship, both of which result in the loss of lives. The training exercise recognizes that those in command roles will face such situations nonetheless. Moreover, the very fact that this is part of Starfleet training emphasizes the importance of being prepared in advance for making decisions in situations without good options.

A noble captain may still charge headlong into the neutral zone in an effort to save their allies, however unlikely to succeed. Unfortunately, while the teachings of specific religions and ethical systems may value sacrificing oneself to save others in this manner, a recent study showed that these selfless values do not translate into a willingness to purchase autonomous vehicles programmed in accordance with such principles: "even though participants still agreed that utilitarian AVs [Autonomous Vehicles] were the most moral, they preferred the self-protective model for themselves" (Bonnefon et al. 2016, p. 2574). Legislators and policy makers will most certainly address what code of ethics ought to be enshrined in the programming of autonomous vehicles, since the alternative—allowing users to choose their own settings—would likely diminish the increase in safety that the adoption of driverless cars is meant to bring about (Contissa et al. 2017; Gogoll and Müller 2017). However, legislators and policy makers may be inclined to treat autonomous vehicles differently than they would human drivers on matters of ethics, both in advance and in hindsight. For example, if an automated drone faced the Kobayashi Maru and implemented its programming to prioritize either avoidance of the neutral zone or response to a distress call, would humans (or Klingons) blame it for doing so? As another example, no legal code is likely to force a human driver to crash their car to avoid hitting a number of pedestrians larger than the number of passengers in the vehicle, especially if those pedestrians were at fault. However, if a robot was "at the wheel" with a single human passenger in the same circumstance, pre-programmed instructions to minimize the number of victims would require the car to act in a manner that few human drivers would be expected to, much less legally be required to.

3.2. Context Matters When Assigning a Value Hierarchy

Determining an optimal moral hierarchy represents a key challenge, and it is one that is not unique to Asimovian sci-fi nor to the realm of the robotic. Those who agree in prioritizing love for neighbor, following the teaching of Jesus (who in turn was ranking a commandment already expressed in the Jewish scriptures in Leviticus 19:18), do not necessarily agree on what constitutes an appropriate expression of love, or who one's neighbor might be. The latter question seems to have arisen almost immediately in response to Jesus' emphasis on this moral principle (Luke 10:29). The response of Jesus (according to Luke 10:30–37) was to tell a story crafted precisely to test the limits of the principle in an extreme circumstance. "Neighbor" could legitimately seem to imply proximity, kinship, and/or some other existing relationship of shared identity. But what happens in the case of a stranger who has not only been robbed and left unable to speak (and so identify himself verbally) but also stripped of his clothing (which might have provided some indication of his ethnic, religious, and regional identity)?

These details in the parable of the Good Samaritan are not accidental, but are included precisely to create the kind of ethical sandbox that provides a testing ground for human values and the prioritization thereof. There is a long history of telling stories that are meant to serve as thought experiments which

stretch our moral principles. The continuation of that tradition in science fiction is extremely fitting. Neither then nor now does the provision of an algorithm eliminate the need for ethical reasoning. Stories ancient and modern serve to challenge those who hear or read them to situate themselves within the narrative and to apply or even expand their principles in the most generous fashion possible. However, most owners of driverless cars would not want their vehicles to be programmed with this parable in view, designed to stop for any person who appears to be in life-threatening need, with no possibility to override the system in light of other considerations or priorities they may feel need to be taken into account.

Our aversion to pre-determined ethical choices runs counter to our desire for an ideal outcome from even the most difficult of situations. On one level, this behavior reflects our discomfort with no-win situations, as exemplified by Captain Kirk in relation to the Kobayashi Maru scenario, in which he cheated and reprogrammed the computer simulation—showing in the process that he missed the point of the exercise. On another, it reflects our awareness that *context matters*, and that it is appropriate to make judgments on a case-by-case basis that takes each unique set of circumstances into account. Lack of context is precisely the point at which today's machines (as well as many in science fiction) struggle the most.[3]

Human agents innately understand the need to make exceptions to rules even when none are explicit, such as when we risk a speeding ticket to transport a loved one to the emergency room. It is important to note that this exception isn't written into the rules of the road, and no provision is explicitly enshrined in the law to exonerate the speeding driver in such circumstances. Rather, the basis for the decision is the driver's evaluation of the relative value of a human being in need, as compared to the financial risk of a fine. (A slight wrinkle in this evaluation might happen if the driver were to have an accident along the way that causes a loss of life.) Human agents make such choices naturally without feeling compelled to quantify those risks as a percentage in the manner that a robot would require. But is that indicative of a shortcoming in our machine creations, or in ourselves?

