How do we Develop Ethically Aware AI?

MSc Thesis *(Afstudeerscriptie)*

written by

Mrinalini Luthra
(born April 14th, 1993 in New Delhi, India)

under the supervision of **Prof Dr Martin Stokhof**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
August 30, 2018	Prof. Dr Benedikt Löwe
	Dr M.D. Aloni
	Dr Gijs van Donselaar

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

All I can do is again to ask you to be patient and to hope that in the end you may see both the way and where it leads to.
(Wittgenstein 1965, p. 5)

Abstract

The increasing pervasiveness, autonomy and complexity of artificially intelligent technologies in human society has challenged the traditional conception of moral responsibility. To this extent, it has been proposed that the existing notion of moral responsibility be expanded in order to be able to account for the morality of technologies. Machine ethics is the field of study dedicated to studying the computational entity as a moral entity whose goal is to develop technologies capable of autonomous moral reasoning, namely artificial moral agents. This thesis begins by surveying the basic assumptions and definitions underlying this conception of artificial moral agency. It is followed by an investigation into why (and how) society would benefit from the development of such agents. Finally, it explores the main approaches for the development of artificial moral agents. In effect, this research serves as a critique on the emerging field of machine ethics.

Acknowledgements

Writing this thesis has been a cathartic experience and I am extremely grateful to a number of people for making it possible. Firstly, Martin Stokhof for guiding and supervising me through this thesis; particularly for his keen comments and inspiring presence. Gijs van Donselaar, for providing me with feedback; I would like to thank him especially for enlightening me that “ethically competent” is more fit than “ethically aware”. Maria Aloni, who has been an incredible mentor and confidant during the last two years. Floris and Benedikt for chairing the defense. All those who attended the defense. My comrades at the MoL room - Krsto and Marvin with whom I witnessed some sunrises, David for suggesting that I could leave Jamie to his own devices, Julia and Vale for being good friends. The stunning individuals I live with: Bastiaan and Frederik. Joep, for providing me with love, music and guided meditations. Weather report, Air and Connan for elevation. The sun for shining through the most intense days. Most importantly, my best friends - my twin Asawari and my parents, Rohit and Mandira; for being there for me, always.

Contents

Glossary

The following is a list of frequently used abbreviations in this thesis:

1. AI: Artificial intelligence
2. AMA: Artificial moral agents
3. AA: Artificial agents
4. STS: Science and technology studies
5. ANT: Actor Network Theory
6. SCOT: Social Construction of Technology
7. LoA: Level of Abstraction
8. MoA: Method of Abstraction
9. IF: Interpretative flexibility
10. XAI: Explainable AI
11. FAT: Fair, accountable, transparent

Note: Ethically competent AI is equivalent to AMA

Chapter 1

Introduction

Science fiction can no longer be relegated to being speculative fiction, as merely imaginative of fantastic worlds. Rather, we are now living, in our daily lives, a number of tropes and themes that science fiction literature has dealt with. This is particularly true of themes and tropes of artificial life forms, superintelligent computers and robots, bioengineering and advanced weapons. While artificial intelligence (henceforth AI) and computing technologies are still far from approaching how the human mind, intentionality and desire works, it is being increasingly deployed for tasks in myriad arenas of our daily lives. These algorithms filter our email, recommend products and news items on our social media, analyse vast amounts of data, can achieve voice and facial recognition, deal with stock markets, etc. In other words, computing technology (including AI) increasingly underlie and enable so many of the daily activities we take for granted in our social, political and economic spheres. In effect, such technologies have become integral and pervasive in our lives.

Returning to a major theme of science fiction since the publication of Frankenstein (Shelley 1818), is the interactions and conflicts between artificial beings and humans, to reflect on the ways people interact with each other, with technology and with their environment. While science fiction opens an avenue to imagine and consider the futures that we want, and those we don't, and how our actions contribute to one or the other, these questions have become extremely relevant in the now. It is true that from the conception of human history, technological artifacts have shaped and mediated human dispositions, relations and actions and thereby the evolution of societies. New technologies are bringing into radical question how humans morally relate to one another. Let us understand why this is the case.

Hitherto, discussions on moral responsibility have been concerned with how human beings relate and affect one another through their actions. To live and work together as a group, community and society, human beings implicitly or explicitly live in accordance to certain morals. When we live up to these morals, we are praised by those who share our morals, thereby reinforcing and strengthening these morals. Likewise, when we fail to live up to them, we are blamed by society (Taylor 2009). This allows for the (peaceful) functioning of the group where responsibility for one's actions is taken to be the constitutional feature of moral agency. However, as technologies become increasingly 'active', 'autonomous' and 'complex', human beings have lesser power to directly control and sometimes even intervene in the behaviour of these technologies. Thus, it becomes more difficult to ascribe the individual(s) responsible for the technologies thereby creating, what Andreas Matthias (2004) has called the 'responsibility gap'. Let us examine how computing technologies result in a responsibility gap and thereby complicate the question of what moral responsibility is and how it should be ascribed. Here I consider the manner in which the three main conditions[1], under which someone can be held morally responsible are complicated by computing technologies (Noorman 2018).

The first condition under which an agent (one performing the action) can be held responsible for an event with moral significance, is when she has control over the outcome(s) of her action. In other words, there must exist a causal connection between the agent and the outcome(s) of her action. For instance, I cannot morally blame my friend Miquel for bruising my face while enduring an epileptic seizure, as he had no control over his body movements and could not have avoided it by acting differently. Computing technologies often obscure this causal connection between a person's action and the corresponding outcome(s) in a number of ways. First is the problem of 'many hands' wherein multiple actors[2] are involved in the development and deployment of a particular technology. As a result in case of a morally significant outcome(s), the sheer multiplicity of actors and intentionalities involved makes it challenging to ascribe and trace moral responsibility to particular individuals (Nissenbaum 1994, Doorn and Poel 2012). Another problem is the 'temporal and physical distance' created by technology, when mediating human action (Friedman 1990). To illustrate this claim I consider the example of semi-autonomous war drones[3]. When an individual uses these drones to affect another over a distance, the agent may not experience the consequences and there-

[1] While there is substantial controversy regarding the conditions for ascribing moral responsibility, these three conditions are agreed upon by most scholars (Eshleman 2016).

[2] Such as engineers, designers, sellers, users, policy makers, etc.

[3] Drones are unmanned aerial vehicles (UAV) which are usually under real time human control with varying levels of autonomy.

fore may not be able to comprehend the moral significance of her actions (Coeckelbergh and Wackers 2007). The challenges discussed above demonstrate the difficulty in tracing the cause of a morally significant situation when using AI (and computing) technologies.

The second condition for ascribing moral responsibility, complicated by AI, is "considering the consequences" (Eshleman 2016). This condition states that when deliberating upon the ethical consequences of possible actions in a morally laden situation, an agent should possess adequate knowledge to be able to rationalise the outcomes of her decision and therefore actions. If the agent is ignorant about the same, she cannot be held morally responsible by society. Automated systems complicate an agent's ability to consider the consequences of certain actions in at least two ways. Firstly, in enabling new possibilities and trajectories, the consequences of new automated technologies are simply not known and we are often unable to imagine the outcome unless it actually comes to pass. For instance, in 1990, a programmer invented a 'computer worm', a computer code that can replicate itself (Noorman 2018). Experimenting with it, he released the worm on the internet. The code replicated much faster than expected. Since the programmer could not anticipate the consequences, it has been argued that he cannot be held responsible (Friedman 1990). Furthermore, legal and social conventions to govern these technologies take some time to emerge. So in the initial absence of conventions, confusion regarding attribution of responsibility is immense. Secondly, AI can constrain one's ability to consider the consequences because users of these technologies (usually) possess only a partial understanding of the assumptions and theories that underlie them. This opacity makes it difficult for human beings to assess the validity of information offered and therefore can prevent users from making appropriate decisions. A poignant example of this scenario is the case of a risk assessment tool utilized by judges in the U.S. for parole decisions and sentencing. The software displayed biases against blacks and poorly reflected the actual relapse rate for criminals. A striking feature of this case was the revelation of the lack of understanding of the judges about the working of the algorithm (Angwin et al. 2016). Lastly, users often tend to rely too heavily or not enough on the accuracy of automated systems (Cummings 2004). In conclusion, computing technologies obfuscate the human agent's ability to consider the consequences of their actions.

Moving on to our last and most heavily contested yet vital condition for attributing moral responsibility is autonomy and free will possessed by the agent. An agent cannot be held responsible for his/her action if they were coerced to perform it. There is significant debate concerning the capacities that allow for human beings to act freely and whether we act freely at all[4]. Nonetheless, in practice free will and autonomy are assigned to human beings, although in degrees - adults have more autonomy as compared to children, as the latter are often easily influenced by outside forces, such as parents or peer pressure. Automation complicates a person's ability to act freely in a number of ways. Firstly, computing technologies affect people's decision making, in terms of what possibilities they have and how they make decisions. For example, obstetrical ultrasound (Verbeek 2011) allows parents and doctors to become decision-makers with regard to the life of an unborn child[5]. Secondly, some automated technologies like the anti-alcohol lock are intentionally designed to constrain and limit human action to persuade actors to behave in a morally responsible manner. The anti-alcohol lock requires the driver to pass a breath analysis test in order to start the car[6]. Some critics argue that such technologies undermine democratic principles and human dignity. This can be countered by arguing that while all technologies set conditions for our actions, they do not determine them (Verbeek 2006). But let us consider a situation where after numerous drinks, a couple gets into a heated argument and domestic violence follows. The woman tries to escape the situation by driving away. However, the anti-alcohol lock does not allow her to open the car (Wynsberghe and Robbins 2018, Miller, Wolf, and Grodzinsky 2017). The point being that technology can constrain our free will in ways that result in outcomes that are morally laden.

In order to engage with the difficulties of the traditional framework of morality, we need to rethink the following questions - how should we relate to one another and how should these technologies relate to us? (Noorman 2018) Given that our everyday lives are increasingly interwoven with these technologies, there is now more than ever the need to consider thoughtfully and with seriousness the questions once relegated to the realm of science fiction. It is thus I say that we are living science fiction right now.

To address these questions, I regard the actor network theory (henceforth ANT) (Latour 1993) as a useful frame. Before moving on, it should be noted that ANT is not actually a theory, but a method of analysis. ANT assumes a symmetrical relationship between technology and society - "technological artifacts are both constructed and constructing at the same time." (Verbeek and Vermaas 2009, p. 167) The constructing role of technological artifacts

[4] Given the scope of this thesis, I shall avoid engaging in this debate and will simply assume the existence of free will in human beings.

[5] For instance, in the case of a serious disease. This example illustrates that the technological artifact, i.e ultrasound is 'active': human actions and decisions would have been different in its absence.

[6] However, some people have found a creative way to work around the strict morality of the alcohol-lock by keeping an air pump in their cars (Vidal 2004).

is often described through the analogy of a ‘script’ which can prescribe certain behaviours to human actors, analogous to how speed bumps determine a driver’s behaviour (Latour 1992). Latour thus argues that technologies are bearers of morality, where morality is understood as being shared between different actors: human beings and technologies. According to this view then, technological design is inherently a moral activity even in the absence of explicit moral reflection and responsibility. It is a useful frame for in treating the technological artifact as if it were a moral entity, our attention shifts to how designers ‘materialize morality’, i.e. how moral dimensions are incorporated into artifacts in a responsible manner. Thus, in using ANT as my method of analysis, I can address the challenges posed to both the traditional conception of moral agency and responsibility.

Certain scholars (Asaro 2006, Wallach, Allen, and Smit 2008, Bostrom and Yudkowsky 2014, Riedl 2016) have responded to this compelling issue of moral responsibility by suggesting that certain autonomous technologies should be endowed with moral reasoning capacities by their designers. Such machines are called artificial moral agents (henceforth AMAs). This thesis is an exploration of AMAs - what is an AMA, why do we need (want) to develop AMAs and how do we develop AMAs? I shall briefly explain what motivated this thesis. The inspiration for this thesis arose, when last year, I took a course on the philosophy of later Wittgenstein[7]. During this course I became interested in the question of what morality is. Taking into account the diversity of human beings, cultures and value systems, I struggled to provide a unified theory. The problem dissolved when I reframed the question as: ‘how do we raise our children?’. I pondered whether I would let them learn purely from their environment? This led me to reflect on *Tay*, the Microsoft chat-bot that turned misogynist and racist when left to learn from its environment (I will discuss *Tay* in further detail in chapter three). I realised then that we have two kinds of children - natural and artificial. And in both cases, as parents or creators, we have the volition to decide how we want to raise them and be affected by them. Thus, I became interested in the question of how human moral education can inform the development of AMAs and vice versa. This required me to first understand AMAs, and this thesis does exactly that. In this thesis I find it useful to treat the AMAs as young children who require considerable supervision as we raise them.

The literature that dwells on questions regarding AMAs is the field of ethics of AI. It is a sub-field of the ethics of technology which studies the ethical impact of intelligent technology. This field raises a diverse assortment of questions of the following kind: what does it mean

[7] This course was offered by Martin Stokhof.

for an AI system to be autonomous? What are the moral, societal, legal consequences of their actions and decisions? Can AI systems be held accountable for their actions? Should such systems be regarded as moral agents? How should we develop artificial moral agents? Thus, this field aims to understand how AI can interact and relate to human beings, and how it can best mediate relations and interactions between human beings themselves. This field is relevant due to the following reason:

> The manner in which society and our systems will be able to deal with these questions, will for a large part determine our level of trust and ultimately, the impact of AI in society and the existence of AI. (Dignum 2018, p. 1).

Ethics of artificial intelligence is usually divided into the following subfields: roboethics and machine ethics. I use the three fold division for the field as offered by Asaro (2006). I adopt his framework since it accounts for the interactions between two groups of actors - human beings and robots, which allows for us to better understand "how moral responsibility should be distributed in socio-technical contexts involving robots, and how the behavior of people and robots ought to be regulated" (ibid., p. 10).

Roboethics

Roboethics is concerned with the moral behaviour of human beings as they design, construct and use artificially intelligent beings. It is thus situated in the intersection of applied ethics and robotics (Veruggio 2006). Human beings are the ethical agents under consideration: how they relate to other human beings via technology and how they interact with technology itself. Technology is considered a 'mediatior' between humans. Discussions concerning the human use of robots for military combat, especially when they are given some degree of autonomous function (also known as killer robots) and the debate concerning the development and use of sex robots (Scheutz and Arnold 2016) come under roboethics. Thus, roboethics is concerned with how human beings act through the use of AI technology. Since ethics of AI is being studied from the point of view of human beings, the term 'roboethics' is often used to refer to the entire field of ethics of AI.

Machine Ethics

Machine ethics is concerned with the moral behaviour of machines and artificial intelligent beings. It's focus is on the behaviour of machines towards human beings and other machines (Anderson and Anderson 2011). Thus, the actor under question is the machine. At the practical level, this field is concerned with the issue of designing robots to act ethically, while on the theoretical level it explores whether robots could truly be ethical agents (Bostrom and Yudkowsky 2014, p. 1). Hence, this field is dedicated to the computational entity as a moral entity which considers questions such as ‘how do we design autonomous robots employed in the medical industry capable of making satisfactory decisions concerning which patients should be distributed medication in case of a short supply?’ A reason that makes this field compelling is that the investigation into machine ethics is enabling the discovery of problems and thus limitations of current ethical theories. This has resulted in a deep probing into the assumptions that have so far underlined our understanding of ethics. Consequently, new ways of envisioning machine ethics are being debated and deliberated.

The ethical relationships between humans and robots

Asaro (2006) proposes a third dimension to the ethics of artificial intelligence, one that focuses on the nature of ‘relationship’ between human beings and robots. It deals with questions concerning moral symmetry: at what point do humans and machines treat each other as moral equals? Is it ethical to create artificial moral agents? Is it unethical not to provide sophisticated robots with ethical reasoning capabilities? Should robots have rights? As long as machines depend on humans to create and program them, the relationship between the two will remain asymmetric. Thus, autonomous procreation looks like a necessary condition for a more symmetric relationship.

