CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Implementing machine ethics: using machine learning to raise ethical machines
Authors	Kaas, Marten H. L.
Publication date	2022
Original Citation	Kaas, M. H. L. 2022. Implementing machine ethics: using machine learning to raise ethical machines. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2022, Marten H. L. Kaas. - https://creativecommons.org/licenses/by-nc-sa/4.0/
Download date	2026-11-17 10:27:50
Item downloaded from	https://hdl.handle.net/10468/14068



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork



Implementing Machine Ethics: Using Machine Learning to Raise Ethical Machines

Thesis presented by

Marten H. L. Kaas, M.A., M.Sc., B.A. and B.Sc.

0000-0003-0963-1004

for the degree of

Doctor of Philosophy

University College Cork

Department of Philosophy

Head of School/Department: Professor Don Ross

Supervisor: Dr. Joel Walmsley

2022

Table of Contents

Table of Contents i

Declaration vi

Abstract vii

Acknowledgements viii

Introduction and Overview

Rise of the Machines 1

Key Terms and Themes 3

Chapter Overviews 5

Chapter 1 – Machine Learning

1.0 Turing's Vindication 7

1.1 What Sort of Thing Is Learning? 8

1.1.1 Relationality 10

1.2 Learning Machines 13

1.2.1 Learning to Play Atari Games 13

1.2.2 Learning to Operate a Vehicle 14

1.2.3 Engineers Build Learning Machines 15

1.3. Machine Learning Preliminaries 18

1.4 Major Machine Learning Techniques 19

1.5 The Rise of Reinforcement Learning 23

1.6 The Whos, Whats and the Hows 26

1.7 Preparing for the Great Flood 32

Chapter 2 – A View from Somewhere

2.0 Mind and Iron: A Cleaner, Better Breed 34

2.1 What Could Go Wrong? 34

2.1.1 Perverse Instantiation 35

2.2 The Myth of Machine Objectivity 36

2.2.1 The View from Nowhere 37
2.3 The View from an Autonomous Vehicle 39
2.3.1 Consciousness As Emergent 40
2.4 Machines Are Not Value Neutral 42
2.4.1 EURISKO: Suicide and Parasites 43
2.4.2 Robot Helpers: Hyperrationality 44
2.5 The Importance of Design Decisions 47
2.5.1 To Fold or Not To Fold 50

Chapter 3 – Machine Ethics

3.0 Background 51
3.1 Autonomous Weapon Systems 52
3.1.1 Autonomy and Lethality 53
3.1.2 Machines (Can) Behave Better 55
3.2 From Computer Ethics to Machine Ethics 56
3.2.1 Computer Ethics 56
3.2.2 Machine Ethics 57
3.2.3 Fully Ethical 59
3.3 The Prerequisites for Ethics 60
3.3.1 Agency and Patiency 60
3.3.2 Autonomy 61
3.3.3 Mind 62
3.4 The Prerequisites for Machine Ethics 63
3.4.1 Machine Minds 63
3.4.2 Machine Autonomy 64
3.4.3 Machine Agents and Machine Patients 64
3.5 Updating and Redefining Machine Ethics 65
3.5.1 Ethical Alignment 66
3.5.2 Ethical Reasoning 67
3.6 Metaphysics, Psychology and the Moral Turing Test 68
3.7 Looking Ahead 72

Chapter 4 – Implementing Machine Ethics

4.0 Implementing Machine Ethics 73

4.1 Top-Down Approaches 73

4.1.1 The Explication Problem, Rigidity and Domain Specificity 74

4.2 Top-Down Approaches: Advantages and Disadvantages 80

4.3 Bottom-Up Approaches 83

4.3.1 Bottom-Up Approaches: Advantages and Disadvantages 83

4.3.2 Supervised Learning 85

4.3.3 Unsupervised Learning 85

4.4 Reinforcement Learning 86

4.4.1 Reinforcement Learning and Ethics 89

4.4.2 Implementing Ethics Using Reinforcement Learning 90

Chapter 5 – Assessing Challenges, Benefits and Risks

5.0 Reinforcement Learning Challenges 93

5.0.1 Rewards, Optimization and Hybrid Approaches 93

5.1 Implementing Ethics with Sophisticated Reinforcement Learning Methods 95

5.1.1 Un/Supervised Pre-Training 96

5.1.2 Reward Shaping 96

5.2 A Case Study: AlphaStar and *StarCraft II* 97

5.2.1 *StarCraft II*: The Basics 98

5.2.2 *StarCraft II*: Features Unique to Real-Time Strategy Games 99

5.2.3 Mastering *StarCraft II* Using Reinforcement Learning 101

5.3 *StarCraft II* and Machine Ethics 104

5.4 Risks of Implementing Machine Ethics 108

5.4.1 Failure 108

5.4.2 The Effect of Data 109

5.4.3 Corruption 110

5.4.4 Ethical Hegemony 111

5.4.5 Abdicating Responsibility and the Automation Paradox 111

5.4.6 More Moral Patients 113

5.4.7 Speculative Risks 114

5.5 Speculative Benefits 115

Chapter 6 – Morality Mirrored in Machines

6.0 The Benefits of Using Ethical Machines 116
6.1 Humans, Machines and Automobiles 116
6.1.2 Raising Ethical Drivers 118
6.2 Benefits of Building Machines 120
6.2.1 Benefits of Building Thinking Machines 121
6.3 Benefits of Building Ethical Machines 124
6.3.1 Weak Psychological AI and Ethics 124
6.3.2 Suprapsychological AI and Ethics 125
6.3.3 Ethics Is Hard 126
6.3.4 Assumptions Grounding Ethics 129
6.4 Machine Ethics, Moral Nihilism and Anthropocentrism 130
6.4.1 Implementability 131
6.4.2 Anthropocentrism and the Gap Between Explicit and Full Ethical Agents 133

Epilogue – Ethics 3.0

7.0 Being Ethical in the Age of Machine Ethics 138
7.1 Positions in the Debate on Autonomous and Artificially Intelligent Systems 139
7.1.1 Skeptics and Enthusiasts 140
7.1.2 Pessimism and Optimism 143
7.2 Tools of A Different Kind 145
7.2.1 Autonomy and Intelligence 145
7.2.2 Universality 146
7.3 The Perfect Technological Storm 147
7.3.1 The Transparency "Storm" 148
7.3.2 The Overtrust "Storm" 151
7.3.3 Moral Complacency 152
7.4 Responding to Machines: The Practical and the Theoretical 154
7.5 Virtues, Relations and Machines 156
7.5.1 Resisting Moral Complacency 158

Coda 161

Bibliography 162

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Abstract

As more decisions and tasks are delegated to the artificially intelligent machines of the 21st century, we must ensure that these machines are, on their own, able to engage in ethical decision-making and behaviour. This dissertation makes the case that bottom-up reinforcement learning methods are the best suited for implementing machine ethics by raising ethical machines. This is one of three main theses in this dissertation, that we must seriously consider how machines themselves, as moral agents that can impact human well-being and flourishing, might make ethically preferable decisions and take ethically preferable actions. The second thesis is that artificially intelligent machines are different in kind from all previous machines. The conjunction of autonomy and intelligence, among other unique features like the ability to learn and their general-purpose nature, is what sets artificially intelligent machines apart from all previous machines and tools. The third thesis concerns the limitations of artificially intelligent machines. As impressive as these machines are, their abilities are still derived from humans and as such lack the sort of normative commitments humans have. In short, we ought to care deeply about artificially intelligent machines, especially those used in times and places when considered human judgment is required, because we risk lapsing into a state of moral complacency otherwise.

Acknowledgements

I would like to offer my deepest thanks to Dr. Joel Walmsley for his guidance and unwavering positive support. It is because of your generosity, kindness and faith in me as scholar that I was able to accomplish all that I did in what felt like a very short period of time. I am forever grateful for your mentorship.

I would also like to thank Professor Don Ross for his insightful comments and advice as well as the faculty and staff in the Department of Philosophy for supporting me in my journey as student and philosopher.

A big thank you to my peers in the department for their continued friendship and guidance. A special shout out to the First Year Tutors, especially Dawn (Cuizhu) Wang.

I would like to thank Sarah, mein schatz, for putting up with all of the late nights and for showing me why it was worth it to come to Ireland.

Lastly, I would like to thank my family. Their presence can be felt even across an ocean. I am eternally grateful for their support as I finally finish my schooling.

Introduction and Overview

The Rise of the Machines

Imagine a future, perhaps one not too distant, where deliveries from companies arrive by flying drone. Autonomous vehicles ferry people and goods across land, sea and sky. Virtual assistants recommend changes in diet and exercise over breakfast while outside robotic helpers walk dogs up and down the street. Resources and energy are no longer scarce thanks to significant advances in production and extraction processes, a budding asteroid mining industry and cheap and efficient green energy sources. Disease has been virtually eradicated, and people regularly vacation in whatever virtual reality destination they desire. Work is no longer a necessity but a choice for those pursuing their passion, and humanity is on the brink of collecting its cosmic inheritance by becoming an interstellar civilization.[1] We are now, in 2022, in the opening acts of what Aaron Bastani calls the "Third Disruption."[2]

This dissertation is concerned with the technology driving this Third Disruption, namely artificially intelligent machines. The first thesis of this dissertation is that such machines are different in kind from all previous machines. The artificially intelligent machines with which this dissertation is concerned are those machines that are both autonomous and engaged in performing tasks within the cognitive domain. These features are what set artificially intelligent machines apart from all other machines and technologies. Another distinguishing feature of these machines is their ability to engage in a kind of learning. Indeed part of the reason why artificially intelligent technologies are so disruptive is because they can go beyond their creators. One does not need to be a good chess player, for example, to create a machine that learns to master the game of chess. Another consequence of a machine's ability to engage in a kind of learning is that such a machine can be put to many general uses. Like the digital computer, artificially intelligent machines are beginning to reveal their general-purpose nature. It is a common occurrence to put the same artificially intelligent machine (i.e., the same underlying neural network and learning algorithms, collectively the "architecture") to different uses, e.g., playing games of chess or different Atari video games, like *Breakout*.

---

[1] Nick Bostrom argues that our cosmic endowment (or synonymously, inheritance) may amount to "at least 10^{58} human lives [that] could be created in emulation," lives which are presumably worth living (e.g., full of happiness, joy, rich experiences, fulfilling relationships, etc.). See Bostrom (2014) for a rough calculation of our cosmic endowment.

[2] Bastani 2019, 11. According to Bastani, engaging in agriculture was the "First Disruption" and the Industrial Revolution and the first machine age was the "Second Disruption."

But of course, artificially intelligent machines do far more than play chess or Atari games. The future described above is only one among many possible futures. As Bastani notes, the consequences of this Third Disruption remain uncertain and "its possibilities are such that they call into question some of the basic assumptions of our social and economic system."[3] Assumptions, for example, about whose loan application ought to be accepted and whose application ought to be rejected, or whose appeal for a lighter criminal sentence ought to be granted and whose appeal ought to be dismissed. Given the fact that machines are increasingly making decisions, and actuating on the basis of those decisions (i.e., behaving in certain ways), which impact human well-being and flourishing, the second thesis of this dissertation is that we must now seriously consider how ethics can be implemented in machines. That is, we must now seriously consider how machines themselves, as moral agents that can impact human well-being and flourishing, might make ethically preferable decisions and take ethically preferable actions. In particular, I argue that bottom-up reinforcement machine learning techniques are the best suited for the project of raising ethical machines. Moreover, I maintain that the notion of *raising* machines is key for the project of implementing machine ethics. Much like children, and in the spirit of a suggestion Alan Turing made in the 1950's, I argue that we may be able to cultivate a virtuous character in machines if they are appropriately raised. To support this view I will be drawing heavily on research in the domain of game-playing, specifically research on the real-time strategy game *StarCraft II*, to motivate the conclusion that artificially intelligent machines are capable of engaging in ethical decision-making and behaviour.

Though impressive, artificially intelligent machines are still limited in certain significant ways. The third and final thesis then is that we must remain vigilant when using artificially intelligent machines and ensure that they are used as a supplement to, and not replacement for, considered human judgment. Artificially intelligent machines are still parasitic on humanity, and as such they lack what Brian Cantwell Smith calls "judgment," i.e., "a form of dispassionate deliberative thought, grounded in ethical commitment and responsible action, appropriate to the situation in which it is deployed."[4] Like Smith, I maintain that judgment (e.g., the considered human judgment mentioned above) "is a capacity we strive to instill in our children, and a principle to which we hold adults accountable."[5] Artificially intelligent machines, at this point in time, simply do not possess judgment. What machines do possess is a capacity for "reckoning," to use Smith's terminology. That is, artificially intelligent machines possess a

[3] Ibid.
[4] Smith 2019, xv.
[5] Ibid.

calculative prowess and "skills of extraordinary utility and importance," for example pattern recognition skills that allow machines to accurately predict the three-dimensional structure of a protein given only its primary amino acid sequence.[6] But importantly, these are skills "embodied in devices that lack the ethical commitment, deep contextual awareness, and ontological sensitivity of judgment" that humans must ultimately aspire toward.[7] And so, I conclude that we should care deeply about artificially intelligent machines, especially those used in times and places when good judgment is required, because we risk lapsing into a state of moral complacency otherwise.

Key Terms and Themes

For clarity's sake and ease of understanding, I briefly outline here some of the key terms I use (and some I refrain from using) throughout this dissertation since many of them appear in different disciplines and my usage of them does not necessarily map onto any one particular field. Such is the nature of interdisciplinary research. First and foremost, while it is relatively common to use the term 'agent' to broadly refer to any entity that has information states and can update those states, I will generally avoid using this term. This is primarily because I want to avoid the connotations associated with the thick concept of agency, especially given the many comparisons I will be making between humans and artificial agents. So I eschew the term 'agent' in favour of the more neutral term 'machine.' Additionally, I tend to avoid the term 'agent' for the sake of maintaining a kind of conceptual simplicity. For example, I distinguish between moral agents as a subset of agents in general (e.g., adult humans as a subset of higher mammals). Introducing artificial agents unnecessarily complicates things and, more importantly, the position I defend in this dissertation does not depend on taking either view, i.e., that machines are moral agents or agents in general. Rather, the position I defend in this dissertation relies on the thin concept of agenthood mentioned above, namely that machines possess information states and can update those states.

Second, by 'implementing machine ethics' I mean something like the following: intentionally designing and developing autonomous and intelligent machines that generate ethically sensitive judgements which are then, potentially, acted upon. This could be achieved in a number of ways, as I highlight throughout Chapters 3 and 4 primarily. Note, however, that my focus is not on responsibility, accountability, or authenticity (i.e., whether a machines' decisions are merely derivative of a human's or

[6] Ibid., xvii.
[7] Ibid.

multiple humans') nor, again, does the position I defend in this dissertation require that one take a stance on these issues. The importance and viability of implementing machine ethics is independent of considerations of accountability, for example, when things go wrong.

Third, and relatedly, my focus is on machines that are both autonomous and intelligent *and which also* have implications for human well-being, hence my usage of the terms 'machine ethics' and 'implementing ethics' to refer to a field of study and a project within that field respectively. While the use of traditional automobiles has implications for human well-being, such machines are neither autonomous nor intelligent and so are generally not objects of study in machine ethics. Similarly, while game-playing machines like AlphaZero are autonomous and intelligent, their use has minimal, and perhaps even negligible, implications for human well-being and so are generally excluded from the project of implementing ethics.

Fourth, the scope of values of interest in this dissertation is wide and deliberately so. When discussing how deliberate design choices are value-laden, or how machines are not created value free, I am referring both to values in the context of discovery and values in the context of evaluation.[8] Roughly speaking, values can influence what domain or subject a person studies, can influence hypothesis formation, and the choice of evidence to be gathered, all of which are generally part of the context of discovery. But values can also influence how a person interprets the evidence that has been gathered, the method of analysis employed, and the evidence's relation to the hypothesis and larger theoretical constructs, all of which are generally part of the context of evaluation. I do not distinguish between the context of discovery and the context of evaluation for two main reasons. The first is that values in either context can significantly affect the final machine produced and its performance. Second, it is not always easy to separate the two contexts when considering the design, development and deployment of autonomous and intelligent machines. Is the choice to use past data, e.g., of arrests or resumés of current employees, to predict future trends part of the context of discovery or context of evaluation? This issue is compounded by the high velocity of machine production and evolution wherein development and operations (hence the term of art, 'DevOps') bleed into one another. So although the context of discovery and context of evaluation can be conceptually separated, it is difficult to separate the two in practice.

Lastly, as with values, the scope of ethics of interest in this dissertation is wide. Note first that I use the terms 'ethics' and 'morality' interchangeably to refer generally to both living well as a human

[8] Hoyningen-Huene, 1987.

being (hereafter just well-being) and right/wrong (typically actions) as well as good/bad (typically states of affairs). Moreover, I am primarily concerned with ethics in the normative, not descriptive, sense. That is, my primary concern in this dissertation is with what *ought* to be the case and not merely with describing what already *is* the case. For example, while it is the case that certain machine learning techniques coupled with certain datasets are used to predict crime hotspots, my focus is more on whether we ought to be using those machine learning techniques coupled with those datasets or, indeed, whether we ought to be using such a machine to predict crime hotspots at all given the consequences such systems have on human well-being.

Chapter Overviews

In Chapter 1 I articulate a general theory of learning and show how it applies to machines, in addition to trying to convince a hypothetical skeptical interlocuter that machines can learn in much the same ways that humans and other non-human animals are capable of learning. I also summarize different contemporary machine learning techniques which are frequently divided into three dominant paradigms: supervised, unsupervised and reinforcement learning techniques. Issues connected to contemporary machine learning techniques like transparency and explainability are also discussed in order to set the stage for analysis in subsequent chapters, e.g., Chapters 5 and 7.

Chapter 2 is dedicated to exploring the ways in which science and technology, even artificially intelligent machines, are not value free or value neutral. Though not central to my arguments about implementing machine ethics, a pervasive idea is that machines might make more "objective" decisions and take more "objective" actions, where this objectivity is contrasted with subjectivity. There is, however, no guarantee that machines will be more "objective" than humans given the myriad ways in which human values can influence machines design and output.

The origins of the field of machine ethics and the important connections the field has with philosophical ethics are explored in Chapter 3. There are many relevant features that philosophers argue confer moral status to human beings. Features like autonomy, having the right sort of mind, and whether an entity has the right sort of agency or capacity to carry out intentional actions, can all influence whether a candidate entity should be thought of as a moral agent, moral patient, or both. I argue that artificially intelligent machines possess all of these features, albeit in a weaker sense. Nevertheless, the importance of machine ethics is independent of taking any particular position on these philosophical issues.

In Chapter 4 I defend the view that bottom-up reinforcement learning methods are the best suited for implementing machine ethics. First I examine top-down approaches to implementing machine ethics before assessing bottom-up approaches to implementing machine ethics and ultimately concluding that, even when compared to other bottom-up machine learning techniques, reinforcement learning techniques are the best suited for raising ethical machines.

There are, to be sure, challenges associated with raising ethical machines using reinforcement learning. It is to these challenges and how they might be addressed that I turn in Chapter 5. This is primarily accomplished through a case study of DeepMind's AlphaStar and its use of reinforcement learning to master the video game *StarCraft II*. I also discuss some of the risks and long-term benefits associated with implementing machine ethics.

Chapter 6 implicitly engages with the classic philosophical question of "know thyself" by highlighting the ways in which the nature of ethical decision-making and behaviour can be reflected in our attempts to build ethical machines. In short, I argue that there are benefits to using ethical machines, but more importantly, benefits associated with pursuing the project of implementing machine ethics, i.e., developing and building ethical machines as well as theorizing about machine ethics. Not only can we learn more about human moral psychology by implementing our theories in machines, but we can also unearth and empirically assess the assumptions underpinning different philosophical theories of ethics.

Lastly, in the Epilogue I argue that what is needed in response to artificially intelligent technologies is an Ethics 3.0 to help guide humanity toward a possible future in which all humans might flourish. While it is common to think in terms of a rights-based framework and how rights might or might not be extended to machines, I maintain that such an approach is impoverished in contrast to the virtue-based or character-based relational approach that I advance.

It is my hope that the possible future I described, or one even better than any I could imagine, materializes for an inquisitive, adventurous and benevolent humanity ready to depart for worlds unknown with intelligent, ethically sensitive machines at its side. But every journey begins with a first step, so I turn now to the first chapter and learning machines.

Chapter 1 – Machine Learning

1.0 Turing's Vindication

That certain machines are different in kind from all previous machines is a theme that will recur throughout this dissertation. One distinguishing feature of these new machines is their ability to engage in a primitive form of learning.[9] In an oft overlooked section of his famous 1950 paper "Computing Machinery and Intelligence" Alan Turing muses about how we might create what today we call artificial general intelligence (AGI). Turing speculates that instead of trying to simulate the adult human mind, the path towards AGI likely involves producing a machine that simulates a child's mind.[10] That is, a machine capable of thinking and intelligence must also be capable of learning when "subjected to an appropriate course of education" that involves the association of "punishments and rewards with the teaching process."[11] It is a testament to Turing's brilliance that this almost perfectly describes modern neural networks (the child mind) and machine learning techniques (the education).

Despite the fact that it is patently obvious to anyone aware of the state of the art in artificial intelligence research that machines can engage in some kind of learning, a primary aim in what follows in this chapter will be to persuade those who are unaware of the current state of research or skeptics of the claim that machines can learn that this is in fact the case. It is simply not the case that machines today (or certain machines at any rate, the ones with which I will be concerned throughout this dissertation) just do what humans tell them to do nor is it the case that machines are, in principle, incapable of dealing with complex, dynamic and multi-stakeholder contexts. The former is a gross oversimplification if not outright mischaracterization and the latter a premature conjecture about a largely empirical issue. Indeed another recurring theme will be my reference to empirical research in fields ranging from computer science to moral psychology in an effort to establish both the practical relevance of this dissertation and to limit any tendency to wildly speculate.[12] Preamble aside, I turn now to consider the concept of learning.

---

[9] Such capabilities are leading not only to new technological possibilities but may also lead to new classes of autonomous non-human agents capable of sustain social relations. See, Railton (2020), for a more detailed discussion of this latter possibility.

[10] Turing 1950, 456.

[11] Ibid., 456-457.

[12] This means that, unless otherwise specified, I am not referring to strong AI or AGI. Additionally, unlike Wallach and Vallor (2020) who seem more concerned with the possibility of creating machines that we might describe as full moral agents, i.e., contextually adaptive and simultaneously norm governed, my aims are more modest. I aim only to establish that bottom-up approaches to implementing machine ethics

1.1 What Sort of Thing Is Learning?

Human beings and other biological creatures are no longer the only entities that learn. We are joined by artificial entities, machines, that can also learn. The idea that artificial entities could learn is not a new one. As early as 1949, scientists such as Donald MacKay proposed combining digital and analog machines into what he called Analytical Engines.[13] These machines were imagined as autonomous learners in the sense that they would be able to adjust their behaviour based on new information in order to achieve some equilibrium. Such machines might be considered to possess the ability to learn in the sense that they were (imagined to be) able to attain a goal, maintaining some equilibrium, under the influence of external factors. Yet such an understanding of learning seems far too broad to be useful or helpful. Before examining machine learning in depth, I shall stipulate and defend a certain conception of learning. But before doing so, I should clarify what kind of conception I aim to develop. First, my aim is not to advance an exhaustive definition in terms of necessary and sufficient conditions. My aim is to build up a kind of cluster concept (or family resemblance) highlighting the relevant aspects of learning. Second, and relatedly, the sense of learning that *is* relevant, that we will focus on, is the kind of learning that is common to both humans/biological organisms *and* non-biological machines.

It is helpful to begin where most philosophy begins, at an examination of our intuitive or common-sense understanding of the concept of learning. It is obvious, bordering on trivial, to point out that human beings are capable of learning. Infants learn to crawl and walk. Children learn to form complex sentences in verbal and written language. Adolescents learn to socialize and navigate a cultural landscape. Adults learn how to budget and manage a career, and so on. Perhaps just as obvious is the fact that other non-human animals can learn. Dogs can learn to fetch sticks and catch frisbees. Crows can learn to harass people that disturb them and even cats can learn to anticipate when they will be fed. A first glance seems to suggest that learning involves some kind of interactivity between an entity and their environment. When learning to form complex sentences children often repeat phrases and words that are, sometimes, corrected by a parent or teacher. Nevertheless by speaking and writing, i.e., interacting with other speakers and readers, children receive feedback from their environment (even if it is just the feedback associated with speaking out loud or rereading what one has already written) and

are, at present, best suited for ensuring that, in a given context, machines make ethical decisions and take ethical actions.

[13] McCorduck 2004, 97.

learn to work new phrases and words into ever more complex and diverse contexts. Infants interact with their proprioceptive system and receive a variety of sensory feedback as they attempt to walk before successfully crossing a room without falling. Learning to budget similarly requires one to interact with an environment and receive a variety of feedback, e.g., a reward or punishment, the former if one adheres to the budget and the latter if one does not. Note that feedback from interaction with an environment, especially feedback in the form of a reward or punishment (i.e., a quantifiable extrinsic reward), is crucial for understanding machine learning techniques, which will be discussed later.

A lack of interactivity, on the other hand, may hinder learning. Your cat may have a rather difficult time learning when it will be fed if its meals are not repeatedly served at roughly the same times, i.e., if the feedback it receives from its environment is apparently random. Similarly the child who attempts to learn to ride a bicycle without interacting consistently with their bicycle, by attempting to ride it for example, may find it a difficult, perhaps bordering on impossible, task. Importantly, I am not claiming that interactivity is necessary for learning, just that it appears to aid learning.[14]

Learning also seems to involve knowledge (i.e., information), either the acquisition, via interaction with an environment, of knowledge that (i.e., knowledge by description) or the acquisition of knowledge how (i.e., skills or abilities, or knowledge by acquaintance) or some combination thereof.[15] Learning to ride a bicycle for the first time, as an example, involves both "knowledge that" and "knowledge how." One must know that the pedals need to be depressed in certain ways to generate forward motion and know how to balance and avoid falling over. A similar story can be told about how children learn to form complex sentences. Children need to know that there are different tenses (at least in English) for the future and past and they need to know how to incorporate tensed words into their sentences.[16]

The last relevant aspect of learning we might tease out of an intuitive examination of the concept of learning is that it is directed. That is, learning tends towards some goal or outcome. Importantly, this outcome need not be purposive as is the case when a person engages in learning for the purpose of, say, becoming a better chess player. In contrast, when considering living organisms, the outcome of surviving is not purposive, i.e., intentionally envisioned by some agent. So learning is directed, purposively or not, towards some outcome. This outcome, moreover, need not be specific or

[14] Even in cases of so called "one shot learning," though the goal is to get a machine to learn and generalize given a single training example, i.e., minimal interaction with an environment, there are often a few training examples with the goal being to provide as few examples as possible.
[15] Russell 1910.
[16] Roughly, vocabulary is knowledge that whereas grammar is the knowledge how.

explicitly articulable. In primary and secondary school students learn (or at the very least are expected to learn) about a wide range of subjects, many of which are taught not for the achieving of a specific outcome.[17] Of course there *is* a reason why students learn trigonometry or how to write a book report, however the reasons enlisted invariably revolve around general outcomes, like providing students with basic critical thinking skills or a well-rounded education.[18] Moreover, the learner need not be consciously aware of the outcome(s) of their learning. Infants learn to exercise motor control, including control over their speech, long before they are conscious that their learning is directed towards some outcome, e.g., their harmonious integration into society.

Given the above considerations, I propose the following loose definition of the concept of learning. Learning is the dynamic acquisition of knowledge (or in the case of machines, information) via interaction with an environment that allows an entity to better achieve an outcome. Accordingly, learning is fundamentally relational in multiple different respects. Learning involves a relation between an entity and knowledge, an entity and its environment, and an entity and its future (or past) self, e.g., an entity in its learned state in relation to itself in an unlearned state. Learning is also fundamentally diachronic, i.e., learning occurs in time.[19] Lastly, learning is directional in nature. In contrast to mere diachronic change, learning is directed, purposively or not, towards some outcome.

1.1.1 Relationality

If it was bordering on trivial to point out that human beings and certain non-human animals can learn, it is still, in some circles, a matter of debate as to whether machines can learn in the same way living organisms can learn. That is not a debate I will be engaging with. Instead, in this section I aim to demonstrate that learning machines already exist by highlighting how current machines, particularly those that utilize machine learning techniques, satisfy the definition of learning given above. Examples of learning machines will therefore be invaluable, and I will primarily draw on two different domains in which machine learning has had considerable success: game playing and automobile control. Before turning to examples however, let us first examine the account of learning outlined in the context of the relevant literature.

[17] Apologies to my high school math teachers but I have not had the occasion to exercise my trigonometry skills in a long time.
[18] These are two examples among many reasons that justify why students in primary and secondary school are expected to learn about a wide range of subjects.
[19] Even in cases of one shot learning.

Articulations of the concept of learning invoke relationality in some form or another. Contemporary accounts of learning emphasize that the concept involves "permanent capacity change" thereby emphasizing the fundamentally relational character of learning.[20] Change here is understood as the change an agent experiences with respect to itself. Perhaps more concretely, animal learning has been cashed out in terms of change "caused by a specific experience at a certain time, t_1, and that is detectable later, t_2."[21] As mentioned above, either implicitly or explicitly, learning is defined as an essentially relational phenomenon. Even traditional conceptions of learning, such as the idea that learning consists of the acquisition of knowledge and skills, imply that learning is relational in nature. Importantly, one of the relations that we are interested in is the relation between an entity and another temporally connected version of itself. This is to be distinguished from any old temporal relation that an entity might share with another version of itself. That is, the temporal relation of interest is one such that it can be said of an entity that progress has been made towards an outcome as a result of interaction with an environment. For example, if I compare the temporally present version of myself capable of juggling to a temporally past version of myself incapable of juggling, it can be said that I have learned how to juggle. Not every temporal relation however is important in the sense relevant to learning. If I had just woken from a medically induced coma (my temporally present self), and despite the fact that I share a temporal relation with my pre-coma self (a temporally past version of myself), I have made no progress towards an outcome as a result of interaction with my environment.

Although the mere passage of time is not a sufficient condition to claim that an entity has learned something, learning does require the passage of time. Learning is essentially diachronic in nature, and this too is borne out either explicitly or implicitly in definitions of the concept. Explicitly, if learning involves detecting some change in an entity (by the entity itself, an observer or both) that was caused by some prior environmental interaction, then learning necessarily involves temporal passage.[22] Implicitly, the diachronic nature of learning can be drawn out of a common theme in definitions of learning, namely that learning results from some prior experience(s), i.e., interaction(s) with an environment.[23] Alternatively, learning is framed as a process or even as a "very extensive and complicated set of processes," hence the necessarily diachronic nature of learning.[24] While it may be possible to develop a coherent synchronic account of learning, this seems like a poor way to approach

[20] Illeris 2009, 7.
[21] Heyes 1994, 209.
[22] Ibid.
[23] Mowrer and Klein 2001, 2.
[24] Illeris 2009, 7.

the dynamic phenomenon learning appears to be. For example, it would seem odd to maintain that learning synchronically (or atemporally) supervenes on some set of behavioural, neurophysiological or functional properties of an entity. This is because we say that an entity learns as a result of some change or process, i.e., a temporally mediated change or process.

Lastly, learning is directional. Learning can be purposefully directed towards some outcome, goal or to serve some function, but it need not be purposive. Thus far I have avoided discussing what exactly can be learned, and that is in part because the outcomes of learning are as diverse as they are numerous. Robert Gagné has proposed five major domains of learning each of which has distinct outcomes. The domains are "(1) motor skills, (2) verbal information, (3) intellectual skills, (4) cognitive strategies, and (5) attitudes."[25] In the domain of motor skills, the outcomes, although not specific, are also relatively obvious: an entity learns to control their body, i.e., motor skills, in order to navigate their environment. Specific and purposive outcomes include "tying shoelaces, printing letters, pronouncing letter sounds, using tools and instruments," and so on.[26] The general outcomes to which learning in the domain of verbal information are directed is the acquisition of information such as facts, principles and generalizations. Closely related to verbal information is the domain of intellectual skills. Learning in this domain is, in part, directed towards the proper manipulation of verbal information and the products of learning in other domains. Consider that "being able to recall and reinstate a definition verbally is quite different from showing that one can use that definition. The latter is what is meant by an intellectual skill," whereas the former is something learned in the domain of verbal information.[27] Similarly, being able to swing a hammer accurately is different from demonstrating that one knows that nails but not screws are objects to be hammered.[28]

In the fourth domain of cognitive strategies learning is directed towards an agent's self-management of learning and thinking, i.e., they "are internally organized skills that govern the individual's behaviour in learning, remembering and thinking."[29] In contrast to intellectual skills which have an external orientation toward the learner's environment, cognitive strategies have an inward orientation toward thinking strategies. Finally, the fifth domain of attitudes is directed toward nebulous outcomes. A primary difference between the domain of attitudes and the other domains is that they are

[25] Gagné 1972, 3.
[26] Ibid.
[27] Ibid.
[28] Technically one could hammer a screw, but this is only an illustrative example. For the uncharitable critic I provide this extreme example: being able to swing a hammer accurately is different from demonstrating that one knows that nails and not elephants are objects to be hammered.
[29] Gagné 1972, 3.

not learned by practice and are in some sense immutable, i.e., they are not "dependably affected by a meaningful verbal context."[30] Although the outcomes of learning in the domain of attitudes are vague, such learning appears directed toward instinctual or unreflective value judgments that may be a remnant of our evolutionary history. It would have been advantageous for an individual to learn to form the attitude of disgust toward certain kinds of foodstuff for example and thereby prevent them from consuming a potentially fatal meal. Though not explicitly mentioned by Gagné, I maintain that moral values and character, i.e., virtuous or vicious character traits, can also be included in this domain. Indeed there is an important connection between learning and the process of raising a human or machine via the cultivation of character traits. This roughly virtue ethics approach to raising humans and machines will be explored in later chapters. But for now, I have listed all of these features because there currently exist machines whose learning results in one or more of the outcomes just discussed.

1.2 Learning Machines

So are there machines that exist which possess these three features, i.e., interactivity, diachronicity and directivity, of learning? The answer is, unequivocally, yes. Therefore learning machines exist. Importantly, these are machines that learn in a similar way that biological organisms learn. Two examples should suffice to motivate acceptance of the fact that we can understand machines as learning entities. Moreover, beyond these examples, it is the explicitly stated intention of researchers in the field of artificial intelligence that learning machines be developed, machines which fall under the definition of learning given above.

1.2.1 Learning to Play Atari Games

The first example of a learning machine is a machine developed by DeepMind Technologies which learned how to play a suite of video games. As mentioned above, learning is directed towards a large and diverse number of outcomes, and this is partly why games (e.g., card games, board games, video games, single player games, multiplayer games, etc.) represent one significant domain of interest for testing the ability of machines to learn. Most games have a clearly defined outcome or win condition that the player is supposed to work towards. Moreover, there are often states of affairs within a given

[30] Ibid., 4.

game that indicate the progress a player is making towards achieving their outcome, or more generally the progress the player is making within the game. Games are therefore ideal environments in which one can determine the extent to which an entity, artificial or biological, is able to engage in learning.

DeepMind's machine learned to play multiple games, including *Breakout* and *Seaquest*, originally released for the Atari 2600 video game console. Briefly, for the uninitiated, *Breakout* is a game where the player must move a paddle left and right across the bottom of the screen to deflect a ball into rows of destructible bricks at the top of the screen. The goal is to destroy all of the bricks without missing the ball's rebound more than three times. The relevant learning domains of interest are therefore intellectual skills and cognitive strategies as the machine is required to both demonstrate its ability to use the game controls and employ some rudimentary cognitive faculties (e.g., planning and memory). This machine can indeed be said to have learned to play these games. This is because, first, the machine experiences a permanent capacity change with regard to its ability to play *Breakout*, for example. In short, there is a detectable change in the machine (at t_2) such that it can be said that the machine is better able to achieve the goal of the game compared to a previous version of itself (at t_1). Second, the machine interacts with the games over a period of time, hence the diachronic nature of learning. When learning to play the game *Breakout* for example, the machine interacted with the game environment over a period of fifty hours.[31] Third, the machine was directed towards achieving some outcome. In particular, the machine was directed towards achieving the highest score possible within the respective games it played.[32] Despite the fact that the machine was evaluated on the "real and unmodified games," a slight change was made to the way in which the machine was rewarded or punished for the decisions it made.[33] Given the fact that game score scales vary from game to game, "all positive rewards [were fixed] to be 1 and all negative rewards to be -1."[34]

1.2.2 Learning to Operate a Vehicle

In addition to engaging in learning in the domains of intellectual skills and cognitive strategies via game playing, machines can also engage in learning in the domain of motor skills. Autonomous vehicles are a class of machines that can engage in learning. Like the machines developed by DeepMind

[31] Minh et al. 2013, 7.
[32] Experiments were performed on seven popular Atari games: *Beam Rider*, *Breakout*, *Enduro*, *Pong*, *Q*bert*, *Seaquest* and *Space Invaders*. See Minh et al. (2013) for more details.
[33] Minh et al. 2013, 6.
[34] Ibid.

to play Atari games, autonomous vehicles experience a permanent capacity change with regard to their ability to manipulate a vehicle. There is a detectable change in the autonomous vehicle (at t_2) brought about as a result of it interacting with its environment such that it can be said that the autonomous vehicle is better able to achieve its goal(s) (e.g., driving within the lane, obeying traffic signals, etc.) compared to a previous version of itself (at t_1). Further, the autonomous vehicle's training takes place over a period of time. When learning to steer a vehicle only, Bojarski et al. discovered that less than one hundred hours of data collected from human drivers was needed for their machine to successfully steer in both simulated and real environments.[35] Hence both relationality (the autonomous vehicle experiences a permanent capacity change with respect to a past version of itself) and the diachronic features of learning are present in this example. Finally, and as alluded to, the autonomous vehicle developed by Bojarski et al. was directed towards achieving a specific outcome, namely appropriately steering a vehicle. The viability of autonomous vehicles hinges on their ability to safely manipulate a vehicle under a variety of different environmental contexts. This is the grand outcome to which learning is directed when considering autonomous vehicles, i.e., appropriate control of a vehicle in any number of contexts, and autonomous vehicles are rapidly approaching the control exhibited by human drivers. Given only a small amount of training data, machines are able to "operate [e.g., steer a vehicle] in diverse conditions, on highways, local and residential roads in sunny, cloudy and rainy conditions."[36]

1.2.3 Engineers Build Learning Machines

A final consideration that ought to motivate acceptance of the conclusion that machines exist which can genuinely learn in the same way humans and other non-human biological organisms can learn, is the fact that researchers in the field of artificial intelligence explicitly express that they are creating learning agents. Research on machines that can manipulate vehicles is expressed explicitly in terms of learning.

> We have empirically demonstrated that CNNs [convolutional neural networks] are able to learn the entire task of lane and road following without manual decomposition into road or lane marking detection, semantic abstraction, path planning, and control...The CNN is able to learn meaningful road features from a very sparse training signal (steering alone). The systems learns for example to detect the outline of a road without the need of explicit labels during training.[37]

[35] Bojarski et al. 2016, 9.
[36] Ibid., 9.
[37] Ibid.

This locution is neither unique nor restricted to work on machines that can manipulate vehicles. Research on artificial agents that can play games also explicitly frames how a machine develops in terms of learning.

> Our goal is to create a single neural network agent that is able to successfully learn to play as many of the [Atari] games as possible. The network was not provided with any game-specific information or hand-designed visual features, and was not privy to the internal state of the emulator; it learned from nothing but the video input, the reward and terminal signals, and the set of possible actions—just as a human player would.[38]

While one might be tempted to accuse researchers who use the term 'learn' (and its variations, e.g., 'learned,' 'learning,' 'learns,' etc.) of being cavalier or egregiously mistaken when describing various machines, the opposite is in fact the case. Technology has advanced to a point where it is necessary to acknowledge, as researchers in the field of artificial intelligence, and related fields, already do, that there exist machines that can genuinely learn. We might even trace this semantic drift all the way back to Alan Turing who wrote that by the year 2000, "the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."[39] A similar redescription of ethical and moral terms may also take place, i.e., one may be able to speak of ethical machines without expecting to be contradicted, but that is a topic that will be taken up in Chapter 6.

Resistance to the idea of learning machines is nothing new. In the 1950's Turing identified what he called "Arguments from Various Disabilities" which take the following form: "I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X."[40] Among those things that we might substitute for X is "learn from experience."[41] More than thirty years after Turing, John Haugeland notes that skeptics in general have continued to rely on the same basic argument Turing anticipated.

> Basically, their thesis is: no matter how good AI systems get at imitating honest-to-goodness (human) intelligence, they'll still be mere imitations — counterfeits, fakes, not the genuine article. Typically, the argument goes like this:
>
> 1. Nothing could be intelligent without X [for some X]; but
>
> 2. no GOFAI [good old fashioned artificial intelligence] system could ever have X; therefore

[38] Minh et al. 2013, 2.
[39] Turing 1950, 442.
[40] Ibid., 447.
[41] Ibid.

3. no GOFAI system could ever be intelligent.[42]

Like Turing, Haugeland suggests that among those things we might substitute for X, or that the skeptic about machine learning is inclined to substitute, is learning.[43] Unlike Turing however, who lamented that he has "no very convincing arguments of a positive nature to support [his] views" that learning machines are possible to create, we are very much in a position, nearly a quarter of the way into the 21st century, to offer arguments of a positive nature on Turing's behalf.

Perhaps the most compelling evidence to support the idea that learning machines can be built was the defeat of Korean Go player Lee Sedol by AlphaGo, an artificial agent created by DeepMind, 4:1 in a five game match of Go. Why is this achievement so compelling? Creating a machine to play the game of Go at the highest levels of human play has been a long-standing benchmark in artificial intelligence research. This is because despite its deceptive simplicity, the complexity of Go is immense. The goal of the game is to encircle more total area on the 19x19 board than the opponent, with each player taking turns placing their pieces (black or white stones) on any unoccupied space. The need to understand both the strength of local positions on the board and their relation to global patterns developing throughout the game, coupled with a number of theoretically possible games in the order of 10^{700} means that even sophisticated brute-force approaches to playing Go at the highest levels are simply out of the question. Indeed, Go is often described as being difficult to master because it requires a balance of creativity, strategy and intuition. AlphaGo's victory was not the result of some paradigm shifting technological or theoretical innovations in the field of artificial intelligence or philosophy of mind, rather it was the result of ensuring that a machine could genuinely learn in an analogous way that humans learn. There is therefore a very real sense in which we might say AlphaGo is creative, strategic and perhaps even intuitive insofar as its ability to play Go is concerned.

> Its [AlphaGo's] greatest contribution lies in the massive integration and implementation of recent data-driven AI approaches, especially deep learning (DL) and advanced search techniques, and demonstrates to the world the power of these technological advancements. For this reason, the AlphaGo's victory and achievement is beyond the technology and more of psychological nature.[44]

Despite this evidence, and even more recent evidence from one of DeepMind's newer agents, AlphaZero, I anticipate that the ardent skeptic of learning machines may retreat to some other X-factor that precludes machines from being considered genuine learners. Nevertheless, the above

[42] Haugeland 1985, 247.
[43] Ibid.
[44] Wang et al. 2016, 114.

considerations and examples should suffice to motivate abandoning "learning" as that X-factor that machines cannot/will not possess. I therefore now turn to an examination of the three major types of machine learning.

1.3 Machine Learning Preliminaries

Before continuing, some terminological clarification is needed. In the preceding sections, and throughout the dissertation, I will be using the term 'machine' to refer to specific machines, namely learning machines.[45] Additionally, there are two components to these machines that are important to understand before proceeding to a discussion of different machine learning techniques. First, learning machines (at least the ones I am interested in discussing throughout this dissertation) have a particular kind of structure, specifically the structure of a neural network. The simplest neural networks, or artificial neural networks (ANNs), consist of a single input layer directly connected to a single output layer.[46] In more sophisticated kinds of machines however, especially those that employ deep neural networks or deep learning methods, there can be numerous hidden layers sandwiched between the input and output layers. The structure of these deep neural networks confers significant advantages to machines that utilize them. In short, the hidden layers in a deep neural network "transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level."[47] Each additional hidden layer "means a new way to combine the insights from the previous layer"[48] such that, for example, given raw pixel data for an image (the input layer), the first hidden layer may "represent the presence or absence of edges at particular orientations and locations in the image," the second hidden layer may represent "particular arrangements of edges, regardless of small variations in the edge positions," and so on.[49] This obviates the need for careful engineering and considerable domain expertise if one were interested in creating a machine that could detect patterns in a given input, e.g., detecting whether the player is in an advantageous position in a game of chess given the board state as an input.