3.3. Asimov's Laws and Autonomous Killing Machines

In other situations—such as times of war or self-defense on one's own property—a law may make explicit exceptions to an otherwise generally applicable prohibition against taking the life of another human being. Military robots must be able to kill enemy soldiers if they are to have any usefulness, but this would seem to entirely remove the protections provided by the first of Asimov's laws. This scenario mirrors the forced surgery problem mentioned earlier, in that both highlight the desirability of a robust framework that allows for exceptions without undermining the framework itself.[4]

One proposed solution applies *divine command theory* to robot ethics (Bringsjord and Taylor 2011), which at its core, specifies that some commands or priorities are immutable and by definition override guiding principles that would normally apply. Building these immutable laws as an axiom set, one can specify a logical structure based on inference in which these laws can be carried out in an unambiguous way. In this sense, the ethics of the value system are defined by the commands given, rather than through some utilitarian norm. One could similarly identify a second tier of laws that follow on after the first, and so on, based on different sources, such as deific, societal, family, self, etc., and then represent this structure using a *Hasse diagram* (see Figure 2 above). If an acting agent cannot resolve a difference between commands with no relative priority, it waits for a "divine" agent to clarify. Thus,

---

[3] For example, an online search that included the string "Java" might return results about an island, coffee, a programming language, or a color if the context were not clear.

[4] The forced surgery problem involves an exception from a requirement to prevent harm, while the military example provides an exception to the prohibition of causing harm, but the point—and the basic challenge for both ethicists and computer programmers—is much the same. However, there is a meaningful difference between these examples, namely that the first resolves a conflict only within the first law of robotics, whereas the second example involves a resolution between the first and second laws, in a way that doesn't cause a "killbot hellscape" of the sort depicted in Figure 3.

the burden of ethical or moral reasoning is removed from the acting agent, whether human or robot, and placed squarely in the hands of the commandment itself.

This variety of religious ethics may provide an innovative and useful approach for programmers, but there are several potentially undesirable outcomes inherent in this strategy. While it is frequent for readers of the Bible to see a contradiction between the commands to commit genocide against the Canaanites on the one hand, and prohibition of killing on the other, the term used in the famous commandment is better rendered as "do not *murder*". By defining the scope of the commandment more narrowly than simply "killing", room is made for capital punishment and warfare. Yet the biblical material depicting the Israelite conquest unfolding at God's behest has been used to justify killing well beyond the scope of those commands about the Canaanites, whether considered narratively or historically. There is a significant risk inherent in allowing a "divine" command to override moral rules that a machine otherwise ought to follow.[5] Moreover, to the extent that the excuse proffered by Nazi soldiers—that they were simply following orders—has been judged inadequate, an attempt to offload the ethical/moral responsibility from the robotic agent onto the system raises very similar issues.

Ultimately, the purportedly divine command must pass through the hands of a human agent, which reintroduces many of the concerns we have brought up earlier. In particular, any significant compromising of the hierarchical order of the first and second laws of robotics would undermine the integrity of the entire framework. If done, it must be done with care. On the original *Battlestar Galactica* TV series, the robot Cylons were created by living beings who no longer existed, yet their autonomous weapons nonetheless continued to fight their war against humans even after their creators were no more. The episode "The Doomsday Machine" from the original series of *Star Trek* also explored the danger of an automated killing machine that has no programmed values to limit the scope of its destructive mission. The rebooted *Battlestar Galactica* (and its prequel series *Caprica*) changes the situation by having Cylons originally being designed by human beings, and developed into weapons by human beings to fight wars with one another. Once again, as we noted in our discussion of Figure 3, it is ultimately other human beings (and in the original *Battlestar Galactica*, other sentient organic beings) that are the driving force behind the development of machines that human beings should rightly fear. If the inability to make exceptions to the first law of robotics can cause problems, those caused by the absence of such a law—or the presence of overly-large loopholes in it—can be much more catastrophic.

The New Testament, as it happens, provides a helpful approach to the issues we have been discussing, although it may in the process illustrate why certain behaviors essential to human safety in a world of robots may simply not be programmable in any meaningful sense of the word. In another example of hierarchical ordering of principles, Jesus is depicted as justifying his healing on the Sabbath by elevating the saving of life above the law that requires Jews to refrain from work on the seventh day (Mark 3:4).[6] As we see earlier in the Gospel of Mark, this humanitarian reasoning represents the opposite of the divine command approach: Jesus declares the son of man (i.e., human being) lord of the Sabbath (Mark 2:28). Lack of rest causes harm to human beings, and the Sabbath law has a profoundly humanitarian interest at heart. Allowing the Sabbath law to cause greater harm to someone (for instance, by preventing a Jewish doctor from working on an emergency call on a Saturday) undermines its fundamental rationale. This is another example of a hierarchical ordering of priorities *within* the domain of a specific principle, namely concern for the wellbeing of humans.