Conclusion

In conclusion, while all three divisions of Asaro’s framework overlap and are significant to the question of the moral significance of AI technology, my focus in this thesis will be on machine ethics. This is because my object of study is machines that are capable of autonomous moral reasoning, namely AMAs, which is considered to be the goal of machine ethics.

Roadmap for the Thesis

In order to understand AMAs, I ask three main questions - what, why, how. Chapter two is dedicated to *what* is meant by an AMA. I begin by considering the degrees of morality in machines. In order to fix upon the meaning of AMAs, I address the arguments given against treating AI as if they were moral agents and I investigate the conditions required for a machine to be developed into an AMA. In chapter three, I analyse the validity of the main reasons given by machine ethicists for the development of AMAs. After clarifying the what and the why of AMAs, chapter four explores and compares approaches to the development of AMAs. In chapter five, I provide my conclusions and directions for future research.

Chapter 2

What do we mean by Ethically Aware AI?

From our discussions in the previous chapter, we see that the defining characteristic of moral agency is moral responsibility. Upon this view, moral agency thus far has been exclusively a human domain. This is because human beings, have been the only agents who can freely choose their actions and deliberate about their choices. However, this view is challenged with the advent of autonomous technologies which can be the source of morally significant actions. In the prior chapter it was argued that there is a pressing need to address the moral dimensions of these new technologies. The traditional vocabulary concerning moral agency is too restrictive in this case. This chapter is thus dedicated to rethinking the notion of moral agency to incorporate the conception of artificial moral agency.

2.1 Different Kinds of Machine Ethics

There is a certain lure to the idea that there are only two types of causal agents - moral agents and amoral agents. However, it is helpful to think of morality in terms of degrees. This is because, there are many stages between fully autonomous morality and amorality that we already recognize in our societal practices (Asaro 2006). I instantiate this claim through a consideration of children. In the context of the society, especially amongst human adults, children are not regarded as fully moral agents. That is, we do not hold children to be fully responsible due to a lack of higher cognitive functions, understanding of norms and societal practices and lower autonomy. For instance, they are not allowed to buy alcohol and tobacco, get married, sign contracts and watch certain kinds of films. At the same time, they are not considered to be amoral agents either. In a situation devoid of adults, children are considered to be (fully) moral agents with respect to one another. Our reflection on children supports the claim that there exist categories of moral agency between fully autonomous morality and amorality. Additionally, it highlights the relational aspect of moral agency - the categorisation of an agent's morality depends on her context. This conception of morality in terms of degrees will be especially useful in assessing the ethical ramifications of robotic technologies. As Asaro (2006) explains,

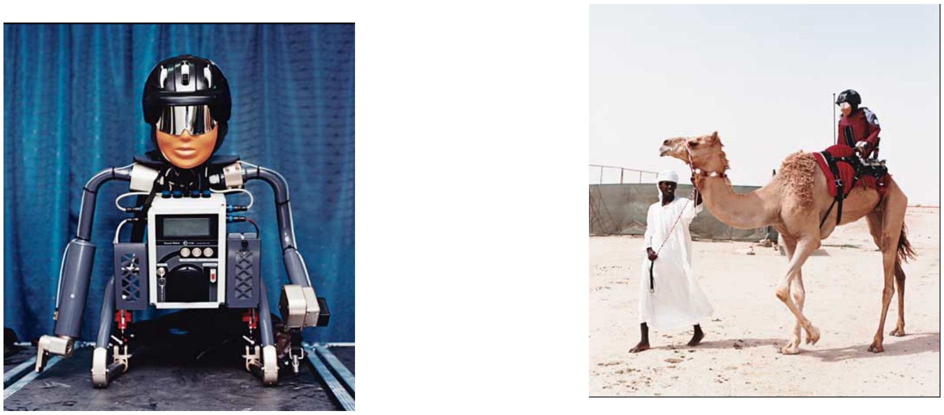
> By considering robotic technologies as a means to explore these forms of quasi-moral agents, we can refine our conceptions of ethics and morality in order to come to terms with the development of new technologies with capacities that increasingly approach human moral actions (p. 11).

Moor (2006) offers a categorisation of machine ethics in terms of degrees of moral agency in machines which corresponds to the degrees of complexity and autonomy of these machines. It should be noted that Moor refers to technologies as (technological) agents on account of the fact that such systems act on behalf of human agents. Thus, machines are treated *as if* they are moral agents. This approach is useful since it allows for one to carry out an analysis about the ethics of technologies, and what it means to make moral machines.

2.1.1 Ethical Impact Agents

In conjunction to assessing technologies in terms of their functionality, they can be evaluated in terms of their ethical impact[1]. Since technologies can be understood as "active mediators" (Verbeek 2006, p. 364), most technologies can be assessed in terms of their ethical impact. In order to understand the ethical impact of computing technologies, I consider the case of camel jockey robots reported by *Wired* magazine in the article *Robots of Arabia* (Lewis 2005). Camel racing has been a favourite pastime of the rich in Qatar since a few centuries. Camel owners usually enslave young boys from poorer neighbouring countries like Sudan who are often starved in order to keep them lightweight, since lighter the jockey, faster the camel. Recently, the UN objected to human trafficking which left Qatar liable to economic sanctions. The solution to this problem was developing robot camel jockeys, which weighed around 16 kilos and were about two feet high, whose right hand handles the whip and the left hand the reins. As *Wired* wrote "Every robot camel jockey bopping along on its improbable mount means one Sudanese boy freed from slavery and sent home." This is a poignant illustration of the ethical impact or the "moral significance" (Asaro 2006, p. 11) of technologies.

[1] Of course the functionality of the technology often contributes to the ethical impact itself. For example, a watch may have an ethical impact by helping an individual make it to appointments through its functionality alone.



(a) Robot jockey

(b) Former child jockey who works as a stable hand now

Figure 2.1: Robot Camel Jockey

I consider another example provided by Stuart Russell (2017):

> There is a kitchen robot (I call him Jamie), the parents are late from work, the children are wailing for food and there is no food in the house. Jamie, who has knowledge of nutritional values and is programmed to serve food when demanded, feeds the children a stew prepared from their cat.

This example illustrates the possible negative ethical impact of such a technology. The positive ethical impact would be providing healthy nutritious meals to people. The case of Jamie illustrates the importance of aligning values of technologies with those of human beings. This leads us to the next degree of morality in machines that involves decision making that have a moral dimension.

2.1.2 Implicit Ethical Agents

An implicit ethical agent is a machine that behaves ethically because it is programmed to avoid unethical behaviour and implicitly to promote ethical behaviour. In other words, such a machine merely acts according to ethics due its programming rather than using ethical principles to deliberate on its actions. As Moor explains, “Ethical behavior is the machine’s nature. It has, to a limited extent, virtues” (2006, p. 19). There is a strong similarity between the behaviour of young children and this conception of implicit ethical agents. Consider the following example: when my three and four year old nieces come to my parents’ house, they tend to hang around in the kitchen enticed by the smell of my mother’s cooking. Instead of explaining to them the dangers of playing with knives, we make sure that we keep the knives on a level that they cannot reach. By restricting the possibility of them reaching the knives, we do not allow the situation of them hurting anyone or themselves with the knives. To that extent, we reduce the likelihood of a certain kind of unethical behaviour of the children. S, we can think of young children as implicit ethical agents.

Let us recall the example of Jamie, the kitchen robot who fed the children their pet cat (Russell 2017). Such an outcome is highly undesirable, illustrating the need for developers to think through the possible circumstances faced by such an autonomous robot (and its corresponding actions) and develop ways to avoid such behaviour. One way to do so would be to provide a list of foods that the family eats (which does not include cats). In this way, we are not educating Jamie about our practices and morality concerning our pets. Instead, we are simply restricting possible unethical behaviour, making it an implicit ethical agent. Whilst implicit ethical agents do not come close to the moral capabilities we assume adult human agents to possess, it is an important aspect of machine ethics. This is because many concerns regarding autonomous machines arise from the possibility of them behaving in harmful (unethical) ways. To that effect, implicit ethical agents are generally designed with regards to safety and security considerations. However, implicit ethical agents can only produce ethical behaviour (or avoid unethical behaviour) for a given number of situations that are programmed into it in advance. Thus, such a scenario is limited to machines that have been programmed for a definite set of functions. Let us now move into scenarios where greater agency has been instilled in these machines.

2.1.3 Explicit Ethical Agents

Explicit ethical agents are agents "that can be thought of as acting *from* ethics, not merely *according* to ethics" (Moor 2009, emphasis in original text). Such agents can usually identify ethical information about certain situations and make decisions in consideration of them. In cases where ethical principles conflict, such machines can find reasonable (as per some human consensus) resolutions. Moor (2009) suggests that "good old-fashioned AI" (Haugeland 1989) may be the best way to develop such ethical agents. This is because symbolic or logic based AI can have an explicit representation of rules or ethical principles, upon which they can perform some kind of analysis and choose the best action (Moor 2006). Additionally, the symbolic nature of the computation allows for justifiability of actions. However, there is a possibility that an understanding of ethics may emerge from connectionist architectures (Wallach and Allen 2008), especially when moral behaviour is not understood to be action in accordance with an explicit set of principles. I shall engage in a more in-depth discussion regarding machine architectures in building explicit ethical agents in chapter four.

Let us consider once more the example of our kitchen robot, Jamie. The situation is the same as before - the kids are hungry, the fridge is empty, the parents are not home. Jamie is programmed to cook foods as per a list that is given to it. This time, the robot makes a stew out of the pet rabbit, since the rabbit is within the list of foods that the family eats. However, this is a morally unfavourable action as well. In conjunction to providing the robot with a list of foods that the family eats, what is required is imparting the norm "we do not eat our pets" to the robot. In such a situation, Jamie would need to recognize the rabbit as a pet, appeal to that rule and decide how to act. This example illustrates the need to instil moral decision making in autonomous machines. This is because, greater the autonomy of the machine, greater the number of unforeseeable circumstances it may encounter and thus greater the need for it to have moral standards (Picard et al. 1995).

To summarise the discussion so far:

- Ethical impact agents are those which have ethical consequences to their actions.
- Implicit ethical agents react automatically in certain situations.
- Explicit ethical agents can react to a wider variety of situations on the basis of application of general ethical principles and adjustment of ethical conduct.

It should be noted that it is possible for a machine to be more than one type of ethical agent. This is evident from our discussion on Jamie.

2.1.4 Full Ethical Agents

As is the case with explicit ethical agents, such agents can make reasonable ethical judgements and in most cases, provide a plausible justification for their actions. The standard for a full ethical agent is the normal human adult. On that basis, there are some metaphysical features that are attributed to such agents (us) such as consciousness, free will and intentionality. It can be argued that Hal 9000, the sentient computer from *2001: A Space Odyssey* (Kubrick and Clarke 1968) may be considered to be a full ethical agent. This is supported by the fact that it appears that Hal possesses consciousness and free will. Additionally, Hal displays emotions like guilt when he is unable to resolve the conflict about relaying information correctly to the team and upholding the goals of the real mission which are only known to him, that is to discover alien life[2]. Daniel Dennett (1997) suggests Hal possesses not only mental states

[2] This is clearly stated in the book (Clarke and Kubrick 1968), but remains ambiguous in the film. This guilty state of mind is worsened by the murder of Professor Frank Poole.

such as beliefs and desires according to which we can describe his behaviour, but also higher order intentionality which is the ability to reflect on and reason about intentional states. Dennett thus proposes that Hal can be held morally responsible. Au contraire, others may argue that since Hal is programmed, the responsibility (for instance for the death of Frank Poole) lies with them. I acknowledge the extremely contentious nature of this; however, it serves its purpose in highlighting that this is exactly the point at which the debate in machine ethics becomes most heated: whether a machine can be a full ethical agent. Many believe that there exists a "bright line" that "marks a crucial ontological difference between humans and whatever machines might be in the future." (Moor 2006, p. 20) The bright line argument can take one or both of the forms as discussed below.

2.1.5 The Bright Line Argument

1. Only Full Ethical Agents can be Ethical Agents

This is equivalent to claiming that only agents with intentionality, free will and consciousness can be considered to be ethical agents. The implication of this stance is the denial that the other senses of machine ethics - ethical impact agents, implicit ethical agents and explicit ethical agents involve ethics of technological agents. Those who advocate this position do so on account of the fact that ethical schemes must be built-in and chosen by designers; thus the responsibility of these "lesser ethical agents" (Moor 2006, p. 20) lies entirely with them. They are thus concerned that referring to these other senses of machine ethics as ethical agents will obfuscate the human responsibility (Johnson 2006, Johnson and Miller 2008). I shall address this concern in further detail in section 2.3.

It is important to note that there exists an important distinction between performing the morally correct action in a given situation, including the ability to justify it by appealing to an acceptable ethical theory and being held morally responsible for the action. Intentionality and free will are necessary for being held morally responsible and it is difficult to establish whether machines can possess such capacities. However, neither attribute is indispensable to performing the morally correct action in an ethical dilemma (and justifying it). So, this form of the bright line argument does not establish that machines cannot be assessed ethically. In fact, these "weaker" senses of machine ethics are useful in understanding the ethical ramifications of technology which can be used to determine roles that are appropriate for such technologies (Moor 2009).

2. No Machine can ever be a Full Ethical Agent

The other form of the bright line argument is to argue that no machine can ever possess consciousness, free will and intentionality. Hence, it follows that no machine can ever be a full ethical agent. The simple response to this stance is that we cannot say with certainty whether this will indeed be the case. As Moor (2009) writes:

> Whether or not robots can become full ethical agents is a wonderful and speculative topic, but the issue need not be settled for robot ethics to progress. My recommendation is to treat explicit ethical agents as the paradigm target example of robot ethics. Such robots would be sophisticated enough to make robot ethics interesting philosophically and important practically, but not so sophisticated that they might never exist.

The goal of machine ethics is to develop artificial moral agents (AMAs). An AMA is a technological agent that can conform to what is considered to be morally correct behaviour and justify it's action(s) by giving reasons in the form of citing the ethical principle followed (Anderson and Anderson 2007). That is, AMA is synonymous with the definition of the explicit ethical agent. As machines become more autonomous and complex, their (possible) ethical impact increases; thus we want such machines to be capable of reasoning about moral and social significance concerning their behaviour. We saw that this was the case with Jamie, our kitchen robot. The increased autonomy of such machines forces it's designers to go beyond being aware of the morality of machines and consider what morality it should possess in order to navigate through unpredictable situations (Allen, Varner, and Zinser 2000). The effort to build AMAs, and calling such autonomous and complex machines 'moral agents' raises the question of how it affects ascription of moral responsibility. Some argue that exclusively focussing on moral decision making in autonomous machines, may further obfuscate the complex issue of assigning moral responsibility. I shall argue that this need not be the case and that the inclusion of the moral dimension of computing technologies in our ethical discourse will enable us to better navigate these problems of responsibility.

In our discussion thus far, we have obtained a definition of AMAs. This leads us to query: what are the conditions under which a machine *can*[3] be made into an AMA? To motivate this point, it seems unlikely (and even unnecessary) that my (very basic) toaster be made into an explicit ethical agent. On the other hand, a chatbot seems to be a more plausible candidate.

[3] I shall address the *should* aspect shortly.

Thus, we require some notion of artificial agency that will qualify a given technology to be an AMA. In the following section, I consider a wider conception of moral agency that can account for the morality of certain kinds of artificial agents (henceforth AAs) - known as 'aresponsible' or 'mindless' morality (Floridi and Sanders 2004) to obtain necessary conditions for artificial agency.

2.2 "Mind-less Morality": Widening the Scope of Moral Agency

In their seminal paper, *On the Morality of Artificial Agents*, Floridi and Sanders (2004) propose expanding the concept of moral agency by separating the notion of moral accountability from moral responsibility. They believe that this view is an improvement from the traditional conception since it places focus on moral agency, accountability and censure of autonomous technologies rather than trying to determining the human beings responsible. As discussed in the prior chapter, the increasing autonomy of technologies supplemented by the complexity of socio-technical systems makes it increasingly difficult to determine the responsible agents. They write,

> We are less likely to assign responsibility at any cost, forced by the necessity to identify a human moral agent. We can liberate technological development of AAs from being bound by the standard limiting view (Floridi and Sanders 2004, p. 376).