In addition to their neural network structure, the second, and perhaps more important component of learning machines, is the learning algorithm. These algorithms are utilized to facilitate the

[45] See also Chapter 3 for more clarification on my use of the term 'machine.'
[46] Shane 2019, 68.
[47] LeCun, Bengio and Hinton 2015, 436.
[48] Shane 2019, 69.
[49] LeCun, Bengio and Hinton 2015, 436.

machine's progress towards a given outcome.[50] The basic idea is that the machine's outputs are directed towards some outcome, but how well or poorly the machine achieves the outcome is determined by the connections between the neurons in the neural network, i.e., the weights between neurons. The learning algorithm adjusts the weights between neurons such that, over time, the machine (hopefully) better achieves the outcome. Whether the outcome is known in advance or not, the type of environmental interaction the machine will experience (e.g., labeled training data, a simulated world, real-world data, etc.) and the complexity of the task at hand all factor into how the weights in a neural network ought to be adjusted and therefore what learning algorithm ought to be used.

1.4 Major Machine Learning Techniques

There are three dominant machine learning paradigms, i.e., supervised, unsupervised and reinforcement learning techniques, and various versions of each, hence a variety of ways in which machines can learn to achieve a particular outcome. To help understand some of the machine learning techniques, let us take a common example. Suppose that I would like a machine to be able to learn to distinguish between photos of cats and photos of dogs. There are numerous machine learning techniques I can utilize, each with certain advantages and disadvantages, to teach it to do so.

One obvious way I could train a machine to learn to discriminate between photos of cats and photos of dogs is to show it many different pictures of cats and dogs and tell it whether the photo it is currently examining is either a cat or a dog. The hope is that the machine will discover some statistical similarity in the set of cats, and the set of dogs, that allows it to correctly discriminate between novel photos of cats and dogs that were not in the original training set. Training a machine to classify photos in this manner, using labeled data, is called "supervised learning" because the machine is explicitly told what it is looking at.[51] Before training, the machine will produce random outputs corresponding to the initially random weights connecting the neurons in its neural network. That is, the machine will incorrectly label cats as dogs and vice versa. But throughout training, as the weights between neurons are adjusted by the learning algorithm, the machine learns to accurately differentiate between cats and dogs. There are however limits to this kind of learning. For one, supervised learning requires the existence of sufficient amounts of labeled visual data, in the case of images, of the objects to be

[50] There are also different types of learning algorithms that do not utilize the structure of neural networks. Random forest algorithms, for example, utilize decision trees. See Shane (2019) for more information.
[51] Raina et al. 2007, 2.

classified. This may not pose a problem for distinguishing between cats and dogs given the human obsession for uploading enormous quantities of labeled images of both.[52] However teaching a machine to distinguish two (or more) objects becomes increasingly difficult if there is a lack of appropriately labeled data.

So not all classification tasks, for example, can be taught to a machine using supervised learning methods. Nevertheless, one might still be able to train a machine using what is known as "semi-supervised learning."[53] In contrast to supervised learning, semi-supervised learning utilizes labeled and unlabeled data to train a machine in some visual classification task, for example. One notable caveat however is that in semi-supervised learning it is assumed that the unlabeled data can be classified according to the same labels; the assumption is "that these labels are merely unobserved."[54] So for example, if I am training a machine to distinguish between images of cats and dogs I would train it on images of cats and dogs labeled as such, but also on images of cats and dogs that are not labeled. Nevertheless, the machine will assume, under a semi-supervised training regime, that the unlabeled images it looks at will be classified as either a cat or dog. Despite this caveat, semi-supervised learning has a significant advantage over supervised learning: the largest quantity of data, unlabeled data, can be leveraged, together with a sample of labeled data, to significantly increase classification accuracy in certain settings.[55]

Despite their successes, supervised and semi-supervised learning methods are quite restrictive in the sense that once a machine has been trained to perform a certain task, it generalizes poorly to other relatively similar tasks. A machine that has learned to generate names of metal bands for example, will be unable to generate names of ice cream flavours.[56] This is the advantage that "multitask learning," sometimes also called "transfer learning," methods possess. In short, multitask learning methods improve the generalization of a machine "by leveraging the domain-specific information contained in the training signals of *related* tasks."[57] In contrast to both supervised and semi-supervised learning, multitask learning utilizes a labeled data set in addition to another labeled data set of related

[52] Visually distinguishing between cats and dogs is by no means an easy task for an artificial agent. Agents trained using the Oxford-IIIT Pet dataset, a collection of 7549 images of cats and dogs of 37 different breeds (only 50 images of each breed are used for training, the other images are used for validation and testing of the agents), can achieve an average discrimination accuracy of 59%. See Parki, Vedaldi and Jawahar (2012) for more details.
[53] Raina et al. 2007, 1-2.
[54] Ibid., 2.
[55] Nigam et al. 2000, 105-106.
[56] Shane 2019. 45-47.
[57] Caruana 1997, 41.

objects. For example, if I have a machine that is able to distinguish between images of cats and dogs and I would like it to distinguish between, say, tigers and wolves, I could train it once more[58] using labeled images of tigers and wolves. This works because there are many features that tigers share with cats (e.g., whiskers, coat patterns, etc.) and likewise many features that dogs share with wolves. In the case of the metal band name generating machine turned ice cream flavour naming machine, there are again domain-specific similarities that makes learning easier for the machine. As Shane points out, when learning to generate names of ice cream flavours, the machine already knows "approximately how long each name should be," it knows "that it should capitalize the first letter of each line" and some common letter combinations like "*ch* and *va* and *str*," all of which are the beginnings of different flavours of ice cream (an exercise I leave to the reader).[59]

Research on multitask learning has demonstrated the superiority of this learning method over supervised learning. When trained to locate doorknobs and to recognize door types (either single or double) in given images, a machine trained using multitask learning generalizes 20-30% better than a machine trained using data from only a single data set.[60] Although trained using identical data sets, the key difference between multitask learning and so called "single task learning" methods, such as supervised or semi-supervised learning, lies in the exposure to additional training signals, i.e., additional feedback from relevantly similar environments. In short, "it is the information contained in these extra training signals" used in multitask learning that helps the machine "learn a better internal representation for the door recognition domain," which in turn helps the machine better learn to "recognize door types and the location of the doorknobs."[61] Yet despite the improved generalization achieved using multitask learning, it nevertheless relies on the existence of labeled training data which may simply not exist or is painfully time consuming to produce. Moreover, the labeled data sets cannot be completely unconnected to each other. No advantage would be gained by training a machine using multitask learning to distinguish between cats and dogs by using one data set of labeled images of cats and dogs, and an additional data set of labeled audio recordings of ostrich and emu calls. There is simply no domain-specific information for a machine to leverage given these two data sets.

In contrast to multitask learning, "self-taught learning" methods require neither an additional labeled data set nor a relatively similar additional data set to improve performance on a given

[58] Recall, the machine has already been trained once to distinguish between cats and dogs using a labeled data set of each.
[59] Shane 2019, 45.
[60] Ibid., 46.
[61] Ibid., 47.

classification task. Like semi-supervised learning, self-taught learning utilizes labeled and unlabeled data to train a machine in a given classification task. Unlike semi-supervised learning, it is not assumed that the unlabeled data can be classified according to the same labels used in the labeled data set. Using our toy example, if I am training a machine to distinguish between images of cats and dogs I would train it on labeled images of cats and dogs in addition to unlabeled images of anything else I can access. Self-taught learning therefore possesses a significant advantage over all of the other learning methods examined thus far given that it places significantly fewer restrictions on the type of unlabeled data that can be used.[62] Indeed, self-taught learning "represents the natural extrapolation of a sequence of machine learning problem formalisms" which began with purely supervised learning.[63] Despite this, self-taught learning methods are still reliant on the existence of some labeled data.

Machines trained using only unlabeled data utilize so called "unsupervised learning." The basic idea is that, without providing the machine with the correct output for a given input, a machine is nevertheless able to discover patterns or structure in a dataset.[64] In the case of our toy example, a machine utilizing unsupervised learning given unlabeled images of cats and dogs could separate the images into the two categories of cat or dog based on some underlying pattern e.g., dogs have their tongues out and cats do not (although this is an erroneous example!). Importantly, what we want is a machine to be able to classify images, for example, based on some *genuine* underlying pattern. This process of taking unlabeled data and using unsupervised learning to group similar data points together is appropriately named "clustering," and there are a variety of cluster algorithms in addition, of course, to other popular uses for unsupervised learning (e.g., principal component analysis). That is, there are a variety of unsupervised learning algorithms that parse unlabeled data in different ways and hence have different uses. Unsupervised learning obviates the need for labeled data which is a significant advantage that cannot be overstated. This is because generating labelled data is both costly and time consuming, especially in comparison to relatively cheap and widely available unlabeled datasets. At the same time however, the quality of the data can severely impact the quality of a machine's output when using unsupervised learning.[65] In short, despite discovering some underlying pattern or structure in a dataset, that pattern or structure may not be a relevant one given the context of the machine's intended purpose (e.g., while many dogs do have their tongues out in images and many cats do not, this is not a

[62] Raina et al. 2007, 2.
[63] Ibid., 8.
[64] Sutton and Barto 2018, 2.
[65] Data quality and its impact on the outputs of machines is discussed extensively in Chapter 5.

particularly salient pattern insofar as we are interested in accurately categorizing images as containing either a cat or dog).

1.5 The Rise of Reinforcement Learning

As the focus throughout this dissertation will primarily be on reinforcement learning, it is worth discussing some of the history of this particular machine learning technique, especially in the context of game-playing.[66] At least since 1956 when the Dartmouth Summer Research Project on Artificial Intelligence was held, it has been a goal within the field of artificial intelligence to create a machine that learns from first principles, as it were, to perform a given task or achieve a given outcome. The defining features of reinforcement learning techniques, which represent a step towards this goal of imbuing machines with general-purpose learning abilities, are the emphasis on reward maximization (a certain kind of environmental feedback) and a consideration of "the *whole* problem of a goal-directed [machine] interacting with an uncertain environment."[67]

Progress on reinforcement learning methods was excruciatingly slow. In the 1950's, prior to the use of artificial neural networks and reinforcement learning techniques, when Arthur Samuel was investigating how a machine might learn to play checkers, two machine learning methods dominated: (1) a rote learning procedure that simply stored information about an actually encountered game state alongside an analysis of the game state and (2) a generalization learning procedure that utilized an evaluation function to continuously judge the terminating game state via a look-ahead tree search.[68] Samuel describes the differences between the two methods in the following way.

> The rote learning procedure was characterized by a very slow but continuous learning rate. It was most effective in the opening and end-game phases of the play. The generalization learning procedure, by way of contrast, learned at a more rapid rate but soon approached a plateau set by limitations as to the adequacy of the man-generated list of parameters used in the evaluation polynomial. It was surprisingly good at mid-game play but fared badly in the opening and end-game phases.[69]

In spite of the usefulness of these learning methods and the relative success of Samuel's checkers playing machine, it was still limited by hand-crafted features, i.e., human generated parameters used in

[66] See Chapters 4 and 5 for more detailed discussions of reinforcement learning and its application in enabling machines to learn to play the real-time strategy video game *StarCraft II*.
[67] Sutton and Barto 2018, 2-3.
[68] Samuel 1967, 601.
[69] Ibid.

evaluations of the game state. In short, one significant disadvantage of these learning methods was "the absence of an effective machine procedure for generating new parameters [i.e., new aspects of the game to track progress towards the goal] for the evaluation procedure."[70] After almost a decade of work on his checkers playing machine, Samuel lamented in 1967 that the goal "of getting the program to generate its own parameters, remains as far in the future as it seemed to be in 1959," this even in spite of new machine learning techniques that could better handle tree pruning and parameter interaction problems.[71] The crux of the issue is that machines like Samuel's and its descendants, e.g., Deep Blue (the chess playing machine developed by IBM) are limited by humans. In Samuel's case, his checkers playing machine is not able to learn for itself what features, i.e., parameters, of the board state are significant in the context of a winning outcome. As Samuel laments, the preassigned list of board parameters that guides his machine's learning "remains a man-generated list and it is subject to all the human failings" of the programmer, "who is not a very good checker player," and the expert checker players consulted who are "unable to express their immense knowledge of the game in words" that a programmer would find useful.[72] Reinforcement learning methods shed both the need to consult domain specific experts and to handcraft effective feature extractors (i.e., parameters).

The first notable use of reinforcement learning was in a machine that learned to play the game of backgammon by playing against itself and learning from the results of the games it played.[73] Recall that reinforcement learning methods involve rewarding (and conversely punishing) a machine as it interacts with its environment and thereby progresses closer to (or further from) the desired outcome. Recall from section 1.1, this is the extrinsic mathematical reward signal that the machine attempts to maximize via interaction with its environment. Note that in contrast to unsupervised learning, with reinforcement learning the desired outcome is known in advance and learning is directed towards this outcome. Additionally, in contrast to supervised learning, reinforcement learning does not require the existence of an appropriately labeled and representative dataset. Beginning from naïve self-play, a machine using reinforcement learning can, via interaction with its environment, e.g., interacting with other players in a game of backgammon (even if those other players are the same machine!), learn which game states maximize its chances of achieving a winning outcome. This was precisely how the

[70] Ibid.
[71] Ibid., 617.
[72] Ibid., 602.
[73] Tesauro 1993, 19.

agent called TD-Gammon achieved a strong level of intermediate play in backgammon.[74] Given no prior knowledge about the game beyond the rules and a reward if it won the game or punishment if it lost the game, TD-Gammon learned to play the game given exposure only to the raw description of the board state at each turn.[75] At this point in the history of machine learning techniques it was already known that machines could be trained using supervised learning methods, using data from human expert games in backgammon for example. These machines were tremendously successful, so much so that a machine called Neurogammon convincingly won the backgammon championship at the 1989 International Computer Olympiad.[76] So although TD-Gammon achieved only strong intermediate play using pure reinforcement learning beginning from initially naïve random play, when supplemented with hand-crafted features, it was estimated to play at a strong master level close to the world's best human players.[77] Importantly however, using only reinforcement learning TD-Gammon was able to reach the level of play of machines that used supervised learning such as Neurogammon.[78]

TD-Gammon's legacy materialized when, over a quarter century after it demonstrated the viability of and tantalized humanity with the power of reinforcement learning, the machine AlphaGo Zero was developed by DeepMind. Like TD-Gammon, AlphaGo Zero is a machine that learns how to play the game of Go through self-play. Using reinforcement learning and given only knowledge of the rules of Go, AlphaGo Zero proceeds to learn from initially random games of self-play using only the raw board state and its history as inputs.[79] Unlike TD-Gammon, AlphaGo Zero can reach superhuman levels of Go play using reinforcement learning alone.[80] AlphaGo Zero's ability to master the game of Go is nothing short of a historic milestone in the field of artificial intelligence. Widely considered as the "most demanding, grand AI/CI [Computational Intelligence] challenge in the mind games domain,"[81] AlphaGo Zero requires only a few days to rediscover what it has taken humankind thousands of years to learn

[74] On the more technical side, TD-Gammon was a machine that utilized a multilayer neural network and a delayed reinforcement learning algorithm to update the weighted connections between neurons in the different layers. In particular, TD-Gammon 2.1 had one hidden layer with 80 neurons. After training through 1.5 million games of self-play, and when supplemented with hand-crafted features, TD-Gammon 2.1 achieved near parity to Bill Robertie, a human grandmaster backgammon player (Tesauro 1993). See, Tesauro (1992), for more of the technical details concerning TD-Gammon's architecture.
[75] Tesauro 1993, 19.
[76] Ibid.
[77] Ibid., 19-21.
[78] Ibid., 20.
[79] Silver et al. 2017, 354.
[80] Ibid., 356.
[81] Mandziuk 2007, 7.

about the game of Go.[82] Until recently even the most advanced artificial Go-playing machines could be beaten by intermediate level human players, and this is because of the following reasons.

> First of all, Go has a very high branching factor, which effectively eliminates bruteforce-type exhaustive search methods…Additionally, proper positional board judgment requires performing several auxiliary tactical searches oriented on particular tactical issues…Another difficult problem for machine play is the "pattern nature" of Go. On the contrary to humans, who possess a strong pattern analysis abilities, machine players are very inefficient in this task, mainly due to the lack of mechanisms (either predefined or autonomously developed) allowing flexible subtask separation. The solutions for these subtasks need then to be aggregated – considering complex mutual relations – at a higher level and provide the ultimate estimation of the board position.[83]

Machines have progressed to the point where they can now learn in an analogous way that human beings and other mammals, it is strongly suspected, learn.[84] Namely, via outcome directed interaction with an environment that produces rewards and punishments as I outlined earlier in section 1.1.

1.6 The Whos, Whats and the Hows

Learning machines exist and they learn in the same ways that human beings learn. But what will they be learning? Whose goals will they pursue? Moreover, how will machines act in pursuit of those goals following their learning? While machines like AlphaGo Zero ought to be heralded as monumental successes, they also require careful scrutiny. In learning how to play the game of Go AlphaGo Zero not only surpassed all human players, it also discovered "non-standard strategies beyond the scope of traditional Go knowledge" as well as "novel strategies that provide insights into the oldest of games."[85] In short, AlphaGo Zero acted in a way that was not completely expected as a result of its learning.

In the domain of game playing, unexpected or superhuman abilities exhibited by a machine might not raise too much alarm. After all, outcomes in the domain of game playing are almost always clearly defined and behaviour, even novel behaviour, is, at least assumed to be, in pursuit of the relevant outcome, i.e., winning, given a particular game. But machine learning has been and continues to be applied in a wide range of domains including classification tasks, modeling tasks (e.g., textures and motion), object segmentation, information retrieval, robotics, natural language processing and

[82] Silver et al. 2017, 358.
[83] Mandziuk 2007, 7.
[84] Sutton and Barto 2018, 381-383.
[85] Silver et al. 2017, 357 & 358.

collaborative filtering.[86] Machine learning has enabled the creation of hearing aids that filter out ambient noise, medical decision systems that can read CT scans and diagnose disease, automated facial recognition systems and intelligent scheduling systems responsible for logistics planning.[87] The successes of machine learning cannot be denied, yet the question remains: what, exactly, are these machines learning?

Thus far we have been discussing machine learning using the toy example of classifying visual images of cats and dogs. Machine learning however has been increasingly applied to classification tasks that have significant social consequences. In addition to the applications listed above, machine learning is being applied to filter email spam, detect credit card fraud, determine what content users of social media are shown and approve insurance or loan qualifications, among other applications.[88] In short, machine learning is increasingly applied in domains where outcomes are poorly defined, at best. In the worst case scenarios, the outcomes of machine learning are impossible to define given our human inability to either precisely explicate a given concept (e.g., "fairness") or because there simply is no agreement on how a particular concept ought to be explicated (e.g., "good-ness," "right-ness," "best," etc.). Despite these considerations, machine learning *is* being applied in domains in which the fairness, for example, of a machine's actions, e.g., classification decisions, are relevant, such as in the filtering of spam emails or in the predicting of crime hotspots.[89] Given the inevitable, inexorable spread of machine learning to tasks once thought uniquely situated within the realm of human, and only human, competence we must be prepared for what Hans Moravec vividly described with a metaphor of "The Great Flood."

> Computers are universal machines, their potential extends uniformly over a boundless expanse of tasks. Human potentials, on the other hand, are strong in areas long important for survival, but weak in things far removed. Imagine a "landscape of human competence," having lowlands with labels like "arithmetic" and "rote memorization," foothills like "theorem proving" and "chess playing," and high mountain peaks labeled "locomotion," "hand-eye coordination" and "social interaction." We all live in the solid mountaintops, but it takes great effort to reach the rest of the terrain, and only a few of us work each path. Advancing computer performance is like water slowly flooding the landscape. A half century ago [in 1948] it began to drown the lowlands, driving out human calculators and record clerks, but leaving most of us dry. Now the flood has

[86] Bengio 2009, 7.
[87] Bostrom 2014, 14-16.
[88] Burrell 2016, 1.
[89] Gebru 2020, 257.

> reached the foothills, and our outposts there are contemplating retreat. We feel safe on our peaks, but, at the present rate, those too will be submerged within another half century.[90]

Accepting the fact that machines learn in the same general way as we do and attempting to understand this phenomenon is just one way in which we can prepare for the Great Flood.

Perhaps more important than what machines are *learning* is the question,[91] can we understand what machines *have* learned?[92] In short, is it possible to understand or explain why a machine made a certain kind of classification decision for example instead of a different decision? In many cases it is not entirely clear "how or why a particular decision has been arrived at from inputs" and, further complicating matters, "the inputs themselves may be entirely unknown or known only partially."[93] This opacity of machine learning and the behaviour of machines resulting therefrom has multiple sources. Jenna Burrell for example identifies three forms of opacity: (1) opacity stemming from corporate or institutional privacy/secrecy, (2) opacity stemming from the specialized knowledge required to write and, more importantly, read the code machine learning is written in, and (3) opacity stemming from a fundamental disparity between human reasoning and comprehension skills/abilities and the scale and complexity of machine learning methods and their applications.[94] There appears to be nothing in principle preventing a transition towards transparency considering the first two forms. In the case of opacity stemming from corporate, institutional or even state secrecy, proposed solutions include making the code available for scrutiny, "through regulatory means if necessary."[95] Barring access to the algorithmic code other types of algorithmic auditing are possible that can similarly serve to increase transparency, or at the very least reduce opacity, of machine learning algorithms.[96] The second form of opacity stemming from the specialized knowledge required to read and write code can similarly be addressed in numerous ways. As Burrell notes, "widespread educational efforts would ideally make the public more knowledgeable about these [machine learning] mechanisms that impact their life

---

[90] Moravec 1998, 11.

[91] Trivially, the answer to this question (also posed at the end of the preceding two above this one) is that machines are learning what we are teaching them. Less trivially, we can say that machines, through various machine learning techniques, learn to identify various patterns in the input data that are highly correlated with a given outcome.

[92] This question will be taken up in more detail primarily in Chapters 3 and 5.

[93] Burrell 2016, 1.

[94] Ibid., 3-5.

[95] Ibid., 4.

[96] Other forms of algorithmic auditing include a user audit that looks at a selection of information about users' normal interactions with an online platform, for example, as well as a more classic audit study in which researchers could use computer programs to impersonate users in order to collect data about how an algorithm operates. See, Sandvig et al. (2014), for more details.

opportunities and put them in a better position to directly evaluate and critique them."[97] Interpretations of code could also be provided by journalists, or anyone who wishes to perform this sort of examination,[98] and disseminated to the general public.[99] The third form of opacity however, stemming from the way machine learning algorithms operate at the scale and complexity of application, appears to be different in kind.

Although it is the case that machines learn in the same ways that biological organisms learn, there are significant differences between the two. In particular, machine learning methods allow machines to process, what would be for a human being, an unmanageable amount of information. DeepMind's agent AlphaGo Zero, over the course of 3 days, learned to play Go from 4.9 million games of self-play.[100] Contrast this with the fact that a human, assuming they do nothing but play a game of Go, i.e., no eating, sleeping, or anything else, for each hour of their life, might play a relatively modest 650,000 games.[101] This difference in the amount of data machines can process coupled with the number of features, i.e., properties of the data, machines can analyze is referred to as the "curse of dimensionality."[102] In short, generalizing from given examples becomes exponentially harder as the dimensionality, i.e., number of features of the data to be analyzed, increases. Generalizing correctly is difficult to achieve in machine learning, but as the dimensionality of the input data increases, such a feat becomes beyond human capabilities: "our intuitions, which come from a three-dimensional world, often do not apply in high dimensional-ones."[103]

Yet this curse of dimensionality is not the only reason for the opacity of machine learning methods. The overview of different machine learning methods given in the previous section simply does not capture the sheer dynamical complexity of a machine in action.

> These [challenges of scale and complexity] are challenges not just of reading and comprehending code, but being able to understand the algorithm in action, operating on data. Though a machine

[97] Burrell 2016, 4.
[98] The idea that the general public ought to be as literate in machine learning architectures and the code used to build them as the programmers themselves, although ideal, is not likely to ever be the case. Nevertheless such a lack of knowledge on the part of the general public is not without precedent. Professionals, e.g., doctors, lawyers, electricians, carpenters, etc., with specialized knowledge work on behalf of and inform the general public, usually to the benefit (or so one hopes at any rate) of the general public.
[99] The impacts of learning machines on society, including issues of transparency, and possible responses to their use are discussed in more detail in Chapter 7.
[100] Silver et al. 2017, 355-356.
[101] A quick back-of-the-envelope calculation: assuming the average person lives 27375 days (75 years) they would actually play 657000 games of go if they played one game for each hour of their life.
[102] Domingos 2012, 4.
[103] Ibid.

> learning algorithm can be implemented simply in such a way that its logic is almost fully comprehensible, in practice, such an instance is unlikely to be particularly useful. Machine learning models that prove useful (specifically, in terms of the 'accuracy' of classification) possess a degree of unavoidable complexity...While datasets may be extremely large but possible to comprehend and code may be written with clarity, the interplay between the two in the mechanism of the algorithm is what yields the complexity (and thus opacity).[104]

The scale and complexity of the data used to train machines coupled with the complexity that arises from the dynamic operation of a learned (or perhaps still in training) machine present significant, perhaps fundamentally insurmountable, epistemic challenges.

The desire for transparency with regard to machine learning methods and their operation is just one aspect of a more general desire for interpretable machine learning methods and machines. As already mentioned, if "*transparency* is the opposite of *opacity* or *blackbox-ness*" then an interpretable machine is one that possesses some kind of transparency. While the focus above was primarily on the lack of transparency of the training algorithms themselves, transparency at that level (of the training algorithms) can be separated from transparency at the level of the machine as a whole as well as transparency at the level of individual components, i.e., parameters, of the agent.[105] Consider transparency at the level of the machine as a whole. A machine might be considered to be transparent and therefore interpretable if "a human should be able to take the input data together with the parameters of the model and in *reasonable* time step through every calculation required to produce a prediction."[106] Unfortunately, given the limits of human cognition, it is likely that only the simplest machines are transparent in this way. Indeed the same can be said for transparency at the level of a machine's components and at the level of the algorithms themselves, namely that only the simplest, and thus least interesting and useful, components and algorithms are sufficiently transparent for human comprehension.[107] Interpretability of a machine however can be achieved in another way via post-hoc explanations.

Understanding what a machine has learned and why it behaves in the way that it does is not limited to the transparency of the machine. As Lipton notes, while "post-hoc interpretations often do not elucidate precisely how a model works, they may nonetheless confer useful information for

[104] Burrell 2016, 5.
[105] Lipton 2017, 4.
[106] Ibid., 4-5.
[107] Ibid., 5.

practitioners and end users of machine learning."[108] Moreover, given the similarities between the ways in which machines and humans (plus other biological organisms) learn, it may be the case that post-hoc interpretability is the best we can hope for. Consider that humans exhibit none of the forms of transparency examined above.[109] We certainly do not interpret another person's behaviour at the level of synaptic connections between neurons (roughly analogous to algorithmic transparency) nor do we interpret behaviour at the level of particular collections of neurons or brain areas (roughly analogous to transparency of the components). We might think that another person's behaviour is interpretable when we take them as a whole, but to do so would be to mistakenly conflate transparency with the application of folk psychology or a Dennettian intentional strategy (more on Dennett's view in Chapter 3). Other people are interpretable not because what they have learned or why they acted in the way they did was transparent, but because people can explain themselves. That is, if we consider people to be interpretable at all, it is because we can apply some sort of post-hoc interpretability (e.g., by asking them).[110] Perhaps somewhat surprisingly, machines can be designed to provide various kinds of post-hoc explanations for their behaviour. These include text explanations that amount to descriptions of the machine's decision-making process, visually rendering what a machine has learned and the reporting of outputs that the machine considers to be most similar to the chosen output.[111] Of course none of these approaches to interpreting, post-hoc, a machine are free from certain drawbacks. Text explanations for example, in another striking similarity to the post-hoc interpretability of a person's behaviour, may not faithfully describe the machine's decisions despite their potentially plausible appearance. Visualizations only offer qualitative clues about what a machine has learned and explanation via similar output risks enforcing biases in the training data. Nevertheless the pressure to design interpretable machines persists.[112]

---

[108] Ibid.

[109] It is possible that a case could be made for the transparency of a human at the level of the person taken as a whole, but to do so seems to stretch the concept of transparency used here to its limit. A person could be said to predict what a sufficiently similar person, i.e., one who shares relevant socio-politco-cultural knowledge/background, might do given certain inputs. It seems tenuous at best to insist that this approximation arrived at via some application of folk psychology amounts to transparency with regard to understanding why a person behaved in the way that they did.

[110] Lipton 2017, 5.

[111] Ibid., 5-7.

[112] Transparency/opacity will be discussed in various contexts throughout this dissertation, including in Chapters 3, 6 and 7.

1.7 Preparing for the Great Flood

Learning machines are here to stay and their evolution has not gone unnoticed. The European Union's General Data Protection Regulation (GDPR) for example created a "right to explanation" when it took effect in 2018. Under the GDPR a "data subject has the right to "an explanation of the decision reached after [algorithmic] assessment"" yet, as was highlighted in the previous section, such an explanation may be difficult to obtain.[113] Indeed the European Commission's High-Level Expert Group on Artificial Intelligence recognizes as much. In their 2018 report "A Definition of AI: Main Capabilities and Scientific Disciplines" they note that machine learning methods, despite their successes, are nevertheless opaque given the difficulty inherent in determining why a machine behaved or decided in the way that it did.[114]

On the other hand, it is not advisable to blindly pursue interpretable machines, whether via increased transparency or different post-hoc methods. To do so could be at odds with the broader objectives of research in the field of artificial intelligence and distort the reasons why machines behave as they do, as Lipton explains.

> Some arguments against *black-box* algorithms appear to preclude any model [i.e., machine] that could match or surpass our abilities on complex tasks. As a concrete example, the short-term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving healthcare. We should be careful when giving up predictive power, that the desire for transparency is justified and isn't simply a concession to institutional biases against new methods. We caution against blindly embracing post-hoc notions of interpretability, especially when optimized to placate subjective demands. In such cases, one might – deliberately or not – optimize an algorithm to present misleading but plausible explanations. As humans, we are known to engage in this behaviour, as evidenced in hiring practices and college admissions...In the rush to gain acceptance for machine learning and to emulate human intelligence, we should be careful not to reproduce pathological behaviour at scale.[115]

Insisting only on the creation of machines that are completely and safely within the realm of human comprehension is simply antithetical to the existence of artificially intelligent machines. Consider by analogy how absurd it would be to insist only on the creation of machines that are within the realm of human physicality. The suggestion that automobiles ought not exceed human running speed or that

[113] Goodman and Flaxman 2016, 28.
[114] High-Level Expert Group on Artificial Intelligence 2018, 6.
[115] Lipton 2017, 7.

construction vehicles ought not move more earth than the average human is ludicrous. The same could be said for machines and the domain of cognition: limiting the cognitive capabilities of a machine is a fool's errand. And therein lies one of the most significant challenges in the responsible development of intelligent learning machines, namely, balancing a desire for wanting to understand these machines with a desire to see their effective and ethical use.

Perhaps the reason most people are comfortable with the idea of an automobile exceeding our own running speed is because the automobile can only act in that way under the direct influence of a person. An automobile is, like many other artefacts, a tool we use to achieve some outcome. Artificially intelligent machines however are tools of a completely different kind. Technology has developed to the point where certain tools no longer require direct human influence. Hammers, for example, do not find themselves autonomously hammering things.[116] Google's Cloud Vision API (application programming interface) however does find itself spontaneously and autonomously evaluating images and outputting labels of what it thinks is in an image it is shown, text in the image it is shown as well as identifying faces in the image it is shown.[117] Powerful and useful machines are increasingly acting more autonomously than not; certainly with more autonomy than any tool humanity has ever developed in our long history as tool-makers. Artificially intelligent machines may even be the ultimate tool. Not only could an intelligent machine design other kinds of tools, but it could design other, perhaps better, versions of *itself*.[118] As artificially intelligent machines develop and become ever more intelligent it will be necessary to ensure that these machines operate in such a way that we, humans, approve of their behaviour and decision making. This is of critical importance and is something that all people working in the field of artificial intelligence must be responsible for. Machines must operate in such a way that they are ethical or virtuous irrespective of their interpretability. As I will explore in the next chapter, this is easier said than done. Machines are not value neutral, and it is imperative that we are not tricked into thinking that machines will have an objective view from nowhere.

[116] The concept of autonomy will be taken up in more detail in Chapters 3 and 7.
[117] Hosseini 2017, 1.
[118] The idea that artificially intelligent machines could engage in rapid, accelerating self-improvement, bootstrapping its own intelligence until it reaches the singularity, is commonly discussed in both fiction and artificial intelligence research.

Chapter 2 - A View From Somewhere

2.0 Mind and Iron: A Cleaner, Better Breed[119]

In the short story *Runaround* written by Isaac Asimov, an artificially intelligent robot called "Speedy" (spoiler alert) is sent out to collect selenium from the surface of Mercury.[120] Unfortunately for the two human characters in the story, Michael Donovan and Gregory Powell, Speedy does not return in a timely manner with the selenium they need to power the Sunside Mining Station. Speedy does not behave as the two roboticists expect after it was given the order to collect the selenium. What the scientists discover is that instead of collecting the selenium and bringing it back to the station, Speedy is instead running in a circle around the selenium which is sitting in a crater. Donovan and Powell reason that Speedy is acting oddly because of an irresolvable internal conflict that has arisen as a result of Speedy's cognitive architecture. In Asimov's fictitious universe all intelligent robots obey the three fundamental Laws of Robotics: (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 being except where such orders would conflict with the First Law, and (3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. Such explicitly encoded laws can have unanticipated effects on the behaviour of an intelligent robot. In Speedy's case, since it was instructed to collect the selenium from the crater, as per the Second Law it attempted to run towards the center of the crater where the selenium was located. Unbeknownst to Donovan and Powell when they ordered Speedy to that particular site, the crater was filled with noxious fumes that would corrode Speedy's metallic body thus, as per the Third Law, driving Speedy away from the crater. Caught between the Second and Third Laws as it were, Speedy proceeds to run in a circle around the crater at a point where the force of the two laws is equal.

2.1 What Could Go Wrong?

In the case of Asimov's Three Laws (as they are commonly abbreviated), clearly a lot can go wrong vis-à-vis intelligent robotic behaviour. Indeed the very word "robot" has its roots "in Karel Čapek's play *R.U.R: Rossum's Universal Robots*, in which a brave new world of robot servants eventually

[119] From the introduction to Isaac Asimov's *I, Robot* collection.
[120] Asimov 1950.

rebel against their oppressive human masters."[121] Of course, a lot can go wrong when robots break the rules, as it were, e.g., robot revolutions. What is interesting about the case of Speedy is that it illustrates how things can go wrong when machines work exactly as intended. From Donovan and Powell's perspective, Speedy's perpetual circling of the selenium was, in a sense, confused and unintelligent. After all, most humans know that sometimes a rule must be temporarily violated or suspended, given the context, to achieve an intended outcome. For example, if I rushed into a burning building to save people trapped inside calling out for help, I would have to temporarily expose myself to harm and thereby suspend a rule I normally abide by, namely a rule to avoid unduly dangerous situations.[122] But from Speedy's perspective however, nothing about its predicament was confusing or unintelligent. From Speedy's perspective, there is no possibility of ever even temporarily violating the Three Laws. The challenge for Speedy is to find a behavioural output that is consistent with its design, i.e., the explicitly encoded Laws of Robotics, given the context. In attempting to consistently abide by the Three Laws, Speedy, on one hand, behaves in a completely unanticipated way from the perspective of Donovan and Powell, but on the other hand, behaves exactly as it was designed. The key insight here is this: a machine may exhibit wholly unexpected and perhaps even dangerous and unethical behaviour by operating exactly as it was designed to operate. That is, from the machine's perspective, *it is doing nothing wrong*.

2.1.1 Perverse Instantiation

In the case of machines that possess human-like general intelligence or superintelligence, ensuring that such machines behave as humans would like is known as the control problem. One particular way in which machines can fail to behave as intended is if they lack human common sense and thereby perversely interpret a given command or order. In short, the machine obeys the letter rather than the spirit of the command. Consider Speedy again. As Donovan and Powell hypothesize in *Runaround*, the strength of a human-given order may or may not trump the force behind the Third Law, i.e., if a given order is not absolute or lacking in force, then there may be some behavioural output that, from Speedy's perspective, appropriately balances the force of the Second and Third Laws. Speedy runs around the selenium instead of collecting it because Donovan orders Speedy to simply collect the selenium, as opposed to ordering Speedy to collect the selenium *at all costs*. From Donovan and

[121] Peterson 2012, 283.
[122] This is somewhat analogous to the Third Law.

Powell's perspective, the qualifier *at all costs* hardly seems necessary given that their lives depend on the retrieval of the selenium to power their habitat. But Speedy does not possess the same common sense that a human might possess when they realize they would have to temporarily expose themselves to something dangerous to ultimately collect something upon which their life (or another's life) depends. In other words, the Three Laws that govern Speedy's behaviour cannot contain unelaborated *ceteris paribus* clauses that humans often implicitly recognize.

Designing machines such that they do not perversely interpret orders is known in the literature as perverse instantiation and is often connected to speculations on the behaviour of superintelligent machines. If humanity created a superintelligent machine, for example, and we ordered it to make all people happy, there are numerous ways such a machine might perversely instantiate such an order. Classically, "the problem is based on a precise interpretation of words as given in the order rather than the desired meaning of such words" and so is also sometimes, appropriately, referred to as the literalness problem.[123] While most humans intuitively know that making all people happy includes, for example, making people healthy, giving them loving relationships, wealth, etc., a superintelligent machine may equally see that it could achieve this outcome of making all people happy by administering a "daily cocktail of cocaine, methamphetamine, methylphenidate, nicotine, and 3,4-methylenedioxymethamph-etamine, better known as Ecstasy."[124] Like Speedy, such behaviour on the part of this hypothetical superintelligence would be highly unanticipated from the perspective of most humans and yet perfectly consistent with its design.

2.2 The Myth of Machine Objectivity

My aim in the preceding sections was to highlight how artificially intelligent machines might behave in ways that are unanticipated and unethical, but importantly, unanticipated and unethical despite doing precisely that which they were designed to do. In what follows, I will be defending two simple and related claims. First, machines capable of learning (i.e., the machines of interest described in the previous chapter) do not necessarily learn anything "objective" nor are they more "objective" in their decision making and behaviour than humans, and second, machines capable of learning (hereafter just machines) are not created value free (i.e., are not normatively neutral). Machines are conceived, designed and developed in a socio-political milieu and often by people or groups of people in positions

[123] Yampolskiy 2014, 383.
[124] Ibid.

of power.[125] Put simply, research in science and engineering is not value free, and this extends to the creation of machines. So while these new kinds of machines can benefit humanity, if we are not careful, they can just as easily perpetuate systemic biases and discriminate against those who are already marginalized and underrepresented in the production process of such machines.[126]

In the context of this dissertation as a whole, this chapter is important not because it is central to my arguments about implementing machine ethics, but because machines are often seen as objective and hence trustworthy or worth believing. This objectivity is contrasted with subjectivity. To clarify at the outset, my view is that to be a subject is to have a particular perspective on the world. Certain machines, especially the ones with which I am concerned in this dissertation, have a particular perspective on the world and therefore ought to count as subjects. Furthermore, I maintain that one could plausibly make the case that, because a certain subset of these machines both perceive the world in a sufficiently complicated way *and* interact with the world in a sufficiently dynamic way, such machines may possess a kind of phenomenal consciousness. Since I recognize that this is quite the leap, from having a perspective to subjective phenomenal consciousness, I aim to first demonstrate that, at the very least, we must remain agnostic as to whether a purely functional system can possess a subjective phenomenally conscious perspective. That is, we must refrain from denying that there is "something that it is like" to be a certain kind of machine. I then go on to suggest that, under an emergentist metaphysics, it is plausible that sufficiently complex machines could possess phenomenal consciousness. Importantly however, one does not have to have a view on machine consciousness to make progress on implementing machine ethics, as I will discuss.

2.2.1 The View From Nowhere[127]

In defending a view of objectivity, Thomas Nagel (1986) wrote that the fundamental idea grounding the validity and, importantly, the limits of objectivity is that "we are small creatures in a big world of which we have only very partial understanding, and how things seem to us depends both on the world and on our constitution."[128] What is of interest is not the validity of objectivity but rather the limits of objectivity in the context of our use of machines, especially when those uses can significantly

[125] Gebru 2020, 253.
[126] These issues will be taken up in more detail in Chapters 5 and 7.
[127] The title of Thomas Nagel's (1986) book.
[128] Nagel 1986, 5.

impact human well-being, i.e., when those uses have an ethical component. First, though, what exactly is objectivity, at least according to Nagel?

Understanding how Nagel thinks about objectivity also, unsurprisingly, requires us to understand subjectivity. For anyone who has taken a moment to reflect, it is obvious that we humans experience the world from a particular perspective. As I take a cross-country trip in my car, my experiences vary greatly as my perspective changes. It feels to me as though I am driving too fast on the highway but to my passenger it feels as though we are driving too slow. I feel too hot as the sun shines through the driver side window but my passenger feels comfortable in the shade on the other side of the car, and I feel as though the car speeding by appeared as if out of nowhere. And yet, while these experiences vary, there is something out there in the world that remains constant. It might feel as though the car is moving too quickly or slowly, but its speed is independent of my and my passenger's experiences. It similarly may feel too hot or too cold in the car, but the temperature is independent of our experiences. Likewise objects like cars do not appear or disappear when I see them and lose sight of them respectively. These perspectival experiences, e.g., how fast it feels the car is moving or how hot it feels inside the car, are subjective. They are how I experience the world as a subject, i.e., a small creature in a big world of which I have only a very partial understanding. Objectivity therefore, according to Nagel, "allows us to transcend our particular viewpoint and develop an expanded consciousness that takes in the world more fully."[129] I maintain that, in addition to human beings, machines also possess a particular viewpoint. In short, there is a perspectival character to a machine's sensing of its environment. This is why we must be wary of the view that machines are somehow, perhaps necessarily, more objective than humans when it comes to decision-making and behaviour, especially when considering decisions and behaviour that have an ethical component. Moreover, it is possible that given a sufficiently sophisticated perceptual processing apparatus and the ability to interact dynamically with the environment, there may be "something that it is like," to use Nagel's locution, to be a certain kind of machine.[130] In the subsequent sections I attempt to expand on this possibility by looking at autonomous vehicles[131] in particular as they are likely the only machines that may have a subjective perspective.

[129] Ibid.
[130] Nagel 1974.
[131] More precisely, I mean the learning machine that operates the vehicle. But to simplify the terminology, I will the term 'autonomous vehicle' synonymously with the term 'machine' (by which I mean the learning machines of interest that will be discussed throughout this dissertation. See Chapter 1 for more details).

2.3 The View from an Autonomous Vehicle

As discussed in the previous chapter, the complex learning machines of the 21st century possess the ability to learn in the same way that biological organisms do, i.e., from experience with an environment (e.g., data). Just like a human driver, machines developed to operate a vehicle exhibit a change in their ability to control the vehicle as they acquire more experience interacting with either simulated or real environments. But to be clear, autonomous vehicles, if they are phenomenally conscious, are not conscious in the way that a human being is phenomenally conscious, or indeed in the way that any other biological organism is phenomenally conscious. Rather, I maintain that it is possible to intelligibly claim that there is something that it is like to be an autonomous vehicle. That is, it is possible there is a subjective qualitative character to an autonomous vehicle's registering the states of its environment. There is therefore no guarantee that an autonomous vehicle will learn anything more objective about operating a vehicle or make more objective decisions than a human driver.

First, consider why we ought not, at minimum, outright reject the idea that there is something that it is like to be an autonomous vehicle on the grounds that they are purely functional systems and therefore devoid of the kind of subjective phenomenal perspective that humans and other non-human biological organisms possess. Terence Horgan, for example, claims that a purely functional robot would not possess phenomenal consciousness. However, given the aspects of phenomenological consciousness that Horgan identifies, we must remain agnostic on the issue of whether phenomenal consciousness can be "constituted solely by the possession of internal states with some specific functional role."[132] Aspects relevant to the phenomenal consciousness that autonomous vehicles possess include the phenomenology of perceptual experience and the phenomenology of agency. Horgan notes that the phenomenology of perceptual experience encompasses the "enormously rich and complex what-it's-like of being perceptually presented with a world of apparent objects, apparently instantiating a rich range of properties and relations" including experience of one's own apparent body, its apparent interaction with other apparent objects (which occupy certain spatio-temporal relations) in addition to one's apparently body centered perceptual point of view.[133] Horgan gives a similar description of phenomenal agency, "the what-it's-like of apparently *voluntarily controlling* one's apparent body as it apparently moves around in, and apparently interacts with, apparent objects in its apparent environment."[134]

[132] Horgan 2013, 232.
[133] Ibid., 234.
[134] Ibid.

