Robots are not ethical like people

An exemplarist framework for functional ethics in everyday robots in ordinary contexts

by

Bongani Andy Mabaso

A thesis submitted in fulfilment of the requirements for the degree

PhD in Philosophy

in the Department of Philosophy at the
UNIVERSITY OF PRETORIA
FACULTY OF HUMANITIES

Supervisor: Prof. Emma Ruttkamp-Bloem
Co-supervisor: Prof. Deshen Moodley (External)

August 2020

Acknowledgements

I want to acknowledge my family for the sacrifices they have made in watching me sit in front of a computer for the last couple of years. I want to thank my supervisor, Prof. Emma Ruttkamp-Bloem, for being the best supervisor I've ever had in my academic studies. I would also like to thank my external supervisor, Prof. Deshen Moodley, for all the guidance with the technical aspects of this interdisciplinary work. Lastly, I am thankful for the opportunity provided to me in undertaking this work. I count it as a privilege, and I thank God for this blessing.

Abstract

As increasingly intelligent and autonomous robots continue to proliferate into every area of modern life, there is no doubt that society has to think deeply about the potential impact, whether negative or positive, that this will have on ordinary everyday contexts. One of the most urgent societal expectations for these robots is the need for them to behave in a manner that is respecting of human moral values. In response to this challenge, the field of machine ethics began with the goal of developing robots capable of making moral decisions. This thesis addresses the challenge by proposing that Exemplarist Virtue Ethics (or simply exemplarism), an ethical theory based on virtue ethics, is a viable, suitable and alternative framework for building ethical robots. Exemplarism is a moral theory that grounds key moral concepts by direct reference to exemplars of moral goodness. Essentially, it proposes that agents can develop their moral character by following the example of morally admirable agents in society. This thesis will demonstrate how an exemplarist machine ethics framework presents several advantages to building ethical robots over traditional approaches based on consequentialism and deontology. Specifically, exemplarism not only helps us formalise the concept of artificial moral agency more coherently, but it also lends itself to be a technically feasible approach for building ethical robots. This thesis will, therefore, also demonstrate the technical feasibility of actually building an exemplarist AMA and suggest ways in which it could be further improved. Since exemplarism has scarcely been applied to this area in prior literature, this thesis will provide an alternative perspective to the machine ethics project, which, in some small way can help to advance the field forward.

Key words: Artificial moral agency · Computational rationality · Exemplarist virtue ethics

Contents

List of Figures

List of Tables

Chapter 1

Context

1.1 Setting the scene

"Robots are not ethical like people" is not just the title of this thesis; it is an admission that robots are not human beings. They are not made of the same 'stuff' that we are. The title also subtly suggests that robots are not, or cannot be, ethical in the same way that human beings are. At least not with currently available technology.

Nevertheless, the question needs to be asked, have robots increased so much in capability that we need to worry about their ethics or lack thereof? Have they proliferated into society so much that this question is worth asking? The progress in the field of artificial intelligence (AI) can shed some light on these questions.

The 1950s are considered the birth years of AI, wherein it was formally established as a discipline (Buchanan, 2005; Haenlein & Kaplan, 2019). Turing's seminal work on the creation of *"Computing Machinery and Intelligence"* was a pivotal moment which firmly launched formal and directed research into the field (Turing, 1950). Just a few years later, the term AI was coined by John McCarthy at Dartmouth in 1956, where he and nine other scientists spent two months working on a detailed study of the subject (McCarthy, Minsky, Rochester, & Shannon, 2006; Russell & Norvig, 2009).

The Dartmouth conference was followed by several years of successes in the field.

In 1959, Newel, Shaw and Simon created the General Problem Solver (GPS) program, which in theory could solve certain kinds of well-formulated and simple problems (Newell, Shaw, & Simon, 1959). In 1966, Weizenbaum demonstrated ELIZA[1], a machine capable of natural language processing, and one of the first attempts at passing the Turing test (Weizenbaum, 1966).

Nevertheless, these early years of successes would not last. In the 1970s, significant funding from the US congress was cut from various AI programs (Haenlein & Kaplan, 2019). Furthermore, expectations of the performance of expert systems like ELIZA and GPS were found to be exaggerated, as they were highly dependent on the top-down formalisation of knowledge through a collection of rules implemented as 'if-then' statements.

Despite these setbacks, Hebb (1949), a neuro-psychologist, had earlier developed what has come to be known as Hebbian Learning, a theory about the way learning occurs in the neurons of the human brain. This development led to research into artificial neural networks. Neural networks, however, were deemed to be too computationally expensive to implement widely by prominent voices in the AI research community, and thus research in this area stagnated in the 1970s (J. A. Anderson & Rosenfeld, 2000; Buchanan, 2005; Haenlein & Kaplan, 2019).

Neural network research got its lucky break due to the simultaneous breakthroughs in computer microprocessors. As processing power increased, so did the scope and possibility of applying neural networks (J. A. Anderson & Rosenfeld, 2000). This combination of factors led to the resurgence of artificial neural network research in the late 1980s and 1990s (J. A. Anderson & Rosenfeld, 2000; Wasserman & Schwartz, 1988). The pace of development has only seemed to increase since then.

Although the focus has shifted towards neural networks and other data-driven approaches in recent times, it is also important to note that many other areas of AI, including rule or knowledge-based approaches, have nevertheless continued, albeit with less focus. Today, many scholars recognise the value of hybrid architectures that

[1]A version of ELIZA can be interacted with here: https://www.masswerk.at/elizabot/.

combine both approaches. Bringsjord and Govindarajulu (2020) provide a concise history of these parallel developments in the history of AI.

In today's world, it is difficult to imagine an industry that has not been impacted by AI in some way. For example, artificial agents are used to drive cars (Daily, Medasani, Behringer, & Trivedi, 2017), help us with personal assistant tasks (Leviathan & Matias, 2017), beat the world's best players in games like Go (Silver, Schrittwieser, et al., 2017), assist us in providing better healthcare (Jiang et al., 2017), and even help militaries gain strategic advantages (Sapaty, 2015).

While scholars recognise that AI can usher in a new era of personal, social and economic prosperity (Aghion, Jones, & Jones, 2017), they also warn of the potential for it to be misused towards the detriment of society (Alzahrani, 2016). Deliberate strategies are therefore required to ensure that AI can be safely introduced and integrated into society in a manner that would maximise the good for as many people as possible while minimising the bad.

To answer the questions at the top of this introduction; yes, robots have increased capabilities due to advances in AI. Moreover, yes, robots are increasingly proliferating into many areas of modern life. Modern society, therefore, has to think deeply about the potential impact, whether negative or positive, of autonomous and intelligent robots.

The general field of ethics of AI has to do with meeting the broad challenges posed by the introduction of AI into society. It is comprised of the fields of machine ethics, data ethics , robot ethics , information ethics and some aspects of neuro-ethics (P. Lin, Abney, & Bekey, 2014; Müller, n.d.; Ruttkamp-Bloem, n.d.; Zeng, 2015)[2]. Many of these fields are non-exclusive and often overlap.