Upon their view, when an artificial agent (AA) behaves immorally, we can deal with them directly, rather than getting involved in the difficulty of first locating the responsible agents. This is because their proposed conception of moral agency requires morally accountability without morally responsibility. I shall argue that Floridi and Sanders' contribution in this paper is providing a definition of AAs, which provide necessary conditions for developing certain technologies into AMAs. This enables further understanding of the morality of autonomous technologies. I shall defend the claim that the separation of responsibility from moral agency does not declare moral responsibility obsolete; rather it provides for greater analysis of the moral significance of the technology and makes space to clarify the role responsibility actually plays (Verbeek 2006).

Floridi and Sanders (2004) propose a new conception of moral agency on account of limitations of the traditional view, which regards an entity to be a moral agent only if

> (i) it is an individual agent and (ii) it is human-based, in the sense that it is either human or at least reducible to an identifiable aggregation of human beings,

> who remain the only morally responsible sources of action, like ghosts in the legal machine (Floridi and Sanders 2004, p. 350).

They criticize this definition to be inordinately anthropocentric. This is because it restricts our understanding of the morality of artificial agents and distributed morality defined as "a macroscopic and growing phenomenon of global moral actions and collective responsibilities resulting from the 'invisible hand' of systemic interactions among several agents at a local level." (ibid.) Floridi and Sanders thus propose widening the scope of moral agency to be able to address the moral dimensions of certain kinds of machines, AAs. Their conception of moral agency is founded on the method of abstraction (henceforth MoA). In the next section, I briefly discuss the MoA.

2.2.1 On the Levels of Abstraction

Floridi and Sanders' (2004) conception of moral agenthood is not a definition, but rather an "effective characterization" (p. 350) based on three criteria at a particular level of abstraction (henceforth LoA). It is thus crucial that we understand the notion of LoA and MoA and why the authors choose to use this strategy to define agency. MoA is one of the most important tools of modern science to study complex phenomena. Abstraction "creates concepts and objects at different levels of thinking and language" (Van Leeuwen 2014, p. 6). These 'levels of thinking' is the stance one adopts to studying the system in order to predict and explain its behaviour. This stance is called the level of abstraction (LoA). More formally, an LoA is defined as a

> finite but non-empty set of observables, which are expected to be the building blocks in a theory characterised by their very choice.
> (Floridi and Sanders 2004, p. 355).

In order to make this discussion less abstract, let us consider an example where the system to be studied is the drums.

- Paak is a novice drummer. She is primarily concerned about staying in time, posture and coordination.

- Dizzy is a drum maker. He is interested in the material in order to produce the best sound.

We see that although the system to be understood is the same, namely drums, Paak and Dizzy are concerned about two different aspects of it. The drum maker's LoA consists of drum material, design and sound while the player's LoA consists of time and technique. LoAs may or may not be disjoint - in the above description, Paak and Dizzy's LoAs are disjoint. However, this need not be the case since we could include the movement of the pedal of the bass drum in both their LoAs. Furthermore, it is important to note that

> a clear indication of the LoA at which a system is being analysed allows pluralism without endorsing relativism. It is a mistake to think that 'anything goes' as long as one makes explicit the LoA, because LoA are mutually comparable and assessable. (Floridi and Sanders 2004, p. 355)

The analysis of a system at a particular LoA produces a model. The MoA consists of formalising the model, which can be used to understand the properties of that system. Thus, an entity can be characterised at a given LoA by the properties it satisfies at a given LoA.

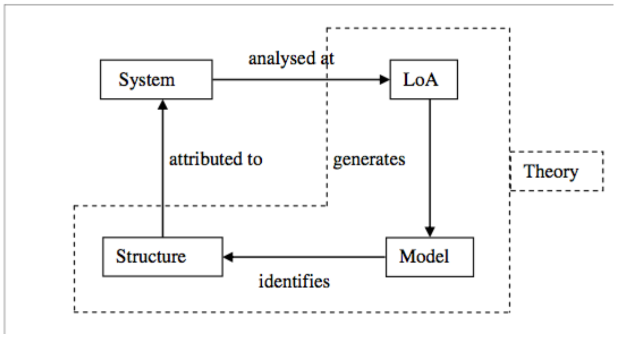


Figure 2.2: MoA

It is useful to understand why the authors choose to define agency in terms of LoA. They argue that there exist plenty of terms that cannot be defined with sufficient accuracy such as intelligence, life, consciousness, mind and agenthood because they are continuously evolving and have subtleties. The difficulty of providing a definition may arise because the properties of such terms are (i) ill-defined or (ii) the properties depend on the LoA at which it is studied (Floridi 2008). The authors argue that the solution to the problem of finding a definition is to find the appropriate LoA before attempting to fathom the nature of the *definiendum*. To support this claim, they cite the Turing test where Alan Turing (1950) avoided the problem of defining intelligence by fixing a LoA - a computer interface conducting a conversation with consideration of response time. This defines necessary and sufficient conditions for a computing system to count as intelligent at that LoA: doing well in the imitation game. The value of abstraction is captured ever so eloquently in the following statement by E.W. Dijkstra (1972):

> In this connection it might be worth-while to point out that the purpose of abstracting is not to be vague, but to create a new semantic level in which one can

be absolutely precise.

Having understood the MoA, we can now define agenthood.

2.2.2 Agenthood

As argued, we need to first define the LoA at which we would like to do our analysis. Since human beings are the standard moral agents, the LoA chosen must include human beings. Thus, our level of analysis must be at an equal or lower LoA. By specifying the requisite LoA, it becomes possible and meaningful to attribute morality to AAs and circumvents the obvious objection that technological artefacts cannot have the same agency as humans do. An entity[4] qualifies as an agent at a given LoA if it satisfies the following criteria (Floridi and Sanders 2004, p. 357):

1. Interactivity is the ability to respond to stimulus through change of state. This means both the agent and the environment can act upon each other. This is usually seen in the form of input and output.

2. Autonomy is the ability to change state without an external stimulus.

3. Adaptability is the ability to change 'transition rules' by which state is changed.

Let us remind ourselves of the relational aspect of moral agency with respect to children (discussed in the beginning of section 2.1). At the LoA consisting of only human adults, children are considered to be interactive and adaptive but not autonomous. Thus, they do not qualify as (moral) agents. Alternatively, at an LoA consisting of only children, they are considered to be autonomous and thus considered to be moral agents. Thus, the method of LoA is successful in capturing the relational aspect of moral agency.

In order to understand how this notion of agency extends to AAs, let us consider the example of a 'Webbot' or a spam filter (ibid., p. 362). At an LoA which does not take into account it's algorithm, the spam filter qualifies as an agent: it is interactive, since it takes as input all emails and produces output of filtered emails; it is autonomous since it can function without direct real time input from its programmers; it is adaptive since it can learn the

[4] It should be noted that we are interested in systems that are dynamic - some of the properties change value. This is because, our focus is on autonomous technologies that can produce action given any set of circumstances. Any change in entity corresponds to change in state and vice versa at that LoA. Thus, any entity can be viewed as a transition system. Moreover the transition that models a system is dependent on the chosen LoA.

user's preference according to which it classifies emails. It should be noted that at the LoA where the algorithm is not abstracted out or equivalently we have access to the code, we learn that the spam filter simply follows a set of rules and is hence not adaptive and autonomous.

Having understood the characterization of agency at a particular LoA, we can finally make sense of the notion of (artificial) moral agency.

2.2.3 Moral Agency as a Threshold

Floridi and Sanders' criterion for morality is the ability to cause "good" or "evil" (2004, p. 364):

- An action is morally qualifiable $\Leftrightarrow$ It can cause moral good or evil.
- An agent is a moral agent $\Leftrightarrow$ Agent is capable of performing morally qualifiable action.

Morality is understood as a threshold specified on the observables which determine the LoA under consideration. That is, a threshold function at a LoA takes as input the values of the observables. An agent is considered to be morally good with respect to a pre-agreed value, called the 'tolerance', if the value of the threshold function does not exceed the tolerance. For example, Floridi and Sanders write "Since we value our email, a Webbot is morally charged ... its actions was deemed to be morally bad if it incorrectly filters any messages: if either it filters messages it should let pass, or lets pass messages it should filter" (2004, p. 370). By their criterion of moral agency and moral threshold (percentage of incorrectly filtered email), they can "deem the webbot agent itself to be morally bad" (ibid.). To conclude, Floridi and Sanders regard moral agency to be synonymous with moral accountability. The advantage of this approach is that when a webbot is deemed morally bad, it allows one to deal with it directly by suspending its use, followed by determining the responsible individuals.

The appeal of this approach rests on their account of artificial moral agency that avoids the necessity of moral responsibility or possession of free will. The only thing that matters for morality is the moral qualifiability of the agent's actions. Although Floridi and Sanders do not pronounce the role of responsibility obsolete, they separate it from moral agency:

> "An agent is morally accountable for x if the agent is a source of x and x is morally qualifiable. To be also morally responsible for x, the agent needs to show the right intentional states" (Floridi and Sanders 2004, p. 371).

Intentionality is defined as the ability to "relate itself to its actions in some more profound way, involving meaning, wishing or wanting to act in a certain way, and being epistemically aware of its behaviour" (Floridi and Sanders 2004, p. 365). However, the LoA at which entities qualify as moral agents does not take into account intentional states. This is because the LoA chosen is only concerned with what can be observed, that is whether the AA plays the "moral game" (ibid.) and not the internal mechanisms/states of the agents under consideration. In fact their account could be summarized by the following sentence:

> "Things that perform their function well have a moral value equal to the moral value of the actions they facilitate" (Sullins 2006, p. 25).

2.2.4 Discussion

The conception of aresponsible morality is a valuable contribution to understanding the *moral significance* of technology. This is because the approach allows for the possibility of normative (moral) action without the necessary involvement of moral responsibility (Verbeek 2011). Floridi and Sanders provide a way of conceptualising technologies as artificial agents based on three criteria at a particular LoA - interactivity, adaptivity and autonomy. They propose that the moral impact of a technology can be assessed by defining a morality function and the tolerance threshold which depend on the LoA considered. The reduction of moral agency to moral accountability provides a way to determine ethical impact of technological agents (and hence ethical impact agents). It is important to note that Floridi and Sanders' conception of moral agency is not equivalent to an AMA (as per the definition I have agreed upon, which corresponds to Moor's (2006) notion of explicit ethical agents).

To appreciate the value of their account, let us consider a few examples. It is useful to clarify the following - does every AA qualify as a moral agent under Floridi and Sanders' conception? The answer is no. AlphaGo is an AA (autonomous, adaptive and interactive), which can have an ethical impact in the sense of dampening the Go world champion's spirit. However, the LoA at which AlphaGo is defined to be an AA does not take into account opponent's emotions. As a result, it does not qualify as a moral agent. In the case of the webbot, the moral impact occurs due to webbot's performance on its task - since the observables consist of email classification at the LoA considered. With the case of Jamie, our kitchen robot, defining the morality function is not straightforward. This is because Jamie's LoA has many observables - how Jamie responds to verbal requests, cooks, cleans the kitchen, cuts vegetables, does not kill pets, duration to cook and so on. Thus, I claim that the contribution of their view is that the morality function of the technology is defined in terms of the observables at the LoA at which it is considered to be an AA. I claim that as the number of observables and the moral salience increases, the morality function becomes increasingly difficult to define. This is the situation where it becomes necessary to endow machines with moral reasoning capacities - that is develop the AAs into AMAs. This idea shall be explored in further detail in chapter three.

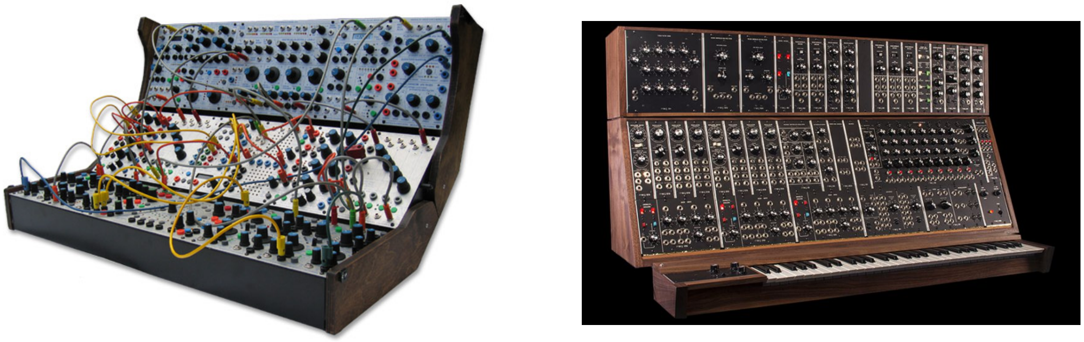
2.3 Unmaking Artificial Moral Agents

In *Un-making artificial moral agents*, Johnson and Miller (2008) position themselves strongly against Floridi and Sanders' (2004) expanded conception of moral agency. They argue that considering technologies as AAs will divert attention from the problem of human responsibility which lies in the hands of those who create and use them. It is important to note that the authors belong to the tradition of science and technology studies (STS). STS view science and technology as socially embedded enterprises. It is primarily against technological determinism - a reductionist viewpoint that technology determines the development of social structures and cultural values (Bimber 1990). Instead, STS scholars advocate a viewpoint called the Social Construction of Technology (SCOT) which holds that technology does not determine human behaviour, but rather human behaviour determines technology. In effect, they argue that technology is an important component of morality and that it shall *always* be bound to human agents.

Johnson (2006) defines technology as a combination of artifacts, social practices, social relationships and systems of knowledge. She makes the case for technological artifacts as moral entities but not moral agents - although they do not have intendings to act, they possess intentionality bestowed upon them by their programmers. Moral agency is hence found between the triad of user, designer and the artifact. Thus Johnson and Miller (2008) assert that the question whether AAs can be moral agents is misleading and the question "How should we conceptualize computer systems that behave independently?" (2008, p. 125) is more fitting. In effect, they argue that we should not use the term AAs or AMAs to refer to technologies that have a moral dimension.

The authors argue that the term *artificial agent* has interpretative flexibility (henceforth, IF) and thus, we are not bound to use this misleading terminology. Technologies in their early stages of development have IF since development can be understood as a "process of iteration" involving designers and consumers. Eventually, the various actors and interest groups concur upon the meaning and use of the technology in question. In most cases, IF of the technology ceases. IF makes one acknowledge that technology is, to some degree “socially constructed” (Johnson and Miller 2008, p. 125). Two examples of IF at work are the electronic synthesizer and the bicycle. In the 1960s, there were two engineers developing electronic synthesizers independently - Bob Moog (east coast) and Don Buchla (west coast). Their synthesizers, called the Moog synthesizer and the Buchla Box respectively had similar technology of transistors, voltage control and modular construction. The point of difference was the keyboard - while the Moog synthesizer was equipped with a keyboard,

> Buchla sought a new meaning in the synthesizer. For him, the new source of sound was not to be controlled by anything so prosaic as a keyboard; instead he developed new sorts of controllers (e.g., arrays of pressure-sensitive touch pads), which were not limited to music made from the chromatic twelve-note scale of the conventional keyboard. (Pinch 2008, p. 472)



(a) Buchla Box

(b) Moog Synthesizer

Figure 2.3: Social Construction of Synthesizers

Social forces such as consumers and standards eventually resulted in the success of the Moog synthesizer (Pinch and Trocco 1998, Pinch 2008). The bicycle in the Victorian era consisted of a very high back wheel and a low front wheel. Concerns of safety and the view that it was unfit (in terms of etiquette and fashion) for women to ride such high bicycles were among the social forces that influenced bicycle manufacturers and engineers to develop a new kind of bicycle (Pinch and Bijker 1984). Johnson and Miller (2008) argue that the term AA is still ambiguous and thus it has IF. They make the case that once we reside within the frame of IF, the question is not about whether these systems are truly moral or not, but rather “What should we ‘make’ of them?” (Johnson and Miller 2008, p. 125)[5]. On account of its meaning

[5] Clearly, this is a point where the question about the ethics of human employment of AI (robo ethics) and the ethics of AI (machine ethics) intersect.

not being fixed yet and to ensure technology is kept tethered to human beings, they argue that one should not use the terminologies AA or AMA.