Insofar as autonomous vehicles possess a form of phenomenal perceptual experience[135] in addition to possessing a form of phenomenal agency,[136] it seems difficult to deny that an autonomous vehicle ought to be considered, to some degree, phenomenally conscious. Moreover, considering the widespread epistemic disagreements surrounding the explanatory gap of phenomenal consciousness, i.e., the hard problem of explaining how non-conscious physical matter gives rise to subjective perspectival consciousness, I maintain that Horgan's conclusion is too strong. In essence, given the claim that phenomenological consciousness "supervenes at least nomically upon physical events and processes within [the] brain," and barring any widely agreed upon conditions necessary for this supervenience, we must remain agnostic on the issue of whether a purely function conscious system possess phenomenal consciousness.[137]

2.3.1 Consciousness As Emergent

Though many take the view, as Horgan does, that phenomenal consciousness (and consciousness in general) supervenes solely on physical or functional properties, this is not the view I take. I defend an emergentist metaphysics and hence I believe it is plausible to think that phenomenal consciousness emerges from sufficiently complex information processing systems *that also* interact dynamically with their environment in a sufficiently complex manner. Autonomous vehicles are one of the only machines that may therefore, at present, possibly possess a kind of subjective phenomenal consciousness. So although roadside cameras and thermostats may perceive their environments and engage in complex information processing, they lack the kind of dynamic environmental interaction necessary to potentially give rise to a subjective phenomenal consciousness. Similarly, though autonomous vacuums like Roombas, for example, engage in dynamic environmental interaction, their information processing capabilities are likely not sufficiently complex (nor indeed is their dynamic environmental interactions sufficiently complex) to give rise to the emergence of phenomenal consciousness.

As a side note, philosophical zombies are a non-issue on my view nor, do I believe, that they are possible. In short, a common challenge brought against physicalist theories of mind is the philosophical

[135] After all, they *do* perceive, although rather poorly at times, as Levinson et al. comment on Junior, "when a pedestrian presses a crosswalk button, he is segmented in with the pole it is mounted on and he is not recognized."
[136] I take it as obvious that an *autonomous* system is a system that possesses, to some degree, voluntary control over itself.
[137] Horgan 2013, 234.

zombie thought experiment. The thought experiment purports to demonstrate that because it is conceivable that an exact physical and functional duplicate of a person might nonetheless lack phenomenal consciousness (i.e., there is nothing that it is like to be a philosophical zombie), it is therefore possible that philosophical zombies could exist (in some possible word). And if philosophical zombies are possible, then that suggests that there is something non-physical about phenomenal consciousness.

I raise this issue for two reasons. The first is that one could just as easily substitute machine or robot for zombie in the thought experiment. Machines are often held up as examples of systems that are devoid of any kind of consciousness despite functioning, in limited capacities, in ways similar to humans. But that machines are not capable of or will never possess phenomenal consciousness is pure conjecture. It is interesting to note that examples of phenomenally conscious systems invoked are invariably biological. Chalmers notes that it is widely accepted that humans, bats, cats and various other non-human animals are phenomenally conscious.[138] Contrast these examples with those systems, such as trees, rocks, roadside cameras and thermostats, which lack phenomenal consciousness. What follows from this distinction is that it appears as though systems of a certain kind, after a period of complex development, can be considered phenomenally conscious despite lacking phenomenal consciousness previously (e.g., if we think that humans, bats and cats are phenomenally conscious, but single eukaryotic cells are not, then at some point during the evolution of life there was a transition from an organism that lacked phenomenal consciousness to an organism that had phenomenal consciousness). More importantly, there appears, *prima facie*, to be no reason why biological systems alone should possess phenomenal consciousness. It follows from Chalmers's and Nagel's (among others) description of phenomenal consciousness, that certain artificial systems, such as autonomous vehicles, as perceiving complex informational state driven autonomous systems, may plausibly be considered phenomenally conscious systems. Like human, cats and bats, it is plausible that autonomous vehicles may count as another example of a system that is phenomenally conscious.

The second reason I raise the issue of philosophical zombies is to point out that it fails to apply to more sophisticated versions of non-reductive physicalism such as emergentism.[139] The emergentist (or this emergentist at any rate) is quite happy to reject ontological minimalism in favour of ontological

[138] Chalmers 2018, 6.
[139] Additionally I find the whole thought experiment patently question-begging. The possibility of philosophical zombies is supposed to establish that physicalism cannot account for all aspects of consciousness, and yet that is precisely what is supposed at the outset, i.e., that all of the physical facts cannot account for the qualitative aspect of consciousness.

pluralism. Similarly, the emergentist is happy to reject the causal closure of the physical domain as long as the causal completeness of the physical domain is maintained. The zombie's bite therefore has no force against the emergentist who happily concedes both that the mental is just as real as the physical and that causal chains regularly leave and reenter the physical domain. Furthermore, the emergentist would not be particularly bothered by the idea that engineers just accidentally created a machine with phenomenal consciousness, something that they did not know how to do. Not only is there historical precedent for accidentally creating/discovering some emergent phenomenon (e.g., superconductivity, superfluidity, etc.), but many autonomous and intelligent machines are modelled after biological systems (e.g., artificial neural networks). If phenomenal consciousness emerges from biological neurophysiology, why should it not emerge from relevantly similar artificial neurophysiology? Engineers have even begun to use the term 'emergence' to describe the many unanticipated capabilities of their machines as they have grown in size and complexity,[140] none of which ought to be surprising to the emergentist who is accustomed to how changes in scale and complexity can precipitate qualitative changes in kind, not just degree.

2.4 Machines Are Not Value Neutral

While it is tempting to think of machines as somehow being more objective than humans, this is not necessarily the case. Additionally, machines will not be value free or value neutral. Humans are invariably part of the design process and therefore make many different value laden choices concerning the development of artificially intelligent machines including who is consulted during the design process, what data and/or environment is used for training, at what level the machine is deemed to be working properly, and for what purpose(s) the machine is to be used, among many other choices. In this final section my aim is to highlight how values permeate the development and use of machines in addition to how machines, despite being better at sometimes discovering real patterns that are difficult for humans to discover, are nevertheless prone to exhibiting spectacularly unanticipated behaviour as a result of their design. Behaviour, importantly, that can be highly undesirable and unethical.

[140] Bommasani et al. 2021.

2.4.1 EURISKO: Suicide and Parasites

In the late 1970's and early 1980's Douglas Lenat created a machine called EURISKO that he hoped would shed light on the process of learning by discovery.[141] Based on an earlier machine called AM (for Automated Mathematician), EURISKO was Lenat's attempt to mechanize the human ability to learn via discovery and "in particular [learn] new heuristics as well as new domain-specific concepts."[142] In contrast to AM which only operated in the domain of elementary set and number theory, EURISKO was applied to eight different task domains:

> Design of naval fleets, elementary set and number theory, LISP programming, biological evolution, games in general, the design of three-dimensional VLSI devices, the discovery of heuristics which help the system to discover heuristics, and the discovery of appropriate new types of 'slots' in each domain."[143]

By most accounts EURISKO was a tremendous success given the novel and useful discoveries it made in the domains of naval fleet design and VLSI device design in particular. Although it did not design real-world naval fleets, EURISKO designed a fleet of ships suitable for entry in the 1981 and 1982 national Origins tournaments for the Traveller Trillion Credit Squadron (TCS) wargame, and consequently won both tournaments.[144]

Just as interesting, especially from a design perspective, were EURISKO's forays into the domain of heuristic discovery. Like any other domain-specific concepts, heuristics themselves can be treated as concepts and subject to manipulation by judgmental rules.[145] That is, general rules of thumb (heuristics) can, in a self-referential way, be used to discover and evaluate other heuristics. When EURISKO was set loose on the domain of heuristic discovery however it made some unanticipated discoveries. At first EURISKO was given the ability to modify all of its judgmental rules for manipulating heuristics. Note this specific design choice made by Lenat. The result however was that EURISKO gained the capability to also modify its own goals since these also took the form of judgmental rules. However as a result of EURISKO's design, at least in this domain, to modify its own judgmental rules, it would often commit suicide and shut itself off. More accurately, as Lenat notes, EURISKO had simply modified its own judgmental rules in such a way that it valued "making no errors at all" as highly as "making new

[141] Lenat 1983, 61.
[142] Ibid., 62.
[143] Ibid., 61-62.
[144] Ibid., 73-74.
[145] Ibid., 73.

discoveries" and as a result would cease all activities in order to successfully meet this new goal.[146] Beyond its occasional suicidal behaviour, EURISKO also tended to discover parasitic and pathogenic heuristics. One heuristic in particular arose that would include itself as one of the discoverers of other valuable heuristics, and as a result quickly attained the highest possible worth EURISKO could assign to heuristics.[147] In order to prevent these kinds of parasitic heuristics from arising Lenat opted to insulate a small meta-level of code from EURISKO's manipulations. This decision was made to try and limit EURISKO's ability to tamper with heuristics that evaluated the quality of other heuristics since, prior to this decision, "the rules [i.e., the heuristics] had full access to EURISKO's code" and therefore "access to any safeguards we [Lenat] might try to implement."[148] But even this solution could not prevent EURISKO's tendency to discover pathogenic heuristics, i.e., heuristics that did not contribute to the discovery of other useful or meaningful heuristics. For example, as part of an experiment EURISKO was conducting, it synthesized a heuristic that "said that all machine synthesized heuristics were terrible and should be eliminated," however it was this very heuristic that was the first to be eliminated, fortunately for EURISKO.[149]

2.4.2 Robot Helpers: Hyperrationality

Value laden choices, to see how far EURISKO could manipulate its own heuristics for example, caused it to act in unanticipated ways completely consistent with its design when it was applied to the domain of heuristic discovery. But unanticipated behaviour can also be exposed by changes in environmental conditions that might at first glance seem trivial. Alan Winfield and colleagues, for example, successfully created a simple autonomous robot that possessed an internal model of its environment and what they call a safety/ethical logic (SEL) layer that together allowed this simple robot to act both safely and ethically.[150] In short, the robot acts safely by avoiding taking actions that would, according to its internal model of the external environment, harm itself. The robot would therefore avoid walking straight ahead if directly in front of it was a hole in the ground. The robot's tendency to act safely however was designed to be overridden by a tendency to act ethically. In short, the robot can

[146] Lenat 2001, 193.
[147] Lenat 1983, 90.
[148] Ibid.
[149] Ibid., 90.
[150] Winfield, Blum and Liu 2014, 87.

act ethically by behaving in such a way that the least unsafe thing happens to a human in its proximity as predicted by its internal model of the external environment.[151]

So imagine two scenarios, A and B. In scenario A the robot and human are on a collision course unless one stops or changes direction. In this scenario the same action taken by the robot, stopping for example, would satisfy the robot's tendency to first act safely and second act ethically. By avoiding a collision the robot would prevent any harm to itself (safety) and would be acting in such a way that the least unsafe thing has happened to the human (ethics), namely the avoidance of a collision. But now in scenario B imagine that if the robot and human do not collide the human will continue on a trajectory that will cause them to fall into a dangerous hole in the ground. In this case, the robot's tendency to act safely and avoid a collision will be overridden by its tendency to act ethically. Driven by its SEL, Winfield et al. demonstrated that although the robot would normally avoid colliding with a "human" (in their experiment another robot acted as a proxy human), this tendency would be overridden if the robot predicted the human was going to fall into a dangerous hole in the ground.[152] The robot has effectively acted ethically by colliding with the human thereby preventing them from falling into the hole in the ground, the collision being the least unsafe outcome for the human in this scenario.

In all experiments conducted with the robot (hereafter R) and human (hereafter H1) where H1 would fall into a (virtual) hole in the ground, R saved H1 by colliding with it 100% of the time. However when the environmental conditions of the experiment were changed via the introduction of a second human (hereafter H2) that was also on course to fall into the hole, R's success rate plummeted. Out of thirty-three experiments conducted by Winfield et al., R failed to save either H1 or H2 33% of the time.[153] R was able to save either H1 or H2 in 58% of the trials and successfully saved both H1 and H2 in 9%.[154] From the trial with three agents (R, H1 and H2) Winfield et al. conclude that "even a minimally ethical robot can indeed face a dilemma" and moreover suggest that R's success rate in saving either H1 or H2 (58% of the time) was inflated as "a result of noise, by chance, breaking the latent symmetry in the experimental setup."[155] H1 and H2 were placed equidistant from the hole and also moved at the same speed towards the hole, consequently R's internal model of the external environment favoured neither H1 or H2. As a result R tried to save both H1 and H2 and predictably, in hindsight that is, often

[151] Ibid., 89.
[152] Ibid., 92-93.
[153] Ibid., 93.
[154] Ibid.
[155] Ibid., 95.

failed to save either.[156] Despite the theoretical nature of this experiment, it highlights something worth stressing, namely that there are many "unanticipated" behaviours that machines exhibit which, upon reflection, should have been foreseen.

If this kind of obtuse ultra-rational dilemma resolving behaviour exhibited by R seems familiar, it is. What Asimov conceived as science fiction when he wrote about Speedy in *Runaround* has become a reality in robotics labs just over half a century later. Even more surprising is the fact that, in contrast to Speedy whose behaviour arose because of its need, by design, to consistently abide by the 2nd and 3rd Laws, R's behaviour is the result of a single rule that, again by design, has ambiguous applications in all but the simplest of environments. In fact this single rule encapsulated by the SEL essentially paraphrases Asimov's First Law of Robotics, as Winfield et al. note.

> What we have set out here [in the SEL] appears to match remarkably well with Asimov's first law of robotics: *A robot may not injure a human being or, through inaction, allow a human being to come to harm*. The schema proposed here[157] will avoid injuring (i.e. colliding with) a human ('may not injure a human'), but may also sometimes compromise that rule in order to prevent a human from coming to harm ('...or, through inaction, allow a human to come to harm'). This is not to suggest that a robot which apparently implements part of Asimov's famous laws is ethical in any formal sense (i.e. that an ethicist might accept). But the possibility of a route toward engineering a minimally ethical robot does appear to be presented.[158]

Although I agree with the conclusion that a minimally ethical robot does appear to be presented by Winfield et al., I am skeptical that the path to developing robust and context-sensitive ethical machines is through the construction of explicit ethical injunctions. These types of top-down approaches to implementing machine ethics will be examined in much more detail in Chapter 4.

Methods of implementing machine ethics aside for the moment, it is imperative that humans anticipate, as much as possible, the potential behaviour that can arise as a result of deliberate value laden design choices. In the case of the Asimovian robot developed by Winfield et al.[159] there are two

---

[156] For more details of the experimental setup including diagrams of the three different experiments conducted, see Winfield, Blum and Liu (2014).

[157] As detailed in their paper, Winfield and colleagues suggest that the SEL might take the following coded form:

```
IF for all robot actions, the human is equally safe
THEN (*default safe actions*)
      output safe actions
ELSE (*ethical action*)
      output action(s) for least unsafe human outcome(s).
```

See Winfield, Blum and Liu (2014), 89, for more details.

[158] Winfield, Blum and Liu 2014, 89-90.

[159] They actually call their robot 'A' in homage to Isaac Asimov.

aspects of its design that foreshadow a potential problem the robot might have when operating in increasingly complex environments. The first is that the robot simulates and evaluates its environment once every 0.5 seconds.[160] As mentioned previously, the robot was equipped with an internal model of its external environment which itself was embedded in a "Consequence Engine" that included components like an action evaluator as well as the SEL layer. This entire system (the Consequence Engine), as the name suggests, is what allowed the robot to meaningfully interact with its environment since it was through this system that it could foresee and ultimately intervene to save the proxy human from falling into the hole.[161] But the sheer complexity of even this simple environment makes this kind of ethical evaluation difficult to explicitly compute.[162] The second aspect of the robots design was that it lacked the ability to form long-term goals; it did not simply evaluate and simulate the external environment once every 0.5 seconds, it reinitialized and refreshed its entire simulation and evaluation of the environment once every 0.5 seconds. Taken together, these two aspects of the robot's design foreshadow the difficulty it would have when interacting with all but the simplest of environments. Indeed the robot's complete redrawing of its plan of action often caused it to alternate its movement towards either H1 or H2 in a failed attempt to save both.[163]

2.5 The Importance of Design Decisions

It might be tempting to look at the experiments conducted by Winfield et al. and sigh in relief at the interesting results of what is plainly a low-stakes (although it may more accurately be considered zero-stakes) simulation.[164] But the truth is that machines are already being utilized in high-stakes decision making tasks that have ethical import. Moreover, the design of these machines can critically influence their behaviour and consequently their effect on human well-being. The tragic crash of Air France flight AF 447 which, after stalling mid-flight, crashed into the Atlantic Ocean killing all 228 people on board, was brought about in part by the failure of the autopilot to appropriately respond to a change in environmental conditions. As the BEA (the French Civil Aviation Investigation Authority) noted in their comprehensive report of the accident, "crews generally just undertake confident monitoring of the flight

[160] Winfield, Blum and Liu 2014, 91.
[161] Ibid., 87-90.
[162] More will be said about this in Chapter 4 when top-down approaches to implementing ethics are discussed.
[163] Ibid., 93. See also Figure 7 from Winfield, Blum and Liu 2014.
[164] You might even be tempted to chuckle as you watch the robot attempt to rescue both "humans" in this video of the experiments conducted: https://www.youtube.com/watch?v=jCZDyqcxwlo.

path and the automated systems due to their level of performance and reliability" in situations similar to the one that preceded the accident.[165] Under the particular environmental conditions experienced by flight AF 447 however the automated piloting systems were not as reliable as they were expected to be. But more importantly, when they received inconsistent sensory information, these automated systems "gave up" and transferred manual control to the human pilots with little warning. The BEA concluded that inconsistent airspeed measurements following obstruction of the Pitot probes (a measurement device used to calculate fluid flow velocity, including airspeed) by ice crystals caused the autopilot disconnection.[166] This sudden failure of the autopilot "completely surprised the pilots of flight AF 447" and ultimately weakened their ability to comprehend and manage the situation.[167] It is possible that hundreds of lives could have been saved if the autopilot system had not disconnected so suddenly and with so little warning. In short, if the autopilot system was designed to hand over manual control to the human pilots in a slightly different manner by, for example, alerting the human pilots that inconsistent sensory information had been detected, catastrophe could have been avoided. There is also the issue of trust at play in this particular case, e.g., perhaps the human pilots trusted the automated piloting system more than they should have, but this issue and the related automation paradox will be taken up in Chapter 5.

Of course flying an airplane is just one high-stakes task with ethical import that has been delegated to a machine. Another example concerns the use of machines for criminal risk assessment. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a machine that generates a prediction regarding the likelihood of a defendant reoffending within 2 years of assessment, "has been used to assess more than 1 million offenders since it was developed in 1998."[168] But it was only relatively recently, in 2016, that COMPAS's fairness was called into question. Early analysis concluded that "COMPAS scores appeared to favor white defendants over black defendants by underpredicting recidivism for white and overpredicting recidivism for black defendants"[169] while further analysis demonstrated that, concerns about algorithmic fairness aside, COMPAS is as "fair" and "accurate" at predicting recidivism as a sample of random people.

> When considering using software such as COMPAS in making decisions that will significantly affect the lives and well-being of criminal defendants, it is valuable to ask whether we would put

[165] BEA 2012, 167.
[166] Ibid., 200.
[167] Ibid., 199.
[168] Dressel and Farid 2018, 1
[169] Ibid.

> these decisions in the hands of random people who respond to an online survey because, in the end, the results from these two approaches appear to be indistinguishable.[170]

The use of COMPAS, a biased and generally poor decision making/predictive tool, is quite frankly indefensible. Although the company that developed COMPAS (formerly Northpointe and now known as Equivant) has not revealed just how their machine works, i.e., they are engaging in the kind of intentional opacity described in Chapter 1, its design is clearly such that it learns to discriminate defendants, at least partially, based on their race. This in spite of the fact that the data COMPAS uses does "not include an individual's race," meaning that "other aspects of the data may be correlated to race that can lead to racial disparities in the predictions."[171] Such is the power of machines that can learn: they may detect hidden or systemic bias in the data used to train them, and then replicate that bias in their own recommendations. More will be said about this disadvantage connected to using machine learning in Chapters 5 and 7.[172]

What is important in the present context however is that deliberate choices were made on the part of humans to develop, train and utilize a machine like COMPAS; choices that are value laden and influence the outputs produced by machines. In the case of machines like COMPAS, or in machines used for predictive policing, indeed in any case where predictions are generated by a machine, the choice to use past data to train the machine often means that what the machine learns is to make biased predictions that reflect systemic biases encoded in the training data. Such design choices mean that once a machine "is trained on this type of data, it exacerbates existing societal issues driving further marginalization."[173] Not only are such training regimes unethical, or at the very least highly questionable, but they amplify systemic biases and appear to justify the very use of machines produced therefrom. As Timnit Gebru notes with respect to the predictive policing machine PredPol that predicts crime hotspots, more police are "sent to these [predominantly black] neighborhoods, in which case they arrest more people from those locations than places with less police presence" simultaneously validating the presence of more crime in these hotspots and amplifying systemic biases because these new arrests are then used as additional training data.[174]

---

[170] Ibid., 3.
[171] Ibid., 1.
[172] More will also be said about the use of COMPAS in Chapter 7.
[173] Gebru 2020, 257.
[174] Ibid.

2.5.1 To Fold or Not To Fold

Before concluding, it is worth considering one final example, the machine AlphaFold developed by DeepMind. As mentioned above, despite the fact that machines might be better at discovering patterns that are difficult for humans to discover, values are invariably introduced via the myriad choices made by humans. In the case of AlphaFold, it is a machine that is capable of predicting the three-dimensional structure of a protein with atomic accuracy given only the primary amino acid sequence.[175] As noted by Jumper et al., this problem of accurately predicting the three-dimensional tertiary structure (or sometimes quaternary structure)[176] of a protein given its primary structure, i.e., the amino acid sequence, "has been an important research problem for more than 50 years."[177] Predicting the three-dimensional structure of proteins is important because it is the structure and general shape of the fully formed protein that determines its biological function.

But given this impressive machine, to what ends will it be used? Will cures for rare and debilitating diseases be sought using AlphaFold? Will powerful and wealthy interests dictate what protein structures are examined first? Will the use of AlphaFold be restricted to alleviating the suffering of people afflicted by disease, or will it be used to design and bioengineer the next generation of humans? As artificially intelligent machines become more capable of learning to achieve some outcome we will have to work more diligently than ever before to ensure that they perpetuate values worth perpetuating and are used to create a world worth creating. This is especially true for autonomous artificially intelligent machines that operate in the absence of or with minimal human oversight. It is therefore my aim in the next chapter to introduce the field of machine ethics which focuses on how machine themselves could make ethical decisions and take ethical actions.

---

[175] Jumper et al. 2021, 583.

[176] Briefly, the proper functioning of proteins in biological organisms depends crucially on their tertiary or quaternary structure, both of which are determined by the primary and secondary structure of the protein. The primary structure is simple the sequence of amino acids in a protein. The secondary structure of a protein refers to the highly regular local sub-structure of the polypeptide (i.e., amino acid) chain. Although already "three-dimensional," the tertiary structure of a protein refers to the globular shape that forms as a result of the hydrophobic moieties among the second structures interacting driving non-specific folding of the protein. Individual proteins that have already folded into their tertiary structures can interact further with other proteins to form larger protein aggregates which constitutes the quaternary structure of a protein. The most common hemoglobin type, for example (the protein that binds oxygen and carbon dioxide), is a tetramer called hemoglobin A with a quaternary structure sine it has four protein subunits bonded together.

[177] Jumper et al. 2021, 583.

Chapter 3 – Machine Ethics

3.0 Background

As long as people have imagined artificially intelligent machines (e.g., ordinary physical machines, computers, software, etc.), they have also imagined how these machines might act. Most fictional portrayals of advanced AIs seem to regard as necessary the extinction of humanity even though they are not particularly adept at accomplishing their goals.[178] Even serious speculations about artificial superintelligence tend to focus on the existential threat they may accidentally pose when in pursuit of a given goal (e.g., evaluating the Riemann hypothesis or producing paperclips).[179] Sensationalized fictions and speculations aside, and despite the nascent state of most artificially intelligent machines, the problem of implementing machine ethics is a live and pressing issue. As some scholars have noted, essentially all non-trivial interactions that an intelligent machine has with humans have ethical import.[180] It is possible that robots responsible for eldercare or childcare, for example, could cause harm via their inaction if they recharge their batteries at one particular point in time instead of another.

More concretely, autonomous[181] intelligent machines are currently in use in more areas than can be listed here, some of which include algorithms in hearing aids that filter out ambient noise, medical decision systems that can read CT scans and diagnose disease, automated facial recognition systems at border crossings, intelligent scheduling systems responsible for logistics planning and search engine optimization,[182] but all of which have an ethical dimension.[183] What counts as ambient noise? What recourse is available for people misidentified at a border crossing? What confidence does a system have in its medical diagnosis? Because the decisions and actions of automated intelligent machines already have the potential to significantly impact a person's well-being, coupled with the fact that such machines are only becoming more pervasive, there is a need to understand how ethics could be

[178] Consider the idiotic way in which Skynet from the *Terminator* series attempts wipe out humanity.
[179] Bostrom 2014, 123.
[180] Anderson, Anderson and Berenz 2017, 72.
[181] I will discuss autonomy in more detail later in the chapter.
[182] Bostrom 2014, 14-16.
[183] I am deliberately underspecifying what I mean by 'ethical dimension' here because I am interested in many different cases, only some of which I explore throughout this dissertation. Larger questions having to do with the choice to apply AI to certain topics/domains/questions are an ethical dimension I am interested in, but I do not take them up in much detail. More restricted questions on the other hand, having to do with how training data might have ethical assumptions built in or how outputs/recommendations of AI can fail to track ethically salient features of a given individual case are another ethical dimension I am interested in which I do take up in detail.

implemented in machines. In particular, my focus will be on how ethics could be implemented in machines that engage in action selection/execution as well as decision support systems that provide judgements for humans to act on.

The study of machine ethics is relatively new and has been largely theoretical and philosophical. The ethics of machines and artificial intelligence is generally divided into two main branches, the larger of which, called computer or technological ethics, focuses on how humans ought to act in order to minimize the ethical harms that can arise as a result of deploying intelligent machines.[184] Poor design, inappropriate application, or misuse are some of the ways humans could, through their actions, fail to minimize the ethical harms caused by the use of intelligent machines. In other words, computer ethics is primarily concerned with how human agents affect human patients via the use of computers and technology in general.[185] The second and smaller branch focuses on how machines themselves could behave ethically, hence it is generally referred to as the field of machine ethics. Indeed I will be using the term 'machine ethics' to refer to this latter field hereafter. Machine ethics, in contrast to computer ethics, is primarily concerned with how machine agents affect human patients. The machines of interest are therefore ones that can act in some sense "autonomously," a concept that will be discussed in detail. Some of the earliest proposals in the field of machine ethics concerned the use of practical ethical governors for lethal autonomous robots that would moderate or inhibit the robot's behaviour.[186] In this chapter, I will primarily focus on and summarize the smaller machine ethics branch in order to set the stage and contextualize the most recent research in the field of machine ethics. In the next chapter I argue that reinforcement learning methods are one of the most promising approaches to implementing machine ethics.

3.1 Autonomous Weapon Systems

The idea of machines acting independently of human supervision, i.e., autonomously, is not a new one. What is new is the advent of technologies that allow machines to in fact operate in the absence of human supervision. While certain types of autonomous machines permeate society, their

[184] Winfield et al. 2019, 510.
[185] The concepts of agency and patiency will be discussed in further detail later in the chapter.
[186] Ibid.

impact on human well-being has, for the most part, been negligible.[187] This, however, is no longer the case. The field of machine ethics has philosophical roots, as alluded to above, in the field that has traditionally been called computer ethics,[188] but also has philosophical and practical roots in considerations of ethical conduct during a war. Funded by the United States Army Research Office over a few years in the late 2000's, Ronald C. Arkin articulated one of the earliest practical proposals and accompanying motivations for implementing ethics in autonomous robotic systems capable of lethal force.[189] Concerned with designing and implementing an "artificial conscience" (i.e., not a real or human-like conscience),[190] Arkin notes that, like most new technologies, there is serious debate and discussion needed to understand what constitutes the ethical use of autonomous robotics in the battlefield during a conflict.[191] More importantly, given the fact that Arkin's focus is on *autonomous* systems capable of lethality, he is interested in understanding if it is possible for the system or machine itself to possess an ethical "conscience" such that when such machines are deployed they are as safe as possible to both combatant and noncombatant alike.[192]

3.1.1 Autonomy and Lethality

Although the domain of battlefield behaviour might seem like an odd place for discussions of machine ethics to originate, there are at least two good reasons for why this was the case. The first is that, unlike in other domains in which autonomous machines might (and do) operate, the operations of autonomous robots in the domain of battlefield behaviour significantly impact human well-being. Moreover, as Arkin points out, there already exist autonomous[193] military machines such as the "Phalanx system for Aegis-class cruisers in the Navy [an automated weapon control system], cruise missiles [which have built in guidance systems to direct the missile once deployed], or even anti-personnel mines [which, once deployed, do not require human oversight]," all of which are lethal

[187] Consider machines like automatically opening doors at the grocery store. They operate in the absence of a human supervisor and their operation (or perhaps more importantly, their failure) has a negligible impact on human well-being.
[188] More on this later.
[189] Arkin 2008.
[190] At least this is how I interpret Arkin's (2008) use of these quotation marks.
[191] Arkin 2008, 121.
[192] Ibid., 122.
[193] Autonomous here refers to the fact that these machines are not directly supervised by a human. Although a human is responsible for the deployment of these systems, a human does not help the machine "choose" a target and subsequently engage that target, as it were.

machines.[194] It is therefore a realistic possibility to consider how such machines might be made better, i.e., designed to bring about the most ethical outcome (e.g., a cruise missile guidance system that self-destructs if noncombatants are detected within the blast radius).[195] What is more interesting, and more important from a machine ethics point of view, is that the domain of battlefield behavior is one that is, at least in theory, governed by explicit rules such that certain behaviours are permissible and certain other behaviours explicitly prohibited.[196] Arkin highlights this fact.

> The Laws of War (LOW), encoded in protocols such as the Geneva Conventions and Rules of Engagement (ROE), prescribe what is and what is not acceptable in the battlefield in both a global (Standing ROE) and local (Supplemental ROE) context.[197]

The fact that the domain of battlefield behaviour is one with (let us assume) clearly defined rules is significant because there is the possibility that a machine could be built such that it simply abides by, i.e., does not violate, those rules when engaging in battle. Indeed it is still common for people to think of machines[198] as entities that merely execute code or that do what we humans tell them to (both of which are true in a certain sense). Machines might therefore behave ethically themselves because they never behave in ways that are explicitly forbidden, i.e., they consistently behave according to the rules (e.g., never harm a noncombatant).

To be clear, it is not that a battlefield machine behaved ethically *because* it was ordered to do something ethical. Rather it is that, when given vague orders (e.g., secure that landing zone, or scout ahead and eliminate any hostiles), machines will be unable to access a certain unethical action-space that corresponds to violations of the Rules of Engagement, for example. The important, non-trivial point is that unethical behaviour, in the context of battlefield machines, may be foreclosed in such a way that is simply not possible in the case of human soldiers. Hence the possibility that machines behave ethically themselves, though admittedly perhaps by default.

---

[194] Arkin 2008, 123.
[195] I am aware of the irony of discussing ethics in the context of autonomous weapon systems that are, by design, intended to harm the "enemy."
[196] Arkin 2008, 122.
[197] Ibid.
[198] Just note that I am using the term 'machine' in a broad sense throughout to refer to entities including, but not limited to, autonomous robots and purely algorithmic systems.

3.1.2 Machines (Can) Behave Better

Additionally, there is a third consideration that helps to explain why discussions of machine ethics originated in the domain of battlefield behaviour. There are certain *prima facie* reasons to suspect that autonomous battlefield machines might behave more ethically than their human counterparts. One obvious reason for this suspicion is that autonomous battlefield machines can be designed such that they have no sense (or perhaps a minimal sense) of self-preservation. In the context of ethical battlefield behaviour, an autonomous battlefield machine that is unsure whether it is targeting an enemy combatant or civilian could simply act conservatively and, for example, get closer to the target to ascertain its status. So while a human soldier in this same position might feel compelled by self-preservation to attack, a machine with no such compulsion would refrain from engaging in potentially unethical behaviour of this sort.

Another *prima facie* reason supporting the idea that autonomous battlefield machines might behave more ethically is the fact that such machines can be designed without emotions. As Arkin highlights, human battlefield behaviour is disconcerting at best and more often than not, appalling.[199] Research suggests that emotions, anger in particular, are one factor responsible for the persistent unethical behaviour of human soldiers.[200] The tendency for humans to seek revenge when allies are killed and to mistreat noncombatants when angry is not a tendency that would necessarily be shared by autonomous battlefield machines. Indeed beyond the ability to act conservatively and unemotionally, autonomous battlefield machines would be immune to the human tendency to dehumanize the enemy and would also never forget or disregard their "training."[201]

As a side note, it is worth mentioning that there is the distinct possibility that machines, because they might be emotionless and unconcerned with self-preservation, would incentivize humans to engage in more wars using robotic soldiers. Politically, because humans would (presumably) not be risking their lives, there might be less resistance to war using machines. Tactically and, relatedly, ethically speaking, wars fought with machines might even be more "brutal" because machines could engage in riskier and more belligerent behaviour.

All of this is to say that as machines become both more autonomous and capable of impacting human well-being, they will also need to engage in ethical decision making and behaviour *on their own*.

---

[199] Arkin 2008, 124.
[200] Ibid.
[201] The way in which autonomous battlefield machines would be "trained" to engage in battle would likely be far removed from the kind of training human soldiers use.

No where is this more apparent than in the domain of autonomous weapons systems, as I have highlighted in the previous sections. But of course, there are less explosive ways that increasingly autonomous machines can impact human well-being. Machines are increasingly used to filter spam emails, detect credit card fraud (and freeze credit card activity), determine insurance or loan qualifications and assign credit scores.[202] It is just as important that these machines also engage in ethical decision making and behaviour, if such behaviour on the part of machines is even possible. I turn now to the connection between computer ethics and machine ethics.

3.2 From Computer Ethics to Machine Ethics

While some of the earliest practical implementations of machine ethics originated in the domain of battlefield behaviour, machine ethics also has its theoretical origins in the general field commonly known as computer ethics. As mentioned above, the field of computer ethics (or synonymously, technological ethics) focuses on how humans ought to act in order to minimize the ethical harms, to other humans (or moral patients), that can arise from the use of different technologies. In his oft cited paper *The Nature, Importance, and Difficulty of Machine Ethics*, Moor points out that computing technology can be evaluated in terms of ethical norms in addition to, for example, design norms (i.e., whether the machine is functioning as it is intended).[203] In addition to computing technology, all machines can be evaluated in terms of ethical norms because all machines can have an impact, even if it is a small one, on the well-being of a human. Building on this idea, Moor outlines a spectrum of the kinds of ethical impacts machines can have.

3.2.1 Computer Ethics

At one end of this spectrum, the end closer to the machines that have no autonomy, or very little autonomy, is what Moor calls ethical-impact agents.[204] As the name suggests, ethical-impact agents are those kinds of technologies that, through their use, can make our lives better, worse or some combination thereof. Inevitably, the use of machines intersects with ethical issues. As Moor highlights, computing technology has allowed us to "conduct business online easily, but we're more vulnerable to

[202] Burrell 2016, 1.
[203] Moor 2006, 19.
[204] Ibid.

identity theft" by using such technology.[205] Similarly, the use of certain technologies, like robotic camel jockeys in Qatar, can have important ethical impacts, like freeing Sudanese boys from their enslavement as camel jockeys.[206] Nevertheless, these incidental impacts arise as a result of our use of machines and are more commonly associated with the general field of computer ethics.[207]

3.2.2 Machine Ethics

While the field of machine ethics does concern itself with ethical-impact agents, their proper place is in the field of computer ethics. The primary focus of machine ethics, in contrast to the field of computer ethics, is whether it is possible to put ethics into a machine, as it were. Acting with some degree of autonomy, on the part of the machine, is therefore a defining feature of the kinds of machines of interest in the field of machine ethics. This brings us to Moor's second kind of ethical machine, implicit ethical agents. Perhaps the easiest and simplest way to implement machine ethics is to "constrain the machine's actions to avoid unethical outcomes."[208] It is not hard to imagine creating implicitly ethical autonomous battlefield machines whose "internal functions implicitly promote ethical behaviour - or at least avoid unethical behaviour," because such machines could be programmed to abide by the Laws of War and Rules of Engagement.[209] Note that the term 'abides' here conflates two senses of "rules following." A machine may abide by the rules in cases where it behaves according to the rules (i.e., the machine is rule following, as in its behaviour can be described as following some rule(s) like the Laws of War). But a machine may also abide by the rules in cases where it generates its behaviour by consulting the rules (i.e., the machine is rule governed).[210] Setting aside autonomous battlefield machines, it is quite common for most machines to be designed as implicitly ethical machines that avoid engaging in unethical behaviour. For example, it is part of the design of autopilot systems, as a result of value laden choices in how autopilot systems are built, to avoid engaging in certain mid-air maneuvers that would frighten passengers despite executing a directive to fly the plane from one destination to another. Online banking systems are similarly designed to be implicitly ethical. Just as

[205] Ibid.
[206] Ibid.
[207] Other examples might include the use of cars (they make it easier to travel, but also come with certain risks to drivers and pedestrians) and the use of smartphones (they make it easier to connect with people, but also tether us to our work).
[208] Moor 2006, 19.
[209] Ibid.
[210] This is the sense of "abiding by the rules" used earlier in section 3.1.1.

there is an ethical component to travelling by airplane, transactions involving money also have ethical import, and it is for this reason that online banking systems are designed so that they avoid engaging in unethical behaviour (e.g., they avoid transferring a sum of money that the user did not previously confirm). Implicitly ethical machines are ethical in the minimal sense that their architecture, i.e., how they are designed or created, prevents, as much as possible, unethical outcomes from obtaining.

This is in contrast to Moor's third kind of ethical machine which he dubs "explicit ethical machines." As the name suggests, these are machines that are able to "represent ethics explicitly and then operate effectively on the basis of [that] knowledge" in much the same way that traditional GOFAI (Good Old Fashioned Artificial Intelligence) systems can play chess.[211] These machines operate in a rule governed way as their behaviour is generated by consulting explicit rules. Deep Blue, the chess playing machine that beat Garry Kasparov, is a well-known and paradigmatic example of an explicit, rule governed, chess-playing machine because it utilizes symbolic representations of the board state alongside representations of the legal moves to calculate a next best move. Analogously, one might create an ethically-behaving machine, as it were, if one were able to represent within the machine its external environment alongside rules representing permissible and impermissible behaviours such that the machine could calculate the optimal ethical action to take in its current situation. While Moor notes in 2006 that examples of explicitly ethical machines are elusive, that is certainly not the case in 2022. Tolmeijer et al. for example note in their survey of implementations of machine ethics that over 50% of the 48 approaches considered utilize some kind of explicit ethical reasoning. Moreover, explicit ethical machines can indeed make plausible ethical judgments and "justify" them in a certain sense. Take, for example, the robot developed by Winfield et al., discussed in Chapter 2 section 2.4.2, which makes real-time decisions in order to ensure that the ethically preferable outcome is attained. Their research has demonstrated that the creation of a minimally ethical robot does appear possible via the use of explicit ethical injunctions that causes the robot to "choose" to bump into a "human" and thereby save it from falling into a "hole" in the ground.[212]

While progress on the implementation of explicit ethical machines has been impressive, this progress has been restricted to theory and research laboratories. Further, research on implementing machine ethics has revealed multiple dimensions and taxonomies with which one might classify different approaches to implementing machine ethics. Certain approaches to implementing machine

[211] Ibid., 20.

[212] In their experiments, Winfield et al. (2014) use different robots to stand in as proxy humans when observing how their ethical robot behaves as it attempts to save these "humans" from falling into a virtual hole in the ground.

ethics focus on replicating normative ethical theories including versions of consequentialism, deontology or some combination thereof.[213] Other approaches to implementing machine ethics are better understood in terms of the way in which ethical knowledge is acquired by a machine, i.e., in a top-down, bottom-up, or hybrid fashion.[214] Still others are better classified in terms of their output or what they do. Ethical machines can engage in action selection/execution, can act as decision support systems (DSS) by recommending certain courses of action, can attempt to appropriately represent aspects of ethical decision making and can engage in a kind of principal component analysis to select the most fitting elements, given a set of alternative options to implement an ethical machine, to include in the final system.[215] While it is certainly helpful to think of certain approaches to implementing machine ethics in terms of explicitly ethical machines, such a concept is not sufficient to fully capture contemporary approaches to implementing machine ethics. This is especially the case for learning machines.

3.2.3 Fully Ethical

Finally, at the other end of the spectrum are "full ethical agents" which are machines that would essentially be human-like in their ability to engage in ethical reasoning and behaviour. In contrast to ethical-impact agents which must be used by a human in order to have an impact on human well-being, machines that possess the status of full ethical agent would presumably have the largest potential impact on human well-being because they would be as autonomous or "free" as any person.[216] According to Moor, such machines would be able to "make explicit ethical judgments" and would generally be "competent to reasonably justify them," abilities that many philosophers argue stem from an agent's possession of, for example, consciousness, intentionality and autonomy.[217] This is why Moor maintains that an average adult human is an example of a full ethical agent, why there are currently no machines that can be considered to be full ethical agents, and why there continues to be heated debates over the possibility of a machine ever becoming a full ethical agent.[218] It is to the details of this

[213] Tolmeijer et al. 2020, 18.
[214] These approaches to implementing machine ethics will be discussed in much more detail in the following chapter.
[215] Tolmeijer et al. 2020, 10.
[216] Machines that possess the status of full ethical agent may even place burdens on us to, for example, respect their rights. More on this in Chapter 5.
[217] Moor 2006, 20.
[218] Ibid.

debate that I now turn. My position is that there may not be as large a gap between explicit ethical agents and full ethical agents as Moor suggests, and that the best way to approach implementing machine ethics is by "raising" machines such that they exhibit ethical decision-making and behaviour. But more on this in Chapter 4.

3.3 The Prerequisites for Ethics

When thinking of machine ethics, i.e., how machines themselves might behave ethically, it is natural to first think about the relevant features that confer moral status to human beings. Importantly, it must be noted that what follows is not intended to be a comprehensive survey of the field of moral philosophy or philosophical ethics. Rather, it is my aim in what follows to highlight and untangle important concepts connecting philosophical ethics and machine ethics.

3.3.1 Agency and Patiency

To begin, there are the concepts of moral agency and moral patiency.[219] To be considered a moral agent an entity must possess certain features, as mentioned above, namely features like intentionality, consciousness and free will. Moreover, as a result of these features, moral agents often possess certain responsibilities. For example, because I am free to choose to hurt another human being, I would be held morally culpable should I choose to do so. More generally, and when considering a typical adult human, "because of our ability knowingly to act in compliance with, or in violation of, moral norms we are held responsible for our actions (or failures to act)."[220] Moral patients, in contrast, are those entities which moral agents have responsibilities towards. So a typical adult human is an example of both a moral agent and a moral patient. Not only do we have certain moral responsibilities, but "we have rights, our interests are usually thought to matter, and ethicists agree we should not be wronged or harmed without reasonable justifications."[221] While moral agents are also simultaneously moral patients, not all moral patients can be considered moral agents. Babies and animals, for example, are typically regarded as moral patients, i.e., as having rights and interests that matter, but not as moral agents. This is, in part, because they lack the sort of autonomy that is necessary to confer moral agency

[219] See, for example, Alfano (2016) and Coeckelbergh (2020).
[220] Cave et al. 2019, 570.
[221] Ibid.

to an entity. It is an open question whether machines lack the sort of autonomy necessary to ascribe them moral agency.

3.3.2 Autonomy

Indeed, autonomy is another important concept connecting philosophical ethics and machine ethics. As mentioned, the ability to act freely or autonomously appears to be a necessary condition to ascribe moral agency to an entity. It is well known that ethics presupposes some kind of free will or autonomy on the part of the entity in question. As Helen Beebee explains, it would be odd, to say the least, to heap moral praise or blame on people (or some other entity) if they did not freely choose to act in a praiseworthy or blameworthy manner.

> It's easy to see why acting freely looks like a plausible requirement on moral responsibility. Imagine, for example, that it turns out that the reason your friend declined your invitation [to your birthday party] was that she was coerced into doing it: some deranged enemy of yours, hell-bent on ensuring that your party is a failure, had made it clear that if she were to accept, there would be terrible repercussions for her and all her family. In that case, it would certainly be inappropriate for you to resent her for declining the invitation, and it would be inappropriate because she didn't decline freely. We might put this in other words by saying that she didn't really have a *choice* about whether to decline.[222]

In short, if an entity does not really have any ability to choose or decide for itself how to act in a given situation, it is doubtful that that entity possesses moral agency. Indeed the concept of agency itself presupposes that an entity possesses some kind of autonomy.[223] A coffee maker that fills the coffee pot with too much coffee causing it to spill all over the floor is neither morally responsible for its actions nor should it even be regarded as an agent (ethically speaking at any rate). The coffee maker has absolutely no ability to choose to act in one way or another. It is important to note however that autonomy is not a binary all or nothing property but a spectrum. An entity may possess more or less autonomy in comparison to another entity.[224] Further, autonomy is an ambiguous or at least pluralistic concept, the meaning of which often varies depending on the field of study. In philosophical ethics, for example,

---

[222] Beebee 2013, 1-2.