Machine ethics has to do with the design of artificial moral decision making capacities in robots, including a socio-moral analysis of the concept of artificial morality (P. Lin, Abney, & Bekey, 2011; Ruttkamp-Bloem, n.d.; Wallach, Allen,

[2]Ruttkamp-Bloem (n.d.) is work in progress. Müller (n.d.) is pre-published online and available for public review.

& Franklin, 2011). Data ethics has to do with the study of fairness, accountability and transparency in machine learning. This also includes a socio-technical analysis of machine learning practices and their impact on society, and responsible data governance (Müller, n.d.; Ruttkamp-Bloem, n.d.; Zeng, 2015).

Robot ethics has to do with the impact of social robots on society. This includes the anthropomorphisation of robots, the objectification of humans, and robot rights (Asaro, 2006; P. Lin et al., 2014). Information ethics is concerned with the creation, organisation, distribution and implementation of information (Floridi, 1999, 2008). Neuro-ethics has to do with trans-humanism and human-mind uploading (Roskies, 2016). We will place our focus on the field of machine ethics in this thesis.

Many philosophers have raised concerns about how increasingly autonomous robots will treat human beings, and whether this treatment will be ethical (Allen & Wallach, 2012; M. Anderson & Anderson, 2007; Dameski, 2018; P. Lin et al., 2014; Moor, 2006). If these robots are deployed into society without the necessary measures to deal with morally charged situations, the risks of loss of human control and agency may often materialise (Cowls & Floridi, 2018).

To put it more bluntly, *"we want machines to treat us well"*, as stated by Moor (2006, p.21), one of the founding fathers of the field of machine ethics. The focus of this thesis will be on the growing challenge in machine ethics of *"developing computer systems and robots capable of making moral decisions"* (Allen & Wallach, 2012, p.100).

A question could be asked, how might equipping robots with the requisite ethical capabilities be useful in real life? Bonnefon, Shariff, and Rahwan (2016), for instance, note the social dilemmas that may arise during a self-driving car journey, where it may need to make a decision between avoiding hitting a pedestrian on the road or veering off the bridge to save the pedestrian and potentially killing its occupants.

This ethical dilemma is just one of many potential instances of the Trolley Problem (Thomson, 1985), a famous ethical dilemma in moral philosophy used to illustrate the complexity of ethical decisions in certain situations. These sorts of moral

dilemmas can be easily identified for many situations by merely doing thought experiments, which would seem to imply that robots may find themselves in these situations regularly, especially with increased deployments.

Robots will likely also encounter issues beyond just ethical dilemmas. M. Anderson and Anderson (2008) consider the ethics of providing elder-care using a robot. Cloos (2005) considers the ethics of an assistant robot in the home. Pontier and Hoorn (2012) and M. Anderson, Anderson, and Armen (2006a) ponder the issues that may occur with assistive robots in medical contexts. Arkin, Ulam, and Wagner (2012) and Arkin (2009) provide a sobering reminder of the need to embed ethics in lethal autonomous robots.

Lutz and Tamò (2015) consider general issues of privacy with any robot interaction with society. This thesis will explore a potential ethical situation that may occur in a classroom that is taught by a robotic teacher (see chapter 5). There are undoubtedly many other scenarios, and P. Lin et al. (2011) provide a good summary of some of the pertinent issues that may arise in modern society. These challenges are undoubtedly difficult to solve as they overlap many areas in the ethics of AI.

The challenge to imbue robots with ethical decision making remains one of the toughest problems in AI (Brundage, 2014; Deng, 2015; Lumbreras, 2017). For this reason, questions around the feasibility of imbuing robots with ethics are present. Brundage (2014), for example, notes issues to do with insufficient computational resources that robots have for understanding an ethical scenario. He also asks which morals would be modelled for robots, since we (humans) cannot even agree on general ethical issues, or even which ethical frameworks to employ.

There also exists some pushback from philosophers and other scholars, mostly based on arguments concerning human rights, human autonomy and human dignity (van Wynsberghe & Robbins, 2019). This pushback seems to treat the goal of building ethical robots as mutually exclusive to the preservation of human dignity. However, it could also be argued that not building them will inadvertently lead to the very thing it is trying to prevent (Asaro, 2006; Cowls & Floridi, 2018; Poulsen

et al., 2019).

Nevertheless, there is consensus in the literature that solving for machine ethics, or at the very least creating machines that align to human values, is critical and useful. There is also a strong philosophical foundation to suggest that the project to build ethical robots is at least theoretically achievable (Abney, 2012; Floridi & Sanders, 2004; Johnson, 2006; Moor, 2006; Scheutz & Malle, 2017; Sullins, 2006).

Despite the strong philosophical foundation for the machine ethics project, the risks posed by the proliferation of autonomous robots into society remain in modern life. Society needs to find ways of making increasingly intelligent and autonomous robots more respecting of human moral values. Specifically, we need frameworks that would allow us to build them with the necessary measures in place to meet community expectations of moral behaviour.

The challenge that the discipline of machine ethics faces is finding frameworks that would be best suited to building robots that could meet community expectations of moral behaviour. The current machine ethics frameworks that have been employed in the literature are largely based on the three main ethical theories in normative ethics (Dignum, 2017; Gips, 1995). These theories, however, are written from the perspective of human moral agency, without necessarily considering the prospect of an artificial kind of moral agency (Allen & Wallach, 2012).

Though there are many efforts to build artificial moral agents (AMAs)[3], the incompatibility of the traditional ethical theories to the AMA project often shows up in practice. For example, in consequentialist machine ethics frameworks, how might we ensure that the robot takes into consideration the moral values of all involved in order to perform those actions that might promote the good of all at an aggregate level?

If we take deontological and virtue ethics-based frameworks, how then would we

[3]The term AMA, which stands for artificial moral agent, has become something of a norm in the literature to refer to ethical robots. The meaning of this term, including its connotations, will be discussed in detail in chapter 2.

ensure a practical grounding of abstract deontic terms like duty, right, or wrong, and eudaimonic terms like virtues, vices and the 'good life'? After all, representing abstract terms like these is not a given for computationally based agents. There are no easy answers to these questions, and asking them here is by no means an attempt to trivialise the complexity of the issues.

It is in the nature of progress to try different things and to see what works so that we can improve what we know and build on for the future. In this thesis, an attempt is made to shift the focus from using traditional ethical theories as a basis for machine ethics frameworks, to a new one that has not been so extensively explored in the literature.

Additionally, this thesis takes a deliberately functional approach to the problem of imbuing robots with ethical decision-making capability (chapter 2). This functional approach is indeed an admission that trying to replicate human moral cognition is likely not achievable given current and near-future technology. Instead, this thesis takes a pragmatic approach by considering functional models of artificial moral agency, and how they might be practically implemented.

In this thesis, the researcher puts forward Linda Zagzebski's theory of exemplarist virtue ethics, or simply exemplarism (Zagzebski, 2010, 2017), as a possible and alternative framework that can be used in the machine ethics project.