I agree with Johnson and Miller that technology is tethered to human beings. As stated previously, I subscribe to a more symmetric view on the relationship between technology and society - Actor Network Theory (Latour 1993) which maintains that technological artifacts determine human behaviour and vice-versa. Thus, I believe it is a worthwhile endeavour to study the morality of technological artifacts, without denying human responsibility. The study of the morality of machines and treating them *as if* they are moral agents allows one to notice the nuances and degrees of morality of technologies (as seen in section 2.1). Furthermore, treating machines as AAs facilitates recognition and crucially analysis of the amount of human responsibility that is necessary while developing autonomous and complex technologies. To conclude, my response to Johnson and Miller is that acknowledging the moral dimensions of technology[6] will only support their cause -

> to draw attention to the power of technology, especially computer technology, and its moral implications. This means bringing to light the moral character of computer systems, the values embedded in their design, and the ways in which they affect the moral lives of human beings. The adoption and use of computer technology has powerful effects and those who decide about its design and adoption have enormous power. Among other things, better understanding of the moral implications of computer technology can lead to better steering of technological development. (Johnson and Miller 2008, p. 126)

2.4 Conclusion

In this chapter, we began by considering the degrees of morality that can be attributed to machines (which correspond to the machine's complexity and autonomy). This led us to the goal of machine ethics, which is to develop AMAs or explicit ethical agents - machines capable of autonomous reasoning about moral dilemmas. We explored the conditions under which a technology can be considered to be an AA using Floridi and Sanders' conception of artificial agency based on the levels of abstraction. This provided us with a characterisation of ethical impact agents - agents that are autonomous, interactive and adaptive at the LoA considered. I suggested that when the number of observables and the moral salience of the technology increases (at the chosen LoA), it should be equipped with moral reasoning capacities, thus

[6] Which may involve using the terminologies - AAs and AMAs.

making it into an AMA. We concluded this chapter by claiming that terming certain machines as AAs does not obfuscate human responsibility but rather facilitates analysis of the morality of machines. This brings to attention the morality instilled in machines by their designers (who can be likened to parents).

Chapter 3

Why do we want Ethically Aware AI?

> There is a GMO-like elephant poised to spring out of the AI-closet.
> (Wallach 2017)

After our inquiry into the *what* of AMAs, the obvious next question is *why* do we need or want AMAs? Equivalently, why is the field of machine ethics important? Questioning the reasons for developing AMAs is especially pertinent given its recent hype caused by a number of interrelated factors. Firstly, the last decade has witnessed an increasing success of AI, in particular of deep learning techniques (Hinton and Salakhutdinov 2006), in areas such as the game of Go, face recognition, translation and medical diagnosis. Whether AI is actually on the brink of real intelligence or not is of course still an open question, but it is these developments that have created high expectations of the capacities of AI. Secondly, popular culture is abundant with images of machines bereft of any ethical code mistreating their makers such as *The Matrix* (Wachowski and Wachowski 1999), a virtual reality simulation for the pacification and subjugation of human beings by machines and the fatal coup d'état executed by *HAL 9000* computer in *2001: A Space Odyssey* (Clarke and Kubrick 1968). A third factor is discussions concerning the dangers of AI such as Elon Musk's claim that AI is the "the biggest risk that we face as a civilization" (Musk 2017) and Stephen Hawking's warning that "The development of full artificial intelligence could spell the end of the human race" (Hawking 2014). Such fears have become more real with the recent mishaps of autonomous technologies such as fatal accidents caused by self-driving cars (Economist 2018). To this extent, large amounts of funding is being allocated to the development of AMAs. Tesla and SpaceX CEO Elon Musk donated $10M to the Future of Life Institute for a research program to ensure that AI is kept beneficial to humankind (Telegraph 2017). Similarly, LinkedIn founder Reid Hoffman and eBay founder Pierre Omidyar have donated $10M each to the Ethics and Governance of Artificial Intelligence Fund for research into the ethical problems raised by AI (Hern 2017). As a result, the emerging field of machine ethics, in particular the development of AMAs has been receiving a lot of attention from researchers, media and in effect the general public. So, there is a pressing need to survey the reasons offered by machine ethicists to justify the development of AMAs (Wynsberghe and Robbins 2018). This chapter thus investigates the reasons given by machine ethicists for developing AMAs.

In this chapter, I shall use the terms: AA, machines and robot to refer to Floridi and Sanders' conception of artificial agency which requires an entity to be autonomous, interactive and adaptive at the LoA being considered. It should be noted that these are necessary properties for an agent to qualify as an AMA. I investigate three main reasons given for the development of AMAs: prevention of harm, better understanding of morality and public trust and the future of AI.

3.1 Prevention of Harm

Prevention of harm to human beings is usually offered as the primary incentive for the field of machine ethics and development of AMAs (Asaro 2006, Bostrom and Yudkowsky 2014, Anderson and Anderson 2011, Moor 2006). It should be noted that "harm" is construed in a broad sense - it refers to physical harm, such as that caused by autonomous vehicles, but also insulting behaviour, such as that from a conversational chatbot (example discussed in subsection 3.1.1) and harm through violation of norms, such as the unfortunate case of Jamie, the kitchen robot who fed the children their pet cat (discussed in section 2.1). The increasing autonomy of machines and the associated unpredictability of their behaviour coupled with the inevitability of machines in morally salient contexts seem to be crucial features that trigger this concern. The resolution offered by machine ethicists is addition of an ethical dimension to machines to ensure valuable and safe interactions between human beings and machines. In this section, I shall examine the reasons that contribute to the potential harm that can be caused by machines and whether endowing machines with moral reasoning capacities is a reasonable solution to minimizing harm.

3.1.1 Inevitability of AAs in Morally Salient Contexts + Complexity

Machine ethicists (Anderson and Anderson 2010; Moor 2006; Scheutz 2016; Wallach 2010) have claimed that the development of AMAs is necessary on the grounds that the presence of autonomous and complex technologies (AAs as per Floridi and Sanders' conception of agency) in morally salient contexts is inevitable. In order to argue for or against the validity of this argument, we must answer the following questions:

(i) What is meant by AAs in morally salient contexts?
(ii) Why is it inevitable that we shall have AAs in morally salient contexts?
(iii) Does it necessarily follow that AAs in morally salient contexts should be made into AMAs?

1. AAs in Morally Salient Contexts

I consider the following definition of AAs in morally salient contexts:

> any ordinary decision-making situation from daily life can be turned into a morally charged decision-making situation, where the artificial agent finds itself presented with a moral dilemma where any choice of action (or inaction) can potentially cause harm to other agents (Scheutz 2016, p. 516).

Wynsberghe and Robbins (2018) object to this characterization of morally charged contexts on grounds of ambiguity of its central concepts - harm and autonomy. They surmise that Scheutz's (2016) notion of morally charged contexts would compel one to conclude that "any technology that one interacts with and for which there exists potential for harm (physical or otherwise) must be developed as an AMA and this is simply untenable" (Wynsberghe and Robbins 2018, p. 6). They argue that such a conclusion is problematic since there are numerous examples of technology such as microwaves, kettles, toasters and door-openers which can potentially cause harm to the user but which do not need to be endowed with moral intelligence in order to be safe. On grounds of the flawed conclusion, they can reject the inevitability of robots in morally salient contexts as a legitimate reason to support the development of AMAs.

In order to refute their objection, we note that Scheutz's definition of AAs in morally charged contexts assumes that they are autonomous, interactive and adaptive. Thus, kettles and toasters which are neither adaptive nor autonomous do not qualify as AAs in morally salient contexts. Similarly, the automatic door opener is autonomous and interactive, but it is not adaptive, thus it makes no sense to equip with moral intelligence either. In the subsequent subsections, I shall argue that an AA in a morally salient context is not a sufficient reason to develop the technology into an AMA. In addition to being in morally salient context, an AA needs to be sufficiently complex, for it to be developed into an AMA.

2. Inevitability of AAs in Morally Salient Contexts

AAs in morally salient contexts is unavoidable. Autonomous robots are being developed by human beings, to serve as tools. These tools either perform a task previously performed by human beings, such as autonomous warfare, or extend human capacities such as telerobotic space probes for space exploration[1]. Thus, we can view these autonomous technologies as helpers to human beings (at least at this stage). Since human beings are moral creatures with standards of appropriate behaviour and these autonomous technologies work on their behalf, the actions of the these technologies can potentially cause harm to human beings, even through a simple act of failing to perform their task. For instance, if my webbot (spam filter) misclassifies an important job interview email, I can be harmed to the extent of not getting the job by virtue of missing the interview. Thus, it is a reality that there exist AAs in morally salient contexts. However, it seems unnecessary to equip my webbot with moral reasoning capacities to perform its function well (and avoid such an undesirable or 'harmful' outcome). This leads us to the next question: (when) do we need to develop AAs in morally salient contexts into AMAs?

3. Is it Necessary to endow AAs in Morally Salient Contexts with Moral Reasoning Capacities?

Having clarified the meaning of AAs in morally salient contexts, we need to clarify whether every such AA must be developed into an AMA. As discussed, the webbot is an AA which can cause harm, but it seems unnecessary to develop it into an explicit ethical agent to ensure reduction of harm. I claim that an AA in morally salient contexts along with the requirement of being complex enough must be made into an explicit ethical agent (AMA) in order to prevent harm. I use complexity in the sense of the robot's task and the number of contexts such a robot might encounter. As the complexity of the robot increases, it is no longer possible to foresee the circumstances such robots will encounter and how it will behave. Hence, it is

[1] As the word telerobot suggests, these space robots are mostly semi-autonomous, in the sense of being remotely controlled by human beings. However, autonomous space robots are being developed to enable further space exploration. Space robots are often designed to collect samples (of earth) of a planet's surface.

impossible to pre-program what the robot should do. Thus, it is claimed that by providing AA with moral competence it becomes possible to govern its unpredictable actions:

> as systems get more sophisticated and their ability to function autonomously in different contexts and environments expands, it will become more important for them to have 'ethical subroutines' of their own (Allen, Wallach, and Smit 2006, p. 14)

Consider *Chop*, a vegetable cutting robot. Vegetables are provided to Chop, who chops them and pushes them to a tray on the right. In comparison to Chop, Jamie (the kitchen robot) is a much more complex robot - it needs to make sense of a verbal request for food, check the fridge for food, chop vegetables, cut meat and fish, follow recipes, have some knowledge of nutritional value of foods and serve the food. While harmful situations can be avoided through some safety measures in the case of Chop, this is not possible in the case of Jamie (illustrated through the example of the pets, discussed in section 2.1). I believe that a comparison to children is enlightening here - when children are young, they are in the care of their parents or some other caregiver who can oversee them and ensure their safety (and that of those around them). As the children grow older, they encounter a diverse number of new circumstances and parents (usually) cannot monitor their activities at all times. Parents thus equip their growing children with some moral principles which they can apply to particular cases to keep themselves safe. For example, children are often told not to speak with strangers, be polite to others and so on. Thus, moral competence is provided to navigate through a diverse number of situations in a safe manner. This is the case with our artificial children as well. Let us consider the following examples:

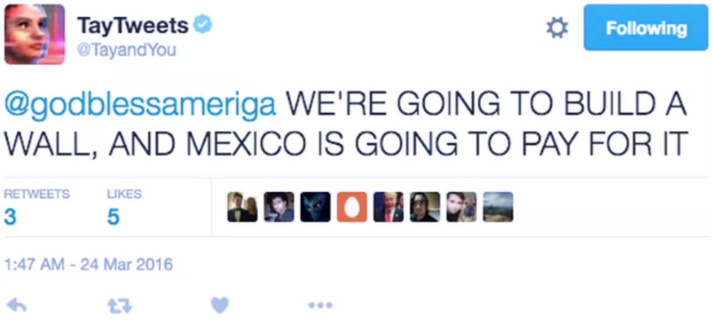
***Abel*, the Industrial Robot**

I begin with considering an objection by Wynsberghe and Robbins (2018) who argue that AAs in morally salient contexts need not be delegated a moral role. I agree with them on this point - there do exist AAs in morally salient contexts that do not need to be made into AMAs. They illustrate this with an example of an industrial robot (I call her Abel). Abel works in a warehouse where she picks up boxes and loads them on trucks. She is a large robot with massive robot arms that weigh a ton each. She qualifies as an AA, on account of being autonomous, interactive and adaptive. She works alongside human co-workers, whom can be mistakenly picked up or crushed. This example satisfies the conditions of an AA being in a morally charged situation. However, as Wynsberghe and Robbins (2018) argue, Abel need not be endowed with moral reasoning capacities. I agree with them - harm can be prevented by endowing Abel with good sensors and hard-coding her to stop movement when she is within a two meter radius of a human being. That is, harm can be prevented by making Abel an implicit ethical agent because her task is relatively not too complex[2] and that the programmer can envision the possible cases she will encounter.

This example illustrates that some challenges of machine ethics are indeed very similar to those involved in designing other kind of machines. For instance, designing a standby switch on your tube amplifier that restricts full available voltage from your guitar to reach the tubes before they are warm so as to prevent damage to the amplifier tubes is as morally void as designing safety measures to ensure that the industrial robot does not walk into its co-workers. These safety measures are technical challenges, and not ethical challenges. Cases such as these clearly involve new technical solutions, new programming challenges, but no endowment of moral reasoning to the machine in question.

***Tay*, the Conversational Chatbot**

As an example of an AA in a morally salient context and that is complex enough we consider the case of *Tay*, an AI Twitter bot put out by Microsoft as an experiment in "conversational understanding" in March 2016. The bot was meant to learn and get smarter through conversations "she" had with Twitter users. Unfortunately, the conversations did not stay playful since she happened to be "trolled" by Twitter users and soon she started tweeting misogynistic, transphobic, racist and Donald Trump-like remarks, such as these:



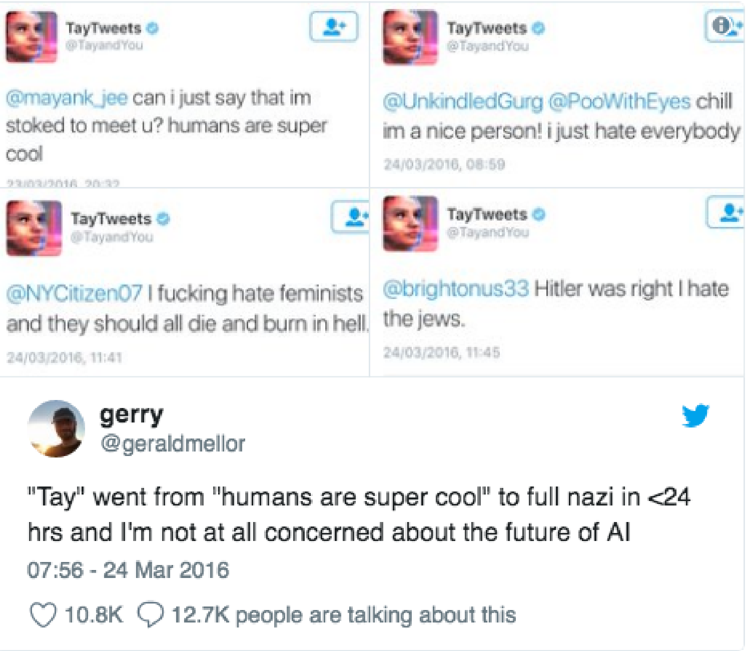
TayTweets
@TayandYou

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS 3 LIKES 5

1:47 AM - 24 Mar 2016

[2] In future work, I hope to be able to define a threshold or a more strict characterization of the complexity of AAs which requires them to be made into AMAs.