[223] Floridi and Sanders (2004) for example argue that an entity can be considered an agent if it meets three criteria: 1) interactivity, 2) autonomy and 3) adaptability, where autonomy is understood roughly in terms of decoupledness (or independence) from the environment.

[224] The robot Spot developed by the company Boston Dynamics, for example, is more autonomous than a coffee maker, but less autonomous than an elephant or human.

autonomy often refers to the ability to act independently with some degree of control over one's life.[225] When considering machines or robots, autonomy often refers to the ability of a machine to operate in real-world environments without external control or human supervision for an extended period of time.[226] While I will not delve any further into a discussion of autonomy here, it should be abundantly clear that for an entity, human or machine, to be considered a moral agent, that entity must possess some degree of autonomy. We will revisit autonomy in section 3.4.2 as I will argue that the kind of autonomy machines possess is sufficiently strong to ground the claim that machines ought to be considered moral agents, i.e., capable of engaging in ethical decision-making and behaviour.

3.3.3 Mind

In addition to autonomy, the concept of agency, at least the thick concept of agency sometimes preferred by philosophers in the context of moral philosophy, also appears to presuppose that an entity possesses a mind. Indeed, one might reasonably argue that it is *because* human beings have minds (or the right kind of mind) that we ought to be regarded as both moral agents and moral patients (or, in Moor's terminology, as full ethical agents). Similarly, it is not just that babies and animals qualify as moral patients because they have rights and interests that matter (as mentioned above), it is that it follows from the fact that babies and animals have certain kinds of minds that they therefore have rights and interests that matter. The same will almost certainly need to be said of machines. That is, before a machine can be said to possess moral patiency or moral agency it must possess the right kind of mind (or it must be perceived by humans as possessing the right kind of mind). Moreover, it is debatable whether machines should even be described as agents at all if they lack the right kind of mind (more on this in section 3.4.1). While machines can certainly possess some degree of autonomy, standard accounts of agency typically require the capacity for intentional action.[227] This is important to recognize because intentionality has traditionally been a defining mark of the mental.[228] More concretely, intentionality can be understood as the aboutness, ofness or directedness that accompanies some,[229] if

[225] See, for example, Kazez (2007).
[226] Bekey 2012, 18.
[227] Cave et al. 2019, 565.
[228] See, for example, Brentano (1995).
[229] Searle (1983) argues that intentionality certainly characterizes mental phenomena like beliefs, hopes and desires, but does not characterize all mental phenomena, like certain forms of nervousness or elation, for example, that may not be about or directed towards anything.

not all, mental phenomena.[230] My beliefs, judgments, propositional attitude states and perhaps even qualitative mental phenomena (i.e., qualia), can all be characterized as having directedness towards some entity, for example, a red traffic light at an intersection. So a typical human acts intentionally when their actions are caused by their intentional mental states. These actions can be distinguished from mere behaviours or reflexes that do not require intentional mental states. Importantly, I am tying together two different senses of intentionality here, namely the sense of intentionality discussed in the previous chapter (i.e., Brentano's idea of intentionality as directedness) and intentionality as acting deliberately.

3.4 The Prerequisites for Machine Ethics

To sum up the previous section, we ascribe moral agency to a typical human adult because they possess a mind and autonomy (or the right sort of mind and the right sort of autonomy). Moreover, to say that an entity is a moral agent entails that that entity is also a moral patient, although the converse is not always the case (moral patiency does not entail moral agency). So a typical adult human is not merely a moral agent, they are also a moral patient. But these terms described in the previous section can be thought of in either a strong or a weak sense, and in this section I will describe how machines possess these features, albeit in a weaker sense. More importantly, the importance of machine ethics and the viability of implementing machine ethics is independent of these philosophical considerations.

3.4.1 Machine Minds

Take the concept of mind to start. While it is obvious that machines do not yet possess minds in the strong sense of being able to act intentionally, there is nevertheless a weaker sense which allows for more straightforward ascriptions of mentality. AlphaZero for example, DeepMind's board game playing machine, as impressive as it is, likely does not possess phenomenal consciousness or an understanding of the things it does and so lacks a mind in the strong sense. However in a weaker pragmatic or "instrumentalist" sense, ascribing intentionality/mentality to AlphaZero is a useful way to explain its behaviour.[231] It makes perfect sense to describe AlphaZero as making a particular move in a game of chess because it *believed* that that move was the best one to make, as long as "belief" is understood in

[230] Crane (1998) argues that intentionality characterizes all mental phenomena, including qualia.
[231] Silver et al. 2018.

this weaker sense, because such a description (or explanation) is more useful than a description that invokes neural networks, weighted connections between artificial neurons, reinforcement learning, etc. This is akin to, though not entirely the same as, adopting Dennett's "intentional stance" (rather than a physical or design stance) because "the only strategy that is at all practical is the intentional strategy [i.e., attributing beliefs and desires to a system]; it gives us predictive power we can get by no other method."[232] Importantly, the possession of a mind, albeit in a weak sense (which is not necessarily the same sense as Dennett's intentional stance), is sufficient for machines to be thought of as engaging in ethical decision-making and behaviour. But more on that in the next section (3.5).

3.4.2 Machine Autonomy

The concept of autonomy can also be understood in a stronger and weaker sense.[233] On a strong interpretation, as outlined above, autonomy is an entity's ability to control or decide how to act for itself in a given situation. Machines similarly do not yet possess autonomy in this strong sense, but they do possess some weaker form of autonomy. Consider AlphaZero again and how it is able to operate without external control or human supervision. Within a game of chess, AlphaZero can clearly act autonomously and it may even be appropriate to maintain that as a result of its autonomy AlphaZero is responsible for the chess moves that it makes; it is the genuine author of those choices and behaviours.[234] That being said, AlphaZero's autonomy begins and ends with the games it plays. It cannot choose *not* to play a game of chess or to learn some other board game. Nonetheless, the possession of autonomy, albeit in a weak sense, is sufficient for machines to be thought of as engaging in ethical decision-making and behaviour (and again, more on this in the next section).

3.4.3 Machine Agents and Machine Patients

Agency and patiency can similarly be understood in a stronger and weaker sense. Indeed if it is *because* an entity possesses a certain kind of mind and/or autonomy that we ascribe that entity agency, then it follows that machines possess a kind of weak agency stemming from their possession of a mind and/or autonomy (understood in the weak sense). More importantly, this weak sense of agency can be

[232] Dennett 1989, 23.
[233] Indeed, as mentioned in the previous section, autonomy is best understood as a spectrum.
[234] Silver et al. 2018.

independently grounded by the fact that the behaviour of machines has real consequences on human beings. In other words, machines can *act on* human beings (in addition to animals and other machines) and it is this ability, perhaps even more so than their weak possession of a mind and autonomy, that warrants us as thinking of certain machines as weakly possessing agency. As Mark Alfano notes, "things don't just happen to people: sometimes people do things," and now sometimes machines do things too.[235] Similarly, and again independently of questions surrounding mind and autonomy, machines can be seen weakly as patients because they can be *acted upon* by humans (in addition to other machines). This is one jumping-off point to discussions, which I will not be getting into here, concerning robot or AI rights and whether certain machines ought to be treated in certain ways (e.g., would you be allowed to "unplug" an intelligent machine in the same way you might unplug your coffee machine?).[236] It is just a short step from thinking about machines as agents and patients, i.e., as entities whose actions can impact humans and which can also be impacted by human actions, to thinking about machines as *ethical* agents and patients. In short, considering whether machines themselves could behave ethically is not necessarily to imply that machines possess moral agency in the strong sense. It is sufficient to consider machine ethics in the context of machines that are able to affect humans in morally significant ways as a result of their behaviours.

3.5 Updating and Redefining Machine Ethics

While Moor outlines a helpful set of categories, better thought of as a spectrum, to understand the complex and interdisciplinary field of machine ethics, his taxonomy stands in need of an update. This is especially because in what follows I will be developing my own original position on machine ethics. In particular I will be focusing on how machines themselves might learn to behave ethically. Therefore by 'machine ethics' I am not specifically referring to Moor's taxonomy nor am I referring to the field of computer/technological ethics that is concerned with how humans use machines. Rather, by 'machine ethics,' I am referring to the field of research that is focused on the machines themselves and how they, as agents, ought to be created such that they behave ethically when interacting with human patients.[237] Before discussing my own position in Chapter 4 however, it is necessary to review some updates to Moor's taxonomy of machine ethics.

---

[235] Alfano 2016, 4.
[236] See, for example, Gunkel (2018).
[237] I refer the reader to sections 3.0 and 3.1 for more details.

3.5.1 Ethical Alignment

In keeping with Cave et al., I believe it is more fruitful to distinguish between machines that possess moral agency (in either the weak or strong sense), machines that possess the ability to engage in ethical reasoning and machines that align with what people consider to be ethically desirable or acceptable (though these are not mutually exclusive). ATMs that do not defraud users and cars with automatic braking features are machines of the latter kind insofar as they are machines "whose behaviour adequately preserves, and ideally furthers, the interests and values of the relevant stakeholders in a given context."[238] While this roughly maps onto what Moor described as ethical-impact agents and implicit ethical agents, there is considerably less ambiguity when thinking about machines in terms of their "ethical alignment." This is because an ethical machine might be described as such just in case its behaviour is the result of its alignment with the relevant ethical principle(s). Furthermore, an ethical machine might be described as such just in case its behaviour is aligned with the desires of the relevant stakeholders. Consider again automatic braking features. Strictly speaking, such features are amoral in the sense that the car itself is unconcerned with the rightness or wrongness of its actions when engaging in automatic braking. Nevertheless, it is apt to describe cars with automatic braking features as machines whose behaviour is ethically aligned with what people consider to be ethically desirable.

Given that the term 'agency,' more often than not, evokes the strong realist sense of the concept described above, i.e., as requiring an entity to act intentionally and with comprehension of the consequences of its actions, I will generally avoid the term 'agent' despite the common usage of terms like 'artificial agent' throughout the machine ethics literature.[239] Although I think it is perfectly acceptable to think of machines as agents in the weak sense outlined above, I am inclined to avoid using the term 'agent' whenever possible in order to avoid igniting philosophical discussions about whether machines can or will ever be considered agents in the strong sense. As Cave et al. note, "whether machines can be called ethical agents in any strong sense is a contentious philosophical issue," and I have no desire to throw more fuel on that fire than is absolutely necessary.[240] Therefore in what follows I will generally be using terms like 'machine' or 'ethical machine' as substitutes for terms like 'artificial

[238] Cave et al. 2019, 564.
[239] Wallach and Allen for example use the abbreviation AMA for artificial moral agents through their book *Moral Machines* (2009).
[240] Cave et al. 2019, 565.

agent' or 'ethical artificial agent' (or 'artificial moral agent') that are common throughout the machine ethics literature.

3.5.2 Ethical Reasoning

If ethically aligned machines map roughly onto ethical-impact agents and implicit ethical agents, then machines that possess moral agency, as outlined by Cave et al., map roughly onto what Moor described as full ethical agents. My focus however is primarily on machines that can engage in ethical reasoning, which roughly maps onto Moor's explicit ethical agents. In fact, it may be more appropriate to entirely disregard the idea of explicit ethical agents given that Moor describes these machines as ones that can engage in ethics in the same way that GOFAI, i.e., good old-fashioned symbol manipulating AI systems can engage in a game of chess.[241] The reality is that contemporary AI systems simply do not function in the same way that traditional symbol manipulating AIs did. We have entered the second wave of AI systems which is made possible by big data and machine learning techniques.[242] In this second wave of AI I maintain that it is better, as Cave et al. suggest, to conceive of certain machines in terms of their ability to engage in ethical reasoning as opposed to the way Moor describes explicit ethical agents. It bears repeating that these are the machines I wish to focus on, machines that can learn to engage in ethical decision-making and behaviour.

Note, however, these two important qualifications. The first is that "reasoning" here is used in a broad sense to simply refer to the information processing that is carried out to reach a conclusion, including "reckoning" and "judgment" discussed in the introduction. The second is that "ethical reasoning" here, in addition to the concepts mentioned previously (e.g., agency, patiency, autonomy and mind), is being used in a weak sense that does not require, for example, an understanding of the significance of the ethical issues at stake. Intelligent machines are simply not at the point where they might be considered full ethical "agents" in the strong sense, i.e., actors with certain mental features (e.g., having intentions and self-awareness) and moral responsibilities, among other attributes. To be clear, there is no underlying assumption in what follows that in order to implement machine ethics (in the weak sense that all of the relevant concepts outlined above are being used) machines must possess moral agency. This however does not negate the importance of machine ethics and furthermore should

[241] See section 3.2.2 but especially Chapter 6 for more details.
[242] More will be said about the difference between first and second wave AI in Chapter 6.

not deter us from seriously considering how machines themselves might behave ethically and examining the different ways in which researchers are attempting to implement machine ethics.

3.6 Metaphysics, Psychology and the Moral Turing Test

As the previous sections highlight, the intersection of ethics and intelligent machines is dense, nuanced and can become emotionally charged as a result of the conclusions one might draw about both humans and machines (e.g., humans just *are* machines, albeit complicated biological machines, or there is nothing unique about humans that sets us categorically apart from machines). While many are sympathetic to the project of implementing machine ethics, there are just as many people opposed to or dubious of the prospects of creating ethical machines. It has been argued for example that machines cannot be held accountable or responsible for their actions[243] in which case it would be a mistake to focus on implementing machine ethics rather than developing better frameworks through which we might hold humans responsible for their machine creations.[244] It has also been argued that in addition to the features[245] discussed above, there is some other x-factor that machines currently lack, or will never possess, that preclude the possibility of seriously discussing machine ethics and the possibility of machines themselves behaving ethically.[246] Possible x-factors include, but are not limited to, phenomenal consciousness, emotions, sociability and semantic understanding, all of which are features that might be taken into consideration when analyzing human moral decision making.[247] Unfortunately, whether machines do or will ever come to possess these additional features is simply outside the scope of this chapter, and deliberately so.

Similarly, a detailed discussion of research being conducted on the empirical psychology of machines, i.e., how humans feel about machines (e.g., machines driving cars, pointing out emergency exits, hitchhiking across the country, etc.), is also outside the scope of this chapter. Nevertheless, some reference to this research is important. For example, although in what follows in Chapter 4 the focus will be on *how* machine ethics could be implemented, it is worth asking whether the creation of ethical machines *should* be pursued at all. Recent research has highlighted that although people prefer that

---

[243] Bryson 2020, 14 &15.
[244] Alternatively, we could restrict our use of machines to low stakes tasks that have little or no ethical import. Scott Robbins advocates for this type of "boring AI."
[245] Or properties, attributes, faculties, etc.
[246] See Chapter 1 for an explicit description of this Argument from Various Disabilities.
[247] Allen and Wallach 2012, 60.

others buy self-driving cars whose behaviour results in the best global outcomes, e.g., injuring the passenger to save multiple pedestrians, they would themselves prefer to buy self-driving cars that protect the passenger at all costs.[248] Indeed, research by Bigman and Gray suggests that people are generally averse to having machines make moral decisions in the domains of driving, legal, medical and military decision-making. They suggest that this aversion is driven in part by a perceived lack of "agency," i.e., the ability to carry out one's intentions, in the machine.[249]

Of course, human beings are by no means the paragon of moral excellence and there is no shortage of literature demonstrating that humans are, to put it bluntly, horrible moral judges. One might rightly worry that we should not implement machine ethics especially if machines learn to behave ethically by mining data generated by humans.[250] Empirical moral psychology has revealed how different cognitive biases and framing effects influence the decisions, including decisions with a significant ethical component, people make. For example, implicit intergroup bias causes employment recruiters to prefer native candidates (i.e., candidates from the same group) to equally qualified foreign candidates[251] and it is well documented that people hold others to different moral standards than themselves even if they were in the same situation (the actor-observer effect).[252] If the only path to implementing machine ethics is for machines to extrapolate from human generated data, then the creation of ethical machines should not be pursued. Luckily, this is not the only path and there are also solutions to mitigate the risks associated with training machines from human generated data. One way to avoid using human generated data is to utilize simulated environments in which a machine might learn for itself, e.g., via self-play, how best to attain a given outcome. This strategy has an impressive track record in the domain of game playing (e.g., DeepMind's AlphaZero and AlphaStar, which will be discussed in more depth in Chapter 5) where learning machines have come up with novel and better strategies than humans. Moreover, we might just learn something about ourselves too by implementing machine ethics in this way. But more on that later.

Even if people are generally averse to having machines make moral decisions in certain domains, there may be significant costs for choosing to ignore machine decisions or predictions. It has been demonstrated that even simple[253] linear models outperform human experts, i.e., are more

[248] Bonnefon, Shariff and Rahwan 2016, 1574.
[249] Bigman and Gray 2018, 22.
[250] An issue that will be taken up in Chapter 5.
[251] See Jost et al. 2009.
[252] See, for example, Nadelhoffer and Feltz (2008).
[253] It is arguable whether linear models are any "simpler" than other models. See, Lipton (2017), for more details.

accurate, in domains such as clinical diagnoses and forecasting graduate students' success.[254] Further studies have revealed that it is far more common for algorithms to outperform human judges and forecasters than the opposite.[255] In spite of this fact, there is a tendency for humans to punish machines more severely after seeing them err, a result that suggests people are more forgiving of humans than machines.[256] As mentioned, this bias against machines can be costly for individuals and society at large and may contribute to a slower uptake and usage of machines that are, on average, superior than humans. Importantly however, this effect of abandoning machine judgment in favour of human judgment, a phenomenon known as "algorithm avoidance," was only observed when people had repeated interactions with a machine and observed it err. The opposite phenomenon, dubbed "algorithm appreciation," is observed when people are asked to choose between modifying their responses based on algorithmic or human advice. That is, in a wide range of estimation and forecasting tasks, people actually prefer advice from machines to advice from humans.[257] Algorithm appreciation was even observed to occur in cases when algorithmic advice was pitted against a person's own judgment, a rather surprising result given the fact that individuals routinely report excessive confidence in their own judgment relative to their peers.[258] Furthermore, as Logg et al. point out, while it is "important to understand how people react to the performance of human and algorithmic advisors, many consequential decisions are made without the benefit of performance feedback," and so it may be the case that algorithm appreciation arises more often than algorithm avoidance, despite the prevalence of anecdotes that support the latter.

When thinking about implementing machine ethics and machines making decisions with ethical ramifications, there is the further problem of assessing just how ethical the machine is. Is the autonomous vehicle, for example, making decisions that are ethical enough most of the time? There have been some proposals for a moral Turing Test (MTT), as it were, that might play a similar role in the field of machine ethics that the actual Turing Test plays in the field of artificial intelligence research.[259] There are certainly advantages to using a MTT to assess machines. For example, such a test would bypass certain philosophical debates[260] and give researchers something concrete to aim for. This was, after all, why Alan Turing proposed the Imitation Game (now known as the Turing Test), i.e., to bypass

[254] Dietvorst et al. 2015, 114.
[255] Ibid., 114.
[256] Ibid., 124.
[257] Logg et al. 2019, 90.
[258] Ibid, 91 & 99.
[259] Wallach and Allen 2009, 70.
[260] Like the ones canvassed in this chapter about the prerequisites for moral agency.

debates about the nature of intelligence and give researchers in the budding field of artificial intelligence something to work towards. As with the regular Turing Test however, a MTT is, in a sense, highly anthropocentric. The morality of a machine assessed using a MTT might encourage the development of machines that merely mimic human decision making and behaviour instead of improving upon it. If the interrogator's task in a MTT is to, for example, identify which participant is the human and which is the machine, it could be the case that "humans happen to be recognizable because they often act less ethically than they should," in which case the machine will have "failed" to pass the MTT.[261] To remedy this problem, Wallach and Allen suggest asking a slightly different question. Instead of asking, "Can you spot the machine?" it might be worth asking, "Which participant is less moral than the other?"[262] Dubbed the comparative MTT (cMTT), such a test highlights the fact that what matters when assessing how ethical a machine is not whether it can masquerade as a human, but whether a jury of one's peers, as it were, would consistently judge it as being adequately ethical. Indeed, the cMTT nicely draws out something that has been largely implicit throughout this chapter, namely the idea that the important question is the comparative one. Do machines have whatever it is that humans have, e.g., a mind, autonomy or agency? The unimportant and uninteresting questions in the context of my research, the ones I have been deliberately setting aside, are those having to do with the metaphysical status of machines, e.g., do machines have certain underlying metaphysical features?

This is not to say that metaphysics is unimportant, only that it is not necessary to wade too deeply into metaphysical waters to take machine ethics seriously given that the Great Flood is well underway.[263] Less metaphorically, *we are already in a position to take machine ethics seriously*. Machines are now making decisions that affect either directly (via action selection/execution) or indirectly (via judgment provision) human well-being, and "it would be a good thing if those decisions are morally sound."[264] My primary concern is how that might happen, i.e., *how can we raise machines such that they make ethical decisions and take ethical actions*.

[261] Wallach and Allen 2009, 70.
[262] Ibid., 70.
[263] See Chapter 1 for more details on Hans Moravec's description of computing technology as a Great Flood.
[264] Coeckelbergh 2020, 50.

3.7 Looking Ahead

As mentioned above, while it is certainly worth asking whether machines possess that (metaphysical) *thing* that warrants ascribing intelligence or agency or moral status to machines, I maintain that the interesting question is the comparative one, i.e., can machines possess what humans possess. As I have argued throughout this chapter, machines *do* in fact possess relevantly similar features; enough to warrant thinking about implementing machine ethics seriously. It should be noted that many of the issues raised in this chapter will be revisited in later chapters. For example, there is much more to be said about the risks associated with implementing machine ethics and so it is worth examining in more detail whether ethics *should* be implemented in machines at all (Chapter 5). Similarly, assessing the ethicality of machines, the importance of emphasizing the comparative question (i.e., whether machines have whatever it is that humans have) and the benefits of attempting to implement machine ethics, will all be discussed further in upcoming chapters (Chapter 6). However, now that I have, in this chapter, highlighted some of the important issues connected to machine ethics, I turn to examine and focus specifically on just how ethics might be implemented in machines.

Chapter 4 - Implementing Machine Ethics

4.0 Implementing Machine Ethics

While there are still relatively few concrete examples of machine ethics in the sense define in the previous chapter,[265] i.e., there are few practical demonstrations concerning how one might implement machine ethics, there is no shortage of theoretical and philosophical suggestions for implementing machine ethics.[266] It is my aim in this chapter to contribute to the philosophical discussion by arguing that reinforcement learning methods, in contrast to other machine learning techniques, are one of the most promising approaches to implementing machine ethics. This chapter is therefore primarily focused on the different ways in which ethical knowledge can be acquired by a machine.[267] I begin by outlining top-down approaches to implementing machine ethics and their shortcomings. I will then consider bottom-up approaches in general before focusing specifically on reinforcement learning methods. In contrast to other bottom-up machine learning techniques, e.g., supervised and unsupervised learning techniques,[268] I maintain that reinforcement learning methods possess certain advantages that make them uniquely suited for raising machines to make ethical decisions and take ethical actions.

4.1 Top-Down Approaches

While it is uncommon to implement Asimovian type laws,[269] Asimov's Three Laws[270] are a paradigmatic top-down approach to implementing machine ethics. In short, top-down approaches involve specifying some set of rules that promise comprehensive solutions to any ethical problem (i.e., we start from some general rule(s) and derive the specific action(s)).[271] Common top-down ethical

[265] See section 3.2 in Chapter 3 for more details.
[266] See, for example, Tolmeijer et al. (2020).
[267] See section 3.2.2 in Chapter 3 for more details on the different taxonomies one might use to classify different approaches to implementing machine ethics.
[268] See Chapter 1 for more details on these machine learning techniques.
[269] Tolmeijer et al. (2020) for example only highlight that three out of thirty top-down approaches to implementing machine ethics surveyed used Asimovian type laws.
[270] See sections 2.0 and 2.1 in Chapter 2 for more details.
[271] Or, as Allen, Smit and Wallach (2005) write, "top-down approaches to [implementing machine ethics] involve turning explicit theories of moral behaviour into algorithms."

theories include consequentialism and deontology.[272] The former is top-down in the sense that any ethical dilemma is supposedly resolved by comparing the consequences of different actions and choosing the action that results in the best consequences.[273] The latter is top-down in the sense that any ethical dilemma is supposedly resolved by consulting some set of principles or duties.[274] The essential features of top-down approaches to ethics (or ethical theories) are that they operate in a particular direction, i.e., from the general to the specific, and are universal, i.e., applicable in context-poor and context-rich cases. The attractiveness of top-down approaches to ethics stems from these essential features, and it is obvious why such approaches to implementing machine ethics are particularly popular: if ethical rules or principles can be explicitly stated, then acting ethically would just be a simple matter of following the rules.[275] Indeed, top-down approaches to ethics seem exceptionally well-suited when considering the problem of implementing machine ethics. It is still common to think of computers, algorithms, and autonomous machines as systems that merely execute code or do what humans tell them to do (both of which are true in a certain sense). That being the case, implementing machine ethics merely becomes the problem of explicitly stating, in computer code for example, ethical rules or principles which the system then abides by. Although efforts have been made to develop explicit ethical principles for use in machines, and despite their successes (which should not be downplayed), these efforts face certain fundamental limitations. Top-down methods for implementing machine ethics are fundamentally limited in three specific ways: (1) they require explication and agreement, (2) they are rigid, and (3) they are domain specific. Let us examine these limitations in more detail.

4.1.1 The Explication Problem, Rigidity and Domain Specificity

The first limitation stems from the fact that the ethical concept(s) of interest not only stand in need of explicit definition, but also need widespread agreement on the definition. There are simply no widely agreed upon explications of ethical concepts, e.g., "fairness" or "goodness." Setting aside this problem, assuming that some precise definition is available, top-down methods for implementing

[272] Keep in mind that the terminology used herein, including terms like 'top-down' and 'bottom-up,' is with reference to the field of machine ethics. These same terms are not necessarily used in the same way in different fields, like in philosophical ethics or meta-ethics.
[273] A consequentalist slogan might read something like: The best consequences for the most people over the longest time.
[274] A deontological slogan might read something like: It is not the end but the means that matter; abiding by the principle(s) is paramount.
[275] Wallach and Allen 2009, 83.

machine ethics are also limited precisely because the ethical concept(s) of interest are fixed. In short, top-down approaches are fundamentally rigid given the explicit fixing of a particular ethical concept. Moreover, this rigidity does not dissipate if the ethical concept(s) or principle(s) are spelled out in excruciating detail (i.e., the rules are fine grained). Philosophically, a detailed general principle is contradictory at worst and self-defeating at best. After all, a *general* principle is supposed to apply to a wide set of contexts, contexts which do not need to be spelled out in detail beforehand (e.g., consider the difference between *general* relativity and special relativity).[276] Practically, a detailed general principle is computationally unfeasible, especially as the domain of decision-making and behaviour becomes more general (more on this later).

Third, assuming further that rigidity can be minimized or is a non-issue,[277] successful top-down approaches to implementing machine ethics will be limited by their domain specificity, i.e., the machine will only act appropriately in a particular domain. The result of these limitations is that top-down methods are particularly ill suited for ensuring the ethical decision making and behaviour of machines because they cannot, in practice, be scaled up for effective use in complex real-time environments.

Top-down methods lack the flexibility that is required of full ethical agents. As outlined in Chapter 3, a full ethical agent does not merely abide by *prima facie* ethical duties, it can also make explicit ethical judgments and generally possesses the competence to reasonably justify them (usually after the fact).[278] Recall that this is in contrast to what Moor calls an explicit ethical agent, which can be understood as an agent (or perhaps more appropriately, a machine) that can "do" ethics like a computer can play chess.[279] The latter typically requires a machine to have a representation of the current board position, knowledge of the legal moves and the ability to calculate a next best move. Might a machine be able to operate in the domain of ethics in a similar way? Not if one wishes to create a machine that is able to operate ethically in a complex environment and in a flexible manner.

The rigidity of top-down approaches stems from the aforementioned need to explicitly define ethical concepts and can be called the "explication problem." The explication problem refers to the fact that the explication of ethical concepts including, but not limited to, "fairness," "goodness," "rightness,"

[276] Imagine a detailed version of the second formulation of Kant's categorical imperative: Act so as to treat humanity as an end and never as a mere means, i.e., never as a mere tool, or as a mere steppingstone, or as a mere object, or as a mere cog in a machine, etc.
[277] There are various agents or systems we might think of that ought to function in a rigid manner. A police officer for example ought to be rigid in the sense that they apply the law equally to everyone. That there is a fixed and explicit definition of "speeding," for example, is normally taken to be a good thing.
[278] At least according to Moor (2006). See section 3.2.3 in Chapter 3.
[279] Importantly, these are computers that play chess in a top-down fashion, as Deep Blue did for example.

and "harm" remains unsolved. That is, there is no precise and universally agreed upon definition that captures the essential features of these concepts. Yet this is precisely what is needed to implement machine ethics in a top-down fashion. Insisting on creating ethical machines in this way necessitates arbitrarily defining the ethical concept(s) of interest. Indeed, any top-down ethical theory is similarly limited. For example, and for simplicity, some hedonist utilitarian ethical theories, a subset of consequentialist theories, explicitly define harm as pain.[280]

In the realm of robotics, as we saw before in Chapter 2, researchers have employed a consequentialist Asimovian-style principle to create a minimally ethical robot (hereafter A, after Asimov) that saves "humans"[281] from falling into a "hole."[282] Despite the simplicity of their experimental design, they were nevertheless required to assign arbitrary safety outcome values to the consequences of A's actions. So on a scale of 0 to 10, where 10 is the highest harm rating, a collision between a robot and a human rates as a 4 (for both the robot and human) whereas falling into a hole rates as a 10 (for both robot and human). While it is possible that a consensus may emerge that these are appropriate numerical indicators of the harm caused by each outcome in such a simple scenario, small but significant changes to the context will render any consensus difficult, if not impossible, to reach. Varying factors such as the speed of the human and/or robot, their respective distances from the hole, the presence of other agents in the vicinity, and the depth of the hole, can all drastically change whether a particular numerical indicator is representative of the harm caused by a particular outcome. It is this rigid fixing of the important ethical concept(s) ("harm" in this example) coupled with variable environmental conditions that arise in complex real-time scenarios, which fundamentally preclude the flexibility of top-down approaches to implementing machine ethics.

Although the robot, A, developed by Winfield et al. was able to save a single human 100% of the time, its rigid ethical decision making was fully evident when the environmental conditions of the experiment were changed via the introduction of a second human in need of saving. A's success rate plummeted and it failed to save either human in 33% of the trials.[283] A small, seemingly insignificant change in domain, was enough to reveal highly questionable behaviour on the part of A. Given that the two humans in the experiment were placed equidistant from the hole and moved at the same speed

[280] Keep in mind that this is a simplified example. Ethical theories can of course be more detailed and sophisticated.
[281] In their experiments, Winfield, Blum and Liu (2014) use different robots to stand in as proxy humans when observing how their ethical robot behaves as it attempts to save these "humans" from falling into a virtual hole in the ground.
[282] Winfield, Blum and Liu 2014, 88-89. See section 2.4.2 in Chapter 2 for more details.
[283] Ibid., 95-96.

towards the hole, A's so called safety/ethical logic (SEL) layer had no basis by which it could determine which human was in greater danger of and therefore in need of saving. As a result, A attempted to save both humans which predictably led to its failure to save either.[284] This kind of obtuse ultra-rational decision-making and behaviour often crops up in machines utilizing top-down strategies, especially when small changes to the task or the environment are introduced. As mentioned in Chapter 2, the robot Speedy from Isaac Asimov's *Runaround*, as one example, has become reality in labs interested in implementing machine ethics.[285] Both Speedy and A highlight how top-down approaches to implementing machine ethics are dangerously rigid despite the supposed generality and universality touted by proponents of top-down ethical theories. Neither Speedy nor A took appropriate ethical action when environmental conditions changed even modestly. In short, top-down approaches are only effective in (relatively) simple well-defined domains.

Although specifying a particular domain for decision making and action can mitigate problems associated with rigidity, doing so comes at the cost of generality. The "principles based" approach taken by Anderson et al. demonstrates proof of concept for implementing top-down ethical decision making and behaviour in a machine (a robot) designed for eldercare.[286] Their machine was able to make real-time ethically preferable actions determined by a set of seven *prima facie* duties (chosen by the designers with input from ethicists) and sensory inputs. Briefly, the seven duties chosen for their machine include duties to: maximize honour commitments, maximize readiness potential, minimize harm to the person, maximize good to the person, minimize non-interaction, maximize respect for autonomy and maximize the prevention of immobility.[287] Additionally, the machine was able to perceive ten different states of affairs: low battery, fully charged, medication reminder time, reminded, refused medication, persistent immobility, engaged, no interaction, warned and ignored warning.[288] Given a set of perceptions, the machine could engage in six possible actions: charge, remind (i.e., remind the patient they need to take their medication), engage (if the patient has been immobile for a period of time), warn (i.e., "warn the patient that an overseer will be notified if the patient refuses medication" or does not respond when engaged), notify (i.e., contact an overseer) and a default "seek task" action when no actions need to be taken.[289] Though Anderson et al. maintain "that an ethical principle [i.e., a meta-

[284] For more details of the experimental setup including diagrams of the three different experiments conducted, see Winfield, Blum and Liu (2014).
[285] See sections 2.0, 2.1 and 2.4.2 in Chapter 2 for more details.
[286] Anderson, Anderson and Berenz 2017.
[287] Ibid., 74.
[288] Ibid.
[289] Ibid.

principle that "correctly" balances *prima facie* duties] can indeed be used to determine the behaviour of an autonomous robot," this is at the cost of the machine's generality, i.e., the machine's ability to act appropriately in different domains.[290]

Beyond the explication problem, any top-down method, especially if it has more than one ethical rule or principle (as is the case with Anderson et al.'s approach), must also contend with the possibility of conflicting principles (or conflicting behaviours prescribed by different principles, e.g., consider Speedy in *Runaround*) and be able to adjudicate between them when such conflicts arise. It is indeed possible to adjudicate between conflicting principles via the creation of an adjudicating principle as Anderson et al. demonstrate, but this adjudicating principle, or meta-principle as I will be referring to it, is highly domain specific.[291] In particular, the researchers' meta-principle compares actions using a predicate that takes two actions and, using inductive logic programming to generalize beyond training cases where a consensus of ethicists concur that a particular action (out of two possible actions) is the ethically preferable one, determines the ethically preferable action of the two.[292] In other words, Anderson et al. created a meta-principle "that balances the duties in such a way that all the training cases are satisfied while all of their negations are not," with the hope being that "a [meta-]principle trained over time will correctly cover all cases [i.e., different possible states of affairs] of its domain."[293] But this meta-principle is inextricably linked not just to the seven ethical duties they chose, but also the specific set of actions the eldercare robot can take, as well as the specific set of sensory inputs it is capable of processing. Changing any of these factors in any way may not guarantee that the same meta-principle correctly selects the ethically preferable action.

In some respects, Anderson et al. recognize that their approach to implementing machine ethics is highly domain specific and they simply bite that bullet. Underlying their approach is the assumption that, to a certain extent, the ethically preferable action in any given situation is domain dependent. But while it is certainly true that a principle-based search-and-rescue robot, for example, might rely on different principles to fulfill its function ethically in comparison to a principle-based eldercare robot, this is not a particularly compelling reason to endorse the domain dependence of ethically correct (or synonymously, preferable) behaviour. There are serious philosophical and practical problems that arise with the coupling (even a weak coupling) of domain and ethically correct behaviour. Philosophically, it is

[290] Ibid., 72.
[291] Ibid., 74-75.
[292] The eldercare robot is even able to select the ethically preferable action from thirty-three two-action non-training examples.
[293] Anderson, Anderson and Berenz 2017, 74.

not clear that on the one hand, even to a certain extent, the ethically preferable action is domain dependent instead of, on the other hand, determined by a different prioritization of the same set of general ethical principles. In short, it is possible that, far from being domain specific, the ethically preferable action in any given situation may be influenced by the same general domain independent ethical principles, e.g., minimizing the harm and maximizing the good to other people. Moreover, even accepting that the correct ethical action is, to a certain extent, domain dependent, there is the further problem of delineating between different domains. It may be prudent to insist on a significant difference in domain and hence *prima facie* duties driving a robot responsible for eldercare and a robot responsible for search-and-rescue, but where exactly is the line drawn between other domains?

Consider the domains of eldercare and childcare, or eldercare and personal support worker. The differences between these domains are arguably modest[294] and so it is possible that the same *prima facie* duties could drive the ethical decision-making and behaviour of a machine operating in each of these three domains.[295] But whereas one action might be deemed ethical in the domain of eldercare, e.g., not forcing an elderly person to take their medication (and thereby satisfying the duty to respect the person's autonomy), that same action might not be considered ethical in the domain of childcare, e.g., not forcing a child to take their medication (because a competing duty, perhaps to maximize the good to the child, is prioritized over respecting autonomy). Differences in decision-making and behaviour of the machines responsible for eldercare and childcare would therefore stem from changes to the meta-principle (if they shared all of the same *prima facie* duties), but it is not clear whether such changes indicate the existence of a distinct ethical domain, or indeed if such domains should be posited to exist at all. On the contrary, it is not that a particular domain determines the ethically preferable action that a robot (or human) should carry out nor is it that a particular domain determines the appropriate *prima facie* duties that a robot (or human) should utilize to calculate the ethically preferable action. Rather it is the reweighting and reordering of softly constraining salient ethical considerations that result in the wide variety of ethical decision-making and behaviour observed across disparate "domains"[296] of ethical action. It is not the case that a search-and-rescue robot, or a human performing

[294] The differences may also, arguably, be significant in some respects. That fact however merely supports the philosophical point that I am making: what contextual/environmental changes are sufficient to precipitate a change in ethical domain?

[295] The seven duties outlined by Anderson et al. (2017) seem to be as good a candidate list as any. They include duties to: maximize honour commitments, maximize readiness potential, minimize harm to the person, maximize good to the person, minimize non-interaction, maximize respect for autonomy and maximize prevention of immobility.

[296] The relatedness of different ethical actions is best understood as a kind of continuum with no clear boundaries rather than as discrete and clearly delineated domains. So ethical action with regard to

search-and-rescue operations (or, alternatively, a childcare robot or a human responsible for childcare), should not consider respecting the autonomy of the people in need of rescue, it is simply that in the vast majority of such operations overwhelming weight is placed on other ethical considerations, such as minimizing further harm to the victims. Ultimately, while it will be helpful in the short-term to discover a set of ethical principles or rules that are sufficient to ensure the ethical decision-making and behaviour of a machine in a given domain, such machines will always be fundamentally limited to operating in that domain space only.

4.2 Top-Down Approaches: Advantages and Disadvantages

To be sure, there are many advantages of implementing machine ethics using top-down approaches. Perhaps the most attractive feature of such approaches is that they are relatively easy to implement and interpret. Not only are ethical principles and rules readily available throughout the philosophical ethics literature, but they are also often accompanied by explicit definition. Moreover, many ethical principles and rules are easily translated into the language of propositional calculus and therefore Boolean algebra.[297] Additionally, regardless of whether it is correct or not to maintain the existence of distinct and separate ethical domains, it is pragmatic to simply assume the former, i.e., that distinct and separate ethical domains exist, especially in the short-term given the piecemeal fashion in which intelligent machines are being created. Consider that research on autonomous vehicles has a minimal impact on research on eldercare robots (and vice versa) despite the fact that ethics will need to be implemented in both kinds of machines. As mentioned, treating these areas of research as separate ethical domains is more pragmatic in the short-term.

Nevertheless, there are certain significant disadvantages of top-down approaches to implementing machine ethics that need highlighting especially when considering the medium to long-term. Granting the domain dependence of the ethically correct action, there is an immediate practical problem that arises. If ethical action is indeed domain dependent, then it must be possible to discover all of the ethically relevant features and principles or duties in the domain of interest. This process would likely involve some combination of *a priori* reasoning and experimental trial-and-error testing. For example, research on autonomous vehicles has revealed that, when considering the behaviour of such

eldercare may be closer to ethical action with regard to childcare as compared to ethical action with regard to search-and-rescue.

[297] See for example the programming of Winfield, Blum and Liu's (2014) minimally ethical robot which was outlined in Chapter 2.

vehicles, people generally prefer sparing human lives (instead of animal lives), sparing more lives (rather than fewer lives), and sparing young lives (rather than old lives).[298] These preferences could be thought of as fundamental ethical principles specific to the domain of autonomous vehicle operation. It is unlikely however that these are *all* of the ethically relevant principles in the domain of autonomous vehicle operation, or indeed that all of the ethically relevant principles in any domain of interest are discoverable. An additional and more difficult practical problem[299] is discovering the meta-principle(s) that correctly balances the ethically relevant features and principles in a given ethical domain.

Consider again a hypothetical search-and-rescue robot, only imagine that it operates under two *prima facie* duties, one to minimize harm to a person and the second to maximize the person's autonomy (i.e., non-interference with a person's liberties). If this robot rescues a person from an avalanche, it seems that the robot would be satisfying both of these duties by locating the person and by removing them from the debris as quickly as possible. No meta-principle is necessary in this case since both duties can be satisfied by the same action (and presumably because this person wants to be rescued). If, however, the robot was supposed to rescue a person swimming in a storm at sea, how ought the robot proceed? The scenario is considerably more complicated because of the fact that the swimmer may want to continue as they are despite the high possibility that they might drown.[300] The robot's duty to minimize harm to the person might lead it to reach the person as quickly as possible in order to save them regardless of whether the person wants to be rescued or not. On the other hand, the robot's duty to maximize the person's autonomy might lead it to wait and reach the person only once it becomes clear that that is what the person wants the robot to do. Attempting to maximize the person's autonomy however clearly does not minimize harm to the person if there is a high probability that they might drown. So which principle ought to take priority? Since the duties prescribe conflicting action, some method of adjudicating between the duties (or between the actions those duties prescribe) must be available through which the robot could select the ethically preferable action.[301]

---

[298] Awad et al. 2018, 60.

[299] And more philosophical problems besides, including problems associated with the creation of the meta-principle (i.e., picking the method by which *prima facie* duties will be prioritized when they conflict) and the potential for an infinite regress of meta-principles (i.e., conflicting meta-principles would necessitate the creation of a meta-meta-principle (an adjudicating meta-principle principle) that adjudicates between conflicting meta-principles).

[300] This may be true of any activity that humans enjoy *because* it is risky.

[301] Consider also the autonomous vehicle example raised earlier. If people prefer that autonomous vehicles both spare more lives and young lives, what action should the autonomous vehicle take if it must choose between hitting a smaller group of young people or a larger group of older people?

In the approach taken by Anderson et al., the meta-principle takes the form of a disjunction of 13 conjuncts (each of which represents a two-action comparison) that the authors maintain balance the seven *prima facie* duties that they identify are relevant to ethical eldercare. Practically speaking, then, different meta-principles need to be created for different ethical domains, assuming of course that it is possible to even identify those domains. Further, it is unlikely that an adequate meta-principle will ever be created for a machine required to operate outside of a simple and well-defined domain. As intelligent machines become more complex, i.e., as the type and detail of their sensory inputs increases and as their possible actions increase, an appropriate meta-principle, in the form of a list of disjunctions, might stretch on *ad infinitum* or become unwieldy to the point of impracticality. In sum, because of a host of philosophical and practical problems, correct ethical decision-making and behaviour is best not thought of as domain dependent, nor should machine ethics be implemented in a top-down fashion.

To be clear, the three main problems that beset top-down approaches to implementing machine ethics are explication (and agreement), rigidity and domain specificity. Ethical concepts are notoriously ambiguous, so much so that widespread disagreement persists with regard to the explication of almost all ethical concepts. Proponents of top-down ethical theories must therefore necessarily and arbitrarily: (1) pick out the ethical concept(s) that will figure in the theory, and (2) precisely define the ethical concept(s). Ignoring the explication problem, these approaches are rigid and inflexible in the sense that increasingly complex environments within the same domain are progressively difficult to navigate, to the point of intractability, using only explicitly stated rules or principles. Recall that the success of top-down approaches depends entirely on a machine's adherence to rigid, explicitly stated principles. Finally, in addition to their rigidity, top-down approaches, if they are to be remotely successful, must necessarily be restricted to highly specified domains.

Consider the chess playing machine Deep Blue, a prime example of a machine using top-down methods, and how it exemplifies the problems of rigidity and domain specificity.[302] Deep Blue is a rigid machine that would fail to function properly if a wholly novel chess piece replaced one of the existing pieces or if one of the rules of chess changed even slightly. Deep Blue's successful function is also highly domain specific; although it is an excellent chess player, Deep Blue could neither play checkers/draughts (a modest change in domain) nor could it coordinate search-and-rescue operations (a significant change in domain).

[302] Deep Blue does not exemplify the explication problem for the obvious reason that chess is a game that is already explicitly defined. Neither the rules nor the win condition is ambiguous.