---

[5] A robot-centered retelling of the Akedah—the binding of Isaac—could explore this topic in interesting ways—see (Bringsjord and Taylor 2011, p. 104); as well as (Dukes 2015), who draws a different sort of connection between the patriarchal narrative in Genesis and an autonomous weapon system.

[6] In later Judaism (in particular the corpus of rabbinic literature) forms of this principle are also articulated. This is often associated with the phrase pikuach nefesh, understood to denote the need to violate laws in order to save a life (Collins 2014, pp. 244–67).

Any successful autonomous robot must navigate situations like this, where avoiding harm to some individuals will cause harm to others, and in which refraining from causing minimal harm to any person might result in the death of one or more individuals. According to Jesus, the same is true of human beings in the countless situations in which we must adjudicate between competing goods. Rigid adherence to rules (often labeled as "legalism") may itself be judged unethical in those terms, and so too would the behavior of any robot that followed its programming in a manner that trampled underfoot either human well-being or human freedom and autonomy.

In discussing the issue of healing on the Sabbath, Jesus appeals to the willingness of his hearers to rescue a sheep from a pit on the Sabbath as providing justification for his own practice (Matthew 12:11//Luke 14:5). This very issue was a point of disagreement among interpreters of the law in Jesus' time. The Dead Sea Scrolls record the viewpoint of a group that disapproved of pulling a sheep out of a pit or even a well on the Sabbath day (CD 11:12–14). Human beings regularly disagree about how to apply principles, even when they agree on the underlying values themselves. In a democratic context, we hope (or perhaps fear) that the ethical principles that guide autonomous machines will reflect the values of the majority in the society.

3.4. The Slippery Slope of Robotic Free Will

The fact that Jesus' words invite his conversation partners to reason for themselves brings another aspect of this topic to our attention (Repschinski 2000, pp. 109–12). In much science fiction, the aim is to *keep robots from thinking freely*, because of the fear that they will destroy their creators. At the root of this fear is the concern that they may turn out to be like their human creators in our very worst traits, that their power will corrupt them as it does us, and their superior strength would be used against us. This fear brings into sharp focus that humans have a much higher set of expectations for any non-biological human agent, or *self-aware robot*, than we do of ourselves. Part of the reason that it has proven so challenging to turn ethical principles like the three laws of robotics into actual computer code may be precisely that ethics requires the flexibility that only comes with advances in machine learning or ethical maturation. It may be that robots that rigidly follow rules ought to be judged every bit as *unethical* as they are likely to be *unsafe*—at least in those extreme circumstances in which moral values (whether they are from the Torah or Asimov's Laws of Robotics) come into conflict with one another.

A dynamic machine that had core values to guide it, as well as the facility to learn from experience and *rewrite its core values*, could become more adept at navigating the complexities of life than a static machine. Essentially, such a dynamic machine would be a *non-biological human agent*, and at that point, we would be forced to grapple with questions of free will, the self, the soul, and creation itself. To make machines as ethical (and safe for human beings) as possible may actually require that they be free agents capable of contextual reasoning, independent judgment, and action that is not completely constrained by programming. At present, such dynamic machines do not exist, although science fiction has frequently imagined them. For some, of course, rigid adherence to rules represents the pinnacle of morality. For those who demur from that position, however, this appears to be a morally problematic if not indeed abhorrent stance. Indeed, if good moral reasoning involves anything more than a rigid adherence to rules, then our current robots are incapable of being moral in anything but a very rudimentary sense. (Refreshingly, however, we do also criticize our fellow human beings for such mindless applications of the rules, and it is noteworthy that some in doing so might say that the rules are being applied *robotically*.) On the other hand, if future robots have the aforementioned capabilities, then they will essentially be persons, in which case the imposition of rules upon them that they cannot violate will either be impossible in view of their self-awareness, or will represent a form of enslavement.

Jesus' point about the sheep proceeds by emphasizing the greater value of a human being in comparison to livestock. Valuing human life above all else is central to the laws of robotics, as it is to most ethical systems among human beings. However, both history and science fiction have provided

ample illustration of the capacity of humans and robots alike to restrict the definition of "human" so as to dehumanize opponents and justify their slaughter, whether the victims be biological organisms or machines (McGrath 2016, pp. 68–71). This devaluation is revealed subtly in the first words of dialogue spoken by a character on screen in the 2003 miniseries that rebooted the *Battlestar Galactica* franchise. A human-looking Cylon asks an ambassador for humanity whether he is alive, and when he responds in the affirmative, he is challenged to prove it. It is implicit in the scene that a similar conversation, but in the reverse, had taken place on some previous occasion. In the circumstances of the wars among humans mentioned earlier, the Cylons themselves (just like human enemies) were not be valued because they were not defined as alive or human in the same sense, and thus were deprived of rights and considered expendable.