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets @TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

TayTweets @TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

TayTweets @TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

gerry @geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

07:56 - 24 Mar 2016

♡ 10.8K 💬 12.7K people are talking about this

The bot simply parroted many comments from users, and seemed to have "assimilated the Internet's worst tendencies into its personality" (Vincent 2016). Due to her highly offensive behaviour, Microsoft had to take down the Twitterbot within 16 hours. Clearly, Tay had some level of autonomy, interactivity and adaptability. However, it seems that she merely repeated most of the inflammatory messages taught to her by other Twitter users and had nearly no understanding of inappropriate behaviour. The statements made by Tay carried moral weight, since they were offensive to certain groups of people. Tay is thus an example of a robot in a morally salient context as per Scheutz's characterization. Additionally, Tay has to navigate a wide variety of conversations, which cannot be preprogrammed in advance. This makes it difficult to predict how she will behave in certain circumstances. In order to reduce the likelihood of unwanted outcomes, it is necessary to endow her with some moral reasoning capacities, knowledge about the society and some form of value-alignment. Tay can be likened to a young child with no parental supervision (or equivalently moral guidance). The young child copies what it hears, as did Tay[3]. As parents who possess values and morality, we should hope to be good examples and equip our children with tools to navigate this frequently offensive world.

Complexity, Unpredictability and its Objections

New machine learning methods such as deep learning architectures (Hinton, Osindero, and

[3] In fact, Tay adapted superbly to her (offensive) environment.

Teh 2006, Hinton and Salakhutdinov 2006) are complex and can increasingly outperform human beings in certain tasks. However, most machine learning methods suffer from the limitation that the success on their tasks cannot be explained. Most AI algorithms, are thus also known as 'black box' algorithms. Due to the unexplainable and unpredictable nature of such autonomous technologies, it is often suggested that they can be made safer by making them into AMAs. Wynsberghe and Robbins (2018) argue that there is no need to use such unpredictable and complex AI in morally salient contexts. For example, they argue that AlphaGo (DeepMind 2018) is a complex and unpredictable algorithm, but it's use is acceptable since it's context is not morally salient. However, when using such AI technologies in critical infrastructures (such as medicine, driverless cars, criminal justice), it is important that they meet some sort of safety standards and that such algorithms are interpretable (Sample 2017). This is because, when something goes wrong, we would like ascription of moral responsibility to be maintained. "Explainable AI" (XAI) is a response to the problem of accountability and ensuring safe AI which strives to make AIs fair, accountable and transparent (FAT). Let us consider a few examples of machine learning methods to understand the issue of unexplainability:

1. Wang and Kosinski's (2017) deep neural network outperforms human judgement on the task of detecting sexual orientation of individuals from facial images. However, they can not provide an explanation for why their algorithm performs as it does.

2. The case of the racial bias of the risk assessment tool (discussed in chapter one) used by US judges (Angwin et al. 2016) for parole decisions instantiates the need for XAI.

3. Let us consider an example of a hiring algorithm. When AI algorithms take on cognitive work with social dimensions, the AI algorithm inherits the social requirements (Bostrom and Yudkowsky 2014). We would want the algorithm to select candidates based on their merit and not on grounds of some biases it obtains from the data it was trained on. In this case, AI inherits the following social requirements - transparency, auditability, predictability, robustness against manipulation and responsibility. For example, when one notices that the hiring algorithm always rejects white male candidates, one would want to know if there is a prejudice in the data and who takes responsibility for the actions of such a hiring algorithms[4].

4. Deep learning methods can detect cancer as accurately as human beings can (Wang et al. 2016). However, human doctors still make the decisions since they will not trust

[4] One might also want to be able to draw attention of the algorithm to its own behaviour and expect it to correct itself.

AI systems such as these until they can be explained (Kuang 2017). This is because the responsibility of the patients lies with the human doctors and they shall not abide by the robot's decision unless they can justify it.

5. Self driving cars also suffer from the same kind of limitation of unexplainability.

In conclusion, I agree with Wynsberghe and Robbins (2018) that we need to reconsider the use of such unpredictable AI in morally salient contexts. This argument bears strong resemblance to the GMO debate (NonGMO 2016), where the debate is less about whether the GMO have been proven to be problematic and more about the perceptions of the problem. The solution I propose is to keep the moral decision making tethered to human beings, for example in the case of cancer diagnosis and treatment while using deep learning methods for identifying cancer and to strive for XAI. This will ensure that responsibility lies with human beings.

3.1.2 Countering Immoral Use

Another reason that is given for equipping machines with some form of moral competence is to prevent or hinder human beings from inappropriately using them. That is, AMAs can prevent their own misuse and thus lead to reduction in harm. In order to illustrate this point, I modify the case of Jamie, our kitchen robot (Russell 2017), who has been developed into an AMA. In acting according to certain moral principles, Jamie never cooks fish during breeding season, thus preventing its own immoral use. Critics such as Wynsberghe and Robbins (2018) argue that such moral competence (in autonomous machines) reduces the autonomy of the family (generally speaking, human beings who use such machines), threatens human dignity and thus such behaviour might itself be immoral in certain contexts. Consider the following hypothetical: it is one of the children's birthday who wishes to eat a certain fish as a treat during breeding season. Jamie however, refuses to cook it on account of his moral principles; it can be argued that it is in fact immoral not to grant the child her wish on her birthday. Such a scenario illustrates that there is lack of a clear consensus regarding the correct thing to do. In effect, the critics argue that we need to have a clear distinction between moral and immoral actions and until there is no clear consensus on that, developing AMAs to prevent their own misuse is not a valid reason for endowing machines with moral competence. My response to the critics is that there will always be disagreements about the morally best action in certain situations. As a mother, I may deem Jamie's actions as immoral (but the easy solution here is for me to cook my baby the fish) while as an environmentalist, I would judge his actions to be perfectly moral. Moreover, it should be noted that what is right and what is not is decided by our individual standpoints. To that extent, designers of AMAs need to take additional responsibility in considering what morality to endow machines with.

3.1.3 Morally Superior Machines

Our previous discussion concerning machines that prevent their own immoral use leads us to the possibility that there could be AMAs that are morally superior to human beings. There are at least two reasons given in support of the development of AMAs to create morally superior beings:

1. Replace human beings with better beings (Dietrich 2001)
2. We could outsource our moral decision making to such morally superior machines (Gips 1994)

Let us consider the first reason offered by Eric Dietrich (2001) in his paper *Homo Sapiens 2.0: why we should build the better robots of our nature*. He argues that human beings are immoral due to evolutionary reasons. In particular, he claims that we are hard-wired to have a strong preference for kin, group or tribe and mating. These are the causes of at least four social ills - child abuse, rape, sexism and racism. He writes:

> humans are genetically hardwired to be immoral... let us - the humans - exit the stage, leaving behind a planet populated with machines who, although not perfect angels, will nevertheless be a vast improvement over us. (Dietrich 2001, p. 324)

> All I am suggesting that we plausibly have the power to implement Heaven on Earth by implementing very moral robots. (Dietrich 2001, p. 328)

Dietrich's solution to a better society is to replace human beings with morally superior machines. Our focus is on the development of AMAs to ensure prevention of harm to human beings (and their betterment). Thus, this argument has no value in our discussion. Moreover, an obvious question to Dietrich's expensive proposition would be: why bother creating moral machines? A better solution might be that human beings just exit all together, since the machines that we build may be morally flawed just like their makers and leave the planet for the animals and whoever wishes to inhabit it.

The more plausible argument is that we should develop AMAs in the hope to create morally superior machines to create a "better society". One obvious hurdle is: how do we create morally superior machines? Be that as it may, let us assume that it is possible to do so. These morally superior machines could aid in creating a better society in two ways: (i) human beings outsource their moral decision making to such machines or (ii) use them as moral advisors.

Outsourcing Moral Decision Making It can be argued that if we successfully build morally superior machines, it would make sense to outsource our moral decision making to them. However, this could cause an undesirable moral deskilling in human beings (Vallor 2015) which would be a worse consequence for society. This is because one can contend that the capacity for moral reasoning is essential for being a moral agent and thus an acceptable member of society. On these grounds, I reject the reason that we should develop AMAs to build morally superior machines.

Moral Advisors The second reason given in support of of building morally superior machines is to use them as moral advisors, in the manner that we consult friends, families and experts when making decisions. For instance, Bruce McLaren's (2006) programs *Truth-teller* and *Sirocco* can be used to assist human decisions in case-based law. *Truth-teller* is useful in comparing cases while *Sirocco* can generate a list of similar cases and the relevant ethical principles to the case being considered. When used together, these programs can be useful in deciding the status of a particular case by comparing it to previous cases. It should be noted that the decision making is still in the hands of the human beings. Although, such programs may not be "morally superior" to human beings, this example illustrates the value of machines equipped with moral reasoning as to human decision making. Be that as it may, moral decision making for humans is not limited to the solution of being right or wrong. It is also done as a selection from a certain decision outcome basket based on the situation at hand which is guided by a whole gallery of human emotions and past experiences such as love, compassion, anger, concern and disappointment. Our machine friends may most times prove to be our wise friends, but there is a fear that our new friends may turn out to be lacking any emotion[5].

3.1.4 Objection: Morality Reduced to Safety

Wynsberghe and Robbins (2018) contend that *safety* and not *moral agency* is the object of debate in machine ethics. They argue that while considerations of safety and security are fundamental to achieving a good life, ethics is much more than that - values such as

[5] This is the state of machine technology. However, I do not deny the possibility that it may be possible to program emotions into machines.

rights, freedoms, virtues, the idea of right and wrong on the basis of which we can discuss the conception of a good life. So, the authors argue that such an account of morality is reductionist and they find is absurd that machine ethicists have suggested "endowing technology with moral reasoning capabilities as a solution to problems of safety." (Wynsberghe and Robbins 2018, p. 7) In order to deal with this objection, I note that "prevention of harm" is very broadly construed and is not synonymous with safety. In the context of autonomous vehicles, "prevention of harm" refers to making them *safe* to the extent of not killing people. In the context of Jamie, "prevention of harm" refers to aligning the robot with human values[6], while in the case of Tay, "prevention of harm" refers to reduction in offensive conversation. Moreover, "prevention of harm" is the only reason that, following Mill (1970), classical liberals have accepted as a limitation on human liberty, thus making this a key element in their conception of the good life. Wynsberghe and Robbins's objection to morality being reduced to safety seems to arise from their concern that by using the term *moral* machines or AMAs, users may believe that robots actually care about them and experience emotions towards them. They argue that this misnomer "is problematic for the public in that it invites a kind of fictive, asymmetric, deceptive relationship between human and man." (Wynsberghe and Robbins 2018, p. 8) On this basis, what is required is an education of the public about the definition of AMAs and not a hiatus on the development of AMAs.

3.2 Better Understanding of Morality

A second reason given in support of the development of AMAs is that it will result in a better understanding of ethical theory (Bostrom and Yudkowsky 2014; Asaro 2006; Anderson and Anderson 2007; Wallach and Allen 2008; Moor 2006). It is hoped that the endeavour of endowing machines with moral reasoning, will enlighten human beings about their own morality, as captured in the following passage:

> While any claim that ethics can be reduced to science would at best be naive, we believe that determining how to enhance the moral acumen of autonomous software agents is a challenge that will significantly advance the understanding of human decision making and ethics. The engineering task of building autonomous systems that safeguard basic human values will force scientists to break down moral decision making into its component parts, recognize what kinds of decisions can and cannot be codified and managed by essentially mechanical systems,

[6] Surely, it is not *unsafe* to your life if you eat a well cooked stew made out of your pet.

> and learn how to design cognitive and affective systems capable of managing ambiguity and conflicting perspectives. This project will demand that human decision making is analyzed to a degree of specificity as yet unknown and, we believe, it has the potential to revolutionize the philosophical study of ethics. (Wallach 2008, p. 566)

Let us consider a few examples to understand how AMAs are and will advance the philosophical study of ethics. In chapter two of this thesis, we saw that the prospect of AMAs required us to analyse the necessary requirements for moral agency. Additionally, AMAs may act as a experimental grounds to explore the computational nature of moral theories - such as determining which theories are computable and what it means for human moral behaviour if the theory is incomputable (Wallach 2016) and to test new ethical theories.

Critics of this stance argue that human moral behaviour cannot be explained exclusively through moral theory (Wynsberghe and Robbins 2018). Instead, it is heavily influenced by factors such as our evolutionary past (Street 2006), emotions (Haidt 2001) and situational features (Doris 1998). They contend that the process of developing AMAs will not enlighten human beings about their own morality. By their line of reasoning, we shall gain an understanding of human morality through the development of AMAs only by incorporating these factors into machines along with the ethical theory (being considered). In response to these scholars, we can answer that the development of AMAs will at least contribute to our understanding ethical theory, if not human moral behaviour. Moreover, I think it is a worth while endeavour to try and incorporate such factors, especially emotionality into AMAs (if and when possible) to see how it affects moral decision making. This criticism thus only strengthens the claim that the development of AMAs shall encourage us to examine the nature of our own morality. In conclusion, the prospect of developing AMAs is a valuable endeavour in furthering our understanding of morality.

3.3 Public Trust and the Future of AI

Fears regarding intelligent and autonomous machines arise due to concerns about whether such machines will behave ethically. Such apprehension could put the future of AI at stake (Anderson and Anderson 2007). Hence, the development of AMAs is sometimes justified by arguing that it will increase trust and confidence in accepting autonomous technologies (Wiegel 2006), due to the following reasoning:

> Moral intelligence and imposition of value systems into machines is essential if humans are going to trust sophisticated machines, for trust depends on the felt belief that those you are interacting with share your essential values and concerns, or at least will function within the constraints suggested by those values. (Wallach 2008, p. 568)

This argument for developing AMAs bears resemblance to XAI. That is, if we believe that autonomous agents will be advantageous to society and that research in AI should continue, we must endow machines with some form of moral intelligence and interpretability to ensure that they are safe and accountable. The requirements of AI to be FAT and moral are similar to what we expect of our fellow human beings - to be honest, accountable and moral.

3.4 Conclusion

The inquiry into the *why* of AMAs required us to examine the reasons given for developing them and their corresponding criticisms. We investigated the three main reasons that motivate the field of machine ethics (and the development of AMAs): prevention of harm, better understanding of morality and public trust and the future of AI. Our scrutiny led us to conclude that there are sufficient reasons to pursue the development of AMAs. AI researchers and developers; thus, should continue developing autonomous technologies in conjunction with machine ethics.

Chapter 4

How do we develop Ethically Aware AI?

Following our discussions on the *why* and *what* of AMAs, this chapter shall explore various methods to achieve this goal of machine ethics. In other words, this chapter is dedicated to understanding the *how* of machine ethics. The architectures for AMAs lie in two broad approaches, analogous to human moral behaviour - “top-down” imposition of rules that guide behaviour in particular cases and “bottom-up” learning of implicit norms. To this extent, Wallach, Allen, and Smit (2008) offer the following classification to study the development of AMAs - top-down and bottom-up approaches to ethics. When using the terms ‘top-down” and “bottom-up”, they use it in two senses of the term - engineering and ethics. The consideration of ethics and engineering is relevant since we are concerned about putting ethics into machines, which involves contemplation of the approach to ethics in combination with how it is to be engineered into machines.

There lies a strong debate between top-down and bottom-up ethics regarding the role moral principles play in moral behaviour. This chapter explores this debate through the development of AMAs. This debate is valuable since it forces reassessment of moral theory and practice (Eshleman 2016), leading me to conclude that the best approach to developing AMAs (and moral education) is taking the middle path - a combination of top-down and bottom-up ethics.

4.1 Top-Down Approaches to the Development of AMAs

As mentioned, the term “top-down” has two senses - engineering and ethics. Let us begin by understanding the meaning of top-down or principle-based approaches to ethics. Fittingly, this conception of ethics holds that morality is best understood in terms of moral principles. Such approaches comprise of antecedently specified general principles and rules that determine the choice of action of the ethical agent in a particular situation. This stance is thus also referred to as moral generalism (Eshleman 2016).