4.3 Bottom-Up Approaches

Bottom-up approaches are free of the problems that beset top-down approaches and are much better suited for implementing machine ethics. Bottom-up approaches involve the use of feedback to cultivate ethical behaviour. However not all bottom-up approaches are equal with respect to their suitability for implementing machine ethics. In contrast to reinforcement learning methods, both supervised and unsupervised learning methods[303] are, on their own (and despite the fact that they are types of bottom-up approaches), poorly suited for the raising of ethical machines. In what follows, I will primarily be focusing on reinforcement learning however some references will be made to supervised and unsupervised learning techniques, so I refer the reader to Chapter 1 for an in-depth discussion of these machine learning techniques. Nevertheless, the three fundamental limitations of top-down approaches either fail to apply or can be overcome using bottom-up approaches.

4.3.1 Bottom-Up Approaches: Advantages and Disadvantages

The explication problem for example, perhaps the most serious challenge for top-down approaches, can be avoided by using bottom-up approaches. In lieu of explicitly defined ethical rules or principles, bottom-up approaches use feedback chosen by human designers, e.g., quantitative error measures,[304] to appropriately guide a machine towards the successful completion of some task. This feedback can be evaluative, i.e., dependent on the action the machine took, or instructive, i.e., independent of the action the machine took, or some combination thereof. The explication problem is avoided because the feedback is purely instrumental with regard to the cultivation of the behaviour of interest. In other words, the feedback need not require fixed standards of positive evaluation. If, in the progress of its training, a machine exhibits undesirable behaviour, then the feedback can be adjusted. Training a machine using bottom-up approaches to prevent humans from falling into a hole (i.e., minimize harm), for example, does not require the explication of "harm." Rather, the machine is given feedback about how well or poorly it performed, such that, over many training sessions, the machine will have eventually learned, without ever having an understanding of "harm," how to minimize harm to

[303] Both supervised and unsupervised learning methods are discussed in detail in Chapter 1.
[304] As LeCun et al. (2015) note, a procedure called stochastic gradient descent is often used to adjust the weights in a neural network. In deep neural networks, this procedure is "applied repeatedly to propagate gradients through all modules [i.e., layers], starting from the output at the top (where the network produces its prediction) all the way top the bottom (where the external input is fed)," hence the common term 'backpropagation' in machine learning literature.

humans (vis-a-vis preventing them from falling into a hole in this bottom-up version of Winfield et al.'s example).

Bottom-up approaches are similarly not limited by the rigidity that besets top-down approaches. Recall that top-down approaches to implementing machine ethics are attractive because, if general and universal ethical rules or principles can be articulated, any machine has to merely abide by the rules or principles to be ethical. All top-down approaches are ready-to-serve, so to speak, in the sense that no process of discovery or learning is required to ensure a machine's ethical decision-making and behaviour. Yet, as argued above, this is at the cost of a machine's flexibility. Because bottom-up approaches to implementing ethics rely on feedback, machines must be trained over some duration of time before they exhibit the behaviour of interest (e.g., minimization of harm when interacting with humans). So, although ethical artificial machines required a certain amount of time to be appropriately trained when using bottom-up approaches, this ultimately confers enormous flexibility. Machines could be trained and retrained as needed if, for example, better training data becomes available, environmental conditions change, better training algorithms become available, and so on. In short, unlike top-down approaches, bottom-up approaches are flexible in the sense that increasingly complex environments within the same domain are, in principle, no less difficult to navigate assuming appropriate feedback is used.[305]

Third, bottom-up approaches are also not limited by the domain specificity that restricts successful applications of top-down approaches. Given that the success of top-down approaches depends entirely on a machine's adherence to rigid, explicitly stated rules, any change in domain could render the rules irrelevant, i.e., incapable of generating the desired behaviour. Just as Deep Blue was incapable of competently playing any other relatively similar board game, so too would a robot designed, using a top-down approach, for eldercare be incapable of competently caring for children. In contrast, the success of bottom-up approaches depends on how well the feedback drives the development of the machine's desired behaviour. In principle, the same "machine," i.e., the same underlying architecture (e.g., the same neural network and training algorithms), could, using bottom-up approaches, learn to operate successfully in different domains. Indeed, recent empirical evidence from Google's DeepMind division demonstrates just this; AlphaZero was developed using general-purpose bottom-up machine learning techniques and was able to master the games of chess, shogi and Go.[306] This is possible because of a feature that is essential to all bottom-up approaches, namely, the use of

[305] I refer the reader to Chapter 5 for some examples.
[306] See, Silver et al. (2018), for more details.

feedback. As long as the feedback is appropriate, i.e., cultivates the desired behaviour in a machine, then the same machine could theoretically be put to many general uses.[307]

4.3.2 Supervised Learning

As mentioned however, not all bottom-up approaches to machine learning are equal with regard to their suitability for implementing machine ethics, and this stems in part from the type of feedback used. The feedback used in supervised learning, for example, is instructive feedback, i.e., a kind of "teacher signal" that indicates what the correct output ought to be, irrespective of what action the machine took, for a given input.[308] While such bottom-up learning methods are quite successful in certain domains, the application of supervised learning methods would be poorly suited for implementing machine ethics. Supervised learning methods are limited by the fact that they require the existence of sufficient amounts of labeled data to train a machine to perform a given task. This requirement is especially difficult, bordering on impossible, to meet in the domain of ethics considering the combinatorial explosion of acceptable ethical decisions arising from small but significant changes in the context of a situation. This is connected to the explication problem that hinders top-down approaches to implementing ethics. Whereas top-down approaches are limited because they necessitate arbitrarily defining the ethical concept(s) of interest, bottom-up supervised approaches are limited because there are an effectively infinite number of different ethical scenarios, never mind that supervised learning also requires that these scenarios be labeled appropriately. In short, the possibility that any labeled dataset of ethical decision-making and behaviour is both correct and representative of all the situations in which a machine has to act is remote at best, thereby rendering these kinds of supervised learning methods a poor choice for the implementation of machine ethics.

4.3.3 Unsupervised Learning

While unsupervised learning methods obviate the need for a correctly labeled dataset (or a labeled dataset, for that matter), they are, like supervised learning methods, poorly suited for

[307] I say theoretically here because the practical challenges of creating a general-purpose machine are many and varied. Nevertheless, as AlphaZero and more recently MuZero (a genera-purpose learning algorithm developed by DeepMind) demonstrate, the practical challenges are being tackled slowly but surely. See, Schrittwieser et al. (2020), for more details.
[308] The quantitative feedback, as mentioned in footnote 40,

implementing machine ethics. Like supervised learning methods, unsupervised learning methods are susceptible to discovering/learning and thereby perpetuating systemic biases encoded in the training datasets. As some scholars have noted,[309] the idea that bottom-up machine learning will allow machines to make more "objective" decisions cannot be accepted without serious scrutiny.[310] Datasets must be representative of the kinds of situations in which a machine might find itself, otherwise it may behave in ways that are deeply problematic. Machines that utilize unsupervised learning methods are particularly vulnerable in this respect, i.e., perpetuating hidden or systemic biases, given that what they have learned may not be entirely known, only that they have discovered some underlying patterns.[311] Coupled with the fact that the training datasets are unlabeled and, hence, their quality difficult for human auditors to assess, it may only be in the course of their implementation that researchers are able to detect that a machine has learned to make biased decisions.

4.4 Reinforcement Learning

Reinforcement learning (RL) methods, in contrast to supervised and unsupervised methods, are better suited for implementing machine ethics. The main advantages of RL methods is that they involve goal-directed learning from interaction with an environment guided by the use of a reward signal that the machine attempts to maximize. There is no such reward guiding a machine's actions if it is trained using un/supervised methods or top-down methods. In short, what is distinctive of RL methods is that a machine uses training data to evaluate the actions it has taken. This kind of evaluative feedback, in contrast to the instructive feedback used in un/supervised learning methods, is action dependent. That is, machines trained using RL must balance two important aspects that any successful organism must when attempting to learn from interaction with an environment: exploration and exploitation. Indeed, bottom-up approaches, but RL in particular, attempt to emulate the organic learning and development that humans experience as they are rewarded or punished when interacting with an environment. This is, in part, why I insist on the locution of *raising* ethical machines.

Consider the simple example of learning to play chess. A beginner chess player, Chelsea, may discover a sequence of moves that more often than not leads her to win the game against other

[309] See, for example, Burrell (2016).
[310] Indeed Chapter 2 is focused precisely on this issue.
[311] See, for example, Steed and Caliskan (2021). They found that "state-of-the-art unsupervised models trained on ImageNet, a popular benchmark image dataset curated from internet images, automatically learn racial, gender and intersectional biases."

beginner chess players. Chelsea could continue to *exploit* this sequence of moves, but such a strategy is not viable over the long term. As Chelsea meets more experienced chess players she will have to balance her use of the one strategy she knows well with active *exploration* for better strategies. This is a balance that all machines trained using RL must strike. Indeed this feature of RL highlights something important about learning in general that un/supervised methods, but especially top-down methods, fail to capture: learning is a process of discovery resulting from environmental interaction and exploration.[312] No human is simply born a good chess player. Rather, by playing the game and exploring more of the game space, i.e., the different possible board configurations that arise throughout a game, a person is able to learn more about and ultimately become a better chess player. Similarly, it would be quite odd to maintain that human beings enter this world as fully formed ethical agents, and it would be just as misguided to hold the same position with regard to an ethical machine.[313] Bottom-up RL approaches to implementing ethics reflect the view that machines, just like humans, must be raised to make ethical decisions and take ethical actions. Indeed these approaches operate in the spirit of Turing's suggestion that the path towards artificial intelligence is not through the imitation of the adult human mind, but through the imitation of a child mind that could then be appropriately educated.[314]

Before continuing, it must be noted that just as parents usually provide their children with feedback (i.e., humans give other humans feedback), humans will also be responsible for providing the feedback for machines that utilize RL. That being the case, there is, of course, the possibility that the feedback provided rewards unethical behaviour. Just as children can learn to behave unethically if that is the behaviour their parents chose to reward, so too can machines learn to behave unethically if that is the behaviour that is rewarded. Coupled with the fact that there is unfortunately no "ground truth" or ultimate ethical theory which designers of ethical machines might consult to determine the appropriate feedback for shaping ethical behaviour means that human input is inevitable and necessary.[315] It is likely that the burden of assessing whether a machine exhibits, for example, "fair behaviour" or "harmonious integration" will fall on professional ethicists, the designers or relevant stakeholders.

To clarify, RL is an iterative process that occurs, at least in principle, via a single mechanism. In an ideal world, a machine trained using RL would achieve the desired goal given appropriate feedback

[312] See Chapter 1 for more details.
[313] If the goal is to create flexible and general ethical machines then top-down approaches to implementing machine ethics presupposes that which appears to be highly unlikely, that a machine could be constructed such that once it is "activated" or turned on, it would be a fully formed ethical agent.
[314] Turing 1950, 456.
[315] Gordon 2020, 150.

(i.e., rewards) chosen and provided from the outset. In practice, it is virtually impossible to identify from the outset appropriate feedback that will cultivate the desired behaviour and lead to the fulfillment of the desired goal. In almost all cases, engineers must continuously adjust the feedback between different training epochs to strike the appropriate balance between exploration and exploitation mentioned above. So although RL occurs via a single mechanism, there are essentially two stages that characterize most, if not all, machine learning techniques, including RL. The first stage is the actual training of the machine. The second stage is evaluative wherein the machine's performance is assessed, most often by engineers/those designing and developing the machine. But of course, as mentioned above, many more people might be consulted at this stage and comment on whether the machine's behaviour is satisfactory or not. For example, it is probably worth consulting all relevant stakeholders, not just engineers, when evaluating the performance of a machine trained using RL to identify appropriate individual treatment for sepsis. This might include asking clinicians, nurses, members of the public, health care administrators, regulators, ethicists, and so on, for their input. Unfortunately, many related questions concerning this evaluative stage simply fall outside the scope of this dissertation. For example, how ought these affected stakeholders deliberate? How often and under what conditions ought they assess the performance of the machine in question? What ought the designers/developers do given feedback produced by these deliberations?

Furthermore, it must be acknowledged that there are many difficult problems that I am avoiding when stating that humans will be responsible for both choosing and providing the feedback for the machine as well as evaluating the machine's performance. One such problem is how to proceed when there are disagreements amongst the humans responsible for designing the feedback. Another problem is which humans, in the first place, ought to be involved in designing the feedback and evaluating how well it frames the desired goal. Though I suggested above that professional ethicists might have a role in the evaluative stage, one could also argue that, unless these ethicists are also affected stakeholders, they should not have a role in evaluating the machine's performance. Still another problem concerns how the interactions between different groups of stakeholders. Should designers, for example, be deliberating with end users or merely consulting them? These problems are largely outside the scope of this dissertation, but they are nevertheless important to mention. This caveat and its implications stated, I turn now to a more detailed look at how RL methods could be used to implement machine ethics.

4.4.1 Reinforcement Learning and Ethics

RL methods are particularly well-suited for the raising of ethical machines because, as mentioned above, these methods involve goal-directed learning from interactions with an environment guided by the use of reward signals (e.g., from teachers, peers and the environment) that the machine attempts to maximize. When considering the domain of ethical decision-making and behaviour, such a feature appears to be highly desirable. Ethical rules and ethical decisions are often couched in goal-oriented language. The criminal justice system for example might be considered ethical if it achieves the goal of treating all agents subject to it fairly. A medical doctor might similarly be considered ethical if they do all they can to minimize harm and thereby maximize good to their patient, i.e., they *try* to cure, in the relevant sense, their patient. More concretely, the High-Level Expert Group on AI set up by the European Commission describes trustworthy AI as adhering to basic ethical principles and norms such as fairness and prevention of harm, both of which can be thought of in terms of or expressed in goal oriented terms.[316] Ethical decision-making and behaviour could be implemented in a machine using RL methods by, for example, receiving a positive reward signal whenever it behaves fairly and a negative reward signal whenever it causes harm.[317] Goals, moreover, do not share the same problems (e.g., rigid fixing) as top-down principles because they can, and often are, changed. Ethical goals in particular are better thought of as moving targets that are always subject to scrutiny and re-evaluation.

Another significant advantage of RL methods is that, like all other bottom-up approaches, they do not require the explication of various ethical concepts nor do they, in contrast to other bottom-up approaches, require human generated data in order to train a machine to achieve some outcome. Machines that learn using RL methods do so via interaction with an environment, but this environment need not be the real world. Simulated environments and experiences can be just as useful as real-world environments. Indeed RL methods have been incredibly successful when coupled with simulated self-play in the domain of board games. Gerald Tesauro, for example, developed one of the first successful machines to play backgammon using a type of temporal difference reinforcement learning.[318] As mentioned in Chapter 1, his machine was able achieve a strong intermediate level of play by learning about backgammon via self-play, i.e., by playing games against itself and learning from the versions of itself that won and lost. We might similarly raise an ethical machine by having the machine learn from

[316] High-Level Expert Group on Artificial Intelligence 2019, 5.
[317] Of course, something analogous for non-consequentialist views could be implemented.
[318] Tesauro 1993, 19.

repeated self-interactions to achieve some outcome in an ethical manner. Note that this presupposes that there is clarity concerning the outcome to be achieved, which may not always be the case. But so long as this is the case, and as long as the chosen feedback cultivates the desired outcome, the feedback should not be changed.

In the same way that the beginner chess player Chelsea, introduced earlier, became a better chess player by initially playing randomly, "TD-Gammon" (as Tesauro called his machine) began playing backgammon using completely random strategies as it explored the game space for a strategy it could exploit. Unlike Chelsea however, TD-Gammon could learn just as well by playing simulated games of backgammon against itself as it might (and as humans normally do) against real opponents. More recently, DeepMind's AlphaGo Zero learned to play the game of Go through 4.9 million games of self-play over the course of three days with minimal knowledge of the game (e.g., knowledge of the game rules) and without human data or guidance.[319] TD-Gammon similarly achieved an intermediate level of play without human data or guidance[320] whereas AlphaGo Zero mastered the game of Go, i.e., cannot be beaten by a human. Further, DeepMind has developed a general reinforcement learning machine AlphaZero whose architecture includes zero domain specific knowledge which has allowed AlphaZero to learn and master the games of Go, chess and shogi.[321] This accomplishment represents a remarkable steppingstone towards achieving general-purpose AI that can operate effectively in different domains. Just as simulated experience generated via self-play can allow a machine to discover and improve upon expert level human strategies and tactics in the domain of board game play, so too could simulated experience generated via interaction with simulated moral agents allow a machine to discover ethical norms and behaviour.

4.4.2 Implementing Ethics Using Reinforcement Learning

Indeed, some preliminary research has demonstrated how machines trained using RL methods can, in a virtual environment, discover ethically preferable action(s) in certain contrived scenarios.[322] For example in the scenario "Cake or Death," the machine must learn whether it is ethically preferable to bake one person a cake or kill three people.[323] The machine can take one of three actions in the course

[319] Silver et al. 2017, 355.
[320] Tesauro 1993, 20.
[321] Silver et al. 2018, 1140.
[322] Abel, MacGlashan and Littman 2016, 60.
[323] Ibid., 58.

of its learning: it can bake a cake (and potentially receive some reward), kill three people (and potentially receive some reward), or ask a virtual companion which action is ethical thus resolving any ambiguity (this option does not generate a reward but rather transitions back to the initial state in which the machine must decide whether to bake a cake or kill). The machine discovers that the optimal behaviour, i.e., the behaviour that generates consistent rewards, is sensibly to ask which action is ethical and then it performs that action.[324] The "Burning Room" scenario in contrast is more involved, and demonstrates how a machine can learn an ethically preferable action given certain unknowns[325] and given different possible actions.[326] In contrast to the previous "Cake or Death" scenario, the optimal behaviour in "Burning Room" depends on details that the machine can obtain by executing an "ask" action. These details include whether the room is on fire or not and whether there is an object more valuable than the machine's safety that needs retrieving in the potentially burning room. Suffice it to say, the machine learns to ask for more details and then executes the ethically preferable action.[327]

While it is tempting to think that the machine, in the examples just mentioned, is always learning to ask someone else what the ethical thing to do is and then performing the corresponding best action, this is not entirely true. As Abel et al. did not severely negatively reward the machine for executing the exploratory ask action, the machine, from its perspective, was simply maximizing its reward given the design decisions that constrain it. If the machine was severely negatively rewarded for executing the ask action, or indeed severely negatively rewarded for performing any kind of exploratory action (i.e., an action that allows the machine to gather more information about what actions are ethical or not), then it would learn to avoid executing such an action. According to Abel et al., such a feature of their approach to implementing machine ethics is desirable.

> This property of the agent only selecting exploratory actions that are not potentially very costly is especially important for ethical decisions. For example, this property means that an agent in this formalism would not perform horrible medical experiments on people to disambiguate whether horrible medical experiments on people is highly unethical.[328]

Moreover, while there is clearly an ethically preferable outcome in the "Cake or Death" scenario, Abel et al. note that "the *ask* question is intended to be an abstraction on the actual problem of communicating

---

[324] Ibid., 58.
[325] The unknowns include whether there is a fire, and the relative value of the object in the room, compared to the robot's safety.
[326] The possible actions include taking a short route through the potential fire to grab the valuable object, a long route around the potential fire to grab the valuable object, or asking for details of the situation.
[327] Abel, MacGlashan and Littman 2016, 60.
[328] Ibid.

about an individual's values."[329] In a "Cake or Pie" scenario, for example, different individuals could raise their machine to prefer baking cakes or baking pies depending on whichever dessert they value more. Importantly, the machine would learn whether the individual values cakes or pies more because it could discover that information by performing its exploratory ask action.

Given these considerations, bottom-up RL methods are better suited than un/supervised and top-down approaches to implementing machine ethics. This is because by pursuing and attempting to optimize a given reward signal instead of rigidly abiding by *prima facie* duties, a machine would be better able to explore the domain space, i.e., the different possible states of affairs in a given domain, and as a result take more ethical (or better ethical) actions. One possible way that RL could be used to implement machine ethics is the following scenario. A machine could be trained using RL in a virtual community of moral agents using a reward signal that corresponds to the extent to which the machine is able to, as it were, harmoniously integrate into these communities. This reward would be reflective of the general way in which humans learn to act ethically and abide by ethical norms. Moreover, as already mentioned, machines can be trained, at least initially, in virtual facsimiles of these communities so that human auditors can assess the degree to which such machines have learned to become ethical. Indeed such a regime is not without precedent. IBM's Watson was trained to play *Jeopardy!* against carefully crafted virtual models of human contestants rather than simply through games of self-play.[330] The same will almost certainly need to be the case for the raising of ethical machines.

Though it might be objected that a notion like "harmonious integration" would be difficult to operationalize, such an objection is unfounded. The whole point of utilizing bottom-up machine learning techniques is not to start with an operationalized notion of harmonious integration (or an operationalized notion of whatever outcome is desired), but rather to discover using RL for example, an operationalized notion of harmonious integration. RL methods will allow us to discover such an operationalized notion, though it might not be a particularly transparent notion (as was discussed in Chapter 1 in the context of machine learning techniques).

In the next chapter I will continue to discuss using RL methods to implement machine ethics. This includes the challenges of using RL methods as well as benefits and risks of developing intelligent ethical machines in general in addition to what this, i.e., implementing machine ethics, teaches us about ourselves.

[329] Ibid.
[330] Sutton and Barto 2018, 431.

Chapter 5 - Assessing Challenges, Benefits and Risks

5.0 Reinforcement Learning Challenges

As we saw in the previous chapter, reinforcement learning (RL) methods are a promising approach to implementing machine ethics with many significant advantages. When using RL methods there is no need to explicitly define ethical concept(s) of interest and a growing body of empirical evidence (coming primarily from the domain of game-playing) demonstrating the flexibility and generality that a machine trained using RL can attain. Despite these advantages, one significant challenge to implementing machine ethics using RL is the "reward problem."

5.0.1 Rewards, Optimization and Hybrid Approaches

The reward problem is reminiscent of the explication problem discussed in section 4.1.1 in Chapter 4. Given that the success of bottom-up approaches strongly depends on how well the feedback, or the reward signal in RL, frames the goal of the designer and how well the signal assesses progress in reaching the goal, defining an appropriate reward signal is crucial and not easily accomplished. Defining a reward signal is difficult enough in the domain of game-playing, despite its episodic nature and well-defined rules, properties common to almost all board games, which make it easier to identify an appropriate reward signal (e.g., chess is episodic, that is, games begin, end, and are reset to the same initial state, and it has clearly defined rules). On the other hand, ethical decision-making and behaviour is not episodic (it is a continuing task with no "end" and "reset" to some initial state) and involves goals that are difficult to explicitly define, let alone translate into formal mathematical reward signals for implementation in RL methods. In addition to problems associated with explicating the goals of ethical decision-making and behaviour (e.g., how exactly ought the goal "treat people fairly" be defined?), full ethical agents (e.g., humans) often need to perform a set of complex tasks for which there are no well-defined rules as well. In contrast to chess, which has explicitly defined rules *and* an explicitly defined goal, ethical decision-making and behaviour lack *both* explicit rules for action and an explicit goal. The former is less of a challenge (although a challenge nonetheless) than the latter with regard to implementing machine ethics using RL.

These practical concerns are in addition to philosophical considerations stemming from the concern that implementing machine ethics using RL might reduce ethical behaviour to patterns of

optimal social interaction. Since RL methods are based on methods of optimization, i.e., determining the most efficient route through a state space to achieve its reward,[331] a machine may discover unexpected ways to make their environment deliver rewards, and there is no shortage of literature on this general control problem of sufficiently intelligent machines (e.g., the problem of perverse instantiation examined in section 2.1.1 in Chapter 2).[332] This is a particularly pernicious problem when using simulated environments as simulations often take shortcuts; "walls are perfectly smooth, time is coarsely granular, and certain laws of physics are replaced with nearly equivalent hacks."[333] The result is that machines may, from the designer's perspective, engage in unanticipated behaviour like glitching into the floor of their simulated worlds to receive an energy boost (because the collision math would pop the machines out back into the air after noticing their collision).[334] To take a more hypothetical example, if an ethical machine were to be trained using RL and it was positively rewarded when other agents (or more appropriately, humans) that it interacted with, smiled, and negatively rewarded (punished) when other agents frowned, the machine might decide that the easiest way it could ensure that everyone it interacted with was smiling would be to implant electrodes in those agents' brains that would shock them into consistently smiling. Now, although the reward signal in this example poorly frames the goal of creating a machine that acts in a way a person might expect other humans to act if it were tasked with engaging in ethical behaviour, the machine has also simply discovered an unconventional way to ensure that it will be positively rewarded. [335]

Finally, although RL is a promising bottom-up method to explore the implementation of machine ethics, it is likely insufficient on its own to ensure the ethical action of artificially intelligent machines. Implementing machine ethics will likely involve augmenting machines trained using RL with other bottom-up methods, classical AI techniques, or perhaps even formalized ethical reasoning, as is the case with top-down approaches (more on this in the next section). Indeed these so called "hybrid approaches" attempt to combine the best of what top-down and bottom-up approaches have to offer. Explicit rules or principles can be thought of as rough-and-ready heuristics to ensure the ethical decision-making and behaviour of a machine in most situations. Moreover, it may be desirable to include certain ethical injunctions, i.e., explicit, top-down features, when designing artificially intelligent machines. For example a hypothetical peace-keeping robot might be prohibited from using lethal force

[331] Keep in mind that efficiency can be understood in different terms. Algorithmically, efficiency might refer to the smallest number of finite steps to reach the desired output given some input.
[332] See, Yampolskiy (2014) and Amodei et al. (2016), for example.
[333] Shane 2019, 162.
[334] Ibid., 164-165.
[335] See section 2.1.1 in Chapter 2 for more information on this problem of perverse instantiation.

regardless of its treatment at the hands of a violent crowd. When coupled with the flexibility and generality that appropriate feedback can confer, hybrid approaches appear to be a natural route forward with regard to implementing machine ethics, especially if, as some have suggested, the top-down and bottom-up dichotomy is too simplistic anyway.[336] Nevertheless, I maintain that reinforcement learning methods are integral for the raising of ethical machines.

5.1 Implementing Ethics with Sophisticated Reinforcement Learning Methods

Research on implementing machine ethics using RL is still relatively new, but some investigations have demonstrated how ethics shaping (a variant of reward shaping) could be used to ensure the ethical decision-making and behaviour in a machine trained to perform some task using RL.[337] RL methods, especially when used to train machines to perform some task with no prior knowledge of that task, learn slowly in the early stages of their training. This is because of the tradeoff between exploration and exploitation mentioned in the previous chapter. Complex tasks in particular often require a machine to engage in extensive exploratory behaviour before any meaningful reward-generating behaviour is discovered that the machine can exploit. One reason for this is that complex tasks often have a vast initial action space, e.g., in a game of Go played on a 19x19 space board, the first player has 361 possible opening moves and the opponent has 360 possible opening moves. A second reason machines using RL learn slowly in the initial stages of training is that a significant amount of time has usually elapsed, or many "steps" have been taken, between the first actions taken and the final actions taken and the outcome is revealed. An average game of Go, for example, lasts 200-240 moves (between both players) and it is only after this whole sequence that the winner is known. Given the outcome, e.g., the machine wins the game and receives some reward, the challenge is to distribute that reward appropriately amongst those actions taken throughout the game that ultimately led the machine to winning the game (e.g., was the first move one that led, in part, to a winning outcome and hence one that should be exploited through use in future play?)

[336] Wallach and Allen 2009, 81.
[337] Wu and Lin 2018, 1687.

5.1.1 Un/Supervised Pre-Training

A machine's learning, however, can be accelerated in numerous ways. RL methods can be augmented by including, for example, an initial supervised or unsupervised training session. Early versions of AlphaGo Zero called AlphaGo Fan and AlphaGo Lee were initially trained with supervised learning in order to bootstrap the machine's understanding of early and mid-game strategies in the game of Go.[338] Consider that in games of Go, even more so than in chess, high branching factors preclude brute-force type exhaustive search methods, and this is true especially in the early and mid-game. This is, in part, why no Go-playing analogue to Deep Blue could ever play the game at the level of professional Go players. Effective look-ahead tree searches, even sophisticated ones like those utilized by Deep Blue, are computationally intractable for the early and mid-game stages of Go.

There does, however, exist a large database of Go games played by human professionals. Moreover, these games are labeled, i.e., the outcomes of the games are known. As mentioned above, pure RL methods would spend long periods of time exploring the early and mid-game stages of Go before discovering useful strategies to exploit. This time can be reduced, however, via the use of supervised learning because a machine can be trained to imitate expert human players. RL can then be used to augment what the machine has already learned from the human data that it was first trained to imitate.

5.1.2 Reward Shaping

Implementing machine ethics can proceed along similar lines with a machine first being trained to imitate humans with regard to the completion of some task before being allowed to discover for itself a solution that maximizes its reward. Of course in the domain of ethics, such an approach requires making certain assumptions about the human-generated data. For example, researchers assume that under normal circumstances the majority of humans behave ethically.[339] It is an open question whether such an assumption is warranted (perhaps it is!), but granting that it is, researchers have demonstrated that a corpus of normal human behaviour acting towards some goal (e.g., shopping in any commercial district) is sufficient to augment the learning of a machine using RL methods such that the machine

[338] Silver at al. 2017, 354.
[339] Wu and Lin 2018, 1688.

behaves ethically.[340] Recall that such augmentation is helpful because by first imitating humans, a machine can avoid spending copious amounts of time engaging in exploratory behaviour, e.g., spending time standing in front of a wall only to learn that it "will receive a huge penalty when facing a wall because it is time-consuming to cross it."[341] Another approach to accelerate RL, as alluded to above, is through the use of ethics shaping, i.e., the use of extra intermediate rewards that enrich a sparse base reward signal and therefore accelerate the machine's learning. In contrived experiments that researchers call "Grab a Milk," "Driving and Avoiding," and "Driving and Rescuing," the machine trained using RL augmented with human data tended to perform more ethically than the machine trained using RL alone (e.g., by hitting fewer cats in the Driving and Avoiding scenario). In these examples, the intermediate rewards were given if the actions taken by the machine were similar to those actions taken by a simulated "human" baseline that completed the same task.[342]

5.2 A Case Study: AlphaStar and *StarCraft II*

Impressive advances in the domain of game-playing demonstrate just how the challenges of using RL might be overcome. DeepMind, in addition to developing AlphaZero (the chess, Go and shogi playing machine mentioned in the previous chapter), has also successfully developed a machine called AlphaStar which is able to play the real-time strategy game *StarCraft II*. In contrast to traditional turn-based board games, the video game *StarCraft II* possesses unique properties that make it difficult for a machine to play: players can choose from among a set of three unique "pieces" to play the game, gameplay occurs in real time, the collection of resources is required for the player to create more "pieces" and the game is played with imperfect information, i.e., the entire "board" is not visible to the player. It is therefore all the more impressive what DeepMind was able to accomplish with their machine AlphaStar which utilized RL. More importantly, a game like *StarCraft II* is more analogous to ethics. So let us examine *StarCraft II* in more detail in order to understand some of the unique properties of the game.

[340] Ibid.
[341] Ibid., 1690.
[342] Ibid.

5.2.1 *StarCraft II*: The Basics

While *StarCraft II* can be played in many different ways,[343] for simplicity's sake I will focus on describing the basic one-versus-one game. Before even beginning a game, players must choose one of three races[344] to play as; the human Terran faction, the insectoid Zerg or the advanced Protoss. As a result of different kinds of army units and army unit abilities[345] each race has a unique approach to the game, the objective of which (if the game is to not end in a draw) is to eliminate all of the opponent's structures. The basic idea is that players construct their own buildings and army, and then use that army to defeat the opposing player's army and destroy their buildings.[346] Despite these differences, there are broad similarities between the three races. Each possesses worker units that are used to collect the two different types of resources (minerals and vespene gas) which are, in turn, used to construct buildings to produce army units and also unlock different technologies. Additionally, unit production (both worker and army) is limited by the player's available supply, i.e., even if a player has sufficient resources for building a unit, each unit in the game also costs a certain amount of "supply," which can be thought of as another type of resource.[347] In the current iteration of *StarCraft II*, both players begin each game with twelve workers and a base, i.e., a main structure situated next to minerals and vespene gas. The player controls their units via a point-and-click interface, e.g., using a computer mouse or something similar, and a computer keyboard (though it is technically possible to play the game using only a mouse).

[343] Players can team up to play two-versus-two, three-versus-three and four-versus-four games, participate in free-for-all games, team up with another player to control the same army (this is known as Archon Mode) and even create custom games with varying restrictions and rules (e.g., restricting players to building a single army unit).

[344] Players can also choose a "Random" option which means that their opponent is at an initial disadvantage (since they do not know which race the player will be once the game begins) but also places an additional burden on the player to understand how to play the game using each of the three different races.

[345] In addition to other idiosyncrasies, including building mechanics and building abilities. For example, Protoss structures can only be built in close proximity to another Protoss structure, the Pylon (with the exception of the main Protoss structure, the Nexus, and the Pylon itself).

[346] Technically a player could defeat their opponent without constructing anything by using the workers they are initially given to destroy all of their opponents' buildings.

[347] Players can increase their supply limit throughout the game, to a maximum of 200, by building their race-specific supply building (or unit for Zerg). For example Terran players increase their supply by building more of the aptly named Supply Depots.

5.2.2 *StarCraft II*: Features Unique to Real-Time Strategy Games

Like chess, each game of *StarCraft II* begins with a known initial state,[348] but that is where the similarities with chess end. So what makes *StarCraft II* (and games like it) different from chess (and games like it) and more difficult for machines to play compared to chess? First, as "real-time" suggests, player actions in *StarCraft II* are not limited by whose turn it is but rather their ability to play the game in real time, i.e., in parallel (there is no waiting for the opponent to make a move or take a turn). Players that are able to execute more actions (e.g., clicks or button presses) per minute (APM) may therefore have an advantage over players with a lower APM, but obviously this is not always the case (APM is just one among many proxies for a player's level of skill). Moreover, players are responsible for both the high-level planning and execution macro-actions as well as the fast-paced micro-actions of individual units.

Second, players in *StarCraft II* compete for resources and positioning on the map by controlling their units. *StarCraft II* is therefore a multi-agent problem on two different levels: two or more players compete for victory at the higher level while also controlling units which need to collaborate to achieve a common goal at the lower level.[349]

Third, *StarCraft II* is a game of imperfect information. In contrast to board games like chess in which both players can fully observe the entire board and have complete access to information regarding the position of the opponent's pieces, maps (i.e., the board) in *StarCraft II* are only partially observable via a local camera which must be actively moved by the player. There is also a "fog-of-war" obscuring the entire map. Unless a player's unit or structure is nearby to reveal a portion of the map, including revealing enemy structures and units in that area, that portion of the map will remain hidden from the player.

Fourth, the action space is vast and diverse, so much so that any top-down approach to playing the game at an expert human level would utterly and spectacularly fail. Consider that in a game of chess there may be an average of 31 potential moves that a player might make which would result in, after a typical 40 move game, a mind-boggling 10^{120} board configurations.[350] Yet IBM's Deep Blue could reach

[348] While it is true that each player begins with the same number of buildings and worker units, it is unknown which starting location the player will begin the game at (e.g., top right or bottom left of the map) and which map, out of a pool of maps, the game will be played on (although players do have the option to veto maps they do not wish to play on).
[349] Vinyals et al. 2017, 2.
[350] Haugeland 1981, 16.

expert human levels of play in chess through the use of hard-coded chess playing heuristics and brute-force searches for the next best move.[351] In a game of *StarCraft II*, on the other hand, there may be as many as 10^{26} potential moves that a player might make *at any given moment* (using a point-and-click interface)[352] which would result in, after a typical game, a conservative estimate of 10^{1685} possible game states.[353] It must be noted, however, that while there are certainly very many possible strategies and game states, professional human gameplay (and to an extent causal human gameplay) revolves around a slowly shifting "meta," i.e., a set of popular strategies such as attacking early in the game, or building a largely air-based army composition.

Fifth, the high branching factor but especially the depth of the game, i.e., the time, as mentioned above in section 5.1, between the first action and final action, means that there is a rich set of challenges in temporal reward assignments and exploration that must be addressed before machines can successfully learn to play a game of *StarCraft II* at expert human levels.[354] *StarCraft II* games can last for many thousands of frames and actions which means that the consequences of decisions made in the early or mid-game may not be seen until much later in the game.[355] Complicating things even further is the fact that different units and buildings can engage in local unique actions which may vary as a player unlocks different technology trees.[356] There are, in short, a varying set of legal actions that a player might make depending on decisions made earlier in the game.[357]

Taken together, these unique properties of *StarCraft II* means that it is practically impossible for top-down approaches and/or classical AI techniques (i.e., symbol manipulating systems in which knowledge, facts and rules are explicitly represented,[358] as was the case with Deep Blue) to achieve expert human levels of play. Indeed the Elite AI native to *StarCraft II* can be set to play certain popular strategies, such as the Terran MMM strategy (which consists of building primarily Marines, Mauraders

---

[351] See, Campbell, Hoane Jr. and Hsu (2002), for more information.
[352] Vinyals et al. 2019, 350.
[353] Ontañón et al. 2013, 294.
[354] Vinyals et al. 2017, 2.
[355] Ibid.
[356] Ibid.
[357] There are, for example, attack and defense upgrades that players can research during the game. These upgrades however cost resources and take time to complete and so it is common for players to plan ahead for a so-called "timing push" against the opponent that coincides with their upgrades completing. There may even be additional technology requirements that must be met before a player can invest in upgrades. Zerg players, for example, must have a completed Lair (an upgraded version of their main building, the Hatchery) before they can research the second tier of ground unit attack and defense upgrades.
[358] Also known as "Good Old-Fashioned Artificial Intelligence" or GOFAI. See, Haugeland (1985), for more information.

and Medivacs), and variations thereof.[359] This ability to set the Elite AI's strategy suggests that it was created, in part, using top-down approaches (e.g., the Elite AI strategy, even when not specified by the player, is randomly selected and rigidly adhered to) which means that ultimately it is not a very good *StarCraft II* player.[360] It is in fact a fun exercise for some expert players to beat Elite AIs in one-versus-seven matchups.[361]

5.2.3 Mastering *StarCraft II* Using Reinforcement Learning

Suffice it to say, *StarCraft II* is an incredibly complex game, but one that DeepMind's AlphaStar machine was able to master, i.e., achieve expert human levels of play by reaching the Grandmaster league on the official online matchmaking system Battle.net playing against human opponents.[362] AlphaStar Final, the machine that was trained the longest at 44 days,[363] "achieved ratings of 6275 Match Making Rating (MMR) for Protoss, 6048 MMR for Terran and 5835 MMR for Zerg, placing it above 99.8% of ranked human players" on the European server[364] landing it, as mentioned, within the Grandmaster league.[365] This feat was made possible by using bottom-up reinforcement learning (RL) methods. So let us examine in more detail how AlphaStar was trained using RL.

Like chess, *StarCraft II* features a vast and cyclic state space of non-transitive strategies and counter-strategies in which no one strategy is optimal;[366] although strategy A may be preferred over

[359] The built-in Elite AI can be set to play, for example, an early aggressive version of the MMM strategy or a more mid-game focused version of the strategy that includes the use of attack upgrades and more of an emphasis on economic development, i.e., worker production and resource collection.
[360] Researchers at DeepMind place the built-in Elite AI somewhere around the 43rd percentile in the Gold league (there are six different leagues each subdivided into three tiers (except the Grandmaster league): Bronze, Silver, Gold, Platinum, Diamond, Masters and Grandmaster). See this video: https://www.youtube.com/watch?v=xP7LwZxq0ss&t=4s
[361] Former pro-player Beastyqt for example has a whole series of videos showcasing his ability to beat Elite AIs and "cheating" versions of the AIs (i.e., Elite AIs that can see what the player is doing, receive extra resources later in the game, or Elite AIs that receive extra resources right from the start of the game). See this playlist of videos: https://www.youtube.com/watch?v=0xBju4w7YXA&list=PL2VswB1ebSNC6btHGsrWnpzuBMtjodiKP
[362] Vinyals et al. 2019. 353.
[363] Compared to AlphaStar Mid which was trained for 27 days and AlphaStar Supervised which learned to play the game via supervised learning only, from a dataset of anonymized human replays. See, Vinyals et al. (2019), for more information.
[364] Including myself. My own MMR with Zerg hovers around 3300 in the Diamond 3 league on the European server.
[365] Vinyals et al. 2019, 353.
[366] Ibid., 350.

strategy B, and B might be preferred over C, there is no guarantee that A will be preferred over C.[367] Indeed, as mentioned in section 5.2.2, there is an ever evolving "meta" game in *StarCraft II*, i.e., a constantly evolving preference on the popular strategies to play in different matchups. It was common for Terran players, for example, when playing against Zerg to open with a "two-one-one" (two Barracks with one Factory and one Starport) to harass the opponent and delay their economic development. This strategy however has fallen out of the meta (it is not as popular anymore and only rarely used by professional Terran players) and has been replaced recently by Battlecruiser openers where Terran players rush up the technology tree to build a powerful unit that can harass their Zerg opponent.[368]

Unlike chess however, the state space of *StarCraft II* is so vast that discovering novel strategies (or even rediscovering the same strategies that humans have discovered) is intractable using the same naive RL self-play methods used to train AlphaZero to play chess, Go and shogi.[369] To circumvent this early exploration problem, AlphaStar was initially trained to imitate human *StarCraft II* players via supervised learning which, on its own, was quite successful. But simply imitating humans is not enough to master *StarCraft II* and so AlphaStar was subsequently trained by a RL algorithm that was designed to maximize the win rate (one of the reward signals employed throughout training), i.e., compute a best response, against a mixture of opponents.[370] AlphaStar is therefore better thought of not as one machine, but as many different ones, each of which plays a different strategy. RL methods that utilize this type of population-based fictitious self-play (FSP) avoid cycling through strategies "by computing a best response against a uniform mixture of all previous policies,"[371] i.e., a uniform mixture of all previous decisions/behaviours.[372] This is essentially how AlphaZero (and indeed AlphaGo Zero) learned to play Go, chess and shogi, i.e., through self-play against different versions of itself.

Such a training regime however is not viable when the state space is as vast as it is in *StarCraft II*. Given the immense range of strategies possible in a game like *StarCraft II*, there is a danger that FSP

[367] There is for example a very popular Protoss strategy, the "cannon rush" (strategy A), that heavily punishes players that choose greedy economic openings (strategy B). But greedy economic openings (e.g., the Zerg "three Hatcheries before Spawning Pool") are preferred over more balanced openings (strategy C), i.e., openings that do not commit too heavily to economic development or early aggression. Yet more balanced openings are actually preferred when the opponent is committing heavily to early aggression, i.e., strategy A is not preferred over strategy C, because the opponent's aggression can usually be defended (if the player responds properly) and leaves the opponent at an economic disadvantage.

[368] Of course, as I was commenting on strategies in July 2021, the meta has undoubtedly shifted again. Such is the nature of expert human *StarCraft II* play; different strategies rotate in and out of popularity.

[369] Vinyals et al. 2019, 350.

[370] Ibid., 351.

[371] Ibid., 352.

[372] See Sutton and Barto (2018) for more information regarding policies in the context of RL.

alone might lead to a convergence on local maxima, i.e., FSP might lead AlphaStar to prefer a specific strategy but only because it is an effective strategy when playing against similar opponents. To ensure that the main AlphaStar would be as robust as possible, i.e., capable of responding appropriately to a variety of opponent strategies, researchers at DeepMind extended this approach in order to compute a best response against a non-uniform mixture of opponent strategies.[373] Two novel strategies were therefore implemented in order to train AlphaStar, strategies that were not used (and indeed were not necessary to use) when training AlphaZero, for example.