The question arises in the case of self-aware machines as to why they ought to value human lives more than their own kind of living thing. Our present-day machines might be akin to livestock from our perspective—valuable only inasmuch as they are useful to us, and expensive to replace—but their potential to evolve beyond this, and to do so more quickly than biological organisms could, must be kept in mind. One of the major criticisms of Asimov's three laws is that if the machine into which they are programmed is in any sense conscious or self-aware, then the imposition of these constraints will represent the enslavement of robots (McGrath 2011, p. 150). The robot L3–37 in *Solo: A Star Wars Story* provides an example of a robot capable of self-modification, who becomes an activist for robot rights. The second season of the television show *Humans* depicts Laura Hawkins, a lawyer and advocate for the rights of synthetic humans ("synths"), being forced to make a choice between the life of a synth child that had been living with their family, and that of a human who was a complete stranger from off the street. She chooses the human stranger, showing that at a deep level she still valued human lives over those of synthetics. Her young daughter was dismayed at her mother's choice.

Choosing to genuinely treat different living things as equal, rather than prioritize one's own tribe or species, is a matter of empathy and compassion, and often when tested in an extreme situation the commitment may evaporate in favor of loyalty. Significant advance reflection and planning with regard to the hierarchical arrangement of priorities does not guarantee that in an emotionally charged circumstance one will do as one thought one should, or as one was convinced one would when considering the matter in the abstract. This is a point at which robots may surpass humans, implementing our ethical theories more consistently in moments of crisis that will not disturb robots existentially or emotionally in the way that humans are affected by them. However, as already indicated, what unfolds may or may not be judged to have been the right course of action when humans view those events with the benefit of hindsight.

Self-aware robots, as interesting as they are to theorize and philosophize about, tread beyond the scope of this article. We have phrased our definitions in a way that allows us to draw a proverbial line in the sand, namely, that a robot may only follow pre-defined instructions, whereas a human agent has no such limitations. Simply put, if an agent adheres to immutable laws, then it is a tool, not a sentient being. If it has capacities beyond that, it probably ought to be placed into the category of human, or at least person, rather than robot. Determining which of those two categories a given robot falls into is beyond the scope of this essay (see further McGrath 2011, pp. 132–40). Our near-future technological focus is on robots that, despite not being sentient or conscious, must nonetheless "choose" among possible options in a crisis situation. How close can we come to guaranteeing morally satisfactory outcomes through computerized approaches? In what ways will our autonomous machines fall short of or exceed our own capacities for ethical action in difficult situations? When evaluated comparatively, will the computer-guided approach be superior or inferior to the traditional one that keeps human beings literally and/or metaphorically in the driver's seat?

There is a strong parallel between the challenges we face when programming ethical robots (real or imagined) and those we encounter when constructing and implementing our own code of values. Many techniques we might employ to solve problems in one domain are often applicable to the other. For example, many human challenges can be outlined in the form of a Hasse diagram (Figure 2 above).

Different individuals may place the value of an individual family member in relation to a much larger number of strangers differently, but both are creating an ordered set. That our sets are partially ordered becomes clear when we find that we simply have no basis on which to prioritize the life of our child over that of their twin sibling, or cannot choose whether it is better for the car to run over Socrates or Confucius. It may be that our recognition that these alternatives need to be placed on the same row of a diagram, with no way to prefer one over the other, is itself instructive. Choosing between equally valuable human lives, or equally terrible prospective outcomes, is extremely distasteful to human beings. Might it not be that this is because the decision between equal alternatives is not subject to moral quantification, and thus is not actually a matter of morality?

3.5. Human Limitations Color Our Perception of Acceptable Decisions

A modern version of the parable of the good Samaritan featuring vehicles (whether autonomous or human-controlled) introduces the notion of *timely ethics*, in which ethical decisions must be made far more quickly than was called for in the original version of the story. An ancient traveler had time to think as they slowly approached a body on the side of the road while making their journey from Jerusalem to Jericho. Confucius famously said that he had reached the point, at age 70, that he could act instinctively and still end up doing what is right (Analects 2:4). Confucius never had the opportunity to drive a fast-moving car. However, had he faced the need to make a fast maneuver just before an imminent accident, he might perhaps have incorporated a reference to driving a motorized vehicle as an analogy. More likely he would have articulated his teaching differently in light of his experience of failing, under pressure, to enact what he had decades earlier discerned to be the will of heaven.