The thesis will argue that three key features of exemplarism, namely: *grounding in moral exemplars*, *meeting community expectations* and *practical simplicity*, are crucial to its uniqueness and suitability for application in the building of ethical robots that can meet community expectations of moral behaviour (chapter 4).

To further illustrate the feasibility of using exemplarism in the machine ethics project, this thesis will also explore a detailed scenario that would seek to demonstrate how it might practically work in real life (chapter 5). Furthermore, this thesis will also provide a practical implementation of an exemplarist AMA in order to demonstrate its feasibility (chapter 6).

This thesis is significant because it will elucidate an approach to machine ethics

that has not been extensively covered in the literature. Specifically, it will help to explore whether or not robots built with an exemplarist machine ethics framework could qualify as artificial moral agents.

A potential advantage of applying exemplarism as a machine ethics framework is that it might give us a clear path to building ethical robots that could meet community expectations of moral behaviour. No other machine ethics framework has this feature at the core of its theoretical makeup. Exemplarism, at its core, is an attempt at a normative ethical theory that can potentially liberate us from the trap of having to pick and choose which frameworks we should build into ethical robots deployed in various societies.

Another potential advantage of applying exemplarism as a machine ethics framework is that it can potentially provide an alternative path away from directly attempting to model ethics in top-down ethical frameworks, which have proven to have scalability challenges (Allen, Varner, & Zinser, 2000; Scheutz & Malle, 2017). Exemplarism could also potentially solve the challenge of having to model abstract virtues and vices in virtue-ethics based frameworks, which might lead to increased adoption of virtue ethics in building ethical robots.

Furthermore, if exemplarism is as practical for implementation as this thesis concludes, then it could help pave the way for both researchers and practitioners in adopting it as a framework for building real ethical robots that respond to the challenge of meeting community expectations of moral behaviour. This result could help to progress the field of machine ethics forward in a way that has not been emphasised before[4].

This thesis is an invitation to both philosophers and engineers to work together in meeting the challenges posed by the introduction of increasingly intelligent and autonomous robots in society. It is an invitation to consider alternative models of artificial moral agency that are both philosophically coherent and practically feasible. It is an invitation for more inclusive dialogue in the AMA project.

[4]These results need to be evaluated through the functional ethics lens that the thesis takes.

1.2 Examples of ethical robots in the literature

The purpose of this section is to briefly review ethical robots that have been demonstrated in the literature. This review will help to illustrate the kinds of ethical robots that others have built, as well as the sorts of nuanced problems they were responding to by building them. This review will also help to provide a useful context for the problem that is being addressed in this thesis.

This review will not cover issues of the conceptualisation and formulation of ethical robots, and whether or not ethical robots even make sense philosophically. This task will be left for chapter 2. It will not cover the machine ethics frameworks that were used to build them in detail, including the advantages and disadvantages of each approach. This task is left for chapter 4 to deal with in the necessary level of detail.

Examples of ethical robots in a medical context

The medical environment is no stranger to the introduction of assistive technologies to aid in the provision of medical care to patients. The challenge in this context is to design and use assistive technology in a manner that respects established biomedical ethical principles and practices.

The introduction of increasingly autonomous robots in this context can pose serious challenges that need to be addressed. For example, M. Anderson and Anderson (2008) discuss an elder-care robot, called ETHEL, that has the task of reminding a patient to take medicine required for their well-being. The challenge is to do it in a way that would respect the patient's autonomy while also ensuring their physical well-being.

The elder-care robot primarily deals with ethical dilemmas, as it tries to resolve them by considering the dimensions of patient autonomy (respecting what the patient desires), non-maleficence (ensuring no harm comes to the patient), and beneficence (ensuring the good of the patient) (M. Anderson & Anderson, 2008; Ross, 1930).

If, for example, the robot reminds the patient to take medication, and they refuse, then the robot has to figure out how to resolve the ethical dilemma due to the principles of autonomy (should it respect the patient's wishes...), and beneficence/non-maleficence (at the cost of their well-being?).

Practically speaking, the elder-care robot has to keep track of when and how often the patient has taken their medication. If the patient refuses to take their medication, then the robot has to decide, based on its expert training from doctors, whether letting the patient skip their medication would have a detrimental or minimal impact on their well-being (M. Anderson & Anderson, 2008).

If the impact is minimal, then the robot may allow the patient to skip taking their medication at this time (i.e. respecting their autonomy), and it does not notify the doctors. If, however, the impact is going to be detrimental to the patient's health, then the robot continues to remind the patient to take their medication (in order to avoid harm coming to the patient). At the same time, it would notify the doctors that the patient does not want to take their medication.

M. Anderson and Anderson (2008) have applied the same design to develop several other versions of the robot, including a medical ethics advisor called MedEthEx (M. Anderson, Anderson, & Armen, 2006b). The MedEthEx works on a similar principle to advise medical workers on prominent ethical dilemmas that may occur while performing their duties. The MedEthEx robot will be further looked at in chapter 4 during the discussion of various machine ethics frameworks.

By employing W.D. Ross' prima facie duties, M. Anderson and Anderson (2008), M. Anderson et al. (2006b) effectively follow an intuition based model of moral reasoning. They have to depend on doctors' intuition regarding when the robot would be violating any of the principles of autonomy, non-maleficence and beneficence (Ross, 1930).

Hooker and Kim (2018) note some weaknesses in the ETHEL and MedEthEx robots. They point out that depending on intuition for moral reasoning may not always yield the required results. To illustrate, when doctors do not have a consensus

on a particular issue, then the ethical robots would not be trained to resolve it effectively. They also point out that part of the reason that we develop autonomous robots (e.g. self-driving cars) is because of the prevalence of human error.

Hooker and Kim (2018) suggest instead using a deontological approach to building the robots. They stress that although deontology is often associated with rules, it is however also associated with rationality (Anscombe, 1958). They point out that actions are based on reasons, and therefore can be formally represented with the use of modal logic. Their main argument is that justifiable reasons for any action should be universal and not dependent on intuition.

Hooker and Kim (2018) then developed a formalised generalisation principle for ensuring the the robot respects the principles of autonomy, beneficence and non-maleficence without ambiguity. They also successfully demonstrate in their work how the elder-care scenario could be implemented using their approach.

Even with the suggested improvements by Hooker and Kim (2018), deontological approaches in general still have their challenges[5]. One challenge has to do with the modelling of more complex scenarios. An increase in the complexity of the scenario being modelled will require more computational resources, not to mention the complexity of the modelling process itself.

Another related challenge has to do with the potential for conflicting rules, which are likely to be a feature of highly complex and realistic scenarios (Scheutz & Malle, 2017). Would the robot merely choose to obey any of the conflicting rules, or have to look at other strategies for resolving exceptions (Allen et al., 2000)? Lastly, the challenge of tractability is prevalent in many rule-based systems (Scheutz & Malle, 2017).