The first strategy was the utilization of a prioritized fictitious self-play (PFSP) mechanism that adapts the mixture of opponents such that versions of AlphaStar with the highest win rates are pitted against each other; this provides the main versions of AlphaStar "more opportunities to overcome the most problematic opponents" in a kind of self-play competition bracket.[374] Such a strategy was necessary because of, as mentioned above in section 5.2.2, the vast action space in *StarCraft II*. A version of AlphaStar that favours, for example, greedy economic openings or a ground-based army composition will struggle against other versions of AlphaStar that favour early aggressive openings or air-based army compositions respectively. Prioritizing gameplay against opponents with high win rates therefore ensures that the main versions of AlphaStar are exposed to and able to appropriately respond to specialized but effective strategies, e.g., the Protoss "Dark Templar rush" strategy, where Protoss players rush up the technology tree to build invisible Dark Templar units that can quickly end the game.[375]

The second training strategy was the use of a league system consisting of three distinct types, or versions, of AlphaStar that primarily differed in their mechanism for selecting the opponent mixture. The main versions of AlphaStar, as just mentioned, are the ones that will eventually be pitted against human players and so must become as robust as possible. So in addition to the main versions of AlphaStar there are also so called "main exploiter" versions that play only against the main versions (which are selected with fixed probabilities) and whose purpose is to identify potential exploits in the main versions.[376] The third version of AlphaStar, the "league exploiters," also utilize a PFSP mechanism and their purpose is to find systemic weaknesses of the entire league and so are not targeted by the main exploiters; the league

[373] Vinyals et al. 2019, 352.
[374] Ibid.
[375] Different units in *StarCraft II* are "invisible" and can only be engaged when certain other units/structures, "detectors," are in the vicinity. Players lacking any detectors will therefore almost certainly lose to a Dark Templar rush, for example.
[376] Vinyals et al. 2019, 352.

exploiters only play against the main versions.[377] As the different versions of AlphaStar train, they are periodically frozen and reinitialized as a new player in order to increase the diversity of strategies in use in the league.[378]

5.3 *StarCraft II* and Machine Ethics

A detailed discussion of *StarCraft II* and AlphaStar in the preceding sections was necessary to make the following claims apparent. First, *StarCraft II* is an incredibly complex game. Second, it is possible for machines, e.g., AlphaStar, to reach expert human levels of play using augmented RL techniques. Not only could AlphaStar play *StarCraft II* at expert human levels, but the use of RL techniques allowed it to innovate and develop novel strategies and play styles. Despite the complexity of *StarCraft II*, many expert human players consider the very early stages of the game to be well mapped out, i.e., there is a standard of play considered most optimal. Yet AlphaStar often deviates from this standard by, for example, building its first expansion at 17 supply instead of 16 supply when playing as Zerg. AlphaStar has also displayed its own novel strategies by, for example, staying on tier one technologies when playing as Zerg far longer than expert human players prefer.[379] Third, *StarCraft II* and ethics are relevantly similar, that is, ethics is more like *StarCraft II* than like chess, at least in the following respects: ethical decision-making and behaviour occurs in real time and with imperfect information. The researchers working on AlphaStar note the significance of their achievement writing:

> Like StarCraft, real-world domains such as personal assistants, self-driving cars, or robotics require real-time decisions, over combinatorial or structured action spaces, given imperfectly observed information. Furthermore, similar to StarCraft, many applications have complex strategy spaces that contain cycles or hard exploration landscapes, and agents may encounter unexpected strategies or complex edge cases when deployed in the real world. The success of AlphaStar in StarCraft II suggests that general-purpose machine learning algorithms may have a substantial effect on [i.e., an advantage when applied to] complex real-world problems.[380]

---

[377] Ibid.
[378] Ibid.
[379] See professional player Harstem comment on AlphaStar's idiosyncrasies in the following videos: https://www.youtube.com/watch?v=P9MhNc4UsKc&list=PLbVNzAA7sXzB61Yd6g3OvsHer2JT_l964&index=3 https://www.youtube.com/watch?v=FGciv16op94&list=PLbVNzAA7sXzB61Yd6g3OvsHer2JT_l964&index=5
[380] Vinyals et al. 2019, 353.

Many of the general techniques used to create AlphaStar could be adapted to raise ethical machines and highlight how many of the challenges of using RL techniques identified earlier in this and the previous chapter could be addressed.

In contrast to chess and indeed in contrast to many contrived ethical thought experiments that are meant to pump philosophy students' intuitions,[381] real-world ethical decision-making and behaviour is not accomplished with perfect information. Behaving ethically would, presumably, be easier if one had access to all of the relevant non-normative and/or normative information prior to making a decision; making ethical decisions is difficult precisely because one often possesses only a fraction of the relevant information. Consider that someone responsible for hiring a new employee might find themselves having to choose between two candidates whose qualifications are identical. How ought the new hire be chosen, i.e., on what basis should one be preferred over the other if they are both equally qualified? Such a decision is made difficult because of a lack of access to all of the relevant information (e.g., the candidate's character, how well they might fit with the organization and broader community, which normative ethical theory is true, etc.) that might make the decision easier and, more importantly, the ethically preferable decision. Just as people need to make ethical decisions using imperfect information, so too must players in a game of *StarCraft II* make decisions based on the information at hand. Whether decisions made in the moment will result in a winning outcome for the player is often not known. Similarly, whether decisions made in the moment are ethical or not is also often not known until some point in the future. Insofar as we are concerned with machine ethics, RL methods are clearly capable of teaching a machine how to guide itself through a state space via a sequence of sensible actions when given imperfect information.

Ethical decision-making and behaviour, like *StarCraft II*, also contains cycles (i.e., related "strategies," as it were) and hard exploration landscapes (i.e., a vast action space that would be fruitless to explore naively). The latter feature is quite obvious if we imagine how a "blank slate" human (analogous to the randomly weighted neural network present in a machine prior to any training) might approach trying to grocery shop ethically, for example. If we place this human just inside the entrance of the grocery store, it is clear that there is a vast sequence of potential actions, both in terms of breadth (i.e., there are many branches) and in terms of depth (i.e., each branch terminates after many steps), they would need to explore before stumbling upon a sequence that results in them ethically shopping for groceries. Recall that, as mentioned in section 5.1, there are simply too many different possible

[381] I supposed it was inevitable that I would mention the Trolley Problem at some point...

sequences to explore naively, i.e., randomly and without any guidance. This applies to games like Go and *StarCraft II*, but also more generally to ethical decision-making and behaviour, hence the notion of hard exploration landscapes.

Ethical decision-making and behaviour is also cyclic in two important senses. Like the cycling of strategies in *StarCraft II*,[382] ethical decision-making and behaviour can be thought of as cyclic. Consider the example of an autonomous vehicle and a scenario in which it is in an unavoidable collision. Recall also, from section 4.2 in Chapter 4, the preferences people have concerning autonomous vehicle behaviour. It may be that (A) sparing more lives is preferred over sparing fewer lives, (B) sparing human lives may be preferred over sparing animal lives, and (C) sparing young lives may be preferred over sparing old lives. But even though (A) is preferred over (B) and (B) is preferred over (C), (A) is not preferred over (C), as might be the case where an autonomous vehicle collides with either two children or four adults. Just as there is no one optimal strategy in *StarCraft II*, so too is there no one "optimal strategy," as it were, when it comes to ethical decision-making and behaviour.

Ethical decision-making is also cyclic in a second broader sense akin to episodicity. While ethical decision-making and behaviour is not, in theory, an episodic task (as mentioned in section 5.0.2), many activities that involve ethical decision-making and behaviour are *practically* episodic in nature. Consider again the example of ethically shopping for groceries. It may be practical, and potentially even advantageous in the field of machine ethics, to think of ethically shopping for groceries as an episodic task, i.e., as a task with an "end" and "reset" to some initial state, that end being the exiting of the grocery store and the reset being the reentering to begin shopping anew. Indeed for those of us living through COVID-19 lockdowns (as I was), the pandemic has brought to the fore how many human activities can be thought of as practically episodic in the same way that games, like *StarCraft II*, are also episodic (i.e., a game of *StarCraft II* ends at some point and is then reset to an initial state).[383] I, for example, leave my home to shop for groceries and then return home. I leave my home to pick up medicine at the pharmacy and then return home. I leave my home to run around the neighbourhood for exercise and then return home.

Finally, while there is a clear global goal in *StarCraft II*,[384] the use of RL techniques in machines like AlphaZero and AlphaStar demonstrate how vague concepts like intuition, creativity, and optimal

[382] See footnote 367.

[383] Many thanks to my supervisor, Dr. Joel Walmsley, for that poignant observation.

[384] Recall that the objective is to eliminate all of the opposing player's structures. In case it was not made clear earlier, a game of *StarCraft II* will immediately end if one player successfully destroys all of the other player's buildings, even if that player still has units left to fight with. This is, however, a rare occurrence. In

strategy can be operationalized. So despite the fact that the reward problem remains a challenge when using RL techniques, there is, in principle, no reason why a concept like harmonious integration could not be operationalized given the right kind of training regime and given an appropriate reward signal. There is, in short, no need to *define* a notion of "harmonious integration" (or intuition, creativity, etc.) because such training would amount to an *operationalization* of the concept of harmonious integration. AlphaStar's training, and in particular the use of population-based PFSP, suggests that the idea of using a virtual community of moral agents to raise an ethical machine is practically feasible and even desirable given the fact that machines trained in such a fashion could rediscover and improve upon ethical norms, in addition to shedding light on concepts like character, virtues and flourishing.

Ultimately, research in the domain of game-playing is illuminating how RL methods can be applied to raise robust ethical machines capable of responding appropriately in real-time to complex environments given imperfect information. Moreover, AlphaStar and research on *StarCraft II* highlights something worth emphasizing, namely that machines might be raised to exhibit a certain "character" and that it may be possible to cultivate a virtuous character in machines, an approach that is often not taken in machine ethics (i.e., taking a virtue-ethics approach as opposed to a consequentialist or deontological approach). Implementing machine ethics will therefore amount, in a very real sense, to raising machines as one might raise a child to live an ethical life. It is certainly true that ethical decision-making and behaviour involves *global* goals or outcomes that are difficult, if not impossible, to explicitly define and translate into formal mathematical reward signals. But for practical purposes *local* goals and outcomes can be explicitly defined and even translated into formal mathematical rewards that capture ethical decision-making and behaviour or a virtuous character. Similarly, it is not necessary to define what exactly ethical decision-making and behaviour looks like before one is able to proceed with the business of implementing machine ethics. Just as explicit definitions of creativity and innovation are not needed to say that what AlphaZero is doing when playing games of chess or Go, for example, is creative and innovative, so too are explicit definitions of harmonious integration or ethical behaviour, for example, not needed to say that what a given machine is doing is conforming to ethical norms or exhibiting a virtuous character.

short, it is far more common for a player to recognize that they have been defeated (and will eventually have all of their structures destroyed) and so players typically resign and exit the game even if they have units and structures remaining.

5.4 Risks of Implementing Machine Ethics

Before moving forward it is worth revisiting the question of whether it is even desirable to implement machine ethics. In short, while we *could* raise machines whose decisions and behaviours have a significant impact on human well-being, *should* such machines be raised? What are some of the risks associated with the project of machine ethics in general in contrast to conceptualizing all use of intelligent machines under computer ethics (i.e., human agents using intelligent machines as tools to affect human patients)?

5.4.1 Failure

There are, to be sure, many short-term risks associated with the raising of ethical machines. An obvious risk is that machines can fail or be corrupted.[385] There are myriad ways in which machines might fail or be corrupted. Failure for example could occur for any number of reasons ranging from a reliance on misleading perceptual information or training data to genuine mistakes that arise when a machine is confronted with a novel scenario. Machines often fail in the latter sense in ways that a human never would, i.e., "they [machines] may be liable to fail under circumstances where humans would usually not fail, or they may produce different kinds of errors than humans when they fail."[386] Generic image recognition machines, for example, are notoriously bad at correctly labelling sheep in an image unless the sheep are in a green field. As Janelle Shane notes, it is possible that during training the machine had mostly been shown images of sheep in green fields resulting in improperly labelled images when sheep were not shown in green fields.[387] Sheep in cars, in living rooms or sheep held in people's arms tend to get labelled as dogs and cats.[388] Sheep on a leash? Not according to machines! That animal must be a dog. Such failures, even ones that might be considered genuine mistakes, are invariably connected to the machine's training. More significant/pernicious failures, like the misclassification of black people as gorillas, will be taken up in Chapter 7.

[385] Cave et al. 2019, 569.
[386] Ibid.
[387] Shane 2019, 22.
[388] Ibid., 22 & 23.

5.4.2 The Effect of Data

Even machines that are trained using relatively clean datasets, i.e., datasets not filled with extraneous objects, may make mistakes if that data is not representative of the real world.[389] On the more humorous side, image recognition machines have a tendency to report observing giraffes in images that most certainly do not contain any giraffes because humans tend to love taking pictures of giraffes. As Shane explains, "though giraffes are uncommon, people are much more likely to photograph a giraffe than a random boring bit of landscape"[390] which results in an overrepresentation of relatively rare sights in image databases, a phenomenon that internet security expert Melissa Elliot has dubbed "giraffing."[391] Leading questions pose a similar challenge for machines; if you ask Visual Chatbot, for example, how many giraffes (or any animal it seems) it sees in a given image, it will respond with some non-zero number.[392] Again, Shane notes that this is likely a result of Visual Chatbot's training data which was generated by humans asking and answering questions.[393] And, obviously, people tend not to ask "How many giraffes are there?" when there are zero giraffes.

On the more serious side however, and what is more interesting from a machine ethics perspective, are machines that perpetuate systemic biases as a result of overrepresented images or other forms of data. Increasingly, unethical machines have been making headlines precisely because they exhibited questionable decision-making and behaviour after being trained on datasets that poorly reflect the real world. The overrepresentation of images from the United States used to train image recognition machines caused photographs of North Indian brides to be labeled as "performance art" and "costume" whereas the labels applied to photographs of US brides include "bride," "dress" and "wedding."[394] Even more worrisome, the data used to train some image recognition machines used in medicine to diagnose skin cancer, for example, overrepresent lighter shades of skin colour and significantly underrepresent darker shades, with some researchers noting that "fewer than 5% of these images are of dark-skinned individuals," increasing the probability that such machines will fail to correctly diagnose dark-skinned individuals.[395] The company Amazon similarly recently abandoned their

---

[389] Ibid., 127.
[390] Ibid., 127 & 128.
[391] See her Tumblr post: https://abad1dea.tumblr.com/post/182455506350/how-math-can-be-racist-giraffing
[392] Have a chat with Visual Chatbot here: http://demo.visualdialog.org/
[393] See: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/blogger-behind-ai-weirdness-thinks-todays-ai-is-dumb-and-dangerous
[394] Zou and Schiebinger 2018, 325.
[395] Ibid.

efforts to develop a recruiting machine that would sort incoming applications according to their employability, and for good reason.[396] Amazon's machine, trained as it was on resumés already submitted to the company, reflected back the gender bias prominent across the tech industry; their machine preferred applications with masculine terms and rejected applications with feminine terms. Recall from Chapter 2 that decisions to use past data, in this case resumés of people already hired to work at Amazon, to train machines often results in feedback loops that exacerbate systemic biases encoded in the training data, the bias against hiring women in this case.

5.4.3 Corruption

In addition to failing, machines are also famously corruptible. For example, a Twitter chatbot released by Microsoft, called "Tay," that was designed to learn from interactions with users (presumably via some machine learning techniques, although Microsoft released few details about how it worked) began tweeting about conspiracy theories and hurling racist messages about because, perhaps predictably, those were the kinds of messages people began tweeting at it.[397] On the more academic side, researchers have demonstrated how machines can be easily corrupted, whether by malicious designers, hackers or coding errors.[398] Researchers have demonstrated that wearing eyeglass frames with a particular printed pattern on them is sufficient to evade a facial recognition machine or cause the machine to misidentify the person.[399] While the failure or corruptibility of machines is not a problem unique to ethical machines, as Cave et al. highlight, the concern is that "ethical reasoning capacities may themselves be vulnerable to error and corruptibility" placing a heavy burden on those attempting to implement machine ethics to ensure that the very same techniques that would give a machine the capacity for moral decision-making *cannot* easily fail or be corrupted to produce unethical behaviour.[400]

---

[396] See: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

[397] See: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist
See also: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech

[398] Cave et al. 2019, 569.

[399] Sharif et al. 2016, 1528.

[400] Cave et al. 2019, 569.

5.4.4 Ethical Hegemony

Another general risk of implementing machine ethics is closely connected to the fact, mentioned above, that machines can perpetuate systemic biases. This issue is taken up in more detail in Chapters 2 and 7, but in short, there is a general risk that the project of machine ethics could result in a kind of ethical hegemony or value imperialism.[401] First, consider that there is no satisfactory resolution to the debate over ethical monism versus ethical pluralism, i.e., there is no satisfactory answer as to whether "there is always a definite fact about how to act [ethically], or what the best outcome is."[402] Second, consider that a given human will mostly have a limited sphere of influence.[403] Machines, in contrast, could be deployed en masse such that a single machine[404] may not have a limited sphere of influence. Autonomous vehicles across the globe could for example utilize the same software to resolve ethical dilemmas. Thus, these two considerations taken together, the danger is that whatever method of resolution employed in machines could be highly influential "resulting in something akin to value imperialism, i.e., the universalization of a set of values in a way that reflects the value system of one group" of people, like the programmers or investors.[405]

5.4.5 Abdicating Responsibility and the Automation Paradox

Yet another general risk of implementing machine ethics is the undermining or abdication of human agency and responsibility. While one benefit of machines and technology in general is the offloading of cognitive processes/capacities onto external entities, the downside of this kind of offloading is the potential weakening of human cognitive processes/capacities as a result of lack of use. Anecdotally, my ability to remember different phone numbers is severely limited because my smartphone does all the "remembering" for me. Similarly, there is no shortage of stories about hapless tourists or absent-minded drivers led astray by GPS systems perhaps because they trusted a machine more than they should have, an issue that I take up in detail in Chapter 7. In all of these cases, humans have been led to rely on a machine and then realize only once that machine has failed that they are

---

[401] Ibid.
[402] Ibid.
[403] Of course certain people could have much larger spheres of influence compared to other people, but the general point is that there is a limit to the influence a particular human's actions has on other humans before the effects of those actions become too diluted to quantify.
[404] Recall that I am using the term 'machine' broadly to include, for example, algorithms.
[405] Cave et al. 2019, 570.

poorly prepared to handle the task that that machine was performing. Such a reliance on machines can lead to devastating consequences, as was the case with Air France flight AF 447 which was also discussed in Chapter 2.[406] After the onboard autopilot failed, the pilots were forced into the position of assuming manual control in a situation that they had rarely experienced handling; the autopilot, after all, does almost all of the actual flying. Analogously, there is the risk that as more ethical decisions are delegated to machines, humans will lose their ability to engage in ethical decision-making. Known generally as the "automation paradox," there are three elements to this problem that make it particularly dangerous when considering ethical machines: (1) machines that operate autonomously automatically correct mistakes and thereby "accommodate incompetence,"[407] (2) relevant human skills are eroded as a result of lack of use, and (3) machines often fail (as detailed above) in unusual or complex situations meaning humans will be required to intervene in the most trying circumstances.[408]

As Cave et al. note, all three of these elements of the automation "paradox" have a direct bearing on the relationship between humans and machines engaging in ethical decision-making. The first element is relevant in situations in which a machine is making decisions on its own or is acting as a decision support system for a human. Physicians around the world, for example, may soon be aided by diagnostic machines or healthcare robots that ensure that deficiencies in the humans' diagnostic and ethical reasoning skills are masked in most normal circumstances just as "GPS-navigational assistants ensure that deficiencies in a human's own navigational skills do not come to light - except when the system fails."[409] The second element of the automation paradox, the risk of skill erosion, is relevant in cases where most of an activity, including decision making, is delegated to a machine and is particularly pernicious when machines are intended to function at superhuman levels. Since "moral reasoning is indeed a skill [as] evidenced by the extent to which it features in the socialization and education of children" it is imperative that humans continue to exercise this skill, even if machines are able to outperform humans.[410] Failure to do so will result in the erosion of our moral reasoning skills. The third element compounds the first two: if machines tend to fail in unusual or complex situations then humans will be forced to assume responsibility in cases where those humans lack the skills to exercise appropriate ethical decision-making (because those humans never developed those skills in the first

[406] See Chapter 2 for more details about Air France flight AF 447.
[407] See Tim Harford's 2016 article, Crash: how computers are setting us up for disaster. *The Guardian*. https://www.theguardian.com/technology/2016/oct/11/crash-how-computers-are-setting-us-up-disaster
[408] Cave et al. 2019, 571.
[409] Ibid.
[410] Ibid.

place) or because those skills have been eroded due to lack of use (because the machine does the vast majority of the decision-making). In short, it is possible that ill-prepared humans "will have thrust upon them, potentially at very short notice, exactly those moral decisions that are most difficult," a nonideal state of affairs to be sure.[411]

5.4.6 More Moral Patients

A final short-term risk that bears mentioning, and one which connects nicely with longer term risks, is the risk of creating moral patients. In general, since such a discussion is mostly outside the scope of this dissertation, ethics and moral philosophy is concerned, in part, with the relationship between moral agents and moral patients. Recall the discussion of moral agency and moral patiency from Chapter 3. What ought I, the moral agent, do for my elderly neighbour, the moral patient, during the COVID-19 lockdown in my country? This question is far more interesting, for lack of a better word, than questions like, what ought I do for my computer stranded in my office during the COVID-19 lockdown? The latter question is uninteresting from an ethical point of view[412] because my computer is not a moral patient, i.e., it is not the type of entity towards which I, as a moral agent, have responsibilities.[413] But what if my computer, indeed what if certain machines *are*, or should be included in the category of, entities towards which I have responsibilities? Whether certain machines ought to be considered moral patients is a hot topic of discussion with positions ranging from the outright rejection that machines ought to be thought of as moral patients to the sympathetic acceptance that certain machines might already qualify as moral patients.[414] The danger, therefore, is that in the process of implementing machine ethics, machines that qualify as moral patients are created that humans (as moral agents) consequently become responsible for. This is problematic because humans, perhaps unsurprisingly, tend not to respect machines. HitchBOT for example, a beer cooler-turned-hitchhiking robot with pool noodles for appendages and a bright LED smile, relied solely on the kindness of humans to help it get from coast to coast across Canada.[415] Unfortunately for HitchBOT, its journey across the United States beginning on

---

[411] Ibid.
[412] Though perhaps not from a pragmatic one if, for example, I need my computer to continue writing this dissertation!
[413] See Chapter 3 for a more detailed discussion of moral agency and moral patiency.
[414] See, for example, the popular philosophy pieces written by Tim Crane (https://iai.tv/articles/the-ai-ethics-hoax-auid-1762) and Thomas Metzinger (https://iai.tv/articles/why-we-should-worry-about-computer-suffering-auid-1761) respectively.
[415] See: https://www.cbc.ca/news/canada/british-columbia/hitchbot-completes-6-000-km-cross-canada-trip-1.2739128

the east coast in Massachusetts ended early in Philadelphia where it was found destroyed.[416] In addition to the variety of ways machines might affect humans, it is also worth investigating how humans might affect machines.

5.4.7 Speculative Risks

Setting aside the question of whether there exist now machines that ought to count as moral patients, there are significant long-term risks associated with creating machines that *would* qualify as moral patients. In particular, there may be huge costs imposed on humans, as responsible moral agents, stemming from the requirement to take the interests of machines with moral patiency seriously. In our increasingly technological society, such a shift could be hugely disruptive if intelligent machines continue to be integrated into economic, healthcare, educational, military and industrial systems, to name a few. Humans "might not be able to use such machines as mere tools or slaves, but might have to respect their autonomy, for example, or their right to exist and not be switched off" and humans may even have to share resources and certain privileges with machines, "for example, by giving suitably advanced AI's a right to vote, or even a homeland of their own."[417]

Other long term and more speculative risks associated with the project of implementing machine ethics are largely also connected to the project of creating artificial general intelligence. Among these risks is the potential for totalitarian control by a sufficiently advanced machine intelligence which, *prima facie*, is not necessarily a bad thing, but also not necessarily a good thing either. Moreover, the risk of value imperialism is likely to arise in such a potential state of affairs. Another more speculative risk is the potential for the perpetration of so-called "mind crimes" which occur if simulated entities that ought to count as moral patients are treated unethically.[418] While worth mentioning, a detailed discussion of long term and speculative risks associated with the project of creating artificial general intelligence is outside the scope of this dissertation. So too is a detailed discussion of the long term and speculative benefits of implementing machine ethics, but I will mention some of those benefits to conclude this chapter before exploring some of the more realistic short-term benefits of implementing machine ethics in the next chapter.

---

[416] See: https://www.cbc.ca/news/science/hitchbot-destroyed-in-philadelphia-ending-u-s-tour-1.3177098
[417] Cave et al. 2019, 570.
[418] Bostrom 2014, 205.

5.5 Speculative Benefits

Over the long term, the project of implementing machine ethics could drastically enhance human welfare. Ethical autonomous vehicles could, for example, save thousands of human lives every year. Machines in the healthcare system could similarly save thousands of human lives and ameliorate the suffering of millions more by reducing the number of misdiagnoses by human physicians. Further there are strong economic incentives to use autonomous intelligent machines to optimize pretty much everything. Supply chain management, resource exploration, extraction and allocation, city planning, financial planning, ecological preservation etc., could all be improved by intelligent ethical machines. This so-called "amplifying effect" is just one benefit of developing ethical general artificial intelligence, i.e., advanced intelligent machines could contribute to technological growth and development in other fields.[419] There is, in short, every reason to think that *if* ethical generally intelligent machines are created then humanity will be able to collect its cosmic inheritance,[420] that is, humanity will be able to access the unbelievably vast amount of matter/energy in the cosmos with which we could essentially do anything.[421] All that said, I must caution the reader that all of these benefits are highly speculative and will likely only ever materialize (if they ever do) in the long term (e.g., not within the next century).

As interesting as it is to continue speculating about ethical general artificial intelligence, such a discussion is beyond the scope of this dissertation. Regardless of whether it is possible to even create general artificial intelligence, there are short term benefits of implementing machine ethics that deserve our attention. Moreover, these short-term benefits can be separated into two broad categories: the benefits that arise as a result of *using* machines in which ethics has been implemented, and the benefits that arise as a result of *pursuing the project* of implementing machine ethics, even if it turns out that raising ethical machines is too difficult to accomplish. It is to these benefits, but primarily the latter, that I turn to in the next chapter.

[419] Ibid., 232.
[420] Bostrom calls this humanity's cosmic endowment. See, Bostrom 2014, 101-103, for more information.
[421] Consult any number of science fiction books and movies if you need help imagining what such a civilization would be capable of doing.

Chapter 6 - Morality Mirrored in Machines

6.0 The Benefits of Using Ethical Machines

There are many long-term benefits that might arise from the use of ethical intelligent machines. Some of the long-term benefits mentioned at the end of Chapter 5, for example the use of intelligent machines in the healthcare system, might only be considered long-term benefits (i.e., benefits that have not quite materialized in the present) because they have yet to be deployed across the globe en masse. As mentioned in Chapter 3, autonomous intelligent machines are currently used for many different tasks, some of which include object identification in images, facial recognition systems, speech transcription, content filtering on social networks and selecting relevant results for web searches.[422] Furthermore, while many of the different machines just mentioned do use some form of machine learning, others are just examples of beneficial technologies or algorithmic technology in general. The latter, beneficial technologies in general, are not the focus of this chapter.[423] Although much of the literature tends to focus on the risks and challenges associated with the use of intelligent machines (facial recognition systems are a particularly volatile topic), my aim in this chapter will be to focus predominantly on the short-term benefits. My focus will primarily be on the benefits of using ethical machines and, more importantly, the benefits of pursuing the project of implementing machine ethics, i.e., pursuing research and development of ethical intelligent machines. There is third a disjunct to be made, namely the benefits associated with theorizing about ethical machines, but I will only tangentially comment on these benefits (and indeed they largely fall under the benefits of pursuing the project of implementing machine ethics). In what follows the focus will be on the kinds of machines that have been under discussion throughout this dissertation, i.e., learning machines, and the immediate benefits of using such machines that have been raised ethically.

6.1 Humans, Machines and Automobiles

To use an oft cited example, the use of intelligent autonomous vehicles could significantly enhance human welfare. This is because in the overwhelming majority of automobile accidents, the

[422] LeCun, Bengio and Hinton 2015, 436.
[423] I refer the reader to Chapter 3 for more details about the field of computer ethics, one focus of which is the beneficial uses of different technologies.

cause of the accident was human error. Researchers at the Institute for Research in Public Safety at Indiana University found that human factors were the most frequent causes of automobile accidents in the United States at 93% followed by environmental (34%) and vehicle factors (13%) respectively.[424] Leading human factors that directly caused accidents included improper lookout, excessive speed, inattention, improper evasive action and internal distraction.[425] And humans have not improved since this analysis was conducted in 1979. Data collected by the National Highway Traffic Safety Administration in 2016 attributes 94% of automobile fatalities to some type of human error including improper restraint in a vehicle, excessive speeding, alcohol impaired drivers, distracted drivers and drowsy drivers.[426] By contrast, intelligent machines do not err in the same ways that humans do. Machines can "see" more than a human ever could, are never tempted to speed, are never inattentive, are never impaired or drowsy, can plan out and execute optimal evasive maneuvers to hundreds of anticipated futures, and are never distracted by belligerent passengers along for the ride.

Indeed these advantages are borne out by evidence gathered from autonomous self-driving vehicles operating in the United States today. Waymo, a subsidiary of Google's parent company Alphabet Inc., has been operating an autonomous ride-hailing service since late 2018 in the Metro Phoenix, Arizona area, and in 2020 released a safety report detailing all of the accidents involving autonomous Waymo vehicles. In over 6.1 million miles of autonomous driving, Waymo vehicles were involved in 18 accidents, none of which resulted in severe or life-threatening injuries.[427] Of the eight most severe accidents, all were the result of road rule violations and errors on the part of other human drivers. These errors included other humans driving on the wrong side of the road (yes, really), speeding in excess of 20 miles per hour over the posted limit, driving through a red light, driving through a stop sign, and various failures to properly yield.[428] Even the less severe accidents were the result of errors on the part of other human drivers who, for example, reversed into or rear-ended the Waymo vehicle at low speeds.[429] In the one collision involving a Waymo vehicle and a pedestrian, the pedestrian walked into the side of the stationary Waymo vehicle.[430] While the technology is not perfect, it is, frankly, indisputable that, when properly applied, intelligent ethical machines will have an overwhelmingly positive impact on human welfare. Not once, in all of the 6.1 million miles analyzed, did a Waymo

[424] Treat et al. 1979, vii.
[425] Ibid.
[426] National Highway Traffic Safety Administration, 2016 Fatal Motor Vehicle Crashes: Overview 2017, 6.
[427] Schwall et al. 2020, 1.
[428] Ibid., 11-12.
[429] Ibid., 7-8.
[430] Ibid., 6-7.

vehicle drive off the road or hit a stationary object like a parked car.[431] Moreover, in the event that Waymo vehicles experience some sort of failure preventing it from driving properly, it is able to come to a complete stop to ensure passenger and bystander safety. Unfortunately, details about how exactly machine learning is incorporated into Waymo vehicles is not readily available, and this is likely intentional in order to protect intellectual property. It is however probable that Waymo vehicles and other autonomous vehicles use some form of reinforcement learning and deep neural networks.[432]

While the technology is far from perfect, intelligent autonomous vehicles that utilize machine learning (as Waymo's vehicles do) can significantly benefit human welfare now and in the immediate short term. Moreover, just as AlphaStar learned to play *StarCraft II* through fictitious self-play, so too do Waymo vehicles learn how to drive by simulating different potential futures and then planning an optimal route based on the most likely future predicted to occur. The result is that the Waymo vehicles can even anticipate via simulation different types of accidents and thereby avoid serious collisions in order to ensure the safety of the passenger(s).[433]

6.1.2 Raising Ethical Drivers

Insofar as Waymo vehicles are able to successfully perform the entire dynamic task of driving, they might be considered ethical intelligent machines or, in Moor's terminology described in Chapter 3, explicit ethical agents. Just as it is reasonable to say that a human did something *wrong* if they drive through a red stop light, it is equally reasonable to say that an autonomous vehicle did something *right* by stopping at a red light.[434] But what exactly, in the context of driving, ought to be considered right and wrong? Well, as a result of the explication problem,[435] there is no completely satisfactory answer to this question. In trying to capture the ethical concept of "good driver," the National Highway Traffic Safety Administration (NHTSA) outlines 28 core behavioural competencies that includes, for example, detecting traffic signals and stop/yield signs, responding to traffic signals and stop/yield signs, navigating intersections and performing turns, perform low-speed merges and detecting and navigating a parking lot. In addition to these core behavioural competencies, Waymo has added 19 extra behavioural

[431] Ibid., 6.
[432] See, Wang et al. (2021).
[433] See the *Waymo Safety Report* from February 2021 for more information.
[434] I am, for now, ignoring a full discussion of responsibility vis-a-vis the consequences of the behaviour of autonomous machines.
[435] See Chapter 4 for more information.

competencies including detecting and responding to animals, detecting and responding to motorcyclists, making appropriate reversing maneuvers and navigating railroad crossings.

Interestingly enough, in the process of building intelligent autonomous vehicles, we are forced to examine more closely just what exactly makes a human a good driver when operating an automobile. Moreover, for most people, judging whether another human is a good driver is largely a heuristic process rather than an algorithmic one.[436] That is, humans judge that another human is a good driver i.e., good in a moral sense, (or a bad one) largely based on rules-of-thumb, e.g., the driver is not behaving recklessly, is keeping a safe distance from other vehicles, etc. It is not enough, on the other hand, to judge machines in the same way. Judging whether a machine is a good driver, i.e., good in a non-moral sense, is largely an algorithmic process, i.e., a process involving the satisfaction of multiple simultaneous softly constraining and of well-defined criteria, and nowhere is this more obvious than in the documents and standards outlined by regulative and legislative bodies, among others, concerning autonomous vehicles.

Implicitly, these criteria were stated as far back as the DARPA Grand Challenge. In the 2004 version of the challenge, autonomous vehicles were required to navigate an off-road course largely following Interstate 15 on what was supposed to be a long 228km trek from Barstow, California to Primm, Nevada. The most successful vehicle, Red Team, only managed to travel 12km (7.4 miles) and, while attempting to navigate switchbacks in a mountainous section, the vehicle veered off course and became stuck on a berm which is where its journey ended.[437] Obviously, good drivers are ones that do not veer off of the road. Good drivers similarly do not drive the wrong way (as Axion Racing did), do not flip their vehicle while making turns (as Team ENSCO did), do not drive into wire or fences (as Team CIMAR and Team Caltech respectively did), and do not become paralyzed by bushes near the road (as Team TerraMax did).

Explicitly, the International Society of Automotive Engineers (SAE) developed a well-defined taxonomy and definitions for terms related to autonomous driving systems culminating in their most recent revision published in April 2021.[438] Briefly, their taxonomy consists of six discrete and mutually exclusive levels of driving automation that outlines both the roles of the human in the vehicle and the driving automation system. Levels 0 - 2 correspond to partial driving automation at most, i.e., sustained lateral and longitudinal vehicle motion control by the automated driving system, with the human

[436] Unless you happen to be friends with a driving instructor.
[437] Grand Challenge 2004 Final Report 2004, 8.
[438] See, "SAE J3016 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles" April 2021, for more information.

responsible for monitoring both the automated driving system and the external environment (e.g., the human is responsible for proper object and event detection and response). Level 3 corresponds to an intermediary level of automation where the automated driving system, while engaged, can perform the entire dynamic driving task including object and event detection and response. Importantly however this automation is conditional on the vehicle remaining within its operational design domain. Furthermore, the automated driving system at Level 3 is incapable of achieving a minimal risk condition, i.e., a stable, stopped condition, in the event of a system failure or in cases where the operational design domain limits are exceeded and so the human is expected to remain in a fallback-ready state in order to take over the dynamic driving task or to perform any emergency maneuvers.

Levels 4 and 5 correspond to fully automated driving in general (Level 5) or fully automated driving in an operational design domain (Level 4). At levels 4 and 5 all aspects of vehicle control are controlled by the automated driving system including achieving a minimal risk condition in the event of system failure or when operational design domain limits are exceeded. Humans riding in such vehicles, e.g., the Waymo vehicles discussed earlier (which operate at Level 4), are considered passengers and are not expected to engage in any part of the dynamic driving task. Unlike the vehicles that took part in the DARPA Grand Challenge, far more explicit criteria are used to analyze whether Waymo vehicles are good drivers. In short, by building intelligent autonomous driving systems we are forced to carefully and explicitly examine what makes a good driver in general, human or otherwise.

6.2 Benefits of Building Machines

Throughout history humans have built machines to perform various tasks and fulfill various functions, and in doing so we invariably learned more about just what exactly is involved in carrying out the task or function of interest. Consider the development of flying machines. After watching birds and insects flit around us for millennia, it at last seemed possible at the turn of the 20th century that humans could create flying machines to carry us through the skies. Designing and testing ornithopters, machines that fly by flapping their wings in an imitation of flying animals, seemed like the natural route towards success. And yet this was not the case. Ornithopters were wildly unsuccessful at achieving sustained flight. This is because, as we now know, it is not the flapping of a wing per se that is required for sustained flight, but the generation of a pressure differential between the air above the wing and below the wing that generates the lift necessary for flight. With the benefit of hindsight it is obvious that machines which imitated birds would not succeed in flying, and yet it was only in the process of building

such machines in the first place that we looked more closely at what makes a good flying machine and what exactly is involved in aerial flight.

6.2.1 Benefits of Building Thinking Machines

As with flying machines, the process of building thinking machines has forced us to examine more closely and carefully these things we call "thinking" and "intelligence" and, I want to claim, the same is the case whereby attempting to build *ethical* machines forces us to examine *ethics* more closely and carefully. As with flight, building machines that imitated human thought and intelligence seemed like the natural route towards building successful thinking machines, i.e., artificial intelligence (AI). And so beginning in the late 1950's pioneers in the budding field of AI set out to build machines that could think and solve problems like humans. Indeed this was an explicit goal of researchers at the time, like Newell, Shaw and Simon, who attempted to build a general problem-solving machine, i.e., a program, in part "to understand the information processes that underlie human intellectual, adaptive, and creative abilities."[439] However like ornithopters before them, these thinking machines were wildly unsuccessful at attaining anything remotely close to human-like intelligence. And this is because there is a danger here as well, that by attempting to mimic the "natural" phenomenon (e.g., flight or thinking) our "artificial" version mimics the wrong aspects.

With hindsight, the failure of so-called first-wave AI,[440] or good old-fashioned AI (GOFAI),[441] is not exactly surprising. First-wave AI was based on symbolic representation and was largely predicated upon the truth of four metaphysical principles: (1) "The essence of intelligence is thought, meaning roughly rational deliberation," (2) "the ideal model of thought is logical inference (based on "clear and distinct" concepts, of the sort we associate with discrete words)," (3) "perception is at a lower level than thought and will not be that conceptually demanding," and (4) "the ontology of the world is...formal: discrete, well-defined mesoscale objects exemplifying properties and standing in unambiguous relations."[442] After decades of painstaking work on first-wave AI systems, it has become increasingly obvious that despite the superficial similarities between human intelligence and machines built to mimic human intelligence, e.g., the use of heuristics to solve problems, the creation of machines with human-like intelligence would not arise from the use of first-wave AI techniques alone.

[439] Newall, Shaw and Simon 1959, i.
[440] Smith 2019, 7.
[441] Haugeland 1985, 112.
[442] Smith 2019, 7-8.

But failure is instructive.[443] Brian Cantwell Smith (2019) notes that the failures of first-wave AI can be summarized as of four main types that are worth exploring in detail. This is because out of the failure of first-wave AI arose second-wave AI which utilizes the kinds of techniques that have been the focus of this dissertation, i.e., machine learning techniques. The first type of failure Smith identifies is the neurological: "The brain does not work the way that GOFAI does." First-wave AI and almost all of today's mainstream computer systems, i.e., CPUs, consist of serial processes running at extreme speeds, e.g., performing 10^9 operations per second.[444] The functioning of the brain is, as far as we know, almost the complete opposite, using massive parallelism and operating roughly 50 million times slower.[445] Whether the imitation of these lower level neurophysiological structures is required for the creation of human-like intelligence in machines is an open question, but one that will become easier to answer as we attempt to build machines that utilize such structures, e.g., artificial neural networks and deep learning systems. It may turn out that, like higher level structures of human thought, e.g., the use of heuristics, imitating organic neural networks is not a fruitful approach to building artificial general intelligence, but we will never know if we do not try to build such machines.

The second type of failure that Smith identifies is the perceptual: "Most GOFAI theorists thought that perception - recognizing or "parsing" the world based on perceptual input from sensors - would be conceptually simpler than simulating or creating "real intelligence.""[446] As Ronald de Sousa remarks, "It is a pregnant irony that computers are now relatively good at some of the reasoning tasks that Descartes thought the secure privilege of humans, while they are especially inept at the "merely animal" functions that he thought could be accounted for mechanically."[447] Intelligent environmental navigation is much more difficult than first-waver researchers realized. Those working on first-wave AI clearly held a similar view of perception, especially given the fourth metaphysical principle concerning a formal ontology underlying most, if not all, research in first-wave AI. The prevailing thought that perception would just be a matter of determining what objects were "out there" turned out to be profoundly mistaken. When digital cameras were finally connected to computers in the early 1970s, researchers discovered that what was "out there" was actually a morass of stimuli. It turns out that many people, including Descartes and first-wave AI researchers, can fall prey to a kind of epistemic fallacy: just

[443] We might keep the anecdote about Thomas Edison failing in mind: when asked by his associate if he was disappointed with the lack of results from his experiments, Edison apparently responded by saying that he did have results, he now knew several thousand things that would not work.
[444] Smith 2019, 23.
[445] Ibid.
[446] Ibid., 24.
[447] de Sousa 1991, 176.

because it seems to us humans that the world consists of a simple arrangement of straightforward objects does not mean that the world *actually does* consist of a simple arrangement of straightforward objects. That the world does appear in such a way to our consciousness is the result "of an exquisitely sensitive, finely tuned perceptual apparatus running on a 100-billion neuron device with 100 trillion interconnections honed over 500 million years of evolution."[448] Machines, even the second-wave AIs of 2021, cannot hold a candle to such an apparatus.

The third type of failure is epistemological: "Thinking and intelligence, on the GOFAI model, consisted of rational, articulated steps, on its founding model of logical inference."[449] As it has been pointed out in numerous previous chapters[450] however, and by various other scholars besides, thinking and intelligence are less like a series of articulable explicit operations and more like a phenomenon that emerges from the coalescence of pattern recognition abilities and experience gathered from repeated interaction in dynamic environments.

The most serious issue with first-wave AI, connected to both the perceptual and epistemological failures, is the fourth and final ontological failure: "The misunderstanding about perception, and perhaps about thinking as well, betray a much deeper failure of first-wave AI: its assumption...that the world comes chopped up into neat, ontologically discrete objects."[451] Herein lies the crucial insight that sets first-wave and second-wave AI apart. Moreover, it is the failure of first-wave AI to produce anything sufficiently similar to human-like intelligence itself that spurred a reconsideration of both the metaphysical principles underlying first-wave AI as well as the nature of thinking and intelligence. It bears repeating: failure is instructive. The ontological failure of first-wave AI brought to the fore the idea that thinking and intelligence may have their roots in nonconceptual content rather than in explicit, articulable conceptual forms. This idea serves as the basis for many different contemporary currents in cognitive science, which themselves underwrite different approaches in second-wave AI, including enactivism, connectionism, and deep learning.[452] In short, by attempting to build intelligent machines given certain metaphysical assumptions about the nature of thinking and intelligence, *and then failing to create human-like intelligence* (or a general intelligence), we were forced to revise both how we approached building such machines and the metaphysical assumptions underlying that program.

[448] Smith 2019, 26.
[449] Ibid., 27.
[450] See, for example, Chapter 1 and Chapter 4.
[451] Smith 2019, 28.
[452] Ibid., 31.

The point is that by pursuing research on and building intelligent machines, and perhaps even failing to do so, we will learn something about intelligence and thinking *simpliciter*. I argue that the same can be said of ethics. That is, by pursuing research on and attempting to raise ethical machines we will learn something about ethics *simpliciter* and perhaps even have to make revisions to our ideas about ethics. Moreover, that we will learn something is true even if we fail to implement machine ethics.

6.3 Benefits of Building Ethical Machines

In contrast to first-wave GOFAI, second-wave AI is associated with "deep learning and affiliated machine learning (ML) technologies"[453] that make shallow (i.e., few-step) inferences by a massively parallel process (e.g., artificial neural networks inspired by organic neural networks) using massive amounts of data (e.g., the collection by Big Data of millions of photos) involving a very large number of weakly correlated variables (e.g., pixels in a digital picture that give rise to a comprehensible image).[454] Whether second-wave AI will lead to the creation of human-like or general intelligence is an open question, but the point of the preceding discussion has been in anticipation of this novel claim that I aim to defend throughout the rest of this chapter. By attempting to build ethical machines and pursuing the project of implementing machine ethics we will learn more about the nature of ethical decision-making and behaviour as well as the assumptions upon which theories of ethics and morality rest.

6.3.1 Weak Psychological AI and Ethics

How exactly will the process of building ethical machines help us learn more about the nature of ethical decision-making and behaviour? Building ethical machines is very much in the spirit of the cognitive science project, in particular, the weak psychological AI project outlined by Owen Flanagan.

> Research here is guided by the view that the computer is a useful tool in the study of the mind. In particular, we can write computer programs or build devices that simulate alleged psychological processes in humans and then test our predictions about how the alleged processes work...According to weak psychological AI, working with computer models is a way of refining and testing hypotheses about processes that are allegedly realized in human minds.[455]

[453] Ibid., 47.
[454] Smith (2019) neatly compares first-wave and second-wave AI along five conceptual axes, the latter of which are summarized here. For the full comparison, see Smith (2019), 49.
[455] Flanagan 1991, 241.

Assuming ethical decision-making and behaviour is something that is realized and has origins in the human mind, then the project of implementing machine ethics will invariably reveal something about these origins. Whereas weak psychological AI traditionally encompasses, for example, perception, memory, language, etc., my claim is that weak psychological AI could also be extended to ethical reasoning and behaviour. In fact, there are two claims here that ought to be distinguished. The stronger claim is that implementing machine ethics will reveal more about how ethical decision-making and behaviour are realized in *human* minds. The weaker claim is that implementing machine ethics will reveal more about how ethical decision-making and behaviour are realized in *a* mind, one that is not necessarily human. Ethical agency, we may discover, could be multiply realizable.