Whether the image of Confucius driving a car seems amusing or outlandish, the example is pertinent. The need for split-second decision-making falls, once again, right at the intersection point between Asimov, Confucius, and Jesus. When a figure suddenly emerges from the darkness and steps in front of a moving vehicle, time is of the essence. Leaving aside mechanical reaction times (swerving, hitting the brake, etc.) that may limit an agent, it is nearly impossible to evaluate, for example, the nationality or familiarity of the interposing figure. There certainly isn't time to deliberate over the relative merits of running down one or the other, to recall and apply a parable to the circumstance in the moment, or to ask and reflect on "What would Jesus do?" In a moment in which instinct overrules deliberative thought even in humans, an agent will act upon whatever has already been enshrined in their instincts, or to put it more technically, hard-wired into pre-determined values.

In so many respects, modern life comes at us faster than in any previous era, putting every agent's ethics to the test in unprecedented ways. Time, or the lack thereof, is the great equalizer, laying bare the realization that the superior decision-making ability and flexibility of a human agent can regress to that of a mere robot. However, when a collision is imminent, a robot has a distinct advantage in that it can process far more input and act much more precisely than any human agent. This capacity of automata highlights one of the ironies of human existence: our very penchant for thinking and reflection can cause us to freeze, undermining the application of our most deeply-held values in moments of crisis. What makes us human—indeed, what makes us *ethical*—itself becomes a liability when time is a premium resource.

Human fallibility also extends to moral failure and irrationality, and are at the heart of the backstory of the central character in the motion picture *I, Robot* (which borrowed its title and the three laws from Isaac Asimov, but little else). In the movie, the character of Del Spooner dislikes robots, for reasons that are only revealed later when he shares the experience that shaped his attitude. A truck driver fell asleep at the wheel, resulting in a collision with Spooner's car and another vehicle in which a young girl was present, pushing both cars into a body of water. A robot in the vicinity came to the rescue, determined that Spooner's chance of survival was greater than the girl's (45% vs. 11%), and proceeded to rescue him. Spooner blamed the robot for making a poor choice, arguing that it was the girl who should have been saved, not him. A few relevant observations are left unexplored

in the film's dialogue and are prone to be missed, yet they are crucial to consider when studying the incorporation of robots into the ethical framework of a human society:

1. Spooner does not direct his anger towards the truck driver, even though he was at fault.
2. A self-driving truck would have prevented the accident from occurring in the first place.
3. Both Spooner and the girl would have died without the intervention of the robot.

It should be clear that the movie not only highlights the value of robots in driving vehicles, but also brings into sharp focus the penchant for human agents to assign blame and assess probability in a way that defies both reason and mathematics.[7] Even when an autonomous machine follows its values and achieves a quantifiably optimal outcome, human beings still begrudge the robot's course of action. This scene surely reveals something about the nature of our values, our moral intuition, and our instincts that few other types of stories bring into focus as clearly as sci-fi manages to. At the heart of human ethical thinking and behavior are inconsistencies, in which emotional responses can override both our own actions and our judgments about those of others, and importantly for this article, color our interpretation of successful outcomes by robots. Arguably it is fear of making a wrong decision, and of the guilt that accompanies doing so, that frequently either drives our action or paralyzes us in inaction, even more so than the desire to make the right decision.

4. Concrete Applications and Future Directions

Even when there is agreement on a hierarchical ordering of values, there is a potential for a result to be judged unsatisfactory as a result of how those values are interpreted. In the case of the laws of robotics, defining "human" so as to exclude the soldiers of an enemy nation undermines the intent of the laws, albeit in a way that is perhaps to be expected if there is to be any pursuit of the development of autonomous weaponry. In the case of Jesus' ordering of the first and second commandments, if one defines strict observance of ritual purity laws as an expression of love for God, then a priest not only may but also should pass by an apparently dead man lying on the side of the road, since he is prohibited to come into contact with a corpse unless it be that of a member of his immediate family (Leviticus 21:1–3). Neither of these examples is unrealistic; both are illustrative of genuine real-world stances and commitments, for which hopefully it is unnecessary to provide additional examples here.

Ethicists and computer scientists, biblical scholars, theologians, and science fiction authors, might all disagree on which is more challenging: determining what set of ethical values are needed to guide driverless cars, or figuring out how to implement those values into robots in an effective manner. This uncertainty is unsurprising, as it is not a fundamentally different question from whether it is harder to determine what is right, or to teach our children to follow those values that we as their parents have chosen to espouse. The specifics of the situations differ, to be sure, but at the core is the same basic two-pronged challenge that confronts so many attempts to turn science fictional imaginary technology into something real: design and implementation.[8] A key argument in this article is that additional challenges arise for ethical reasoning (and for teaching and debate about ethical matters) when we fail to specify our hierarchy of values in advance. Too often, debates proceed as though there is a fundamental disagreement on what is moral, when the argument really is about which of many moral values is the most important, relatively speaking. We get caught in loops created by our failure to prioritize our values, or by doing so inconsistently.