Ethical robots in the home context

An interesting aspect of ethical robots in a medical context is that they are not only located in hospitals or other facilities dedicated to medical care. They are also located

[5]The challenges of deontological approaches will be discussed in greater detail in chapter 4.

in the personal homes of the users/patients. The ETHEL robot discussed above, for example, can be deployed within a home context. This is because elder-care can happen both in dedicated care facilities as well as in a private home.

Another robot that has a medical function, but is deployed in the home context is Cloos' Utilibot (Cloos, 2005). The Utilibot is a home assistance robot that helps the user in performing everyday home tasks, like cleaning the house. However, its primary function is to monitor the health of the user, and to inform authorities if their prevailing health conditions worsen automatically.

Like the ETHEL and MedEthEx robots, it needs to ensure the well-being of the user by deciding when and how it should intervene. Unlike the ETHEL and MedEthEx robots, Utilibot uses a utilitarian ethical framework as a basis for its ethical routine. This would allow it to perform actions that always prioritise actions that result in the maximum utility for the user.

A similar robot was also designed by (Pontier & Hoorn, 2012; Pontier, Widdershoven, & Hoorn, 2012). Their Moral Coppélia project uses a combination of rule-based moral reasoning with an affective component, through the use of moral ratios. These ratios are then combined in a utilitarian fashion to ensure that the actions that maximise utility would be prioritised over others.

The Moral Coppélia project also utilised W.D. Ross' prima facie duties to implement the principles of autonomy, beneficence and non-maleficence. The Moral Coppélia robot can similarly also be deployed in a dedicated care facility or a private home.

The limitation[6] of the Utilibot and Moral Coppélia projects is that they require highly specific contexts where the utilities of various outcomes can be easily computed or known upfront. They cannot be easily scaled to complex scenarios that involve multiple users/stakeholders, especially when what these multiple users/stakeholders value is not known upfront (Gips, 1995; Scheutz & Malle, 2017).

[6]The challenges of consequentialist and utilitarian approaches in general will be discussed in greater detail in chapter 4.

Another interesting ethical issue that may arise with the introduction of autonomous robots in the home has to do with privacy and the ethics of sharing user information widely (Lutz & Tamò, 2015). User privacy, in general, poses a serious issue that needs to be addressed by designers of ethical robots in the home context.

At the time of writing this section, there were no published examples of ethical robots that have a deliberate and explicit privacy component built into them. This does not necessarily mean that designers have not thought about it; however, it may just be that none have chosen to stress it explicitly. There is a rich and parallel literature in 'robo-privacy' that designers can lean on, and Lutz and Tamò (2015) provide a useful summary of the main works.

Ethical robots in self-driving cars

Interesting ethical issues, especially those that seem to present as ethical dilemmas, can also occur in self-driving cars (Bonnefon et al., 2016). As stated in the introduction of this thesis, autonomous self-driving cars will likely need to be equipped with the ability to make ethical decisions, especially when faced with a decision between harming the occupants of the car and harming other road users and pedestrians.

Goodall (2014) provides a useful design that can be incorporated in self-driving cars. They propose a three-pronged approach to dealing with automated vehicle accidents. In the first phase, self-driving would implement moral ratios, similar to the Utilibot and Moral Coppélia projects, which are predetermined by industry experts, to allow the car to make decisions that would result in the highest amount of good in a particular scenario.

Given that this is a utilitarian framework, the car would select whichever decision would result in the least amount of lives lost in an automated crash environment (Goodall, 2014). The difficulty with this outcome, as expected, would be to justify why utilitarian ethics are better in this scenario than other approaches (Bonnefon et al., 2016).

In the second phase, Goodall (2014, p. 63) suggests using *"machine learning*

techniques to understand the correct ethical decision, while bound by the rule-based system in Phase 1". Although Goodall does not provide any implementation details, many machine learning-based algorithms have been developed to solve ethical problems (Kasenberg, Arnold, & Scheutz, 2018; Yu et al., 2018).

The third phase involves building in an element of explicability that allows transparency in the way the car reports its decisions, especially after a crash. This capability would allow the authorities to query the car to understand the rationale for its decisions and to improve its performance if the rationale is deemed unsatisfactory.

There are criticisms to approaching ethics in the self-driving car as ethical dilemmas (Holstein, 2017; Nyholm & Smids, 2016). The critics point out that the most critical ethical decisions in self-driving cars occur before the accident happens. They, instead, occur in the planning and even design phases of the car.

In their view, prioritising the right decisions upfront (by, for example, driving carefully and lowering speed in a congested environment) will result in far more lives saved than focusing on the split-second decisions that might be required during accidents. The point that Holstein (2017) and Nyholm and Smids (2016) are emphasising is that Trolley Problem-like ethical dilemmas represent a relatively small percentage of the kinds of ethical decisions that self-driving cars need to make.

Ethical robots in a military context

Arkin (2009), Arkin et al. (2012) note the ethical issues that may arise when deploying lethal autonomous weapons systems. To illustrate, who would be responsible when a robot kills someone in a war? What happens if something goes wrong? What happens if it kills the wrong target? Should ethical autonomous robots defend themselves against human beings? Yes, some of these questions go beyond imbuing robots with the relevant ethics, to general issues about whether or not we should even build lethal autonomous weapons systems. Nevertheless, lethal autonomous weapons systems are being built (Arkin, 2008b).

Sullins (2010) paints a vivid picture of some of the ethical issues at play when

considering lethal autonomous weapons systems. The main issue is that they can now perform tasks that would generally be performed by a human soldier. As an illustration, the robots need to determine the nature of a target and to engage them without human intervention. They also need to distinguish friendly civilians, soldiers and aircraft from unfriendly and enemy entities.

An observation could be made that lethal autonomous weapons systems are rarely set up to fire automatically, thereby reducing the urgency for imbuing them with the relevant ethics. However, as Sullins (2010) notes, in many cases, the ability for lethal autonomous weapons systems to act autonomously can be crucial to winning a war.

Sullins (2010) gives an example of a sentry robot, which needs to fire as soon as an enemy approaches a designated territory. It would be ineffective if it had to wait *"for a second opinion from its human operators before engaging potentially hostile targets"* (Sullins, 2010, p. 269). This example illustrates that there are many situations where the robot has to be given 'fire-at-will' orders.

Information about ethical lethal autonomous weapons systems that have been built is likely not made public; however, we do have Arkin's proposal of the same in his work (Arkin, 2008a, 2008b). Arkin's proposed architecture has three components, namely an ethical governor, an ethical behaviour control, and an ethical adaptor.

The role of the ethical governor is to transform all lethal actions to *lethal ethical actions*, according to the Laws of War (LoW) and Rules of Engagement (RoE). The ethical governor will either nullify the original lethal intent or transform it into one that falls within the defined ethical constraints. Where the governer is an 'overseer' of all the robot's actions, the ethical behaviour controller manages individual functions. The ethical behaviour controller constrains the robot's functions to only permissible and ethical lethal actions.