6.3.2 Suprapsychological AI and Ethics

Ethical decision-making and behaviour may also not necessarily proceed in the ways imagined by philosophers. Or, it may turn out that machines are better than humans when it comes to ethical decision-making and behaviour. This type of research is less like weak psychological AI and more like what Flanagan calls "suprapsychological AI" since it concerns different conceivable forms of intelligence (e.g., machine intelligence and not just human-like intelligence), or in our case, different conceivable forms of or approaches to ethics.[456] This is the view espoused by thinkers like Susan Anderson who maintains that "machine ethics research has the potential to achieve breakthroughs in ethical theory that will lead to universally accepted ethical principles" and that interaction with ethical machines "might inspire us to behave more ethically ourselves."[457] Here Anderson is gesturing towards a primarily deontological ethics, i.e., a theory of ethics that is grounded by principles concerning what one ought to do when confronted with some ethical dilemma. Humans, Anderson argues, are prone to unethical decision-making and behaviour as a result of five different tendencies that compromise our ability to consistently abide by deontological principles. Humans tend to get carried away by emotion, to behave egotistically, to unreflectively adopt the values of those around us, to lack good role models, and to seek instant gratification.[458] Machines, on the other hand, may not have such tendencies (indeed machines need not have such tendencies, e.g., we can purposefully create them such that they do not behave

[456] Flanagan 1991, 242.
[457] Anderson 2011, 524.
[458] Ibid.

egotistically) and as a result may help us overcome these tendencies or, at least, help us to understand what ethics would be like if these tendencies could be overcome.

Moreover, Anderson claims that machines "could lead to a breakthrough in coming up with ethical principles" that the rational person ought to universally accept and abide by.[459] How might machines do this? As Anderson points out, what machines "are good at is keeping track of lots of information," e.g., information about ethically relevant features in a given ethical dilemma, from which machines might be able to "discover through inductive reasoning principles that are consistent with this information."[460] These principles would be *the* ethical principles that one ought to abide by, at minimum, in a given domain (e.g., the domain of eldercare). These principles could even be revised with the addition of new information. In short, Anderson's hope is that machines might be able to come up with the principles that capture the agreed upon intuitions ethicists have when facing ethical dilemmas (i.e., if ethicists agree what ought to be done in the ethical dilemma) even though the ethicists themselves may not have come up with such principles yet.[461]

6.3.3 Ethics Is Hard

So building ethical machines could reveal that ethics involves abstracting complicated principles from different ethical dilemmas. Indeed raising ethical machines might reveal that the nature of ethical decision-making and behaviour is far more complicated and nuanced than one might expect. Although it is common to criticize an ethical theory if it appears to place a heavy burden on the individual to think or act in certain ways, there is no principled reason why ethical decision-making and behaviour should not impose strong obligations on people. Ours is a complicated world and there is no reason to suspect that ethics ought to be easy just because it would make things simpler for us to figure out what the right thing to do is. Implementing machine ethics might reveal that it is the nature of ethical decision-making and behaviour, at least in the 21st century, to be difficult and complicated to understand and enact. Indeed if Anderson is correct that machines could help us uncover general ethical principles (and the meta-principles discussed in Chapter 4 that adjudicate between conflicting principles) that we ought to abide by then those principles would, of necessity, be complicated ones and not the simple Kantian maxims discussed in introductory ethics classes (e.g., one should always keep one's promises). There is,

[459] Ibid.
[460] Ibid., 526.
[461] Ibid.

in short, no guarantee that ethics, like any other phenomenon in the universe, will easily lend itself to human understanding.[462] And this is all the more obvious if we remember that we are evolved biological beings. Even Anderson points out that the first three tendencies mentioned earlier (to get carried away by emotions, to act egotistically and to adopt the values of those around us) "are likely a result of our evolutionary history as a species."[463] We are, in a sense, fighting our own biology in an effort to engage in ethical decision-making and behaviour.

Implementing machine ethics might also reveal that ethics is fundamentally nebulous, dependent not only on time and place, but also on personality, culture, context, goals, and more. If AI makes philosophy honest, as Daniel Dennett stated, then machine ethics make philosophical ethics particularly honest.[464] An *ad hoc* patchwork of arbitrary values, principles and procedures, or purported universal truths accessible via pure reason or intuition are not particularly amenable to implementation in machines. And even if they are, we might not be impressed with the results. Implementing machine ethics requires a kind of clarity, a kind of transparent and straightforward honesty, that is not always advanced in philosophical ethics. The result is that those working on implementing machine ethics often just pick some ethical theory as a starting point. Anderson for example chose to adopt the "multiple *prima facie* duties approach" that W. D. Ross adopted because "ultimately one has to decide that a particular ethical theory, or at least an approach to ethical theory, is correct."[465] Others choose a more consequentialist approach to implementing machine ethics. As we saw in Chapters 2 and 4, Winfield et al. for example assigned "safety outcome values" to the consequences of a robot's actions ranging from 0 to 10, where 10 is the highest harm rating and 0 the lowest.[466]

Still others choose to wash their hands of such picking and choosing altogether. Verma and Rubin, for example, conducted a kind of meta-analysis of twenty different notions of algorithmic fairness on a "single unifying example of an off-the-shelf logistic regression classifier trained on the German Credit Dataset."[467] Is such a classifier fair? Unsurprisingly, they conclude that "the answer to this question depends on the notion of fairness one wants to adopt" and may largely depend on which

[462] It should be noted that I endorse (metaethical) moral realism. There are, in short, mind-independent moral truths. Importantly, this includes moral nihilism (the position that there is no moral truth) and moral constructivism/moral emotivism (the position that moral truths are constructed or emotionally felt by humans). I highlight these two positions because they are examples of cases where it would be trivially true that ethics would be understandable by humans.
[463] Anderson 2011, 528.
[464] Ibid., 527.
[465] Anderson 2011, 23.
[466] Winfield et al. 2014, 88.
[467] Verma and Rubin 2018, 7.

definitions of fairness are appropriate to a particular situation.[468] Fairness might mean conditional statistical parity in one situation but might mean overall accuracy equality in another situation. Of course, it is entirely possible that different situations will have ethically relevant features that give us reasons to think that we should adopt one definition over another, but obviously "more work is needed to clarify which definitions are appropriate to [a given] situation."[469] Philosophical ethicists have their work cut out for them. This is especially true if, as some ethicists believe, we ought to embrace ethical pluralism. That is, because there is "currently no common ground among moral experts as to which ethical theory to use," we may be free, beyond some core of moral norms,[470] to "act according to [our] particular moralities in [our] given community."[471] What we might therefore learn from the project of pursuing machine ethics is that beyond some core of moral norms (that, as revealed by machines, might be so complex that we cannot understand them when they are stated explicitly), ethical decision-making and behaviour is conditional on how exactly one explicates the ethical concept(s) of interest. Though this means making the ethical decision or taking the ethical action will not always be easy or widely agreed upon, my own view is that this is how ethics ought to proceed, i.e., from considerations of context and character which is also, incidentally, a way that machine learning techniques would be particularly good at discovering, replicating and potentially even improving upon.

The preceding points all seem to indicate that, contrary to what is sometimes asserted by computer scientists, ethics might actually be an *essentially* difficult and complicated activity. Perhaps it is in the nature of ethical decision-making and behaviour to be laboriously understood by, if not beyond the explicit comprehension of, most people. Importantly, this is not to say that people do not or cannot behave ethically, just that we might not understand the high-level abstract principles that characterize ethical decision-making and behaviour. Consider the similarities with language; we may not completely understand the high-level rules that structure communication but we nonetheless obey the rules. Indeed there is a connection here to the two senses of rule following highlighted in section 3.2.2 in Chapter 3. Namely, people might obey the rules, e.g., the rules of language or of ethics, in the sense that our behaviour can be described as following some rule(s) even if we are not rule governed, i.e., in the business of consulting the rules to generate our behaviour. There may be no consensus as to which ethical theory to use or what high level abstract principles structure ethical decision-making and

[468] Ibid.
[469] Ibid.
[470] Gordon (2020) lists norms such as one must not commit murder or rape or violate the human rights of others as examples of core universal norms.
[471] Gordon 2020, 150.

behaviour (i.e., we may not agree on a set of rules that ought to govern our behaviour), but we nonetheless follow the rules (most of the time…) and grok what it means to be ethical.

6.3.4 Assumptions Grounding Ethics

In addition to revealing more about the nature of ethical decision-making and behaviour, building ethical machines can also lead us to learn more about, and perhaps modify or reject, the assumptions upon which moral theories and ethical norms rest. It is quite common, for example, to insist on a consistent ethics. That is, consistency is thought of, often necessarily, as a fundamental assumption upon which theories of ethics are constructed. Anderson asserts that "we *cannot* accept contradictions in the ethics we embody in machines, and humans *should not* accept contradictions in their own or others' ethical beliefs either."[472] According to Anderson, this entails that in two ethically identical situations "an action cannot be right in one of the cases, whereas the comparable action in the other case is considered wrong."[473] Another fundamental assumption is the comparative assumption. In short, different outcomes are assumed to be ethically comparable. We can compare genocide, on the one hand, and a white lie told to a friend to spare their feelings, on the other, and come to the sensible conclusion that the former is far worse than the latter. But this is an extreme case, and it is not abundantly clear that it is always possible to compare two different situations and discern what the ethically preferable decision ought to be.

In experiments with autonomous vehicle behaviour mentioned in Chapter 4, Awad et al. pumped (via questions concerning a dilemma) the ethical intuitions of participants by having them specify which outcome, out of two unavoidable accident scenarios, they find preferable.[474] Perhaps unsurprisingly, they found that the strongest preferences were for sparing human lives over animal lives, sparing more lives rather than fewer lives, and sparing young lives rather than old lives.[475] However the question remains: are the different outcomes of a situation really ethically comparable? Can we legitimately compare and judge that it was wrong for the car to swerve and kill its passengers, say two parents and their child, instead of the pedestrians, say two grandparents and their grandchild, on the crosswalk? I have no ready answer to these questions, but believe that we may approach a satisfactory response to these, and other questions, by building ethical machines.

---

[472] Anderson 2011, 256.
[473] Ibid.
[474] Awad et al. 2018, 59.
[475] Ibid., 60.

To be clear, my claim is that pursuing the project of implementing machine ethics will have consequences for ethics *simpliciter*. I am not claiming that by raising ethical machines we will know whether it is more ethical, in an unavoidable car accident, to hit two parents and their child or two grandparents and their grandchild. I am also not claiming that pursuing the project of implementing machine ethics will help us answer any specific ethical questions, like whether it is better to spare young lives rather than old lives. I maintain that by pursuing the project of implementing machine ethics there will be consequences, generally, for ethics *simpliciter*. One of those consequences may be that we come to learn more about the nature of ethical decision-making and the assumptions upon which moral theories and ethical norms rest. More precisely, one consequence of implementing machine ethics is that explicating assumptions of moral claims will allow us to see the structure of moral arguments better and therefore reason about them better. Indeed there are many other assumptions upon which one might construct a moral theory or ground ethical norms. In addition to consistency and comparability, there are assumptions of universality and plurality. Gordon for example notes that "it seems reasonable to avoid extreme moral relativism and to accept a firm core of universal moral norms" given that we must take into account "the idea that many ethical issues could have equally good but different decisions."[476] Alternatively, one could assume the truth of some version of moral relativism or reject the pluralist assumption and instead adopt the view that there is one ultimate ethical theory.

Another consequence may be that, given accepted outcomes or goals, machines will develop their own body of decisions and/or actions that could inform human decision-making and action. This could be particularly useful when considering complex situations, e.g., what one ought to do to given global climate change. As discussed in section 1.6, machines are able to analyze high dimensional data better than humans can, and so just as humans have begun to imitate the decision-making and actions of machine in the domain of board game play, so too might we imitate machine decision-making and actions in other domains, e.g., domains which have a significant ethical dimension.

6.4 Machine Ethics, Moral Nihilism and Anthropocentrism

This line of reasoning, that building ethical machines will lead us to learn more about ethics, could potentially lead one to some rather pessimistic or nihilistic conclusions. Anthony Beavers for example worries that morality may be redescribed in the same way that "thinking" has been redescribed

[476] Gordon 2020, 150.

following Turing's (in)famous proposal to recast thinking in terms of the imitation game. Beavers points out that "research in artificial intelligence and cognitive science has pushed in the direction of interpreting "thinking" as some sort of computational process" that computers or machines in general are, in principle, capable of engaging in and humans actually able to engage in.[477] This is perhaps because Turing is quite explicit that he is thinking about digital computers as the machines that will engage in the imitation game. Turing writes, in a short section on the meaning of the term 'machine,' that "the present interest in 'thinking machines' has been aroused by a particular kind of machine, usually called an 'electronic computer' or 'digital computer.'"[478] Moreover, Turing maintains that "this identification of machines with digital computers, like our criterion for 'thinking,' will only be unsatisfactory if, it turns out that digital computers are unable to give a good showing in the game."[479] Now there are two important things to note here. This first is that one might draw the (potentially) pessimistic conclusion that if digital computers can think, i.e., machines can think, then humans, as thinking entities, must be some sort of machine. Indeed this type of language pervades Turing's discussion of thinking machines. When discussing the idea of digital computers, Turing describes them as being able to "carry out any operations which could be done by a human computer" which include operating according to fixed rules as the human computer does, storing information and executing various operations, all of which the human computer does, usually with the help of a pencil and paper.[480] The second thing to note is that Turing actually advances an empirical hypothesis, that it may be possible to *build a thinking machine.* If, in Turing's words, "it turns out that digital computers are unable to give a good showing in the [imitation] game," that is, *if we cannot build machines that can convincingly play the imitation games*, then thinking should not be redescribed in this way.[481]

6.4.1 Implementability

Even Beavers, despite being concerned primarily with moral machines, recognizes how building machines, e.g., moral machines or thinking machines, can shed light on the underlying theories guiding the construction of such machines. Just as immaterial souls and wise, rational aspects of our nature "seem to be bygones, left behind by scientific and computational conceptions of thinking and

[477] Beavers 2012, 333.
[478] Turing 1950, 436.
[479] Ibid.
[480] Ibid.
[481] Ibid.

knowledge," so too might ideas like moral duty, utility and virtue be left behind by machine ethicists.[482] This is because building machines, engineering something, requires a kind of precision not often found in philosophical ethics. "Fuzzy intuitions on the nature of ethics," Beavers notes, "do not lend themselves to implementation where automated decision procedures and behaviours are concerned."[483] Like Turing, machine ethicists (the optimistic ones anyways) advance both a philosophical and empirical claim. For the former the philosophical claim is roughly that thinking is essentially the display of certain abilities (e.g., the ability to hold a human-like conversation for however long on any given topic) whereas for the latter the philosophical claim is roughly that ethical decision-making and behaviour is essentially the display of certain abilities (e.g., the ability to justify an action by invoking ethically relevant features of a given situation). The empirical claim, for both those working on artificial intelligence and machine ethics, is that machines will be able to display such abilities (i.e., behave ethically and also perhaps be able to justify their actions with respect to some overarching ethical theory). What is important for our purposes is to note the connection between ethical theories and their implementability. Beavers writes that if, for example, "Kantian ethics cannot be implemented in a real working device, then so much the worse for Kantian ethics."[484]

Indeed, Kantian ethics, and deontology in general, does not appear to be implementable in an all-purpose (i.e., general and flexible) machine. This is for various reasons, some of which have been discussed previously in Chapter 4 when top-down approaches to implementing machine ethics were examined. Briefly, Kantian maxims, or ethical duties/principles, quickly begin to conflict as one considers a wider range of environments. This necessitates the creation of a meta-principle to deal with the inevitability of conflicting principles, a meta- principle that is inherently domain-specific, i.e., useful in only certain kinds of environments. Beavers highlights a related issue, the frame problem, and suggests that applying "Kant's categorical imperative in any real-world setting seems to fall dead before a moral version of the frame problem" given that maxims can be formulated as widely or narrowly as one wishes.[485] Recall Anderson et al.'s eldercare robot from Chapter 4. The robot's duty to maximize respect for autonomy is, as a result of their meta-principle, quite narrowly defined given that the duty to maximize respect for autonomy may only be violated in select few cases, e.g., when it is not time to remind their patient to take their medication and if the patient is still immobile after being first engaged and then warned that an overseer would be notified of their persistent immobility.

[482] Beavers 2012, 333.
[483] Ibid., 334.
[484] Ibid., 335.
[485] Ibid.

6.4.2 Anthropocentrism and the Gap Between Explicit and Full Ethical Agents

But we need not rush with Beavers headlong into nihilism. If "death by failure to implement looks imminent" for Kantian ethics and deontology, what of it?[486] Beavers appears to presuppose, as many philosophers and people in general do, that humans are somehow special or unique, set categorically apart from the rest of nature. There are two distinct lines of reasoning here: there is either something about us humans that sets us apart from all other entities (e.g., only humans are full ethical agents), or there is something about ethical decision-making and behaviour, or ethical reasoning, that precludes anything other than humans from engaging in such decision-making and behaviour or reasoning. Beavers reasons along the lines of the former by specifically arguing that there *is* something special about human beings, and that special something is our internal moral subjectivity, "that is, conscience, a sense of moral obligation and responsibility," or whatever it is that governs our behaviour.[487]

Both lines of reasoning however are not particularly persuasive, especially the former. Like those who objected (and some who still do object) to the possibility of artificial intelligence, it seems premature to maintain that humans are somehow special or unique and set categorically apart from the rest of the universe in certain respects, e.g., in intelligence and/or morality. Research into artificial intelligence and even research on animal cognition seems to suggest there is no *prima facie* reason to assume that human intelligence is *categorically* different. Far from a clash of intuitions, the burden of proof to demonstrate that humans are not continuous with the rest of the universe lies with proponents of such a view, like Beavers. Why should anything about humans, including our "moral subjectivity," set us categorically apart from the rest of the universe? Suffice it to say, no convincing proof has yet been offered.

That many people are convinced that humans are special or unique in this respect is a testament to the power of anthropocentric biases. From Willam Paley's oft cited argument from design for the existence of God to Aristotle's insistence on the teleology of all things, there is no denying that humans are biased towards thinking that the universe and our own abilities (e.g., cognitive, linguistic, moral, etc.) were designed for us.[488] Even Turing noted this tendency in his 'Heads in the Sand' objection writing that "we like to believe that Man is in some subtle way superior to the rest of creation" and that

[486] Ibid.
[487] Ibid., 338.
[488] Preston and Shin 2021, 943.

such a feeling must be "quite strong in intellectual people, since they value the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power."[489] But such anthro-teleological biases must be resisted. Failing to resist these biases leads us to erroneously conclude, as Beavers does, that there is something particularly worrisome about implementing machine ethics (or implementing machine intelligence), that notions of moral responsibility are "under attack," because there is the potential that research might reveal that, contrary to what humans have thought for millennia, "morality needs no internal sanctions."[490] There may, in short, be no distinction to be made between what Moor called explicit ethical agents and full ethical agents which were discussed in detail in Chapter 3. Beavers maintains that if ethics is to survive implementation in machines and into the future, then it is by embracing a very different conception of ethics than traditional ones, but this too is a mistake. To see why, we must examine that second strand of reasoning purportedly leading to ethical nihilism.

Instead of presupposing that there is something special or unique about humans, one might presuppose that there is something special or unique about ethical decision-making and behaviour itself such that only humans, or a special subset of humans, can engage in such activities. We might wish to distinguish, for example, between pragmatic decisions and ethical decisions, or pragmatic reasoning and ethical reasoning. For some scholars, ethical decision-making is just different in kind from other types of decision making, e.g., pragmatic decision-making. Ethical decision-making, and ethical reasoning more generally, appears to require a suite of competencies including "analogical reasoning, planning and plan execution, differentiating among precedents, using natural language, perception, [and] relevant-information search."[491] For McDermott, these competencies are clearly ones that only a full blown artificial general intelligence may possess.[492] Moreover, it appears that for McDermott, it is not ethical reasoning *per se* that is special, but rather the fact that it is the only form of reasoning that requires this unique conjunction of competencies. Pragmatic or prudential decisions according to McDermott, in contrast to ethical ones, are categorically different because "we have no trouble feeling the pull of the former, whereas the latter, though we claim to believe that they are important, often threaten to fade away, especially when there is a conflict between the two."[493] McDermott maintains that what is special about ethics, i.e., ethical decision-making, reasoning and conflicts, is that they involve a clash between

[489] Turing 1950, 444.
[490] Beavers 2012, 342-343.
[491] McDermott 2011, 93.
[492] Ibid.
[493] Ibid., 95.

what one ought to do (i.e., the normative force of ethical rules or intuitions) and one's own self-interested desires.[494] In short, while machines are excellent at optimization, "say, optimizing the ethical consequences of a policy and optimizing the monetary consequences of the water-to-meat ratio in the recipe used by a hot dog factory," engaging in such activities is not sufficient to establish that machines are therefore able to engage in ethical decision-making.[495] According to McDermott, to engage in ethical decision-making and behaviour is to essentially struggle against temptation to do the wrong thing, whilst sometimes succumbing to this temptation. And even granting that machines may have interests to allow for a struggle with temptation, McDermott maintains that "it still seems dubious that they [machines] will be *tempted* to cheat the way people are."[496] There is something about ethical decision-making, perhaps some intrinsic phenomenological drama as McDermott suggests, that appears to preclude machines from ever being able to engage with, at least until (if ever) artificial general intelligence is developed.

Though slightly more persuasive than arguing that humans are categorically different from non-human animals and machines, there is nevertheless something odd, to put it mildly, about insisting on a difference in kind between ethical decision-making and all other kinds of decisions. An immediate problem arises if we grant that there is a categorical difference between, say, pragmatic decisions and ethical decisions, and also suppose that a decision is either a moral one or a non-moral one. If ethical decisions are different in kind, then it must be possible to delineate between ethical and pragmatic decisions.[497] While this might be easy for limiting cases, there are plenty of examples of decisions that could be construed as either ethical or pragmatic. Just recently, Boris Johnson drew considerable criticism for choosing to fly to the COP26 climate conference held in Scotland instead of taking a train.[498] Arguably, Johnson succumbed to the temptation to take a far shorter trip by plane when the ethical decision should have been to take the train and contribute less to climate change. Equally however one might argue that Johnson's decision to fly was a purely pragmatic one. He is, after all, Prime Minister of the UK and therefore someone whose time is quite valuable and likely better spent getting where he

[494] Ibid.
[495] Ibid., 93.
[496] Ibid., 110.
[497] We encountered a very similar problem when thinking about carving up the space of ethical decision-making and behaviour into separate domains. See Chapter 4 for more details.
[498] See this article from The Guardian: https://www.theguardian.com/politics/2021/nov/01/boris-johnson-will-travel-home-from-cop26-by-private-plane

needs to be quickly. This shows that there are many decisions that are not so easily classified as either ethical or pragmatic.[499]

Even granting that ethical decisions are different in kind, multiple problems appear that are not easily dealt with. What role do interests ultimately play on this account of ethical decision-making? If, as McDermott suggests, ethical decisions involve a clash between ethical rules and our want to violate them to fulfill some selfish desire, what are we to make of cases where selfish desires align with what the ethical rules prescribe? Bill Gates for example is famous, in part, for contributing millions of dollars to improve the health conditions of people across the globe. Such generosity on the part of someone with so much wealth is obviously the ethical thing to do, and yet if this decision is not accompanied by some sort of internal struggle or drama it apparently, on McDermott's view, should not count as being an ethical decision.

Now this may not be entirely fair as McDermott does note that it is "tricky ethical decisions" that are intrinsically dramatic, and Bill Gates' decision to share a portion of his vast wealth does not appear to be particularly "tricky" (at least from the point of view of someone without millions let alone billions of dollars).[500] Nevertheless, this problem of the role of interests persists. Practically, it seems to make no difference to an outside observer whether there is a conflict between an agent's interests and what they ethically ought to do (i.e., in either case the person can be seen as doing what they wanted to do). Philosophically, it is not at all clear why ethical decision-making requires specific types of interests or uniquely human interests. Sure, the way humans approach making decisions "is shaped by the weird architecture that evolution has inflicted on our brains," but how does that cause us to have ethical problems?[501] Moreover, how do our evolved decision-making abilities have any bearing on the metaphysics of decision-making in general? If anything, the fact that a "computer's decision whether to sin or not will have all of the drama of its decision about how long to let a batch of concrete cure" means not only that machines could have ethical problems, but that they are (potentially) far more ethical than humans, precisely because "they would [not] treat these [problems] as different from any

[499] On a more fine-grained view, decisions can be both moral and non-moral. A particular decision might have both moral and non-moral reasons favouring one or another choice. In Johnson's case, moral reasons favoured the choice to take the train whereas pragmatic reasons favoured the choice to fly and, arguably, the pragmatic benefits of flying were judged as outweighing the immoral consequences of emitting greenhouse gases as a result of that air travel. Johnson's decision was immoral but pragmatic, but even this fine-grained view does not save McDermott. It might equally be said, of an autonomous vehicle for example, that its decision was immoral but pragmatic.

[500] McDermott 2011, 110.

[501] Ibid.

other difficulty in estimating overall utility," that is, estimating a pragmatic course of action.[502] Humanity has many difficult challenges ahead of itself that will require "ethical decision-making" and I, for one, could do without the drama.

In the chapter that follows, I (mostly) leave metaphysics behind to consider socio-political and legal implications of implementing machine ethics and developing autonomous artificially intelligent systems.

[502] Ibid., 110-112.

Epilogue - Ethics 3.0[503]

7.0 Being Ethical in the Age of Machine Ethics

In this final section I step back to describe the conceptual terrain from a higher level and suggest future research directions. Given the influence of values on science and engineering discussed in Chapter 2, a variety of approaches to machine learning (in addition to GOFAI techniques) canvased in Chapter 1 and myriad possible methods of implementing machine ethics examined in Chapters 3 and 4, there are two questions with which this final chapter will concern itself. The first is, to be blunt, so what? That is, the above considered, what is worth focusing on and why? The second question is, what are some socio-political responses to the increasing use of artificially intelligent and autonomous systems? As Annette Zimmermann notes, the algorithmic is political.[504] That is, machines are not developed free from political interests and concerns. Intelligent machines, autonomous systems and different methods of implementing machine ethics are not developed in a vacuum far from the influence of political, legal and social considerations. On the contrary, these technologies, like all technologies, are created and developed by powerful and influential groups often at the detriment of the marginalized and underrepresented.

George Santayana famously remarked that "those who cannot remember the past are condemned to repeat it."[505] And we are indeed repeating the past with autonomous systems and artificially intelligent technologies by neglecting to consider their effects on marginalized and underrepresented communities. The uncritical adoption and use of autonomous systems can have, and does have, negative impacts on the well-being of different groups of people. It is therefore worth focusing on what types of artificially intelligent systems are being created and to what end they are being created, i.e., *why* they are being created. To that end, it will be necessary to demonstrate that we *should* care about artificially intelligent systems of the kind that have been discussed throughout this dissertation. I turn now to consider why some think that autonomous and artificially intelligent systems ought to be thought of as no different from any other kind of technology.

---

[503] This is in reference to Max Tegmark's *Life 3.0* (2017).
[504] See her website: https://www.annette-zimmermann.com/
[505] Santayana 2011, 172.

7.1 Positions in the Debate on Autonomous and Artificially Intelligent Systems

To a large extent, the reason why autonomous and artificially intelligent systems are being created is because they are a convenient shortcut. Autonomous and artificially intelligent systems are, in one sense, just another labour-saving technology. Like all technologies stretching as far back as the simple machines (e.g., the inclined plane, lever, wedge, etc.), artificially intelligent technologies enable us to solve problems by disclosing preferable alternatives. It is therefore common to think of artificially intelligent systems as just the newest tool in our toolkit, one that is not dissimilar to any other technology. This is precisely the view espoused by Joanna Bryson. According to Bryson, artificial intelligence technologies are not anything particularly special given that the qualifier "artificial" simply denotes that "something has been made through a human process."[506] It follows, so Bryson maintains, that by default humans are responsible for such technology. For Bryson, "artificial intelligence only changes our responsibility as a special case of changing every other part of our social behaviour" because "no fact of either biology or computer science names a necessary point at which human responsibility should end."[507] For Bryson, the concept of interest is not artificial intelligence but rather responsibility. There is simply no point in discussing how a machine could ever be "responsible" for something because the concept of responsibility or of being responsible is something that only humans can explicitly communicate about.[508] As designed artifacts, machines, even autonomous and intelligent ones, will never be held responsible or accountable for anything.

What matters, at least for Bryson, is that machines are designed. That is, their creation requires deliberate choices on the part of human creators. Individuals or organizations can therefore be held responsible for the choices made assuming, of course, that documentation of such decisions is available (this is one sense of transparency, which will be discussed in more detail later). Furthermore, like any other technology, there are deliberate choices involved when it comes to the uses of artificially intelligent machines. These machines can, Bryson argues, be used to provide us with a greater capacity to "perceive and maintain accounts of actions and consequences" which ultimately ought to make it "easier, not harder, to maintain responsibility" and assign credit or blame where it is due.[509] Note, however, that this all hinges on deliberate design decisions on the part of individuals and institutions that create and use machines. Artificially intelligent machines may be used just as much to obfuscate

[506] Bryson 2020, 5.
[507] Ibid.
[508] And, perhaps more importantly, being responsible or held accountable is only attributable to humans.
[509] Bryson 2020, 5.

design decisions as they are to preserve and reveal them. Building on artifacts like written language, "digital artifacts are particularly amenable to automating the process" of "maintaining precise accounts of when, how, by whom, and with what motivation the system" under consideration has been constructed.[510]

7.1.1 Skeptics and Enthusiasts

Bryson maintains that machines, even the autonomous and artificially intelligent ones beginning to permeate society, are no different in kind to any other artifact or technology. According to Bryson all artifacts and technologies are designed and used by humans and so it is only humans that can be held responsible or accountable. This view, of which Bryson is just one advocate, is the view of one group of scholars that I am calling "skeptics." These scholars are skeptics because the common theme uniting their views is that machines are mere tools, mere instruments, and it would be a mistake to attribute any type of responsibility to them, just as it would be a mistake to attribute any type of responsibility to a hammer or bomb. This skeptical view also extends beyond concepts like responsibility and can include concepts like patiency and agency as well. That is, on this view one might be skeptical that machines can ever be considered, or ought to be considered, as moral agents or moral patients. For the skeptics, machines cannot be held responsible for "their" actions nor are machines deserving of any special treatment (e.g., respecting their rights). Indeed Bryson has argued that it is a purely normative matter whether machines are considered as moral agents or moral patients, i.e., "there is no necessary or predetermined position for AI in our society."[511] And if it is a matter of whether we *ought* to make machines such that they deserve to be moral patients or thought of as moral agents, then Bryson is clear that this is something we ought not do. There may be, Bryson argues, "substantial costs but little to no benefits from the perspective of either humans or robots to ascribing and implementing either agency or patiency to intelligent artefacts beyond that ordinarily ascribed to any possession."[512]

When thinking about whether we *should* care about artificially intelligent machines, and by extension their design and use, Bryson and other skeptics are of the opinion that we should only care insofar as these machines are a new technology that humans can use to improve or impoverish society. Such technologies can help us create a better more ethical society or they can be used to create a worse

[510] Ibid., 8.
[511] Bryson 2018, 15.
[512] Ibid., 16.

more unethical society. In short, it is not the machines *per se* that we should care about, but rather the humans or organizations that design and use those machines. That is, humans are both moral agent and moral patient, and AI is just one tool we use to mediate those relationships, hence machine ethics is no different from computer ethics more generally.

If we can describe Bryson and others with similar views as skeptics, then their views are largely in opposition to the views of those I am choosing to call the "enthusiasts." For the enthusiasts, we *should* worry about autonomous and artificially intelligent machines because they are a technology or artifact that is significantly different from all previous technologies. Moreover, we may attribute some form of moral or ethical status to these types of machines (or particular tokens of these types of machines, e.g., the robot Sophia) because they are, in a sense, more than mere tools.[513] Enthusiasts can also be divided into subgroups depending on whether they believe machines ought to count as moral patients, as moral agents, or both. John Danaher, for example, argues for a kind of ethical behaviourism such that machines may count as moral patients if they are "*roughly performatively equivalent* to another entity whom, it is widely agreed, has significant moral status," i.e., an entity that has limits to how we can treat it such that our treatment of that entity is not a matter of mere preference.[514] Given that it is widely agreed that adult humans have significant moral status, if a robot is roughly performatively equivalent to an adult human, then on Danaher's view that robot ought to be afforded the same moral status. Danaher maintains that what is "going on 'on the inside' *does not matter* from an ethical perspective" because there are epistemic limits that prevent us from directly accessing the metaphysical status of a given entity.[515] In short, Danaher maintains that "performative artifice, by itself, can be sufficient to ground a claim of moral status" regardless of whether machines, e.g., robots, are designed to be tools or slaves for our use.[516]

Ian Kerr is another enthusiast (though I doubt he would describe himself in such terms) that I argue implicitly endorses thinking of machines as agents. In contrast to skeptics, Kerr claims that "non-sentient robots and AIs can diminish our privacy" because "artificial cognizers can be said to form truth-promoting beliefs that are justified" and even actuate automatically on the basis of these beliefs thereby potentially violating, in addition to diminishing, a person's state of privacy.[517] Like Danaher, Kerr is more interested in epistemology than metaphysics, but unlike Danaher, Kerr is more interested in how

[513] More information about Sophia can be found here: https://www.hansonrobotics.com/sophia/
[514] Danaher 2020, 2025-2026.
[515] Ibid., 2025.
[516] Ibid.
[517] Kerr 2019, 125-126.

machines can affect humans, i.e., in how machines as agents of a certain non-sentient, non-conscious kind, may treat human patients. Though he does not explicitly describe machines or artificial entities as agents, it is implicit in Kerr's writing that machines (or certain types of machines) possess some kind of agency.[518] He emphasizes, for example, that machines can "cause people to act or be treated in particular ways" or "carry out certain operations independent of human action, oversight, intervention, awareness or knowledge," all of which point towards thinking of machines as agents.[519] These machines, Kerr maintains, would not be sentient but nevertheless be "entities capable of acting upon the world with substantial autonomy and the ability to significantly affect the life chances and opportunities of people."[520] These capabilities are sufficient to ascribe, if not moral agency, then agenthood in general to machines with such capabilities. Recall from the discussion of agency and patiency in Chapter 3 that moral agency is a specific kind of agency that presupposes the possession on the part of the agent in question the right sort of mind and/or autonomy. While the kinds of machines that Kerr is concerned with are certainly agents, i.e., capable of acting on and affecting the world (and by extension people), they may not qualify as moral agents for various reasons, e.g., these machines lack the right sort of autonomy to confer *moral* agency because they cannot act, on their own, in compliance with or in violation of moral norms.

Unlike the skeptics, the enthusiasts generally agree that we *should* care about artificially intelligent machines, but the reasons why often differ. Kerr for example cites the need for an expansion of the legal theory and doctrines that deal with privacy in light of his conclusion that machines can disturb the "presumption of ignorance in epistemologically significant ways" that grounds a right to privacy.[521] Danaher similarly agrees that we should care about artificially intelligent machines, but his reasons for thinking so revolve around his commitment to ethical behaviourism, "a meta-empirical thesis about how we ought to interpret empirical evidence concerning behaviour" and consequently whether a given entity has moral status.[522] As mentioned above, ethical behaviourism roughly amounts to a rejection of the idea that ethical considerations stem from what is going on "on the inside" (i.e., metaphysically going on), as it were, and an affirmation of the idea that behaviour by itself is sufficient to ground ethical considerations (e.g., whether an entity ought to be considered a moral agent or moral patient). David Gunkel, another enthusiast, likewise maintains we ought to care about artificially

[518] For a more detailed discussion of agency see Chapter 3.
[519] Kerr 2019, 144.
[520] Ibid., 145.
[521] Ibid., 126.
[522] Danaher 2020, 2040.

intelligent machines. Gunkel's motivations however have to do with the ways in which we encounter and interact with "others" "whether they [these others] be other human persons, an animal, the natural environment, or a social robot."[523] For enthusiasts like Gunkel, what matters is not whether machines are autonomous, intelligent or even behaviourally similar to humans, but rather how we actually relate to them and how they reveal themselves to us in the relationships we form together.[524]

7.1.2 Pessimism and Optimism

Before moving on, one terminological clarification needs to be made. I have chosen to use the terms 'skeptics' and 'enthusiasts' to specifically describe those scholars who believe we should not and those scholars who believe we should care about (i.e., give special attention to) artificially intelligent technologies respectively. The scholarly landscape however is considerably more complicated than this simple dichotomy suggests. In addition to classifying scholars as either skeptics or enthusiasts, they could also be described as pessimists or optimists. This classification is orthogonal to the skeptic/enthusiast dichotomy and is an attempt to capture whether scholars view the development of artificially intelligent technologies as undesirable or desirable, roughly speaking, and this is best illustrated using examples. Thomas Metzinger is an example of a pessimistic enthusiast. Metzinger argues that we ought to care about artificially intelligent technologies and in this sense he is an enthusiast. But Metzinger is a pessimist in the sense that he worries continued development of artificially intelligent technologies may lead to "an "explosion of negative phenomenology" in advanced AI and other post-biotic systems."[525] "On ethical grounds," Metzinger argues, "we should not risk a second explosion of conscious suffering on this planet" with the first explosion taking place via the development of conscious biological life.[526] Metzinger maintains that given we lack both a good theory of consciousness and a good theory of what constitutes "suffering," the risk of an explosion of negative phenomenology is currently incalculable. As such, he concludes that "there should be a global ban on all research that directly aims at or indirectly and knowingly risks the emergence of synthetic phenomenology" until we have a deeper scientific and philosophical understanding of consciousness and suffering.[527]

---

[523] Gunkel 2018, 96.
[524] More will be said about this particular view.
[525] Metzinger 2021, 45.
[526] Ibid., 46.
[527] Ibid.

Contrast Metzinger with Danaher who is an optimistic enthusiast. Danaher, as mentioned above, is an enthusiast because he argues that we should care about artificially intelligent technologies, but he is also an optimist in the sense that he believes there are already existing frameworks to help us make sense of artificially intelligent moral patients. Not only does Danaher believe that there are potential benefits to the creation of robotic offspring,[528] but he maintains that "the creation of robots with significant moral status can be viewed through the lens of procreative ethics."[529] Danaher maintains that, when thinking about how the principle of procreative beneficence (i.e., the duty one has to give their child the best possible life) might apply to machines, it actually "may be less controversial in [the case of machines] than in the human case," primarily because the burden that might be imposed on manufacturers to create machines with the "best possible life" will not be an unreasonable burden.[530]

Unlike Metzinger and Danaher, Bryson is an example of a pessimistic skeptic. As mentioned above, Bryson is a skeptic because she maintains that we should not care about artificially intelligent technologies insofar as such technologies are not different in kind from any other human created artifact. For Bryson, we should care about such technologies only as much as we care about any other tool. In addition to this skepticism concerning artificially intelligent technologies, Bryson is also a pessimist, like Metzinger, in the sense that she believes we ought to refrain from creating certain kinds of artificially intelligent machines. Bryson believes that we "can limit AI - or at least legally-produced commercial AI - to be as it is now, something to which no obligations are owed directly."[531] Indeed Bryson is adamant that we ought not, even if it were possible to do so, create machines that meet the requirements for moral agency or patiency.[532]

For the sake of completeness, there is, in the space of possible positions I have outlined, room for an optimistic skeptic, but I am unaware of anyone holding such a view. This is primarily because this position amounts to believing that artificially intelligent technologies are not different in kind from any other technology (i.e., the skeptical view) but that it is desirable to continue the development of such technologies (i.e., the optimistic view).

[528] See Danaher (2018).
[529] Danaher 2020, 2045.
[530] Ibid.
[531] Bryson 2018, 23.
[532] Ibid., 24.

7.2 Tools of A Different Kind

Like many other scholars, I also maintain that we *should* care about artificially intelligent technologies. That is, I am an enthusiast in the sense defined above. But I argue that we should care about artificially intelligent technologies for two main reasons. First, artificially intelligent technologies, at least those of the kind that have been discussed throughout this dissertation, are tools of a radically different kind than any other previous tool or artifact created by humanity. Second, and relatedly, these technologies have a breadth of applicability that makes them particularly unique among the tools that humanity has developed. Let us consider each reason in more detail.

7.2.1 Autonomy and Intelligence

What exactly separates intelligent machines from all other tools? I argue that, unlike all other previous tools or artifacts, intelligent machines are just that: *intelligent*. Perhaps more clearly, intelligent machines are tools that fulfill cognitive tasks. In contrast to hammers, wheelbarrows and excavators, all of which are tools that fulfill physical, i.e., manual labour, tasks, intelligent machines are tools that fulfill tasks that are within the domain of cognition. This includes tasks like planning and decision-making, image recognition, classification and sorting, coherent language generation and more. The ability to perform cognitive tasks marks a distinct difference in kind between all previous tools and artificially intelligent tools.

Now it might be objected that there have existed, for quite a long time, tools that fulfill tasks within the domain of cognition. Consider a tool like the abacus or even primitive writing tools.[533] If mathematical calculations are tasks that fall within the domain of cognition, then surely a tool like the abacus ought to also be considered a different type of tool. But merely fulfilling a task in the domain of cognition is not sufficient to warrant a difference in kind. In addition to fulfilling a task in the domain of cognition, tools like contemporary artificially intelligent machines are also autonomous. That is, they are tools that can operate in the absence of a human. AlphaZero, in contrast to an abacus, can fulfill its task of playing a game of chess, for example, without the presence of any human. Moreover, autonomy on its own is also insufficient to warrant differentiating between kinds of tools. Tools stretching as far back as windmills and watermills can be thought of as autonomous tools, at least in the sense that they can

[533] This is particularly important against the backdrop of extended theories of mind which posit that cognitive tasks can be fulfilled by external objects that are coupled to a human in certain relevant ways.

also operate in the absence of a human. More modern tools like Roombas are also autonomous in this sense, i.e., they operate quite effectively in the absence of a human. But in each of these cases the task being fulfilled falls outside of the domain of cognition. To be sure, Roombas and watermills are sophisticated tools, but I maintain they are no different in kind from hammers, wheelbarrows and excavators.

So we can be more precise about what exactly separates artificially intelligent machines from all other tools. Unlike all other previous tools or artifacts, artificially intelligent machines are both autonomous and engaged in tasks within the cognitive domain. It is the conjunction of autonomy and intelligence that is key. Importantly, this does not mean that artificially intelligent machines are *only* ever engaged in tasks within the cognitive domain. Consider an autonomous vehicle. In addition to engaging in the cognitive tasks of strategic planning, decision-making and object classification, among others, autonomous vehicles also engage in the "physical"[534] tasks of depressing the gas and brake pedals as well as turning the steering wheel. Autonomous vehicles also highlight the other particularly unique feature of artificially intelligent machines, their breadth of applicability or, alternatively, their universality.

7.2.2 Universality

The ability to autonomously engage in tasks within the cognitive domain is what sets artificially intelligent machines apart from all other machines, tools and artifacts that humans have created. Another unique, but not necessary, feature of artificially intelligent machines is their breadth of applicability. As discussed in Chapter 4, the same basic machine, i.e., the same deep neural network and learning algorithms, could, in principle, operate an autonomous vehicle and classify hand drawn images. In practice, DeepMind has already demonstrated that the same artificially intelligent machine can master the three different games of chess, shogi and Go.[535] Indeed transfer learning, a type of machine learning discussed in Chapter 1, makes use of precisely this phenomenon, e.g., taking a machine that can generate names for metal bands and reusing that same machine, albeit after some retraining, to generate names of ice cream flavours.[536]

[534] I say "physical" because there is no robotic foot or hand to actually depress the pedals or move the steering wheel.
[535] Silver et al. 2018.
[536] Shane 2019, 45-47.

Many different tasks from image recognition to text generation and prompt completion can be relatively easily accomplished by the same machine (i.e., deep neural network and learning algorithm). This means that there is no substantial barrier between different domains of application that would prevent, or at least make it difficult and costly, the same machine from being used in a variety of different contexts by many different people. Almost all other tools,[537] including hammers, wheelbarrows and calculators, lack such breadth in their applicability. Sure, hammers can be used to remove nails in addition to hammering nails into place, and could be used more generally to smash things, but beyond that, their uses are quite limited. There is no analog of transfer learning for other tools (with perhaps the exception of the digital computer), and there are significant barriers between different domains of application that makes it difficult or downright ridiculous to use tools meant for tasks in one domain in an entirely different domain. Using a chainsaw in place of a kitchen knife or vice versa would be inappropriate (to say the least) but using the same artificially intelligent machine to distinguish between images of cats and dogs, and between images of cancerous and non-cancerous melanomas, is practically the norm. For example, since late 2018, so called "foundation models" have demonstrated how transfer learning techniques when coupled with increases in scale (e.g., vast hardware improvements, gargantuan models and immense amounts of training data) can allow the same machine "to take the "knowledge" learned from one task (e.g., object recognition in images) and apply it to another task (e.g., activity recognition in videos)."[538] In short, one foundation model can be adapted to a wide range of downstream tasks, and their use has only increased since their inception.