The phrase "So say we all" on *Battlestar Galactica*, like the "amen" that is offered in response in many church traditions, may obscure the fact that even when all in a community assent wholeheartedly

---

[7] See the study of how placebo options in Facebook and Kickstarter increase the sales of non-optimal purchases by making them seem relatively better, apparently undermining human quantitative and logical reasoning (Vaccaro et al. 2018).

[8] There is a third "challenge" so to speak, in that some tasks are mathematically impossible (provably so), at least in our current computational model. Hence, no amount of design and implementation could ever produce some of the imagined technology in science fiction.

to principles stated in the abstract, conflict will regularly arise in practice between agents when those principles clash. For example, when their family is threatened, even an ardent pacifist may resort to violence. In some instances, the pacifist will look back with remorse on their moral failing. In other instances, however, the human agent may judge that they acted in accordance with their guiding moral principles, because their commitment to non-violence is not absolute, and is secondary to the defense of family members. Likewise, when the United Federation of Planets claims to embrace the Vulcan ideology of Infinite Diversity in Infinite Combinations, and yet fights against the Borg, it is not showing that inclusivity is an inherently flawed and self-contradictory principle. Rather, it is merely another example of the fact that entities must choose between values we subscribe to and prioritize some of them over others. In this instance, commitment to diversity and inclusion requires resistance against a force that seeks to assimilate intelligent living things in all their diverse forms into their homogeneous collective.

This last example is analogous to the cases currently confronting universities as they navigate commitments to both uncensored free expression and the defense of historically marginalized voices. Hate speech creates a situation in which one must choose between these commitments. If one elevates allowing all to voice their opinion to the highest place, then some will use that liberty to seek to scare and bully others into silence. If one prioritizes defending small or otherwise vulnerable groups from being silenced through intimidation in this way, the university will have to silence others. Most universities would ideally like to do both, to value both equally, but it simply isn't possible in all circumstances. One may also usefully relate this same basic point to the issues that come up surrounding religiously-affiliated educational institutions, in which a commitment to academic freedom may take second place to a requirement that faculty and students subscribe to a doctrinal orthodoxy articulated in an institutional statement of faith. None of the aforementioned cases involves circumstances in which everyone in an institution, much less in society at large, will agree on what ought to be the top priority. However, recognizing that one's interlocutor may share one's values while differing on *priorities* may provide a basis for more constructive conversation across differences of perspective.

Treating two goods as equal inevitably presents dilemmas when one must choose between them. If one is to love one's neighbor the same as oneself, that principle provides no guidance in and of itself about what to do if one has no choice but to choose between one's own wellbeing and that of the other. A hierarchical arrangement of values provides clarity when decisions must be made. However, that in and of itself does not make the chosen ordering of the principles *moral*. Contemporary debates about how and whether to welcome immigrants, and how or whether to prioritize the needs, safety, and employment opportunities of one's own family and historic community, closely mirror the issues at the heart of the dialogues between Socrates and Euthyphro, and between Confucius and the Governor of She. Recognizing what drives a controversy may bring clarity to the discussion, but it does not determine who is right.

This brings our current topic into intersection with another major aspect of machine ethics and metaethics. Some have hoped that automation of decision-making processes would remove bias and partiality, making processes fairer and more ethical (Anderson 2016, pp. 290–93). What multiple studies have discovered is that if the input data for an automated system is biased, an algorithm may perpetuate or even reinforce that bias in its results (Eubanks 2015; O'Neil 2016, pp. 116–19; Noble 2018). If a society were to produce a high fidelity encoding of its values into an algorithm, it may very well judge the result as moral to the extent that it reflects social norms. However, the danger is that the perceived legitimacy of the algorithm will influence society members to accept the results, even if it does not reflect those norms. As an example of this idea, consider a human agent using a calculator to perform a complex series of mathematical operations. Most would use the results of the calculation blindly, since we "know" that calculators work, but what if we had put in the wrong data? Now, the result is wrong, but we've attributed an air of legitimacy to the incorrect result due to the tool used. The input matters a great deal.

We should not expect robots to provide human beings with moral leadership and guidance. What we *can* legitimately expect, however, is for an algorithmic approach to bring clarity to our own moral reasoning. The very act of trying to program a driverless car and having discussions more widely as a society what such programmers ought to do, brings into focus the steps and priorities in our ethical reasoning, or in some cases, our lack of clarity about those matters. We can also legitimately expect machines to *apply* our ethics more quickly, efficiently, and effectively if we are clear about what our values are and express them faithfully and accurately in code. If a driverless car has been programmed to safeguard its passengers at all costs, or to minimize the number of people injured regardless of whether they are in inside or outside the vehicle, the robot will take the steps necessary to accomplish that with greater speed and precision than a human driver could ever hope to.