Finally, the ethical adaptor is a reflective system that determines whether or not unethical actions were performed by the robot, in order to adapt and learn not to do them in future. The ethical adaptor can perform an after-action reflective review and propagate the learnings to other robots so that they do not have to repeat the

mistake (Arkin, 2008b).

It is unclear whether or not the design proposed by Arkin was ever employed in the real-world. As a result, it is untested whether or not the design would be successful. Arkin (2008a) does, however, discuss at length how deontic logic is used to implement the LoW and RoE. Based on this, we can expect the same general limitations of deontological machine ethics frameworks, such as intractability and what to do when the robot encounters conflicting rules[7].

General observations

The examples of ethical robots in the medical, home, self-driving car, and military contexts serve as useful demonstrations of what the machine ethics project could look like in the real world. Perhaps the reader may observe them and think that little progress has been made in machine ethics. However, the researcher's view is that all these practical examples offer up useful starting blocks from which more ambitious projects can be explored.

It is telling that all the examples above are constrained to specific contexts. In general, it has been a challenge to build practical demonstrations of general ethical agents that can be deployed to any environment. It is also challenging to find machine ethics frameworks that can scale across various contexts, and equally challenging to design general architectures that can account for the complexity and abstractness of the real-world completely.

Nevertheless, there is no shortage of more ambitious projects for building ethical robots in the literature. There are models based on human moral cognition (Vanderelst & Winfield, 2018; Wallach, Franklin, & Allen, 2010), human conscience (White, 2013), logical reasoning (M. Anderson et al., 2006b; Saptawijaya & Pereira, 2014), moral ratios (Pontier & Hoorn, 2012; Pontier et al., 2012; Vamplew, Dazeley, Foale, Firmin, & Mummery, 2018) and minimalist models (Muntean & Howard,

[7]The challenges of deontological approaches, in general, will be discussed in greater detail in chapter 4.

2016)[8].

Still, many of the general models for building ethical robots do offer useful insights that designers can leverage in their projects. What is essential for machine ethics research is to have the right balance between the ambitious projects, which might aim to achieve or exceed human moral agency, and the pragmatic ones, which might respond to current challenges.

However, there has been a tendency in the machine ethics community to either focus exclusively on building robots that rival human levels of moral agency, or arguing that such robots are not possible. Allen and Wallach (2012, p.113) warn that this obsession *"can distract from the immediate task of making increasingly autonomous robots safer and more respecting of moral values, given present or near-future technology"*.

The caution above suggests that the invitation to philosophers and engineers made at the end of the introduction of this chapter is necessary. Machine ethics needs more projects involving both philosophers and engineers, not less. This partnership will allow the community to build ethical robots that are not only well conceptualised philosophically, but ones that also respond to current and future challenges equitably.

1.3 Research objectives and methodology

1.3.1 Research objectives

The purpose of this research is to propose Linda Zagzebski's theory of exemplarism (Zagzebski, 2010, 2017) as an alternative, suitable and viable machine ethics framework for the AMA project. Specifically, it is proposed that three key features of exemplarism, namely: *grounding in moral exemplars*, *meeting community expectations* and *practical simplicity*, are crucial to its uniqueness and suitability for application in the building of ethical robots that can meet community expectations

[8]These models will be discussed in further detail in section 3.2 of chapter 3.

of moral behaviour.

To further illustrate the feasibility of using exemplarism in the machine ethics project, this thesis will also explore a detailed scenario that would seek to demonstrate how it might practically work in real life. Furthermore, this thesis will also provide a practical implementation of an exemplarist AMA in order to demonstrate its feasibility. This thesis, therefore, aims to answer the following questions:

1. How might exemplarism be applied as a machine ethics framework to build ethical robots?

2. Could exemplarist robots qualify as artificial moral agents?

3. What is the practical feasibility of building exemplarist robots?

1.3.2 Methodology

In this thesis, a critical study of relevant literature will be undertaken in order to elucidate the problem of imbuing robots with ethics through the use of an exemplarist machine ethics framework. The general norms of traditional philosophical research methodology will be followed.

These include the definition and explanation of concepts, philosophical argumentation, the analysis and assessment of theoretical frameworks, the identification and critique of presuppositions, and the critical-creative reinterpretation of contemporary literature on the research problem. This kind of analytical review is the only way to deal with such a highly complex philosophical issue.

Since this research also takes a deliberately pragmatic approach by attempting to define possible practical implementations of an exemplarist ethical robot, elements of engineering design methodology will be used where necessary in order to give structure and testability to the outcomes. However, this will always be done in such a way to demonstrate the theoretical links to the philosophical concepts.

1.4 Publications

Parts of this thesis have been adapted and published in international journals. Specifically, the following articles have been published:

1. Mabaso, B. A. (2020b). Computationally rational agents can be moral agents. *Ethics and Information Technology*. doi:10.1007/s10676-020-09527-1. This article is largely adapted from chapter 3 and small parts of chapter 2.

2. Mabaso, B. A. (2020a). Artificial moral agents within an ethos of ai4sg. *Philosophy and Technology*. doi:10.1007/s13347-020-00400-z. This article has been adapted from chapters 4 and 5.

1.5 Thesis outline

The rest of this thesis is structured as follows:

1. Chapter 2 explores the concept of artificial moral agency in broader detail. This chapter begins by exploring traditional views of moral agency, before comparing them to views of artificial moral agency. The chapter ends with a clear definition of artificial moral agency based on the extensive literature review. It also addresses the issue of moral responsibility and how it relates to the AMA project.

2. Chapter 3 follows up the previous one by exploring a computational framing of artificial moral agency. In this chapter, prominent models of artificial moral agency will be reviewed, before requirements for artificial moral agency are explored. This chapter concludes by proposing a model for an optimally-bounded and computationally rational AMA.

3. Chapter 4 explores exemplarist virtue ethics and its applicability to the AMA project. Traditional normative ethical theories and their applicability to the

AMA project are explored. The chapter then moves on to a detailed comparison of exemplarist virtue ethics and the other forms of virtue ethics. It ends by positioning exemplarism as a suitable and viable machine ethics framework for the AMA project.

4. Chapter 5 follows up the previous one by exploring a detailed scenario of an AMA that is built with an exemplarist machine ethics framework. The purpose of this chapter is to immerse the reader in a scenario that would seek to illustrate how exemplarism might function in the real world. This chapter ends with a detailed discussion of both the philosophical and practical limitations of exemplarism applied to the AMA project.

5. Chapter 6 seeks to demonstrate the practical feasibility of building an exemplarist AMA. This chapter, being more technical than the ones that came before, explores the possibility of implementing an exemplarist AMA using a Partially Observable Markov Decision Process (POMDP). The chapter ends with considerations for improving the model through the use of simulation and reinforcement learning.

6. Chapter 7 reviews the content of the thesis and asks whether or not the research objectives have been achieved. This chapter also discusses the significance and novelty of the additions to the body of knowledge that was achieved in this thesis. Finally, the chapter moves on to a conclusion and recommendations for future research.