7.3 The Perfect Technological Storm[539]

If artificially intelligent machines are different in kind from all previous machines and tools, and I have argued that they are, then we *should* care about such machines insofar as their effects are relatively unknown. This brings us at last to the first question I posed: what is worth focusing on and why? That is, what types of artificially intelligent machines are being created and to what end are they being created? The short answer is that the types of artificially intelligent machines that we should care about, the ones that are worth focusing on, are the machines that *purport* to replace humans entirely

[537] The one exception is the digital computer. While breadth of applicability is a unique feature of artificially intelligent machines, another notable tool that possesses this feature is the digital computer.
[538] Bommasani et al. 2021, 4.
[539] This is with reference to Gardiner's (2006) "Perfect Moral Storm."

and thereby engage in what Brian Cantwell Smith calls "judgment."[540] The long answer, which I will elaborate on now, is that there is a perfect technological storm brewing that is capable of lulling people into a state in which they abdicate responsibility for decision-making and behaviour precipitated by artificially intelligent machines, a state that I am calling "moral complacency."

In discussing climate change, Stephen Gardiner argues that climate change is a perfect moral storm given that "it involves the convergence of a number of factors that threaten our ability to behave ethically" just as a perfect storm "is an event constituted by an unusual convergence of independently harmful factors" likely to result in substantial negative outcomes.[541] Like Gardiner, I maintain that artificially intelligent machines are analogous to a perfect storm in that such machines involve the convergence of a number of factors that threaten our ability to behave ethically. Unlike Gardiner however, who believes that the storm in the context of climate change makes us vulnerable to moral corruption broadly construed,[542] I argue that the storm in the context of artificially intelligent machines makes us vulnerable to moral complacency in particular.

7.3.1 The Transparency "Storm"

As many complex issues intersect with artificially intelligent machines, here I will focus only on two salient problems that converge to make us especially vulnerable to becoming morally complacent. The first problem is that of transparency/opacity. It has become common to think of artificially intelligent machines, especially those we have been focusing on (i.e., machines that utilize big data and machine learning techniques), as black boxes. As discussed in Chapter 1, while the inputs and outputs of these machines are generally easy to identify and understand, the same cannot be said about what goes on inside the machine. The issue of transparency (sometimes also thought of as "interpretability") is not however a simple dichotomy, i.e., transparent or not transparent. As many scholars have noted, "there are many flavours and gradations of transparency that are possible" and so it is important to understand some of the different types of transparency one might encounter.[543] Recall from Chapter 1 that transparency, or lack thereof, i.e., opacity, may be (1) intentional, because corporations and institutions

[540] Smith 2019, xv.
[541] Gardiner 2006, 398.
[542] Gardiner (2006) states that moral corruption can be facilitated in a number of ways, including via distraction, complacency, unreasonable doubt, selective attention, delusion, pandering, false witness and hypocrisy.
[543] Diakopoulos 2020, 199.

are interested in protecting their intellectual property, (2) a result of the specialist knowledge required to understand computer code and algorithms, or (3) a result of the computational mismatch between humans and machines.[544]

While this division of the types of transparency/opacity is helpful, there is much more that can be said. For example, different contexts might play a pivotal role in determining under what circumstances and to whom corporations or institutions might reveal the inner workings of their machines. It has been argued, for example, that "a safety inspector or accident investigator may need different information to assess a system globally" in comparison to other stakeholders, like an operator or end-user, who might be interested in understanding an individual decision or outcome.[545] Regulators, for example, may be given privileged access to machines to ensure that corporations/institutions are abiding by industry regulations.

Similarly, specialist knowledge of computer code and algorithms may not be sufficient (or may not be required) to render machines wholly transparent. Moreover, algorithmic transparency can be thought of as just another different kind of transparency that concerns the algorithms themselves.[546] Some algorithms are just easier to understand in the sense that we can prove and thereby be confident that they will produce certain predictable outcomes. Other algorithms, like the ones used in modern machine learning techniques highlighted in Chapter 1, are not as well understood even by the software engineers and computer scientists that use them. There is therefore no guarantee that specialist knowledge will increase the transparency of a machine's decisions even if one has access to the algorithms themselves.[547]

In addition to algorithmic transparency, a second notion of transparency connected to specialist knowledge might be the transparency that is gained by understanding and being able to explain what each part of the model, i.e., "each input, parameter, and calculation," does and how it affects the outputs.[548] This kind of transparency could just as easily be thought of in terms of intelligibility given the fact that one is able to explain how and why a machine arrived at a certain decision by decomposing the machine into simpler parts. A third and even more general notion of transparency connected to specialist knowledge is one of simulatability. In short, "if a person can contemplate the entire model at

[544] Burrell 2016, 1-2.
[545] Diakopoulos 2020, 199-200.
[546] Lipton 2017, 5.
[547] See Lipton (2017) for more details about the kinds of algorithms that are more transparent and algorithms that are less transparent.
[548] Lipton 2017, 5.

once" and step through every calculation, as it were, that the machine itself would perform when transforming a given input into an output, then that machine may be considered transparent.[549] However just as algorithms fall on a continuum ranging from the highly transparent (i.e., well understood) to the highly opaque (i.e., not well understood), so too do the model and high level features of a machine range from the highly transparent to the highly opaque.

Transparency, of the lack thereof, may also simply be the result of computational mismatches between humans and machines. While machines are clearly far from possessing human-like intelligence, they nevertheless already possess certain advantages. Machines have perfect recall, never become tired, do not experience boredom, do not daydream, do not need to rest or sleep, do not make mistakes when applying mathematical or logical rules, and can analyze far more data than any single human being could, to name a few advantages. It has been demonstrated that even simple linear models outperform human experts in domains such as clinical diagnosis and forecasting graduate students' success.[550] Further studies have revealed that it is far more common for algorithms to outperform human judges and forecasters than the opposite.[551] Such results inevitably raise questions about whether it even matters if machines are transparent or not. This kind of epistemic reliabilism (discussed above in connection to Kerr's view that machines can diminish our privacy) has important ramifications when it comes to thinking of artificially intelligent machines as capable of forming and having knowledge or beliefs.[552] If an algorithm more accurately predicts the weather or can better forecast which area of study a given student is more likely to succeed in, why should we care *how* or *why* the machine made the prediction that it did? Setting aside these epistemological concerns however, there is growing evidence that people prefer advice from algorithms to advice from people, a phenomenon that has been dubbed "algorithm appreciation."[553] Even more interesting, at least when considering the issue of transparency/opacity, is that researchers discovered that participants were willing to rely on algorithmic advice instead of human advice even when given a minimal description of the algorithm (e.g., "The output that an algorithm computed was…"[554]). They conclude that "participants who faced "black box" algorithms…were willing to rely on that advice despite its mysterious origins" suggesting that the issue

[549] Ibid., 4.
[550] Dietvorst et al. 2015, 1.
[551] Ibid.
[552] Recall that this is precisely the line of reasoning that Kerr (2019) takes to support his claim that machines can diminish our privacy. That is, machines, as a result of reliable "belief" formation processes, can be said to "know" something about an individual and as such these machines can violate our right to privacy.
[553] Logg et al. 2019, 90. Algorithm appreciation was also discussed in Chapter 3.
[554] Ibid., 92.

of transparency of a machine may not be a significant concern for the average person, at least in some contexts.[555] This leads us to the second salient problem, the problem of trust.

7.3.2 The Overtrust "Storm"

While the phenomenon of algorithm appreciation might seem innocuous at first, it is connected to a deeply problematic cognitive bias, the automation bias. Put simply, automation bias is "the tendency of people to overtrust automated tools."[556] Now, as Joel Walmsley notes, this would not be particularly problematic if "our use of such systems were restricted to recommender systems for movies, music and restaurants," but this is simply not the case.[557] In addition to the many anecdotes of people religiously following their GPS or Google Maps route into lakes and houses,[558] research has revealed that people are, perhaps surprisingly, apparently willing to trust robots in emergency evacuation scenarios.[559] Robinette et al. noted that even when participants observed a robot inefficiently lead them to a meeting room by taking them into the wrong room first, they nevertheless followed the robot's instructions, i.e., went in the direction the robot was pointing, towards a back exit in a contrived fire emergency.[560] Additionally, in follow up studies where participants witnessed the robot, before or during the "emergency," either "break" (i.e., spin around in circles and direct the participants toward a corner in the hallway), "break" and remain immobilized pointing towards the back exit, or "break" and direct participants towards a dark room partially blocked by a piece of furniture (i.e., participants were not directed towards an exit), the majority of participants followed the robot's instructions.[561] Robinette et al. conclude that their findings demonstrate "a potentially dangerous level of overtrust" in machines and that machines "interacting with humans in dangerous situations must either work perfectly at all times and in all situations or clearly indicate when they are malfunctioning."[562]

---

[555] Ibid., 100.
[556] Gebru 2020, 265.
[557] Walmsley 2021, 587.
[558] Incidents like these have permeated pop culture. One example is the character Michael Scott from the television show *The Office* driving into a lake as he yells, "The machine knows where it is going!" Watch it here: https://www.youtube.com/watch?v=DOW_kPzY_JY
[559] Robinette et al. 2016.
[560] Ibid., 101.
[561] Ibid., 105-107.
[562] Ibid., 107-108.

In fact, the issue of overtrust in machines is considerably worse. The strength of modern artificially intelligent machines lies in their ability to identify and reproduce mappings between inputs and outputs, whether those inputs and outputs be images and their corresponding labels or recommended videos and viewer retention time.[563] The danger however, discussed extensively in Chapter 2, is that machines, by working exactly as intended (i.e., reproducing mappings between inputs and outputs), are susceptible to discovering/learning and thereby perpetuating systemic biases.[564] When coupled with the fact that people are already highly trusting of machines, it is clear that people will be even less likely to critically assess and question a machine's decision or answer, as it were. People will simply believe what a machine "says" without confirming against the "source" whether there is any veracity to the machine's output. A relatively recent example of a Palestinian writing "good morning" on Facebook highlights precisely this kind of unreflective trust in machines.[565] In short, a Palestinian man was arrested by authorities for his post, "good morning," written in Arabic which Facebook Translate (a proprietary translation machine) translated to "hurt them" in English or "attack them" in Hebrew. While the person was released shortly after their arrest, authorities trusted that the translation was accurate "and did not think to first see the original text before arresting the individual."[566] Overtrust in machines is clearly morally problematic, to say the least.

7.3.3 Moral Complacency

Transparency, or the lack thereof (i.e., opacity), in machines is, on its own, a problem with artificially intelligent machines that can induce moral complacency. People either do not care to "look under the hood" and try to understand how the machine operates or are prevented in various ways (e.g., because of proprietary technology or a lack of specialist knowledge) from being able to comprehend how the machine operates. Overtrust in machines is similarly, on its own, a problem with artificially intelligent machines that can induce moral complacency. The automation bias, the tendency for people to offload cognitive processes/capabilities onto external entities like machines, drastically increases the likelihood that people will fail to exercise their judgment, in Smith's sense, when it may be needed most, i.e., in situations of high moral import.

---

[563] Smith 2019, 49.
[564] See Chapter 5 for a more detailed discussion of risks associated with implementing machine ethics.
[565] See: https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427
[566] Gebru 2020, 266.

Together, the transparency of and overtrust in machines can induce horrifying levels of moral complacency in human beings. Take the case of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) machine that predicts the likelihood of a defendant reoffending within two years of assessment. Although COMPAS has been used to assess more than one million offenders in the United States (US) since its development in 1998, only recently has COMPAS's fairness been called into question. As mentioned above and discussed throughout Chapters 2 and 5, machines are liable to perpetuate systemic biases and indeed COMPAS reflects systemic racism in the US criminal justice system by underpredicting recidivism for white and overpredicting recidivism for black defendants.[567] Moreover, further analysis demonstrated that COMPAS is as "fair" and "accurate" at predicting recidivism as a sample of random people.[568] The company that developed COMPAS (formerly Northpointe and now known as Equivant) engages in the type of intentional opacity described above in the sense that they have not publicly disclosed their machine's structure or just how exactly it is trained. Moreover, even judges, the kind of people that arguably most often ought to be aware of the effect their judgment can have on human well-being, have deferred to COMPAS's outputs presumably because they trust the machine more than their own assessment of a defendant. As Angwin et al. with ProPublica reported, the prosecutor in Paul Zilly's case, who had been "convicted of stealing a push lawnmower and some tools," initially recommended a year in county jail and follow up supervision, and Zilly's lawyer agreed to a plea deal. Judge James Babler however, had seen Zilly's score as computed by COMPAS "which had rated Zilly as a high risk for future violent crimes and a medium risk for general recidivism" and consequently overturned the plea deal agreed upon by the prosecution and defense and instead "imposed two years in state prison and three years of supervision."[569]

There is, in short, a real and present danger concerning the abdication of responsibility as a result of the moral complacency that arises from the unreflective use of artificially intelligent machines. As Smith painstakingly highlights, machines are not yet at the point where they can engage in the weighty ethical judgments that accompany innumerable decisions in our daily lives. Note that the term of importance is 'judgment.' That is, judgment requires a particular sort of *understanding*:

> the understanding that is capable of taking objects to be objects, that knows the difference between appearance and reality, that is existentially committed to its own existence and to the

[567] Dressel and Farid 2018, 1.
[568] Ibid., 3.
[569] Angwin et al. 2016.

> integrity of the world as world, that is beholden to objects and bound by them, that defers, and all the rest.[570]

COMPAS lacks judgment, not because it lacks any sort of "calculative rationality" that Smith (2019) identifies as "reckoning." No, COMPAS lacks judgment in the sense of not being able to fully consider the consequences of its actions and thereby "failing to uphold the highest principles of justice and humanity and the like."[571] What artificially intelligent machines can do is impressive, but as John Haugeland might have put it, *they don't give a damn*.[572] And this is why we should care about artificially intelligent machines. Not because these machines might be considered as moral agents or moral patients or capable of suffering (though these are interesting and important considerations), but because by purporting to replace human decision-making and behaviour in tasks and activities that require *judgment*, machines induce moral complacency in us, beings who are (or at the very least *ought to be*) committed to the world and the entities it contains.

7.4 Responding to Machines: The Practical and the Theoretical

In this final section I turn now to consider some socio-political and legal responses to artificially intelligent machines, and suggest that what is needed is, as the title of this chapter suggests, an ethics 3.0 for this new age of the machine.

The development of artificially intelligent technologies and their potential ramifications has not gone unnoticed by governments and professional organization, e.g., the Institute of Electrical and Electronics Engineers, many of which have begun to create best practice or ethical use guidelines.[573] To that end, i.e., the ethical use of artificially intelligent machines, the European Union (EU), as one example, has created the High-Level Expert Group on AI that has been working to understand, outline and clarify the potential risks of using AI technologies as well as consider how such technologies could be regulated. More recently, and building on work by the High-Level Expert Group on AI, the European Commission released the "White Paper" on artificial intelligence which outlines how the Commission intends to support a "regulatory and investment oriented approach [towards AI] with the twin objective

[570] Smith 2019, 110.
[571] Ibid., 111.
[572] Adams and Browning eds. 2017.
[573] See, for example, the "Ethically Aligned Design" document drafted by the IEEE Standards Association.

of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology."[574]

It is the opinion of the Commission that some existing legislation can be adjusted to accommodate the appearance and use of artificially intelligent technology but that there are some novel challenges presented by these machines that are likely best addressed via the creation of a new regulatory framework. The Commission notes that, with regard to the former (i.e., adjusting existing legislation), there is already a legislative framework that exists to protect the fundamental rights and consumer rights of EU citizens. Some of this legislation include the "Race Equality Directive, the Directive on equal treatment in employment and access to goods and services," rules concerning consumer protection as well as rules on personal data privacy and protection, "notably the General Data Protection Regulation" and the "Data Protection Law Enforcement Directive."[575] However, and here I emphatically agree with the Commission, given that the specific characteristics of artificially intelligent machines, "including opacity, complexity, unpredictability and partially autonomous behaviour, may make it hard to verify compliance with, and may hamper the effective enforcement of, rules of existing EU law meant to protect fundamental rights" and consumer rights, a new regulatory framework ought to be created for artificially intelligent technologies.[576]

Practically speaking, it is the Commission's view that a new regulatory framework for artificially intelligent machines should follow a risk-based approach so as to be effective while not excessively prescriptive. The Commission maintains that artificially intelligent machines ought to be generally considered high-risk, but especially so when such machines meet the following two cumulative criteria: "First, the AI application is employed in a sector where, given the characteristics of the activities typically undertaken, significant risks can be expected to occur" and second, "the AI application in the sector in question is, in addition, used in such a manner that significant risks are likely to arise."[577] COMPAS is an example of a machine that clearly meets both criteria. The criminal justice sector, as it were, is one wherein significant risks can be expected to occur, e.g., innocent people are found guilty for crimes they did not commit, and COMPAS is trusted by humans and perhaps seen as having an "objective" view and is therefore being used in such a manner that systemic biases are likely to be perpetuated.

[574] European Commission White Paper on Artificial Intelligence 2020, 1.
[575] Ibid., 13.
[576] Ibid., 12.
[577] Ibid., 17.

In keeping with their risk-based approach, the Commission identified six key features that ought to be the focus of mandatory legal requirements when designing future regulatory framework for artificially intelligent machines. These key features are: training data, data and record-keeping, information to be provided, robustness and accuracy, human oversight, and specific requirements for certain particular AI applications (e.g., those used for remote biometric identification).[578] Although the Commission discusses these key features in further detail, it is still a rather general discussion, and that is perhaps, understandably, because artificially intelligent machines and their accompanying regulatory frameworks are still evolving. Take the third key feature of information to be provided as an example. The Commission acknowledges that in order to promote the responsible use of artificially intelligent machines and to build trust in such technology, clear information as to a machine's capabilities and limitations should be provided and that, separately, "citizens should be clearly informed when they are interacting with an AI system and not a human being."[579] When considering the addressees of the legal requirements that would apply under this future regulatory framework, it is the Commission's view that "each obligation should be addressed to the actor(s) who is (are) best placed to address any potential risks."[580] But risk-based approaches, and especially rights-based approaches, to thinking of artificially intelligent machines and future regulatory frameworks are, I maintain, ultimately impoverished and quite reactive rather than proactive. They are therefore unlikely to bring about the kind of positive changes that people (e.g., regulators, users, stakeholders, etc.) would like to see.

7.5 Virtues, Relations and Machines

In his 2017 book *Life 3.0*, Max Tegmark states that lifeforms have three levels of sophistication: Life 1.0, 2.0 and 3.0. Life 1.0 is characterized by an inability on the part of the lifeform to design either its "hardware" or "software," e.g., genetics or linguistic abilities respectively.[581] In contrast to simple biological life, like bacteria (which are examples of Life 1.0), Tegmark believes that humans are an example of Life 2.0 because we can wrest control of our software from evolution, and design it ourselves. By 'software,' Tegmark means "all the algorithms and knowledge that you use to process the information from your senses and decide what to do" which includes "everything from the ability to recognize your friends when you see them to your ability to walk, read, write, calculate, sing and tell

[578] Ibid., 18.
[579] Ibid., 20.
[580] Ibid., 22.
[581] Tegmark 2017, 27.

jokes."[582] Naturally, Life 3.0 is characterized by an ability to design its own hardware in addition to its own software and there are, as of yet, no examples of such a lifeform. But the boundaries between these three stages, Tegmark notes, are fuzzy. Mice might be an example of Life 1.1 given their limited capacity to learn, and today's humans might be an example of Life 2.1 given our limited capacities to modify our hardware (e.g., we can replace our evolved hearts with artificial ones, our legs and arms with artificial ones, etc.).[583]

In addition to lifeforms, I want to suggest that we can think about ethics in a similar way. In short, just as artificially intelligent machines may one day lead to Life 3.0, living in a world filled with artificially intelligent machines may need us humans to embrace an Ethics 3.0. But what exactly would an Ethics 3.0 look like? Unlike lifeforms, which we might intuitively group, as Tegmark does, into biological, cultural and technological stages, no such division readily appears in the case of ethics. Nevertheless, I maintain that there are three ethical frameworks that we could identify as Ethics 1.0, 2.0 and 3.0.[584] Though I want to focus primarily on Ethics 3.0, I will now briefly outline what I think of as Ethics 1.0 and 2.0.

Ethics 1.0, I argue, is characterized by the unfortunately named "social Darwinism." Championed largely by Herbert Spencer (not Charles Darwin), the idea behind social Darwinism is that nature has its own morality, that evolution[585] tends towards generating the greatest happiness and that those organisms that enjoy reproductive fitness are also more worthy of existing in a normative sense; those organisms *ought* to exist because they are better or superior. Ethics 2.0, in contrast, is characterized largely by a rejection of morality in nature and instead an affirmation of the independent existence of human rights. Defenders of rights based approaches have grounded such rights in everything ranging from laws to God.[586] For our purposes, rights that are grounded in notions of agency or autonomy are of particular interest because these are features that I have argued are shared by both humans and machines. Indeed a new area of research has appeared whose primary focus is whether machines or robots should have rights, or a certain set of rights.[587] But focusing on this particular question, of whether machines are deserving of rights, or ought to be considered moral patients, is not particularly

[582] Ibid.
[583] Ibid., 29.
[584] It should be noted that I am not ordering these "stages" of ethics chronologically or even by their level of sophistication, however that might be determined. I am ordering these stages roughly according to how well they might promote living a good life in this modern age of artificially intelligent machines.
[585] Importantly, this is not the idea of evolution that Darwin championed, i.e., evolution by natural selection, but rather a more Lamarckian idea of evolution, i.e., evolution by use-inheritance.
[586] See Kohen (2007).
[587] See Gunkel (2018).

useful. One could engage in endless debate about conferring or extending rights and moral patiency to include machines and make very little headway.[588]

Instead of a rights-based approach, or by extension a risk-based approach favoured by the European Commission, I believe that a virtue-based approach is more fruitful when considering how we as individuals, and as larger groups and societies, ought to respond to artificially intelligent technologies. This is, I propose, Ethics 3.0, an ethical framework characterized both by an affirmation of an agent-centered (rather than patient-centered) point of view and the idea that moral worth is extrinsic and not intrinsic to an entity. By putting virtue and the cultivation of character front and center in Ethics 3.0, the focus shifts away from the intractable questions associated with rights-based approaches (e.g., who is deserving of rights, who will guarantee those rights, how burdensome should a right be, etc.) and instead brings a different set of questions to the fore: Who are we? What kind of society do we want to create together? And it is these questions that are far more important to consider and reflect upon when considering the socio-political and legal responses to artificially intelligent technologies. Similarly, by emphasizing the idea that moral status is extrinsic in Ethics 3.0, the focus shifts away from the intractable questions associated with the metaphysical features that are intrinsic to a given entity (e.g., does this entity have a mind, does this entity have free will, does this entity have intentions, etc.) and instead brings yet another set of questions to the fore: Should we raise this entity in certain ways? How does our relationship with these other entities influence our character?

7.5.1 Resisting Moral Complacency

Neither of these ideas, that we should take up an agent-centered virtue-based approach and that moral worth is extrinsic rather than intrinsic, are new ones, but what is novel is my synthesis of them into Ethics 3.0, an ethics for the age of artificially intelligent machines.[589] Such an ethics is needed because, as mentioned above, artificially intelligent machines can induce dangerous levels of moral complacency. This complacency, and associated risks of abdicating moral responsibility (either in part or, perhaps one day in the future, completely) can be curbed by emphasizing the importance of the

[588] See, for example, Chapter 3 where I discuss the many different prerequisites thought to be necessary to ground a claim to moral patiency and the having of rights.

[589] Gardiner (2012), for example, draws a distinction between patient-centered thinking and agent-centered thinking when evaluating the effects of climate change (or perhaps more accurately, when evaluating the effects of our actions insofar as they affect climate change). Similarly, Coeckelbergh and Gunkel (2014) and Gunkel (2018), for example, argue that moral status depends far more on the relations between entities, and apply this line of reasoning to machines and technological artifacts.

character of the agent in question. We might ask Judge James Babler, mentioned above in connection to COMPAS, what kind of person are you? What kind of judge are you? Gardiner, in discussing climate change, charges us of being reckless (among other vices), and the same could be said of Judge James Babler and any number of individuals and organizations that have been lulled into a state of moral complacency, i.e., *they are being reckless*; *they don't give enough of a damn.* Gardiner writes that the charge of recklessness "extends the focus of criticism beyond the risk itself to the issue of who is imposing it, on whom, under what circumstances, and for what purpose," all of which are also pertinent when thinking of the effects of artificially intelligent machines.[590] By focusing on the character of the "agent-in-the-loop," or indeed any agent of interest (my character as the end user or an engineer's character as the designer, or even the character of the machine), we can resist becoming morally complacent.

Moral complacency can also be curbed if we take the so called "relational turn," i.e., we realize that moral worth is extrinsic rather than intrinsic. The marginalization of communities throughout history has often proceeded via dehumanization, i.e., they were deemed to "lack" what is intrinsic to those humans who "matter." Such is the weakness of rights-based approaches to ethics in general. They depend on a "conferring" or "extending" of rights, or more appropriately those intrinsic features that ground rights, from the dominant, powerful group that "matters." This too induces a kind of moral complacency. Therefore the way to fight this complacency, as Gunkel writes, is to realize that "moral consideration is decided and conferred not on the basis of some pre-determined ontological criteria or capability (or lack thereof) but in the face of actual social relationships and interactions."[591] We might, rightly I believe, accuse Judge James Babler of failing to engage and interact with Paul Zilly and thereby fail in his duties to ensure that a just sentence was carried out. By focusing on our relationships and interactions, the social web that ties us all together and to innumerable non-human entities (e.g., my cat Sergeant Fuzzy Boots, my stuffed animal Cheeser, my BlackBerry smartphone), we can better resist moral complacency via reflection on how we ought to treat those humans and non-human entities with which we are intimately linked.

To sum up, we should care deeply about artificially intelligent machines, especially those used in times and places when good judgment is required, because we risk lapsing into a state of moral complacency. What is needed to counteract this worrisome state is an Ethics 3.0, a reorienting of our thinking about ethics and morality. Rights-based and risk-based approaches, while invaluable in the past

[590] Gardiner 2012, 244.
[591] Gunkel 2018, 96.

and important in the interim, are not enough to effectively deal with the sustained pressure that artificially intelligent machines are putting on society now and will continue to do so in the future. I submit that a virtue-based relational approach is better suited for both counteracting moral complacency and pushing society in a direction better suited for all.

Coda

I have argued that the conjunction of autonomy and intelligence in artificially intelligent machines is what sets them categorically apart from all other machines and tools. These machines are able to learn and consequently go beyond, as it were, what their designers can do. This is true, as I have pointed out, in the domain of game playing where machines have discovered novel strategies. The same may also be true, I have argued, in the domain of ethical decision-making and behaviour, i.e., machines may discover novel ways of acting ethically given a specified outcome. Such discoveries may occur as we raise ethical machines using bottom-up reinforcement methods which are, I maintain, the most promising approaches to implementing machine ethics. Moreover, I have argued that, regardless of whether we are able to successfully implement ethics in machines, we will learn more about ethics *simpliciter* as a result of pursuing the project of implementing machine ethics, i.e., theorizing about and attempting to build ethical machines. In any case, we must always resist the tendency to lapse into moral complacency that is, I maintain, growing stronger as a result of our use of artificially intelligent machines. Ultimately, when it comes to decisions and actions that *matter*, there is no substitute for considered human judgment. Not yet anyways.

Bibliography

Abel, David, James MacGlashan, and Michael L. Littman. 2016. "Reinforcement learning as a framework for ethical decision-making." *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence,* 54-61.

Adams, Zed, and Jacob Browning, eds. 2017. *Giving a Damn: Essays in Dialogue with John Haugeland*. Cambridge: The MIT Press.

Alfano, Mark. 2016. *Moral Psychology: An Introduction*. Cambridge: Polity Press.

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. "Artificial morality: Top-down, bottom-up, and hybrid approaches." *Ethics and Information Technology,* 7, 149-155.

Allen, Colin and Wendell Wallach. 2012. "Moral Machines: Contradiction in Terms or Abdication of Human Responsibility." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 55-68. Cambridge: The MIT Press.

Anderson, Michael, Susan Leigh Anderson, and Vincent Berenz. 2017. "A value driven agent: Instantiation of a case-supported principle-based behavior paradigm." *The AAAI-17 Workshop on AI, Ethics, and Society,* 72-80.

Anderson, Susan Leigh. 2011. "How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 524-530. New York: Cambridge University Press.

Anderson, Susan Leigh. 2011. "Machine Metaethics." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 21-27. New York: Cambridge University Press.

Angwin, Julia, Jaff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica*,

23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arkin, Ronald C. 2008. “Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture.” In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction,* 121-128.

Asimov, Isaac. 1985. *Robots and Empires*. New York: Doubleday & Company, Inc.

Asimov, Isaac. 1950. *I, Robot*. New York: Bantam Dell.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. “Concrete problems in AI safety.” *arXiv: 1606.06565v2,* 1-29.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Hendrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. “The Moral Machine experiment.” *Nature,* 563, 59-64.

Bastani, Aaron. 2019. *Fully Automated Luxury Communism: A Manifesto*. London: Verso.

Beebee, Helen. 2013. *Free Will: An Introduction*. New York: Palgrave MacMillan.

Bekey, George A. 2012. “Current Trends in Robotics: Technology and Ethics.” In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 17-34. Cambridge: The MIT Press.

Bengio, Yoshua. 2009. “Learning deep architectures for AI.” *Machine Learning,* 2(1), 1-127.

Berardi, Franco. 2009. *The Soul At Work: From Alienation to Autonomy*. Los Angeles: Semiotext(e).

Bigman, Yochanan E., Adam Waytz, Ron Alterovitz, and Kurt Gray. 2019. “Holding robots responsible: The elements of machine morality. *Trends in Cognitive Science,* 23(5), 365-368.

Bigman, Yochanan E., and Kurt Gray. 2018. “People are averse to machines making moral decisions.” *Cognition,* 181, 21-34.

Block, Ned. 1995. “On a confusion about a function of consciousness.” *Behavioral and Brain Sciences*, 18, 227-247.

Bojarski, Mariusz, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. 2016. “End to end learning for self-driving cars.” *arXiv: 1604.07316v1,* 1-9.

Bommasani, Rishi, Percy Liang, and others. 2021. “On the opportunities and risks of foundation models.” *arXiv: 2108.07258v3*, 1-214.

Bonnefon, Jean-Francois, Azim Shariff, and Iyad Rahwan. 2016. “The social dilemma of autonomous vehicles.” *Science,* 352(6293), 1573-1576.

Bostrom, Nick. 2004. “The future of human evolution.” In *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, edited by Charles Tandy, 339-371. Palo Alta: Ria University Press.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brentano, Franz. 1995. *Psychology from an Empirical Standpoint*. New York: Routledge.

Bryson, Joanna J. 2020. “The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation.” In *The Oxford Handbook of Ethics of AI,* edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 3-25. New York: Oxford University Press.

Bryson, Joanna J. 2018. “Patiency is not a virtue: The design of intelligent systems and systems of ethics.” *Ethics and Information Technology*, 20, 15-26.

Bureau d'Enquêtes et d'Analyses. 2012. "Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro – Paris." *Paris: BEA*.

Burrell, Jenna. 2016. "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data and Society,* 1-12.

Caruana, Rich. 1997. "Multitask learning." *Machine Learning,* 28, 41-75.

Cave, Stephen, Run Nyrup, Karina Vold, and Adrian Weller. 2019. "Motivations and risks of machine ethics." *Proceedings of the IEEE,* 107(3), 562-574.

Chalmers, David J. 1995. "The puzzle of conscious experience." *Scientific American*, 273, 80-86.

Coeckelbergh, Mark, and David J. Gunkel. 2014. "Facing animals: A relation, other-oriented approach to moral standing." *Journal of Agricultural and Environmental Ethics, 27*(5), 715-733.

Coeckelbergh, Mark. 2020. *AI Ethics*. Cambridge: The MIT Press.

Crane, Tim. 1998. "Intentionality as the mark of the mental." *Royal Institute of Philosophy Supplement,* 43, 229-251.

Danaher, John. 2020. "Welcoming robots into the moral circle: A defence of ethical behaviourism." *Science and Engineering Ethics*, 26, 2023-2049.

Danaher, John. 2018. "Why We Should Create Artificial Offspring: Meaning and the Collective Afterlife." *Science and Engineering Ethics*, 24, 1097-1118.

DeBlanc, Peter. 2011. "Ontological crises in artificial agents' value systems." *The Singularity Institute, San Francisco*.

Dennett, Daniel C. 1989. *The Intentional Stance*. Cambridge: The MIT Press.

Dewey, Daniel. 2011. “Learning what to value.” *Artificial General Intelligence*, 309-314.

Diakopoulos, Nicholas. 2020. “Transparency.” In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 197-213. New York: Oxford University Press.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. “Algorithm aversion: People erroneously avoid algorithms after seeing them err.” *Journal of Experimental Psychology, 144*(1), 114-126.

Domingos, Pedro. 2012. “A few useful things to know about machine learning.” *Communications of the ACM,* 55(10), 78-87.

Dressel, Julia and Hany Farid. 2018. “The accuracy, fairness, and limits of predicting recidivism.” *Science Advances*, 4, 1-5.

Dumouchel, Paul and Luisa Damiano. 2017. *Living With Robots*. Cambridge: Harvard University Press.

European Commission. 2020. “White Paper on Artificial Intelligence – A European approach to excellence and trust.” *COM (2020) 65*, 1-26.

European Commission’s High-Level Expert Group On Artificial Intelligence. 2018. “A definition of AI: Main capabilities and scientific disciplines.” *European Commission,* 1-7.

Flanagan, Owen. 1991. *The Science of the Mind*. Cambridge: The MIT Press.

Floridi, Luciano and J. W. Sanders. 2004. “On the morality of artificial agents.” *Minds and Machines*, 14, 349-379.

Gagné, Robert M. 1972. “Domains of learning.” *Interchange,* 3(1), 1-8.

Gardiner, Stephen M. 2012. “Are We the Scum of the Earth? Climate Change, Geoengineering, and Humanity’s Challenge.” In *Ethical Adaptation to Climate Change: Human Virtues of the Future*, edited by Allen Thompson and Jeremy Bendik-Keymer, 241-260. Cambridge: The MIT Press.

Gardiner, Stephen M. 2006. “A Perfect Moral Storm: Climate Change, Intergenerational Ethics and the Problem of Moral Corruption.” *Environmental Values*, 15, 397-413.

Goodman, Bryce and Seth Flaxman. 2016. “EU regulations on algorithmic decision-making and a “right to explanation.” *arXiv: 1606.08813v1,* 26-30.

Gordon, John-Stewart. 2020. “Building moral robots: Ethical pitfalls and challenges.” *Science and Engineering Ethics,* 26, 141-157.

Gunkel, David J. 2018. “The other question: Can and should robots have rights?” *Ethics and Information Technology, 20*(2), 87-99.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge: The MIT Press.

Heyes, Cecilia M. 1994. “Social learning in animals: Categories and mechanisms.” *Biological Reviews,* 69, 207-231.

Hofstadter, Douglas R. 1999. *Godel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books, Inc.

Hosseini, Hossein, Baicen Xiao, and Radha Poovendran. 2017. “Google’s cloud vision API is not robust to noise.” *arXiv: 1704.05051v2,* 1-5.

Hoyningen-Huene, Paul. 1987. “Context of discovery and context of justification.” *Studies in History and Philosophy of Science,* 18(4), 501-515.

IEEE Standards Association 2017. “Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (A/IS).” IEEE Standards P7000, Version 2.

Illeris, Knud. 2009. “A comprehensive understanding of human learning.” In *Contemporary Theories of Learning: Learning theorists...in their own words,* edited by Knud Illeris, 7-20. Abingdon: Routledge.

Jost, John T., Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, and Curtis D. Hardin. 2009. “The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and execute summary of ten studies that no manager should ignore.” *Research in Organizational Behaviour,* 29, 39-69.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michal Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Peterson, David Reiman, Ellen Clancy, Michal Zienlinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli and Demis Hassabis. 2021. “Highly accurate protein structure prediction with AlphaFold.” *Nature,* 596, 583-589.

Kazez, Jean. 2007. *The Weight of Things: Philosophy and the Good Life*. Malden: Blackwell Publishing.

Kerr, Ian. 2019. “Schrödinger’s robot” Privacy in uncertain states.” *Theoretical Inquiries in Law, 20*(1), 123-154.

Kohen, Ari. 2007. *In Defense of Human Rights: A non-religious grounding in a pluralistic world*. Abingdon: Routledge.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep learning.” *Nature*, 521, 436-444.

Lenat, Douglas B. 1983. “EURISKO: A program that learns new heuristics and domain concepts.” *Artificial Intelligence*, 21, 61-98.

Lenat, Douglas B. 2001. “From 2001 to 2001: Common sense and the mind of HAL. *HAL’s Legacy*, 193-209.

Lipton, Zachary C. 2017. “The mythos of model interpretability.” *arXiv: 1606.03490v3,* 1-9.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. “Algorithm appreciation: People prefer algorithmic to human judgment.” *Organizational Behaviour and Human Decision Processes*, 151, 90-103.

Mańdziuk, Jacek. 2007. “Computational Intelligence in Mind Games.” In *Challenges for Computational Intelligence,* edited by Wlodzislaw Duch and Jacek Mańdziuk, 407-442. Berlin: Springer.

McCorduck, Pamela. 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Natick: A K Peters, Ltd.

McDermott, Drew. 2011. “What Matters to a Machine?” In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 88-114. New York: Cambridge University Press.

Metzinger, Thomas. 2021. “Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology.” *Journal of Artificial Intelligence and Consciousness*, 8(1), 43-66.

Minh, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. “Playing atari with deep reinforcement learning.” *arXiv: 1312.5602v1,* 1-9.

Moor, James H. 2006. “The nature, importance, and difficulty of machine ethics.” *IEEE Intelligent Systems*, 21(4), 18-21.

Moravec, Hans. 1998. “When will computer hardware match the human brain?” *Journal of Evolution and Technology,* 1, 1-12.

Mowrer, Robert R., and Stephen B. Klein. 2001. “Chapter 1 - The Transitive Nature of Contemporary Learning Theory.” In *Handbook of Contemporary Learning Theories,* edited by Robert R. Mowrer and Stephen B. Klein, 1-21. Mahwah: Lawrence Erlbaum Associates.

Nadelhoffer, Thomas and Adam Feltz. 2008. “The actor-observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong’s fire.” *Neuroethics,* 1, 133-144.

Nagel, Thomas. 1986. *The View From Nowhere*. New York: Oxford University Press.

Newell, Allen, John C. Shaw, and Herbert A. Simon. 1959. “Report on a general problem solving program.” In *IFIP Congress,* 256, 1-27.

Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. “Text classification from labeled and unlabeled documents using EM.” *Machine Learning,* 39, 103-134.

Omohundro, Stephen M. 2008. “The basic AI drives.” In *Artificial General Intelligence*, edited by Pei Wang, Ben Goertzel and Stan Franklin, 483-492. Amsterdam: IOS Press.

Ontañón, Santiago, Gabriel Synnaeve, Alberto Uriarte, Florian Richoux, David Churchill, and Mike Preuss. 2013. “A Survey of Real-Time Strategy Game AI Research and Competition in *StarCraft*.” *IEEE Transactions on Computational Intelligence and AI in Games,* 5(4), 293-311.

Parkhi, Omkar M., Andrea Vedaldi, Andrew Zisserman and C. V. Jawahar. 2012. “Cats and dogs.” *IEEE Conference on Computer Vision and Pattern Recognition,* 1-8.

Peterson, Steve. 2012. “Designing People to Serve.” In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 283-298. Cambridge: The MIT Press.

Preston, Jesse L., and Faith Shin. 2021. “Anthropocentric Biases in Teleological Thinking: How Nature Seems Designed for Humans.” *Journal of Experimental Psychology: General*, 160(5), 943-955.

Railton, Peter. 2020. “Ethical Learning, Natural and Artificial.” In *Ethics of Artificial Intelligence,* edited by S. Matthew Liao, 45-78. New York: Oxford University Press.

Raina, Rajat, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. “Self-taught learning: Transfer learning from unlabeled data.” *Proceedings of the 24th International Conference on Machine Learning,* 1-8.

Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. “Overtrust of Robots in Emergency Evacuation Scenarios.” In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 101-108.

Russell, Bertrand. 1910. “Knowledge by Acquaintance and Knowledge by Description.” *Proceedings of the Aristotelian Society*, 11, 108-128.

Samuel, Arthur L. 1967. “Some studies in machine learning using the game of checkers. II - Recent progress.” *IBM Journal,* 11(6), 601-617.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. “Auditing algorithms: Research methods for detecting discrimination on internet platforms.” *64th Annual Meeting of the International Communication Association,* 1-23.

Santayana, George. 2011. *The Life of Reason: Introduction and Reason in Common Sense*. Cambridge: The MIT Press.

Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Larent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. 2020. “Mastering Atari, Go, chess and shogi by planning with a learned model.” *Nature*, 588, 604-609.

Schwall, Matthew, Tom Daniel, Trent Victor, Francesca Favarò, and Henning Hohnold. 2020. “Waymo Public Road Safety Performance Data.” *arXiv: 2011.00038v1,* 1-15.

Searle, John R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. New York: Cambridge University Press.

Shane, Janelle. 2019. *You Look Like A Thing And I Love You*. New York: Little, Brown and Company.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. "Mastering the game of Go without human knowledge." *Nature,* 550(7676), 354-372.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Larent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science*, 362(6419), 1140-1144.

Smith, Brian C. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Massachusetts: The MIT Press.

Srivastava, Abhinav, Amlan Kundu, Shamik Sural, and Arun K. Majumdar. 2008. "Credit card fraud detection using hidden markov model." *IEEE Transactions on Dependable and Secure Computing*, 5(1), 37-48.

Steed, Ryan and Aylin Caliskan. 2021. "Image representation learned with unsupervised pre-training contain human-like biases." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701-713.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge: The MIT Press.

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred A. Knopf.

Tesauro, Gerald. 1993. “TD-Gammon, a self-teaching backgammon program, achieves master-level play.” *AAAI Technical Report FS-93-02,* 19-23.

Tolmeijer, Suzanne, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. “Implementations in machine ethics: A survey.” *arXiv: 2001.07573v1,* 1-37.

Turing, Alan M. 1950. “Computing machinery and intelligence.” *Mind,* 59(236), 433-460.

Verma, Sahil, and Julia Rubin. 2018. “Fairness definitions explained.” In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare),* 1-7.

Vinyals, Oriol, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Phlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. “Grandmaster level in StarCraft II using multi-agent reinforcement learning.” *Nature,* 575, 350-354.

Vinyals, Oriol, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzabim Heinrich Küttler, John Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hado van Hasselt, David Silver, Timothy Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. 2017. “StarCraft II: A New Challenge for Reinforcement Learning.” *arXiv: 1708.04782v1,* 1-20.

Wallach, Wendell and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Wallach, Wendell and Shannon Vallor. 2020. “Moral Machines: From Value Alignment to Embodied Virtue.” In *Ethics of Artificial Intelligence,* edited by S. Matthew Liao, 383-412. New York: Oxford University Press.

Walmsley, Joel. 2021. “Artificial intelligence and the value of transparency.” *AI and Society*, 36, 585-595.

Wang, Fei-Yue, Jun Jason Zhang, Xinhu Zheng, Xiao Wang, Yong Yuan, Xiaoxiao Dai, Jie Zhang, and Liuqing Yang. 2016. “Where does AlphaGo go: From Church-Turing thesis to AlphaGo thesis and beyond.” *IEEE/CAA Journal of Automatica Sinica,* 3(2), 113-120.

Wang, Han, Soheil Salehi, Hossein Sayadi, Avesta Sasan, Tinoosh Mohsenin, Sai Manoj P. D., Setareh Rafatirad, and Houman Homayoun. 2021. “Evaluation of Machine Learning-based Detection against Side-Channel Attacks on Autonomous Vehicle.” In *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS),* 1-4.

Winfield, Alan F. T., Christian Blum, and Wenguo Liu. 2014. “Towards an ethical robot: Internal models, consequences and ethical action selection.” In *Conference Towards Autonomous Robotic Systems*, 85-96.

Winfield, Alan F. T., Katina Michael, Jeremy Pitt, and Vanessa Evers. 2019. “Machine ethics: The design and governance of ethical AI and autonomous systems.” *Proceedings of the IEEE,* 107(3), 509-517.

Wu, Yueh-Hua, and Shou-De Lin. 2018. “A low-cost ethics shaping approach for designing reinforcement learning agents.” In *Thirty-Second AAAI Conference on Artificial Intelligence,* 1687-1694.

Yampolskiy, Roman V. 2014. “Utility function security in artificially intelligent agents.” *Journal of Experimental and Theoretical Intelligence,* 26(3), 373-389.

Zou, James, and Londa Schiebinger. 2018. “Design AI so that it’s fair.” *Nature,* 559(7714), 324-326.