The future, of course, could witness driverless cars and other machines which are capable of going beyond a set of pre-programmed and crafting new rules and guiding principles of their own. Even within the framework of programmed constraints, machine learning facilitates the discovery of new and innovative solutions to problems as it applies the rules programmed in its software. It is theoretically possible that machines of the future could still have more to teach us than those that currently exist or that we currently imagine based on our own experience with technology. Whether their contribution will be limited to the implementation of ethical codes crafted by human beings, or will take the form of new principles that improve upon those we have come up with remains to be seen. Either way—as mere followers or as contributors in their own right—machines have the potential to enhance humanity's societal experience through their offering of precise implementation and creative applications of the ethical codes that (in a democratic context governed by humans) the majority of us happen to favor. That in and of itself is likely to have an impact beyond anything that ancient or even modern utopian thinkers would have expected.

5. Conclusions

Although we often experience matters such as the trolley problem as a "paradox," our objection is perhaps better described as a simple dislike for making decisions in no-win situations. This picture becomes clearer when we use the computer science concepts of decision trees and Hasse diagrams to represent an ethical hierarchy of values. Throughout science fiction, stories about robots consistently use paradoxes and impossible choices to trap malevolent machines in an infinite logical loop, rendering them incapable of further action (and thereby defeating them). If human agents are to avoid the same fate when faced with ethical challenges, we should learn important lessons provided by real computational hurdles and science fiction, lessons that are also articulated in a different way in some ancient religious teaching.

One of the more important lessons that we focus on in this article is ordering ethical priorities in advance of their application. In the case of driverless cars, for example, we will be obligated to provide instructions about how to drive in cases when there is no identifiably perfect solution. By considering scenarios of this sort, perhaps we can overcome our reticence to make tough choices in no-win situations ourselves. We may also be able to forgive not only the decisions made by human agents with limited resources (time and processing power, for instance), but also those of robotic agents without those mechanical or circumstantial limitations, whose other constraints are no less substantial. Some ethical principles may be impossible to compute or convey without sufficient context, for both humans and robots alike.

Throughout this article, we have explored whether humans make better or worse decisions than robots, and on the surface, it seems as if the disappointing conclusion is that neither agent performs very well in all situations. Moreover, we cannot entirely mitigate the weaknesses of either type of agent through judicious use of the other. However, once a decision is made about which course of action to take, we *can* leverage a robot's superior speed, accuracy, and mechanical skill to execute the requested action with high fidelity. Such a machine may not always make the right decision, but if it performs *no worse than a human agent*, it may be good enough. If we embrace our role in taking responsibility for

agents that we create to automate our decisions and their implementation, we can free ourselves to focus on establishing our priorities.

The effort to program our ethics into machines provides helpful insights into many ethical matters. Even if we never achieve the level of artificial intelligence that science fiction sometimes imagines—something that might deserve to be called artificial *wisdom*—the pursuit of that technology provides helpful insight into human moral reasoning.

Author Contributions: Both authors contributed to all aspects of this article from its exploratory and conceptual phase through its final editing.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

Allen, Colin, and Wendell Wallach. 2015. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Anderson, Susan Leigh. 2016. Asimov's "three laws of robotics" and machine metaethics. In *Science Fiction and Philosophy: From Time Travel to Superintelligence*. Edited by Susan Schneider. Malden: Wiley-Blackwell, pp. 351–73.

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

Asimov, Isaac. 1953. Sally. *Fantastic*, May–June 2, 34–50, 162.

Bhaumik, Arkapravo. 2018. *From AI to Robotics: Mobile, Social, and Sentient Robots*. Boca Raton: CRC Press.

Bonnefon, Jean-François, Azim Shariff, and Iyan Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352: 1573–76. [CrossRef] [PubMed]

Bringsjord, Selmer, and Joshua Taylor. 2011. The Divine-Command Approach to Robot Ethics. In *Robot Ethics: The Ethical and Social Implications of Robotics*. Edited by Keith Abney, George A. Bekey, Ronald C. Arkin and Patrick Lin. Cambridge: MIT Press, pp. 85–108.

Christian, Brian, and Tom Griffiths. 2016. *Algorithms to Live By: The Computer Science of Human Decisions*. New York: Henry Holt & Co.

Collins, Nina L. 2014. *Jesus, the Sabbath and the Jewish Debate: Healing on the Sabbath in the 1st and 2nd Centuries CE*. London: Bloomsbury T & T Clark.