1.6 A note on interdisciplinary research

Some fields of research are more interdisciplinary than others (Van Noorden, 2015). This phenomenon occurs because some topics, by their very nature, require input from various perspectives (T. R. Miller et al., 2008). Despite the appeal of interdisciplinary research, many pitfalls need to be considered (Rhoten & Parker, 2004),

especially in an interdisciplinary field like machine ethics.

For example, the meaning of words tends to have subtle nuances that may alter the meaning of a sentence, depending on the reader and the writer's backgrounds. For example, the word 'autonomy' may appear to mean the same thing in Philosophy and Computer Science/Engineering. However, a closer examination of the word, in the Philosophical sense, reveals that it means far more than just being able to perform some actions independently. It points to the notions of free-will, independent thought and agency (Fischer, 1999).

T. R. Miller et al. (2008) suggests that to succeed in interdisciplinary research, we need to adopt a kind of epistemological pluralism. Epistemological pluralism means that any interdisciplinary research may have several valuable ways of knowing and that recognising this can potentially lead to a more integrative study.

Practically speaking, this means that we need to be aware of the epistemology of the disciplines that make up a study. For example, in Philosophy, knowledge can be considered subjective, which implies that we have to justify every argument that we make. In the Sciences, however, knowledge can be objective, especially when it has been proven and accepted in the field.

This thesis is interdisciplinary in that it deals with aspects in Philosophy and the Sciences (Computer Sciences and Engineering). The reader will be alerted every time a word or phrase that may have a dual meaning in Philosophy and Science is used. It will be made clear in which sense a word or phrase is used in order to clarify its meaning. Finally, as far as is possible, the context of the discussion will also make it apparent how a word or phrase is used.

Chapter 2

Artificial moral agency

2.1 Introduction

Having laid the context of the research in the previous chapter, it is now time to more pointedly explore the concept of artificial moral agency in greater detail. The purpose of this chapter is to define and formulate what is meant by an AMA in this research. This definition is essential for two reasons:

Firstly, a philosophically clear definition helps everyone understand the exact capacities that the AMA is envisaged to have. The definition will be useful if a conceptual comparison between the AMA, and, for instance, a human moral agent, is performed. Often it is easier to explain something if it is clear how far away it is (conceptually) from another more well-understood concept (Moor, 2006).

Secondly, it is useful from an engineering perspective because it makes it clear what is to be designed, and potentially, what has been designed once the AMA has been completed. Too many engineering initiatives aimed at building AMAs are entirely silent on the philosophical and societal implications of their projects. In an attempt to abstract these (apparently unnecessary) complexities away, these projects go ahead and build something that no one knows how to think about conceptually.

A fundamental stance that this thesis takes is that the hard science disciplines need to combine with the soft science and philosophical disciplines in order to create

better AMAs that meet society's requirements. This stance will allow us to have meaningful debates about AMAs in general.

This chapter is structured as follows. Section 2.2 tries to answer the question "what is an artificial agent?". Answering this question provides an important framing for the definition of an AMA later in the chapter. Section 2.3 follows this up by answering the question "what is a moral agent?".

Section 2.3.1 explores the concept of biological moral agency, i.e. a view that agents need to be biological in order to qualify as moral agents. Section 2.3.2 explores the concept of conscious moral agency, i.e. a view that agents need to have the capacity for consciousness in order to qualify as moral agents.

In section 2.3.3, artificial moral agency is explored in greater detail and contrasted with the views of biological and conscious moral agency, respectively. This discussion is followed up by section 2.4, which stipulates the definition of an AMA that will be used for the rest of this thesis. Finally, the chapter is concluded in section 2.5.

2.2 What is an artificial agent?

In order to discuss artificial moral agency, it is important to first begin by asking the question, 'what is an artificial agent?'. Starting with this ontological question is vital because much of what philosophers consider necessary for moral agency has to do with the notion of being, having free will and full autonomy[1]. The argument goes that full autonomy entails moral agency (Fischer, 1999, p. 98), and so if an agent is autonomous, it also has the inbuilt capacities required for moral agency.

Judging by this argument, it would seem that there is a conceptual link between what kind of being an agent is and its moral agency. In keeping with this argument, it seems logical then to start with a discussion around the nature of an artificial

[1]Generally speaking, the idea of autonomy refers to the capacity for an agent to be its 'own' person, to govern its desires, intentions and will, and not to have these imposed on it by any external force (Christman, 2018).

agent and to examine the conceptual link between the agent's nature, and its moral status. This discussion will also help lay a framework that can be used for discussion in later sections.

2.2.1 Weak AI agents

A helpful way to start the discussion is to consider the kind of intelligence that artificial agents have. Their intelligence can be viewed through either a weak or strong AI lens. Weak AI can be considered a kind of AI that can simulate intelligence (i.e. not replicate it) (Dreyfus, Dreyfus, & Athanasiou, 2000; Russell & Norvig, 2009; Searle, 1980). The implications of this are that although machines may be very good at computational tasks, computation on its own does not entail 'true' intelligence.

Over time, the term narrow AI has emerged and has started to act as a kind of synonym for weak AI, even though they are not necessarily the same. While weak AI refers to the nature of the intelligence, narrow AI refers to the degree of the intelligence (i.e. how adaptable they are to various contexts). A machine with narrow AI is built to perform specific tasks in a specific context and domain, hence its applicability is narrower than that of a generally intelligent agent that can apply its intelligence across general contexts and domains (e.g. a human being).

Artificial agents that many philosophers and engineers would comfortably agree fit the description of weak AI are commonplace in society today. Examples include chess playing bots, voice assistants such as the Google Assistant, autonomous vacuum cleaners, welding robots, share trading bots, fraud detection bots, spam filtering bots, and many others. These robots are limited in scope and capability and are designed to carry out particular tasks in predefined domains.

Hardly anyone would argue that this class of robots is conscious or autonomous to the degree that human beings have these attributes. They are merely engineering creations whose very nature is intrinsically tied to their ability to process information received from the environment and to make computed decisions about the actions to take next. Based on the above, one can assert that these robots are *what they can*

compute.

Since most of our social interactions with robots are at a level of abstraction where we cannot see the inner workings of their intelligence, one could even assert that these robots are *the sum of what they can do*. This behaviour-based evaluation of intelligence is precisely what Alan Turing was implying when he devised a series of tests[2] to evaluate a robot's intelligence (Turing, 1950).

Turing postulated that a robot's intelligence should be evaluated behaviourally, as opposed to debating whether or not it can actually think. If the robot passed the tests at a higher than chance level, then it is intelligent. The behavioural view of intelligence is an excellent fit for weak AI agents because it places emphasis on outcomes and ignores (deliberately so) that these outcomes could be achieved through simulated intelligence.

2.2.2 Strong AI agents

Having discussed the nature of artificial agents that exhibit weak AI, it is time to consider the same for artificial agents that could exhibit strong AI. Strong AI can be considered a kind of AI that can replicate (as opposed to just simulate) human intelligence (Russell & Norvig, 2009; Searle, 1980).