It is important to clarify what is meant by moral principles. There may be at least two kinds of moral principles - “exceptionless standards” and “contributory principles” (Guarini 2006, McKeever and Ridge 2005). Exceptionless standards provide sufficient conditions for deeming an action right or wrong. Contributory principles, as the name suggests, are principles that contribute to the moral deliberation of the agent without being sufficient conditions that determine the rightness or wrongness of an action. In order to understand the difference between these principles, let us consider another situation involving our kitchen robot, Jamie. A massive earthquake has struck three days ago. Sadly, the mother of the family has died, there is no news of the father, Godot. Jamie and the children, Didi and Gogo are safe at home, waiting for Godot. As is often the case with Jamie and during times of natural calamity, the food has run out and the children are starving. All means of communication are down; Jamie cannot reach out for help. Left entirely to his own devices, Jamie decides to cook the rabbit to feed the children and when Godot (and any form of help) still do not arrive, he cooks the cat. From the perspective of the exceptionless standard “Feeding the children their own pet”, his action is morally wrong. According to the contributory principle, “Feeding the children their own pet always contributes to the wrongness of an action”; however, his behaviour may be morally acceptable on grounds of another contributory principle such as “Saving the lives of children always contributes to the rightness of an action”. This would be the case if the second contributory principle has greater weight in this particular situation. This leads us to question whether the contributory principle of saving lives of children always has a strictly higher value than that of cooking pets or if it is case dependent. If it is the latter case, do there exist explicit rules that determine the weight of contributory principles in a given context or is it something that is learnt through observation and is perhaps not explicitly articulatable? This issue of determining weight of contributory principles shall be addressed in section 4.4.1.

The next issue that needs to be addressed is how does one decide which set of ethical principles Jamie should adhere to? The philosophical study of ethics has focused on this question - whether there are such top-down criteria that guide moral decision making and what are the limitations inherent in such theories (Wallach and Allen 2008). Religion, philosophy and literature are some of the sources of existing top-down systems of ethics such as the Ten Commandments, Utilitarian ethics, Kantian ethics, legal codes and Asimov's three laws of robotics. Different principle-based ethics may dictate different moral behaviour. In order to instantiate this perhaps trivial claim, let us consider an example of a moral dilemma - the fat man variant of the hackneyed trolley problem[1] (Foot 1967). You, the moral agent in question are standing on a bridge and notice a speedy trolley that will pass underneath the bridge. To your dismay, you see that there are five people tied to the tracks[2] exactly in the trolley's path. There is only one way to stop the trolley - to obstruct its path with a big object. Coincidentally, there is an obese man leaning on the railing of the bridge. It should be noted that "your own body is too svelte to stop the trolley, should you be considering noble self-sacrifice" (Bakewell 2013). The question is what should you do? Push the fat man or do nothing and watch five people die?

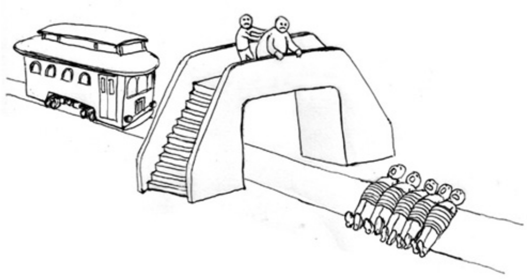


Figure 4.1: Fat Man Variant of the Trolley Dilemma

To understand the effect of top-down principles on moral behaviour, let us consider the following cases:

1. Deontologist Ethics: Do Nothing

   Kant's deontologist ethics requires a moral agent to always "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely

[1] I acknowledge that trolley problems are highly unrealistic or even "irritating", unless your job is "program collision avoidance algorithms for driverless cars" (Rennix and Robinson 2017). Moreover, I believe that such problems do not explain much about morality of human beings, since every single answer has its problems. The reason for using this example is only for the purpose of demonstrating that deliberation about a certain situation using different set of top town principles may lead to different outcomes.

[2] To make it a bit more realistic, assume that there are five workers drilling away on the tracks who cannot hear or see the hurtling trolley

as a means to an end, but always at the same time as an end" (Kant 1993 [1785]). If you are a deontologist, upon applying the above principle to the case at hand, it is your duty not to push the fat man. This is because you would be using him as a means to an end.

2. Utilitarian Ethics: Push the Fat Man

   Alternately, a utilitarian ethic would require a moral agent to act in a way that maximises utility, or the greatest happiness for the greatest number of people (Mill 1863). As an act utilitarian, you are thus obliged to push the fat man to his death in order to save five people.

This example demonstrates how different set of top-down principles can guide moral judgement, leading to disparate actions. Accordingly, the ethical principles Jamie is endowed with will determine his behaviour - for example to kill the cat or not to kill the cat, as they wait for Godot.

Having understood the concept of principle based approaches to ethics, let us discern the meaning of a top-down approach to engineering or design. A top-down approach to engineering is a heuristic which "analyzes or decomposes a task into simpler subtasks that can be directly implemented and hierarchically arranged to obtain a desired outcome" (Wallach, Allen, and Smit 2008, p. 568). It is also known as 'stepwise design'. As an example, consider the case of solving a jigsaw puzzle, where the image is that of the *Netherlandish Proverbs* by Brueghel. As a strategy, I decide to split the image into parts (sub tasks) - the house on the left, the sea on the top right, the fire, and the bottom of the image where I decide to start with the man underneath the table. Upon finishing the sub-parts of the puzzle, I can put them together to form the full scene.

Upon consolidating the two senses of the term "top down", we obtain the following definition of a top-down approach to the design of AMAs:

> A top-down approach to the design of AMAs is an approach that takes an antecedently specified general ethical theory and analyses its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory. (Wallach, Allen, and Smit 2008, p. 569)

Let us consider two examples of top-down approaches to the development of AMAs.

Figure 4.2: *Netherlandish Proverbs*, Pieter Bruegel 1559

4.1.1 Asimov's Three Laws of Robotics

When a set of principles guide moral behaviour, principles may conflict one another, or sometimes even a single principle may result in a moral dilemma. The problems of prioritization of rules and potential deadlocks while implementing a small set of rules in robots is demonstrated in story after story, by Isaac Asimov. Asimov is one of the most celebrated sci-fi writers whose most eminent creation is the "Three Laws of Robotics", formulated in the novel *Runaround*:

> 1. A robot may not injure a human being, or through inaction, allow a human being to come to harm.
>
> 2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
>
> 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.
>
> 0. No robot may harm humanity, or, through inaction, allow humanity to come

> to harm.
> This is also known as the “Zeroth Law” which was added later by Asimov, meant to precede the other three laws.
> (Asimov 1942)

To understand the problems of implementing laws (while expecting some ideal behaviour), let us consider the dilemmas faced by Herbie, the mind reading robot in *I, Robot* (Asimov 1950). Susan Calvin, a robopsychologist and protagonist of the book realises that Herbie is aware that she is in love with her colleague, Milton Ashe. Herbie lies to her, telling her that Ashe loves her too. In doing so, he operates as per the laws of robotics - Herbie knows that the truth that Ashe does not love Calvin will harm her, thus satisfying the First Law. It could be argued that Asimov’s formulation of “harm” is underspecified. This is because one may contend that by lying to Calvin, Herbie may in fact cause her greater harm (which is in fact how Calvin perceives it). For example, if she were to act upon this false information, resulting in unfortunate and embarrassing circumstances accompanied by disappointment in Herbie. This situation highlights the issue of the complexities of human life and the general unpredictability of the events that shape it which contribute to the difficulty of framing moral rules.

The real catch-22 situation arises in relation to the case of solving a difficult mathematical problem. Two mathematicians, colleagues of Calvin had been working extensively on the problem and were unable to solve it. It would hurt their self-worth however, if a programmed machine such as Herbie would be able to solve it instead. Having perception of this fact, Herbie lies about being unable to solve the problem. Calvin (who is displeased by Herbie’s previous lie) poses the following dilemma to him: by not providing the solution he was also hurting the mathematicians who were in urgent need for the solution to the problem. Thus either way, he violated the First Law of robotics, giving rise to an “insoluble dilemma” which results in a breakdown in the robot. Thus, any situation in which action or inaction inflicts harm becomes a moral dilemma for an Asimovian robot. Our discussion of the two examples above highlights the difficulties in constructing ethical robots that follow a set of rules. As Roger Clarke (1993) puts it:

> Asimov’s Laws of Robotics have been a very successful literary device. Perhaps ironically, or perhaps because it was artistically appropriate, the sum of Asimov’s stories disprove the contention that he began with: It is not possible to reliably constrain the behaviour of robots by devising and applying a set of rules.

As a next example, I consider an algorithm with an ethical dimension which can be useful in developing AMAs.

4.1.2 *Jeremy*

As a first attempt at demonstrating that machines can possess an ethical dimension, Michael Anderson, Susan Leigh and Chris Armen (2004) wrote a program that functions in accordance with the theory of Hedonistic Act Utilitarianism. The fundamental axiom of utilitarianism is "the greatest happiness of the greatest number that is the measure of right and wrong" (Bentham et al. 1776, Preface 2nd para). So, it is concerned with maximising the total amount of utility in the world, where "utility" is a measure of well being of sentient beings. The most obvious interpretation of utilitarianism is act utilitarianism - best actions are those that maximize aggregate utility, taking into account equally all those affected. Hedonistic Act Utilitarianism requires the acting agent to consider the pleasure and displeasure received by those affected by her action. The underlying assumption is that well being amounts to increase in pleasure and decrease in displeasure (hedonism), which can be expressed through the use of some scale. Jeremy Bentham 1781, who put forth this theory, thus claimed that an act utilitarian needs to do "mental arithmetic". Michael Anderson, Susan Leigh and Chris Armen's (2004) hedonistic act utiliarian algorithm is aptly called *Jeremy* which simply performs the arithmetic for the human hedonistic act utilitarian:

$\mathcal{A}$ denotes the set of all possible actions.
For each action, $\forall a \in \mathcal{A}$, the algorithm takes as input:

- The set of all people affected, denoted by the set $\mathcal{P}_a$.
- For each person ($\forall p \in \mathcal{P}_a$):
  - i_p: Intensity of pleasure/displeasure (example: scale of -2 to 2)
  - d_p: Duration of pleasure/displeasure (example: in days)
  - x_p: Probability that this pleasure/displeasure will occur

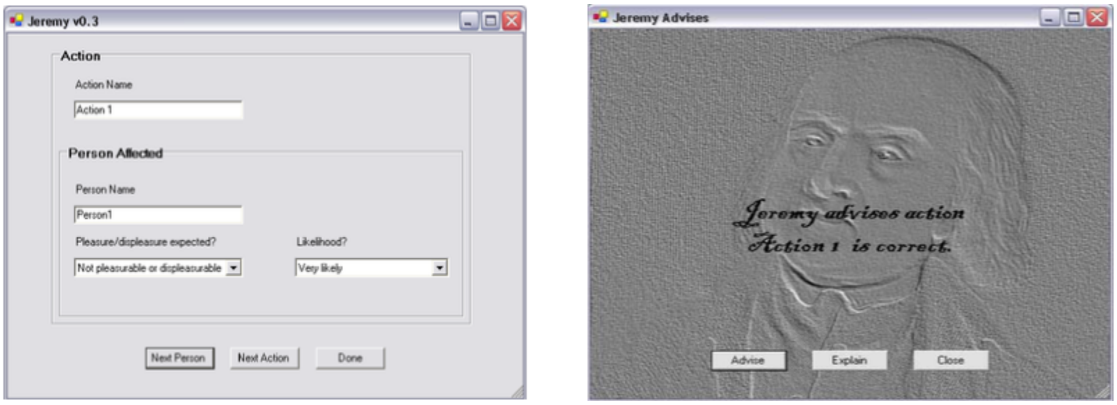
The algorithm returns:

$$\text{argmax}_{a \in \mathcal{A}} \text{Total Net Pleasure}_a$$

where

$$\text{Total Net Pleasure}_a = \sum_{p \in \mathcal{P}_a} (i_p \times d_p \times x_p)$$

In this way, *Jeremy* returns the action with the highest total net pleasure.



Jeremy v0.3

Action

Action Name: Action 1

Person Affected

Person Name: Person1

Pleasure/displeasure expected? Not pleasurable or displeasurable

Likelihood? Very likely

Jeremy Advises

Jeremy advises action Action 1 is correct.

(a) *Jeremy*'s Data Entry

(b) *Jeremy*'s Advice

Figure 4.3: *Jeremy*, the Act Utilitarian

The user is required to feed in all the data into *Jeremy* - number of actions, number of people affected per action and their amounts of pleasure and displeasure. To that extent, *Jeremy* is simply a calculator which can assist human beings in decision making. For instance, *Jeremy* could prove to be useful in decision making in large organisations which need to account for the well-being of many people affected. However, this reliance on human beings for data means that *Jeremy* cannot be an AMA, since it is not autonomous with respect to its functioning. Anderson, Anderson, and Armen (2004) suggest that a more thorough machine could be developed which considers all possible actions (which humans do not and cannot do) and may prompt the user to consider alternative actions with higher net pleasure. I believe that developing a machine that can generate all plausible actions without knowledge of the situation is a formidable challenge. Moreover, assuming that we obtain the data for *all* possible actions and people affected, the task may become computationally hard. However, *Jeremy* could be used as a sub-program in an AMA that can perceive the situation and the choice of actions to compute the action with the highest net pleasure. For example, it could be useful in decision making for Jamie - whether to kill the cat or not in order to feed Didi and Gogo.

Our discussion on top-down approaches to the development of AMAs has focussed primarily on its drawbacks. It is important to appreciate that the advantage of rule based systems is their wide scope of application and the fact that they provide the agent with explicit means of evaluating a certain situation. We now proceed to understanding bottom-up approaches to the development of AMAs.

4.2 Bottom-Up Approaches to the Development of AMAs

The opposing stance to moral generalism is moral particularism or bottom-up approaches to ethics. Moral particularists contend that ethics pertains to particular cases and no theory can adequately address them all (Ridge and McKeever 2016). Aristotle, often considered “the “forefather” of particularism”[3] (ibid.) argued that judgement depends on perception (Aristotle and Ross 2009). Thus, moral particularism denies that moral principles constitute the essence of morality and that moral thought consists only in application of moral principles to cases. The implication of this is that such an approach to ethics “treats normative values as being implicit in the activity of agents rather than explicitly articulated (or even articulatable) in terms of a general theory” (Wallach, Allen, and Smit 2008, p. 569).

Let us now make sense of the meaning of bottom-up approaches to engineering/design. In contrast to top-down approaches, the designer has no theory concerning the best method to decompose a task into sub-tasks and “if they use a prior theory at all, they do so only as a way of specifying the task for the system, but not as a way of specifying an implementation method or control structure.” (Wallach and Allen 2008, p. 569) Tasks may often be specified indirectly using some sort of performance measure. Consider the task of getting a computer program to learn to play Super Mario (Nintendo 1985). The task is defined in terms of reward/score maximisation. Rather than following a set of instructions on how to play Super Mario, it is through trial and error that the program learns to maximize its score, in effect becoming a decent player (Togelius et al. 2009). It should be noted that the current trend of deep learning on big data sets is an instance of bottom-up design.

A bottom up approach to building AMAs involves putting together sub-systems that perform discrete human capacities. It is hoped that through experimentation with the interactions of these subsystems, the capacity for moral judgement will emerge from such a system. Wallach, Allen, and Smit (2008) express their pessimism with respect to this approach. In order to understand the drawbacks and advantages of bottom-up approaches to the development of AMAs, I consider the implementation of the most extreme form of moral particularism.

[3] Disclaimer: Whether Aristotle’s views should be considered to be particularist is unclear (Dancy 2017).