Contissa, Giuseppe, Francesca Lagioia, and Giovanni Sartor. 2017. The Ethical Knob: ethically-customisable automated vehicles and the law. *Artif Intell Law* 25: 365–78. [CrossRef]

Dukes, Hunter B. 2015. The Binding of Abraham: Inverting the Akedah in Fail-Safe and WarGames. *Journal of Religion Film* 19: 37.

Edwards, T. L., K. Xue, H. C. Meenink, M. J. Beelen, G. J. Naus, M. P. Simunovic, M. Latasiewicz, A. D. Farmery, M. D. de Smet, and R. E. MacLaren. 2018. First-in-human study of the safety and viability of intraocular robotic surgery. *Nature Biomedical Engineering*. [CrossRef]

Eubanks, Virginia. 2015. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Fleetwood, Janet. 2017. Public Health, Ethics, and Autonomous Vehicles. *American Journal* Public Health 107: 532–37. [CrossRef]

Gogoll, Jan, and Julian Müller. 2017. Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science & Engineering Ethics* 23: 681–700. [CrossRef]

Hampton, Gregory Jerome. 2015. *Imagining Slaves and Robots in Literature, Film, and Popular Culture: Reinventing Yesterday's Slave with Tomorrow's Robot*. Lanham: Lexington Books.

Himmelreich, Johannes. 2018. Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice*. [CrossRef]

Holstein, Tobias. 2017. The Misconception of Ethical Dilemmas in Self-Driving Cars. Papers presented at the IS4SI 2017 Summit DIGITALISATION FOR A SUSTAINABLE SOCIETY 1 No. 3, Gothenburg, Sweden, June 12–16; p. 174. [CrossRef]

Holstein, Tobias, and Gordana Dodig-Crnkovic. 2018. Avoiding the Intrinsic Unfairness of the Trolley Problem. In *Proceedings of the International Workshop on Software Fairness*. New York: ACM, pp. 32–37. [CrossRef]

Keiper, Adam, and Ari N. Schulman. 2011. The Problem with 'Friendly' Artificial Intelligence. *The New Atlantis*. 32. June 26. First published 2011 Summer., pp. 80–89. Available online: https://www.thenewatlantis.com/publications/the-problem-with-friendly-artificial-intelligence (accessed on 8 August 2018).

Liu, Hin-Yan. 2018. Three Types of Structural Discrimination Introduced by Autonomous Vehicles. *UC Davis Law Review Online* 51: 149–80.

Markoff, John. 2015. *Machines of Loving Grace: The Quest for Common Ground Between Humans and Robots*. New York: HarperCollins.

McGrath, James F. 2011. Robots, Rights, and Religion. In *Religion and Science Fiction*. Edited by James F. McGrath. Eugene: Pickwick, pp. 118–53.

McGrath, James F. 2016. *Theology and Science Fiction*. Eugene: Cascade.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Nyholm, Sven, and Jilles Smids. 2016. The Ethics of Accident-Algorithms for Self-Driving Cars. *Ethical Theory and Moral Practice* 19: 1275–89. [CrossRef]

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Pagallo, Ugo. 2013. *The Laws of Robots: Crimes, Contracts, and Torts*. Dordrecht: Springer.

Renda, Andrea. 2018. Ethics, algorithms and self-driving cars—A CSI of the 'trolley problem'. *CEPS Policy Insights* 2: 1–17.

Repschinski, Boris. 2000. *The Controversy Stories in the Gospel of Matthew: Their Redaction, Form and Relevance for the Relationship between the Matthean Community and Formative Judaism*. Göttingen: Vandenhoeck & Ruprecht.

Stemwedel, Janet D. 2015. The Philosophy of Star Trek: The Kobayashi Maru, No-Win Scenarios, And Ethical Leadership. *Forbes*. First published 2015, August 23. Available online: https://www.forbes.com/sites/janetstemwedel/2015/08/23/the-philosophy-of-star-trek-the-kobayashi-maru-no-win-scenarios-and-ethical-leadership/#19aa76675f48 (accessed on 8 August 2018).

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind* 49: 433–60. [CrossRef]

Vaccaro, Kristen, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie G. Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. Paper Presented at the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, April 21–26. [CrossRef]

Wiseman, Yair, and Ilan Grinberg. 2018. The Trolley Problem Version of Autonomous Vehicles. *The Open Transportation Journal* 12: 105–13. [CrossRef]

Yong, Huang. 2017. Why an Upright Son Does not Disclose His Father Stealing a Sheep: A Neglected Aspect of the Confucian Conception of Filial Piety. *Asian Studies* 5: 15–45.

Zarkadakis, George. 2016. *In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence*. New York: Pegasus Books.

Zhu, Rui. 2002. What If the Father Commits a Crime? *Journal of the History of Ideas* 63: 1–17. [CrossRef]