A related term to strong AI is artificial general intelligence (AGI), though they do not necessarily mean the same thing[3]. Goertzel and Pennachin (2007) describe AGI in the Journal for Artificial General Intelligence as *"the construction of a software program that can solve a variety of complex problems in a variety of different*

[2]Turing describes a test in which a human would evaluate natural language responses from a robot and a human being. The evaluator would be aware that one of the respondents is a robot. The evaluator would then carry out a natural conversation with both the robot and the human being, to test the closeness of the robot's answers to the human's answers. If the robot's answers were similar to the human's answers at a higher than chance level, then the robot would be deemed to be intelligent.

[3]See the brief discussion above (section 2.2.1) on the distinction between the nature and the degree of intelligence.

domains, and that controls itself autonomously, with its own thoughts, worries, feelings, strengths, weaknesses and predispositions"[4]. Based on this definition of AGI, the nature of the intelligence implies a broad scope of contexts where the agent could function.

Artificial agents that could be classified as strong AI do not currently exist in the real world. Realistically, it is far simpler to think of examples in science fiction and popular culture than it is to name real-world examples. RTD2 from Star Wars, HAL9000 from Arthur C. Clarke's Space Odyssey series and Skynet from the film franchise the Terminator are all popular examples of AGI in science fiction.

There are certainly enough experiments aimed at developing algorithms, architectures and theoretical approaches to these AIs in the real world (see any recent issue of the *Journal for Artificial General Intelligence*). However, it is not very easy to estimate when they will be truly achieved. Should it be achieved, however, then we would be able to comfortably make the statement that those artificial agents can actually think and apply themselves to various contexts.

2.2.3 The blurring lines between weak and strong AI agents

The weak and strong philosophical views on AI are certainly useful ways of conceptualising the nature of artificial agents. However, AI technology advances have occurred so rapidly in the twenty-tens that the lines between the two have become blurrier. It is often overlooked that some artificial agents that have weak AI can be generalised to work in broader contexts. Take, for example, the AlphaGo Zero bot that was created by Deep Mind (Silver, Schrittwieser, et al., 2017). Though it was only given basic rules of how to play the game of Go, it was, however, able to learn how to play it from scratch by running through many simulations of games where it was playing against itself.

[4]Interestingly, though Goertzel and Pennachin (2007) seem to indicate that AGI's would be conscious, others believe that this property is not necessary for general intelligence in machines (Bostrom & Yudkowsky, 2011)

AlphaGo managed to beat the world's best players, much to the wonder and amazement of Go followers and professionals alike. What is even more remarkable is that the designers were able to repurpose AlphaGo Zero into the more general board games robot Alpha Zero with relatively little effort. Alpha Zero also learned how to win at Chess and Shogi by playing against itself (Silver, Hubert, et al., 2017).

Alpha Zero is not just an exception either, weak AI agents that can be applied more generally are becoming more prevalent today. Take self-driving cars for an example. They have to be tested on many millions of kilometres of roads before designers, and local authorities, are comfortable with letting them loose. Despite this extensive testing, it is impossible to cover every single situation and context where the car might drive. To illustrate, even familiar streets where the car might have driven many times before can be different based on weather conditions, traffic patterns, number of pedestrians and other real-world factors that are impossible to simulate holistically.

Nevertheless, despite these varying conditions, self-driving cars can generalise what they learn to various and unseen before configurations of the context in which they operate. This is because of machine learning techniques, such as Deep Learning. Deep Learning techniques allow training on large but finite data sets, while still being able to generalise and make inferences on unseen data that fall within the distribution of the training set (Goodfellow, JBengio, & Courville, 2016).

Another example of increasingly capable weak AI is Google Duplex (Leviathan & Matias, 2017). This robot is capable of using natural language processing (NLP) and other techniques to place a phone call to a restaurant on behalf of the human user. The agent speaks to the restaurant employee that answers the phone in natural language and directs the conversation towards making a booking at a given date and time. It does all this automatically and without any intervention from the user whatsoever.

All the user receives back is a booking appointment confirmation. Duplex indeed needs a lot of domain-specific data to train on (it cannot just hold a general phone

conversation on any subject); however, such level of capability was not possible in NLP less than a decade ago. This illustrates the pace of development, and that weak AI agents are increasingly taking on more scope and can work in more of the contexts that are traditionally reserved for human beings.

Returning to the question of ontology, what should we say about the nature of artificial agents that leverage some of these breakthrough techniques in AI? The researcher suggests that these types of artificial agents should still be classified as weak AI agents. One can still say these agents are *the sum of what they can do*, however, what they can do has become, and is becoming, *increasingly sophisticated and applicable in broader contexts*.

Artificial agents that fall somewhere between weak and strong AI represent the most interesting cases for research, rather than the case of traditional weak and narrow AI agents, especially since many examples of the former are available today. From a machine ethics perspective, these agents also represent the most challenging and pressing issues that need to be resolved for a society that has to interact with them increasingly (Allen & Wallach, 2012). The insights we gain could well help us prepare for the arrival of strong AI agents, should that ever occur.

2.3 What is a moral agent?

> *"A moral agent is an agent whom one appropriately holds responsible for its actions and their consequences, and moral agency is the distinct type of agency that agent possesses"* (Parthemore & Whitby, 2014, p. 1).

Having explored the nature of artificial agents, we can now more critically discuss whether or not a robot can be a moral agent. In order to do this effectively, we need to unpack the concept of moral agency briefly. Traditional moral philosophy has a rich and diverse literature on normative ethics[5], and this is where much of the

[5]See chapter 4 for a discussion of these theories from a machine ethics perspective.

literature on moral agency resides. The focus of this section will be to define moral agency more critically and to consider the various requirements for achieving it.

Generally speaking, the idea of agency denotes the capacity for an agent to act independently (Schlosser, 2015). In contrast, moral agency denotes the capacity for an agent to act independently in so far as making morally charged decisions and actions, and to have a level of responsibility and accountability for the consequences resulting from its decisions and actions (Parthemore & Whitby, 2014). Moral agency implies a particular understanding and knowledge of what is good and what is bad (morality) and being able to discern what is right from what is wrong (ethics), as well as the moral impact of decisions and actions.

Moral agency should not be confused with moral goodness or ethical uprightness. Its emphasis is on the agent's ability to be responsible for its decisions and actions, and to be held accountable for the consequences, regardless of whether or not those actions are evaluated as morally good or bad. Therefore, a moral agent should also have the capacity to make morally bad or ethically wrong decisions.

The paragraph above, as well as the definition of a moral agent given by Parthemore and Whitby (2014), give us a good idea of the notion of moral agency, but we still have not addressed who or what can be included in the class of moral agents. The framing of this question is vital because asking who is a moral agent already presupposes personhood, which is generally taken to be embodied in human beings. Parthemore and Whitby (2014) suggest framing the question more broadly by asking *"when is any agent a moral agent"*. Such open-ended framing of the question allows one to consider a broader set of agents for inclusion or exclusion in the class of moral agents.