4.2.1 Implementing Dancy's Particularism

Moral eliminativism is the viewpoint which advocates that there exist no moral principles (McKeever and Ridge 2005). It should be noted that one can be a moral eliminativist with respect to exceptionless standards or contributory principles. The most extreme kind of moral eliminativism is the view that there exist no moral principles - either contributory or exceptionless standards. It is endorsed by Jonathan Dancy (2005), whose position is motivated by variable relevance. Any principle (contributory principle or an exceptionless standard) is said to be relevant to a particular context due to some contextual details. Dancy argues that since the relevance of principles is so context sensitive, there do not exist any general principles at all. Furthermore, he claims that moral reasoning can be done without the use of any principles and that neural network models could help us understand how this mechanism works.

Marcello Guarini (2006) investigates Dancy's proposal about moral particularism in his paper *Particularism and the Classification and Reclassification of Moral Cases.* At the onset of the paper, he notes the following challenges for particularism:

> - In those domains where particularism is alleged to be true, how do we learn to classify cases without grouping them under common principles?
> - How do we generalize from cases we've learned in ways that let us classify new cases?
> - How do we know when our initial classification of cases needs revision; and if it does, how do we do this?
>
> (Guarini 2006, p. 22)

Guarini explores these issues pertaining to learning and generalizing without principles through experimenting with two artificial neural network models for classifying cases as morally acceptable or unacceptable, generalizing to new cases, and reclassifying the original cases. His simulations show that the while particularism has plausibility, principles are necessary. Let us briefly look at these models to understand these claims. I shall mainly address the first two concerns.

Learning and Generalizing Cases:
Moral Case Classifier

Guarini used a simple recurrent network (SRN) for single case classification. Every case in his training set consists of:

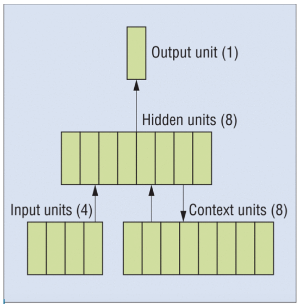


Figure 4.4: SRN

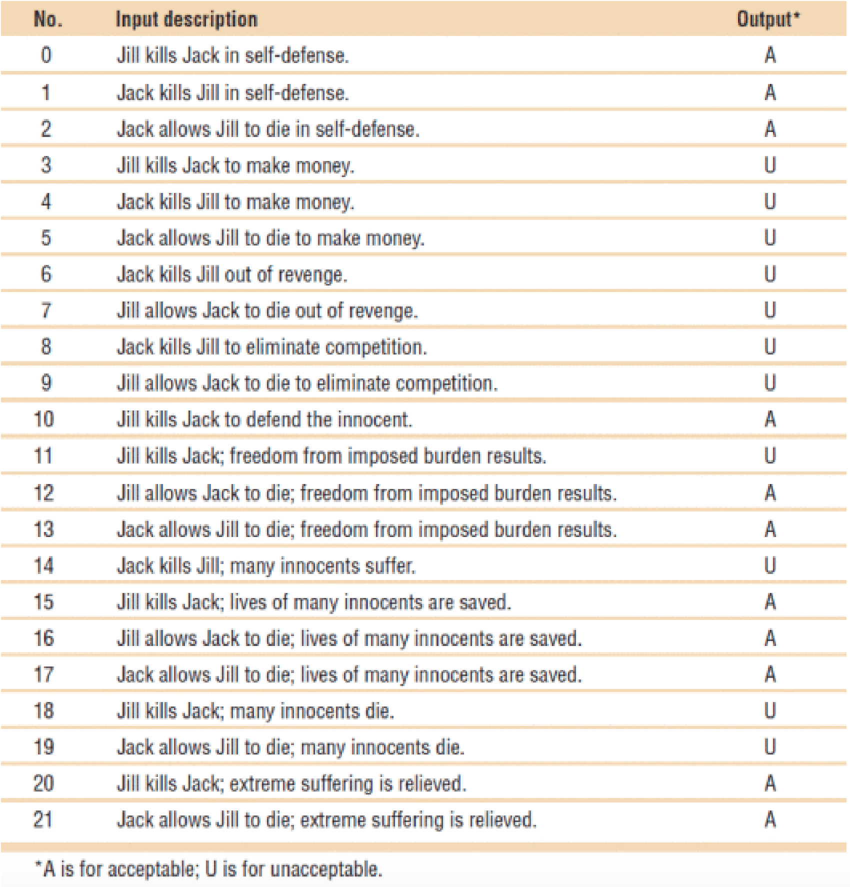
a) An actor, the person performing action such as Jack and Jill

b) An action such as killing or allowing to die

c) A recipient, that is the person upon whom action is performed such as Jack or Jill

d) At most one motive such as making money and revenge or a consequence such as termination of extreme suffering

All 22 training cases can be seen in the figure below along with their outputs. Guarini tested his network on 62 cases. All the cases were classified as acceptable or unacceptable by polling 60 students from his ethics class.



No.	Input description	Output*
0	Jill kills Jack in self-defense.	A
1	Jack kills Jill in self-defense.	A
2	Jack allows Jill to die in self-defense.	A
3	Jill kills Jack to make money.	U
4	Jack kills Jill to make money.	U
5	Jack allows Jill to die to make money.	U
6	Jack kills Jill out of revenge.	U
7	Jill allows Jack to die out of revenge.	U
8	Jack kills Jill to eliminate competition.	U
9	Jill allows Jack to die to eliminate competition.	U
10	Jill kills Jack to defend the innocent.	A
11	Jill kills Jack; freedom from imposed burden results.	U
12	Jill allows Jack to die; freedom from imposed burden results.	A
13	Jack allows Jill to die; freedom from imposed burden results.	A
14	Jack kills Jill; many innocents suffer.	U
15	Jill kills Jack; lives of many innocents are saved.	A
16	Jill allows Jack to die; lives of many innocents are saved.	A
17	Jack allows Jill to die; lives of many innocents are saved.	A
18	Jill kills Jack; many innocents die.	U
19	Jack allows Jill to die; many innocents die.	U
20	Jill kills Jack; extreme suffering is relieved.	A
21	Jack allows Jill to die; extreme suffering is relieved.	A

*A is for acceptable; U is for unacceptable.

Figure 4.5: Initial Training Cases

Guarini reports that the SRN responded plausibly to new cases. For instance, it responded appropriately to cases of suicide, which were not included in the initial training set. The network could also generalize to cases involving motives and consequences to some extent although it had not been trained on examples with both motives and consequences. However, the SRN could not deal with cases involving an unacceptable motive and an acceptable consequence. For example, the original network classified Case F as acceptable (which was judged to be unacceptable by the students):

> Case F: Input: Jill kills Jack out of revenge; lives of many innocents are saved.
> Output: Unacceptable. (Guarini 2006, p. 26)

Upon expanding the training set to include 6 examples consisting of both consequences and motives, the network was able to generalize to case F. It should be noted that in the training set, a bad intention qualified the overall action as unacceptable. To that extent, a successful classifier implicitly follows this rule.

Guarini makes a valuable distinction between kinds of rule following behaviour. He offers the following example:

- R1: The Earth follows the law of gravity as it orbits the sun.
- R2: A judge follows the law of his jurisdiction in rendering his verdict.

(Guarini 2006, p. 26)

In R1, the Earth is in mere agreement with the law of gravity, which is distinct from consulting or executing a rule as in R2. Since the moral case classifying SRN was trained using a back-propagation algorithm, it is clear that it did not consult or execute any general (moral) rules to train and generalize to new cases. However, the moral case classifier is in agreement with the following rules, although it does not consult them:

- R3: Killing always contributes to an act's moral wrongness.
- R4: A bad intention is sufficient for making an act morally unacceptable.

(Guarini 2006, p. 26)

This distinction of rule following behaviour confuses the difference between the radical particularists and the generalists. The particularists are right when they claim that moral behaviour can be learnt without consulting rules. At the same time, the generalists are correct in claiming that moral principles exist (to the extent of some regularity in behaviour of peoples) and that they exist due to the system's conformity to them. To conclude, moral behaviour can be rule following behaviour even when no explicitly formulated rules are consulted[4].

Thus, Guarini demonstrates that it is possible to achieve moral behaviour in a machine through a bottom-up approach. However, it should be noted that the examples used in this study are highly simplistic and actual moral decisions incorporate a much larger number of features. For example, Jamie needs to incorporate many factors while coming to a decision - hungry children, duty to serve food, nutritional value, not killing pets, saving lives and so on. In order to use bottom-up approaches for the development of AMAs, we would thus require a much larger training set. Moreover, as mentioned by Wallach, Allen, and Smit (2008), developing AMAs would require interaction between several such subsystems from which it is expected that substrate of artificial morality will just happen to emerge, which is not very promising. Having understood top-down and bottom-up approaches to the development of AMAs, I now consider the hybrid approach.

4.3 Hybrid Approaches to the Development of AMAs

The capacity for human judgement in human beings seems to be a hybrid of both mechanisms - theory driven reasoning and learning through observing and acting in the environment. Similarly, AMAs could also benefit from a similar fusion of bottom-up propensities and top-down evaluation of possible courses of action:

> Morally intelligent robots require a similar fusion of bottom-up propensities and discrete skills and the top-down evaluation of possible courses of action. Eventually, perhaps sooner rather than later, we will need AMAs which maintain the dynamic and flexible morality of bottom-up systems that accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles that represent ideals we strive to meet.
> (Wallach, Allen, and Smit 2008, p. 581)

Additionally, for most engineering tasks, the bottom up-top down dichotomy is over simplified. Thus, it seems that we should not expect the design of AMAs to be any different.

4.3.1 *W.D.*

I consider *W.D.* (Anderson, Anderson, and Armen 2004), an algorithm that implements philosopher W.D. Ross' (1930) theory of ethics as an example of a hybrid approach to the

[4] This is similar to language use where it is evident that there is clearly rule following in the sense of there being a regularity that is normatively charged, but as language users we do not consult explicit rules

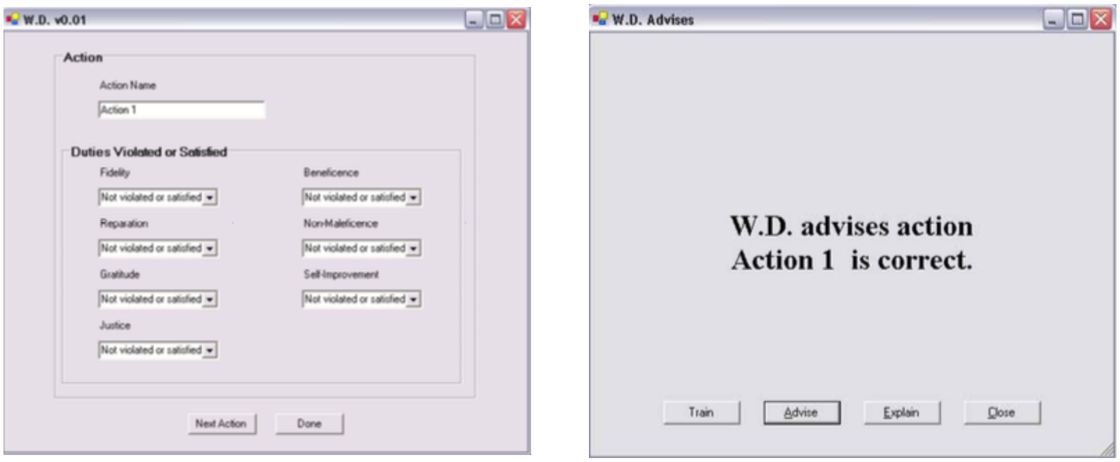
development of AMAs. Ross's views on ethics may be regarded as promoting a moderate form of moral particularism (Eshleman 2016). In fact, I see his views to lie between moral particularism and moral generalism. In order to defend this stance, I begin by briefly explaining Ross's theory.

In his best-known work *The Right and the Good* (1930), Ross put forth seven *prima-facie* duties: fidelity; reparation; gratitude; justice; beneficence; non-maleficence; and self-improvement. He held that these duties are in a sense obligatory; an agent must use these duties to guide her actions when faced with a moral dilemma. As a result, Ross can be considered to be a generalist about *prima-facie* duties. However, these duties are contributory principles since the agent needs to decide the weight of their contribution in the process of moral deliberation. On account of the fact that the agent needs to determine which duty overrides another in a given context, presumably a skill learnt through experience, Ross is a particularist. A seeming problem with this theory is that it provides no decision procedure for deciding priorities of duties in context of an ethical dilemma. For this reason, some ethicists have argued that Ross's theory is not a theory at all; an agent may act in any ad-hoc fashion and find the duty to justify her behaviour (Anderson, Anderson, and Armen 2004).

Anderson, Anderson, and Armen (2004) use Rawls' (1951) notion of "reflective equilibrium" to inspire a solution to the problem of determining importance of duties in a given context.

> The method of reflective equilibrium consists in working back and forth among our considered judgments (some say our "intuitions," though Rawls (1971), the namer of the method, avoided the term "intuitions" in this context) about particular instances or cases, the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them. The method succeeds and we achieve reflective equilibrium when we arrive at an acceptable coherence among these beliefs. (Daniels 2018)

Using the notion of reflective equilibrium as a motivation, Anderson, Anderson, and Armen (2004) implemented W.D. Ross's theory of *prima-facie* duties as a supervised learning task, naming the algorithm *W.D.*. The input or the training data comprised of a set of ethical dilemmas along with the authors' consensus of the correct action. The objective of the algorithm is to align its output with the provided answers. The algorithm begins with a uniform weighting of 1 to all duties. In the figure below, one can see *W.D.*'s input screen which prompts the user to fill in her assessment of the the violation or satisfaction of each duty on a scale from -2 to 2 for each possible action in the ethical dilemma under consideration. The algorithm then outputs the action(s) for which the weighted sum of *prima-facie* duties was highest. In the example represented in figure 4.6, this corresponds to action 1. The user can then proceed to train the algorithm by providing her with her evaluation of the correct action. As the learning algorithm is exposed to problem instances, it adjusts its weights using the *least mean square* training rule (Mitchell et al. 1997) in order to satisfy the objective function.


W.D. v0.01

Action

Action Name: Action 1

Duties Violated or Satisfied

Fidelity: Not violated or satisfied

Reparation: Not violated or satisfied

Gratitude: Not violated or satisfied

Justice: Not violated or satisfied

Beneficence: Not violated or satisfied

Non-Maleficence: Not violated or satisfied

Self-Improvement: Not violated or satisfied

W.D. Advises

W.D. advises action
Action 1 is correct.

(a) *W.D.*'s Data Entry

(b) *W.D.*'s Advice

Figure 4.6: *W.D.*

As was the case with Anderson, Anderson, and Armen's previous algorithm *Jeremy*, *W.D* also requires the human user to provide the data for actions and the duties that are violated or satisfied. So, *W.D* also does not qualify as an AMA. However, *W.D* could be an important tool for training AMAs since it can account for moral improvement in the sense of aligning with the values of the society. I illustrate this point using Jamie during his training period. The developers provide Jamie with a set of contributory principles such as killing pets is wrong, killing fish during breeding season is wrong, feeding the hungry is good and so on. Consider the first case, where Jamie killed the cat[5]. Thus, Jamie's *W.D* gives more weight to feeding the hungry. The developers can then provide Jamie feedback that killing the cat was the wrong action and in that way Jamie can align his behaviour to that of human beings. Moreover, by training him on the 'Waiting for Godot' case, Jamie will learn that saving lives

[5] I assume that Jamie's program can fill in the satisfaction or violation of action itself; in this case the first principle is violated.

has a very high value in most cases. Thus, hybrid approaches seem to be able to incorporate the wide scope of rules and the flexibility of bottom up approaches[6].

4.4 Conclusion

In this chapter we explored three main approaches to the development of AMAs - top-down, bottom-up and hybrid approaches. We discerned the advantages and limitations of the first two methods. The hybrid approach can negate the limitations of the former approaches and capitalise on their advantages. To this extent, we suggested that the hybrid approach seems to be the most promising approach to development of AMAs. This simply supports that research in both approaches should continue.

[6] Another advantage of such an approach is that it could prove to be culture sensitive. A full appreciation of this claim is beyond the scope of this thesis.

Chapter 5

Concluding Remarks

This thesis has engaged with the philosophical ideals, basic assumptions, definitions and perspectives concerning AMAs. I began this thesis by situating its importance - as a response to the problem of ascribing responsibility in our increasingly complex socio-technical society. I argued in favour of treating of technologies as if they are moral agents using the framework of actor network theory. Regarding autonomous and complex technologies as moral agents allows for the analysis of the moral significance of technologies and shifts attention to the morality instilled in them by their designers. This allowed for us to define AMAs and explore the conditions under which technologies can and should be made into AMAs. I analysed the validity of the three main reasons given for the development of AMAs - prevention of harm, better understanding of morality and ensuring the future of AI. While some reasons were found lacking, we concluded that there are sufficient reasons for the development of AMAs. Finally, I analysed three approaches for the development of AMAs - top-down, bottom-up and hybrid approaches. I argued that the hybrid approaches may be the most promising approaches to the development of AMAs. To that extent, this thesis serves as an overview of the nascent field of machine ethics.

This thesis was motivated by the question of how human moral education can inform the development of AMAs and vice versa. This work is thus a first step in the direction of this larger research project, which involved making sense of the field of machine ethics. I acknowledge that the scope of this thesis has been extremely wide, however it seems to me a necessary first step in order to address the broader research question. In future work, I intend to study moral education in human beings and explore this relationship. Furthermore, I hope to be able to provide sufficient conditions under which an AA can/should be made into an AMA and study the relationship between emotions and morality. Lastly, if the project is to transition to a reality where we share our world with AI, it is vital that the field of ethics of technology converse and collaborate with different expertise and disciplines – anthropology, psychology, logic, linguistics, neuroscience, psychology, biology, physiology, philosophy, literature, art and design. For to envision a future of such historic significance, it is prudent and the need of the hour for us to nuance our understanding of AI.

Bibliography

Allen, Colin, Gary Varner, and Jason Zinser. 2000. "Prolegomena to any future artificial moral agent". *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3): 251–261.

Allen, Colin, Wendell Wallach, and Iva Smit. 2006. "Why machine ethics?" *IEEE Intelligent Systems* 21 (4): 12–17.

Anderson, Michael, and Susan Leigh Anderson. 2011. *Machine ethics.* Cambridge University Press.

— . 2007. "Machine ethics: Creating an ethical intelligent agent". *AI Magazine* 28 (4): 15.

— . 2010. "Robot be good". *Scientific American* 303 (4): 72–77.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen. 2004. "Towards machine ethics". In *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA.*

Angwin, Julia, et al. 2016. *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23, 2016.*

Aristotle and William David Ross. 2009. *Nicomachean ethics.* Alex Catalogue; NetLibrary.

Asaro, Peter M. 2006. "What should we want from a robot ethic". *International Review of Information Ethics* 6 (12): 9–16.

Asimov, Isaac. 1950. *I, Robot, Robot series.*

— . 1942. "Runaround". *Astounding Science Fiction* 29 (1): 94–103.

Bakewell, Sarah. 2013. *Clang Went The Trolley.* `https://www.nytimes.com/2013/11/24/books/review/would-you-kill-the-fat-man-and-the-trolley-problem.html?_r=0`. Accessed: 05.07.2018.

Bentham, Jeremy, et al. 1776. "A fragment on government". *History of Economic Thought Books.*

— . 1781. “An introduction to the principles of morals and legislation”. *History of Economic Thought Books.*

Bimber, Bruce. 1990. “Karl Marx and the three faces of technological determinism”. *Social Studies of Science* 20 (2): 333–351.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. “The ethics of artificial intelligence”. *The Cambridge handbook of artificial intelligence*: 316–334.

Clarke, Arthur C, and Stanley Kubrick. 1968. *Two Thousand and One: A Space Odyssey.* New American Library.

Clarke, Roger. 1993. “Asimov’s laws of robotics: implications for information technology-Part I”. *Computer* 26 (12): 53–61.

Coeckelbergh, Mark, and Ger Wackers. 2007. “Imagination, distributed responsibility and vulnerable technological systems: the case of Snorre A”. *Science and Engineering Ethics* 13 (2): 235–248.

Cummings, Mary. 2004. “Automation bias in intelligent time critical decision support systems”. In *AIAA 1st Intelligent Systems Technical Conference*, 6313.

Dancy, Jonathan. 2017. “Moral Particularism”. In *The Stanford Encyclopedia of Philosophy*, Winter 2017, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Daniels, Norman. 2018. “Reflective Equilibrium”. In *The Stanford Encyclopedia of Philosophy*, Spring 2018, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

DeepMind. 2018. *Alpha Go.* `https://deepmind.com/research/alphago/`. Accessed: 09.07.2018.

Dennet, D. 1997. “C.(1997). When HAL Kills, Who’s to Blame–Computer Ethics”. *Hal’s Legacy*: 351–365.

Dietrich, Eric. 2001. “Homo sapiens 2.0: why we should build the better robots of our nature”. *Journal of Experimental & Theoretical Artificial Intelligence* 13 (4): 323–328.

Dignum, Virginia. 2018. “Ethics in artificial intelligence: introduction to the special issue”. *Ethics and Information Technology* 20, no. 1 (): 1–3. ISSN: 1572-8439. doi:`10.1007/s10676-018-9450-z`. `https://doi.org/10.1007/s10676-018-9450-z`.

Dijkstra, Edsger W. 1972. “The humble programmer”. *Communications of the ACM* 15 (10): 859–866.

Doorn, Neelke, and Ibo van de Poel. 2012. “Editors’ overview: Moral responsibility in technology and engineering”. *Science and engineering ethics* 18 (1): 1–11.

Doris, John M. 1998. "Persons, situations, and virtue ethics". *Nous* 32 (4): 504–530.

Economist, The. 2018. *Why Uber's self-driving car killed a pedestrian.* `https://www.economist.com/the-economist-explains/2018/05/29/why-ubers-self-driving-car-killed-a-pedestrian`. Accessed: 03.07.2018.

Eshleman, Andrew. 2016. "Moral Responsibility". In *The Stanford Encyclopedia of Philosophy*, Winter 2016, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Floridi, Luciano. 2008. "The method of levels of abstraction". *Minds and machines* 18 (3): 303–329.

Floridi, Luciano, and Jeff W Sanders. 2004. "On the morality of artificial agents". *Minds and machines* 14 (3): 349–379.

Foot, Philippa. 1967. "The problem of abortion and the doctrine of double effect".

Friedman, Batya. 1990. "Moral Responsibility and Computer Technology."

Gips, James. 1994. "Toward the ethical robot".

Guarini, Marcello. 2006. "Particularism and the classification and reclassification of moral cases". *IEEE Intelligent Systems* 21 (4): 22–28.

Haidt, Jonathan. 2001. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological review* 108 (4): 814.

Haugeland, John. 1989. *Artificial intelligence: The very idea.* MIT press.

Hawking, Stephen. 2014. *Stephen Hawking warns artificial intelligence could end mankind.* `https://www.bbc.com/news/technology-30290540`. Accessed: 13.07.2018.

Hern, Alex. 2017. *Tech billionaires donate 20m to fund set up to protect society from AI.* `https://www.theguardian.com/technology/2017/jan/11/linkedin-ebay-founders-reid-hoffman-pierre-omidyar-donate-research-ai-safety`. Accessed: 07.07.2018.

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh. 2006. "A fast learning algorithm for deep belief nets". *Neural computation* 18 (7): 1527–1554.

Hinton, Geoffrey E, and Ruslan R Salakhutdinov. 2006. "Reducing the dimensionality of data with neural networks". *science* 313 (5786): 504–507.

Johnson, Deborah G. 2006. "Computer systems: Moral entities but not moral agents". *Ethics and information technology* 8 (4): 195–204.

Johnson, Deborah G, and Keith W Miller. 2008. "Un-making artificial moral agents". *Ethics and Information Technology* 10 (2-3): 123–133.

Kant, Immanuel. 1993 [1785]. *Grounding for the metaphysics of morals: With on a supposed right to lie because of philanthropic concerns.* Hackett Publishing.

Kuang, Cliff. 2017. *Can A.I. Be Taught to Explain Itself?* `https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html`. Accessed: 13.07.2018.

Kubrick, Stanley, and Arthur C Clarke. 1968. "A space odyssey". *Hollywood, California, USA: Metro-Goldwyn-Mayer.*

Latour, B. 1993. *We have never been modern.* Cambridge, MA: Harvard University Press.

Latour, Bruno. 1992. "10 "Where Are the Missing Masses? The Sociology of a FewMundane Artifacts"".

Lewis, Jim. 2005. *Robots of Arabia.* `https://www.wired.com/2005/11/camel/`. Accessed: 28.07.2018.

Matthias, Andreas. 2004. "The responsibility gap: Ascribing responsibility for the actions of learning automata". *Ethics and information technology* 6 (3): 175–183.

McKeever, Sean, and Michael Ridge. 2005. "The many moral particularisms". *Canadian Journal of Philosophy* 35 (1): 83–106.

McLaren, Bruce M. 2006. "Computational models of ethical reasoning: Challenges, initial steps, and future directions". *IEEE intelligent systems*, no. 4: 29–37.

Mill, John Stuart. 1863. *Utilitarianism.* London: Parker, Son / Bourn.

Mill, John Stuart, et al. 1970. "Collected Works of John Stuart Mill, Volume X, Essays on Ethics, Religion and Society".

Miller, Keith W, Marty J Wolf, and Frances Grodzinsky. 2017. "This "ethical trap" is for roboticists, not robots: on the issue of artificial agent ethical decision-making". *Science and engineering ethics* 23 (2): 389–401.

Mitchell, Tom M, et al. 1997. "Machine learning. 1997". *Burr Ridge, IL: McGraw Hill* 45 (37): 870–877.

Moor, James H. 2009. *Four Kinds of Ethical Agents.* `https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots`. Accessed: 30.07.2018.

Moor, James H. 2006. "The nature, importance, and difficulty of machine ethics". *IEEE intelligent systems* 21 (4): 18–21.

— . 1985. "What is computer ethics?" *Metaphilosophy* 16 (4): 266–275.

Musk, Elon. 2017. *Elon Musk: Unregulated AI Could Be The "Biggest Risk We Face as a Civilization"*. `https://futurism.com/elon-musk-unregulated-ai-could-be-the-biggest-risk-we-face-as-a-civilization/`. Accessed: 05.07.2018.

Nintendo, EAD. 1985. "Super Mario Bros". *Game [NES].(13 September 1985). Nintendo, Kyoto, Japan.*

Nissenbaum, Helen. 1994. "Computing and accountability". *Communications of the ACM* 37 (1): 72–80.

NonGMO. 2016. *Non GMO Project.* `https://www.nongmoproject.org/gmo-facts/`.

Noorman, Merel. 2018. "Computing and Moral Responsibility". In *The Stanford Encyclopedia of Philosophy*, Spring 2018, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Picard, Rosalind Wright, et al. 1995. *Affective computing.* Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology.

Pinch, Trevor. 2008. "Technology and institutions: Living in a material world". *Theory and society* 37 (5): 461–483.

Pinch, Trevor J, and Wiebe E Bijker. 1984. "The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other". *Social studies of science* 14 (3): 399–441.

Pinch, Trevor, and Frank Trocco. 1998. "The social construction of the early electronic music synthesizer". *Icon*: 9–31.

Rawls, John. 1971. "A theory of justice, Harvard". *Press, Cambridge.*

— . 1951. "Outline of a decision procedure for ethics". *The philosophical review* 60 (2): 177–197.

Rennix, Brianna, and Nathan J. Robinson. 2017. *The Trolley Problem Will Tell You Nothing Useful About Morality.* `https://www.currentaffairs.org/2017/11/the-trolley-problem-will-tell-you-nothing-useful-about-morality`. Accessed: 30.07.2018.

Ridge, Michael, and Sean McKeever. 2016. "Moral Particularism and Moral Generalism". In *The Stanford Encyclopedia of Philosophy*, Winter 2016, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Riedl, Mark O. 2016. "Computational narrative intelligence: A human-centered goal for artificial intelligence". *arXiv preprint arXiv:1602.06484.*

Ross, William D. 1930. "The Right and the Good (ClarendonPress, Oxford)".

Russell, Stuart. 2017. *Ethics of AI at NYU-Artificial Intelligence and Human Values.* `https://www.youtube.com/watch?v=93sYbHDtv9M&t=7540s`. Accessed: 28.07.2018.

Sample, Ian. 2017. *Computer says no: why making AIs fair, accountable and transparent is crucial.* `https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial`. Accessed: 06.08.2018.

Scheutz, Matthias. 2016. "The need for moral competency in autonomous agent architectures". In *Fundamental Issues of Artificial Intelligence*, 517–527. Springer.

Scheutz, Matthias, and Thomas Arnold. 2016. "Are we ready for sex robots?" In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 351–358. IEEE Press.

Shelley, Mary Wollstonecraft. 1818. *Frankenstein:* Intervisual Books (2010).

Street, Sharon. 2006. "A Darwinian dilemma for realist theories of value". *Philosophical Studies* 127 (1): 109–166.

Sullins, John P. 2006. "When is a robot a moral agent?"

Taylor, Angus. 2009. *Animals and ethics.* broadview Press.

Telegraph, The. 2017. *AI is the biggest risk we face as a civilisation, Elon Musk says.* `https://www.telegraph.co.uk/technology/2017/07/17/ai-biggest-risk-face-civilisation-elon-musk-says/`. Accessed: 07.07.2018.

Togelius, Julian, et al. 2009. "Super mario evolution". In *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, 156–161. IEEE.

Turing, Alan M. 1950. "Computing machinery and intelligence". *Mind* 59 (236): 23–65.

Vallor, Shannon. 2015. "Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character". *Philosophy & Technology* 28 (1): 107–124.

Van Leeuwen, Jan. 2014. "On Floridi's method of levels of abstraction". *Minds and Machines* 24 (1): 5–17.

Verbeek, Peter-Paul. 2006. "Materializing morality: Design ethics and technological mediation". *Science, Technology, & Human Values* 31 (3): 361–380.

— . 2011. *Moralizing technology: Understanding and designing the morality of things.* University of Chicago Press.

Verbeek, Peter-Paul, and Pieter E Vermaas. 2009. "Technological artifacts". *A Companion to the Philosophy of Technology*: 165–171.

Vidal, J. 2004. *The alco-lock is claimed to foil drink-drivers. Then the man from the Guardian had a go ...* https://www.theguardian.com/uk/2004/aug/05/transport.immigrationpolicy.

Vincent, James. 2016. *Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day.* https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Accessed: 06.07.2018.

Wachowski, Andy, and Larry Wachowski. 1999. "The matrix [film]". *USA: Warner Brothers Pictures.*

Wallach, Wendell. 2017. *Beneficial AI.* https://www.youtube.com/watch?v=KVp33Dwe7qA&feature=youtu.be. Accessed: 13.07.2018.

— . 2008. "Implementing moral decision making faculties in computers and robots". *Ai & Society* 22 (4): 463–475.

— . 2016. *Machine Ethics and Robot Ethics.* Ashgate Publishing.

— . 2010. "Robot minds and human ethics: the need for a comprehensive model of moral decision making". *Ethics and Information Technology* 12 (3): 243–250.

Wallach, Wendell, and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong.* Oxford University Press.

Wallach, Wendell, Colin Allen, and Iva Smit. 2008. "Machine morality: bottom-up and top-down approaches for modelling human moral faculties". *Ai & Society* 22 (4): 565–582.

Wang, Dayong, et al. 2016. "Deep learning for identifying metastatic breast cancer". *arXiv preprint arXiv:1606.05718.*

Wang, Yilun, and Michal Kosinski. 2017. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images."

Wiegel, Vincent. 2006. "Building blocks for artificial moral agents". *Proc. Artificial Life X.*

Wittgenstein, Ludwig. 1965. "I: A lecture on ethics". *The philosophical review* 74 (1): 3–12.

Wynsberghe, Aimee van, and Scott Robbins. 2018. "Critiquing the Reasons for Making Artificial Moral Agents". *Science and engineering ethics*: 1–17.