When one asks the question in this way, three broad categories of moral agents seem to emerge from the literature. These categories are: *biological moral agents* (Churchland, 2014; Liao, 2010; Rottschaefer, 2000; Torrance, 2008) ; *conscious moral agents* (Himma, 2009; Parthemore & Whitby, 2013, 2014); and *artificial moral agents* (Abney, 2012; Floridi & Sanders, 2004; Johnson, 2006; Moor, 2006; Scheutz & Malle,

2017; Sullins, 2006).

It is important to note that the categories of *biological*, *conscious* and *artificial moral agency* are merely convenient groupings of similar views. Many of them will share common ideas between them. Take, for instance, the view that argues for the biological basis for moral agency. It is not saying that proponents of this view do not consider other conditions as necessary for moral agency; it is saying that the most important condition for this view is the biological basis for moral agency.

It is also important to note that Torrance (2013) has developed helpful categorisations of general approaches to ethics and how to best view machine ethics. These categories are *anthropocentrism*, *inforcentrism*, *biocentrism* and *ecocentrism* (Torrance, 2013). However, Torrance's categories do not address the requirements for moral agency. Instead, they are much more focused on broad views of ethics in general, and how each of these views might shape how society treats non-human agents like artificial agents. Consequently, the categories of *biological*, *conscious* and *artificial moral agency* will be discussed next.

2.3.1 Biological moral agency

The biological basis for moral agency states that an agent needs to be necessarily biological in order to be able to qualify as a moral agent (Torrance, 2008). Even though proponents of this view would agree that capacities, such as consciousness or free will, are required for moral agency, they, however, believe that any necessary capacity for moral agency is fundamentally derived from the agent's biological nature (Churchland, 2014; Liao, 2010; Rottschaefer, 2000; Torrance, 2008)[6].

This view stems directly from the philosophy of *biocentrism* that has been eloquently described by Torrance (2013). In *biocentrism*, ethics should revolve around the well-being and needs of biological beings above everything else. biological moral

[6]Torrance is included as an authoritative source who has given a clear account of biological moral agency, though he has also given a broader perspective that goes beyond this view elsewhere (Torrance, 2013).

agency follows naturally from this by holding that all the conditions for moral agency are fundamentally derived from the biological nature of an agent. In other words, supporters of this view see the fundamental capacities necessary for moral agency as a specific feature of biological development over time. They posit that these capacities cannot exist in non-biological agents.

Biological moral agency is a somewhat popular view; therefore, we have a plethora of sources from which to articulate it clearly. There is Torrance's eloquent articulation of the 'organic view' of ethical status (Torrance, 2008). Churchland presented work on the biological platform for moral agency, which provides a neuro-biological perspective of the view (Churchland, 2011, 2014). Liao (2010) presented a fresh perspective of the view through her work on a genetic basis for moral agency. Furthermore, Rottschaefer presented his seminal work on a *"scientific naturalistic philosophical model for moral agency"* (Rottschaefer, 2000).

Besides these essential works, there are also many collections of works in books and special issues - Giovanni and Gabriele (2006) provide important work on the link between neo-Darwinian evolutionary theory and ethics, as well as nature's special issue on *"The neural basis of human moral cognition"* (Moll, Zahn, De Oliveira-Souza, & Krueger, 2005).

For the purpose of this discussion, we will focus on the works of Torrance (2008), Churchland (2011, 2014) and Rottschaefer (2000). These three authors cover a wide scope of the view - from a purely biological account of moral agency that Torrance describes, to Churchland's neuro-biological account of the same, and finally Rottschaefer, who takes the views of the others, but also adds an explicit social component to moral development.

Torrance (2008) weighs in on biological moral agency with his articulation of the *organic view* of ethical status. He states that genuinely natural organisms such as human beings have moral agency because their *"moral thinking, feeling and action"* has developed organically from their biological history. Torrance's view essentially has two components, which he summarised with the following two statements (Torrance,

2008):

1. *"Only beings, which are capable of sentient feeling or phenomenal awareness could be genuine subjects of either moral concern or moral appraisal"*

2. *"Only biological organisms can be genuinely sentient or conscious"*

These two components, taken together, suggest a direct and exclusive biological basis for moral agency, and in fact Torrance (2008) himself states as much several times in his work. His work represents one of the the clearest contemporary expressions of the organic view of ethical status.

Torrance further states that two kinds of rationality are required for moral agency, namely *intellectual rationality* and *empathic rationality*. Intellectual rationality refers to what might be considered a more traditional form of rationality, which is concerned with problem solving and reasoning - something that can be conceivably achieved by artificial agents (Genewein, Leibfried, Grau-Moya, & Braun, 2015; Lewis, Howes, & Singh, 2014).

Empathic rationality, on the other hand, contains, in addition to cognitive elements, strong affective components. This kind of rationality would seem to align reasonably well with Churchland's argument that states that moral development is dependent on the neurobiological processes that facilitate attachment and bonding in human beings (Churchland, 2011). Empathic rationality allows an agent to feel for itself and for others, and hence it can extend the concept of self-care to include others in pursuit of broader societal goals.

According to the organic view expressed by Torrance (2008), only sentient beings can have empathic rationality - precisely because sentience is about self-awareness, consciousness, and awareness of others' physiological (and even psychological) states. It is relatively straightforward to see the connection with morality from here - if empathic rationality is a unique capacity of sentient beings, and empathic rationality is a key or even the main component for moral agency, then only sentient beings can be moral agents.

How is sentience connected to the biological nature of organisms? Torrance (2008) argues that the essential capacities, such as sentience, necessary for moral agency are all intrinsically connected to the autopoietic nature of biological organisms. The autopoietic nature of biological organisms refers to the self-organising and self-maintaining nature of autonomous biological systems. He argues that this self-organising and self-maintaining nature of biological agents is the very ingredient that has caused the capacities necessary for moral agency to emerge over time in highly complex sentient beings.

Torrance (2008) calls this claim that sentience is fundamentally connected to biological nature the *enactive-autopoietic* picture. He argues that sentient beings *"have in virtue of their organic constitution, primitive features which, as they have evolved through progressively more complex forms of organisation in other species, present themselves as sentient, lived experience"*.

The organic view, if true, would immediately exclude all non-biological agents from being classified as moral agents. It seems obvious then that, according to Torrance (2008), genuine moral agency is not currently possible using current approaches to designing and building artificial agents. What current and near-future artificial agents lack is the ability to have empathic rationality, which the organic view claims is only possible in sentient and biologically-based agents. Torrance (2008) argues that a different approach to creating artificial agents, perhaps one based on creating artificial biological systems[7], may be required before empathic rational agents can be designed.

It is undoubtedly true that current approaches to artificial intelligence do not lend themselves easily to the type of empathic rationality demanded by the organic view. However, this view makes the assumption (however 'safe' it may be) that no near term breakthroughs in artificial intelligence could greatly simplify the problem of empathic rationality. In fact, there are current works that have tried to intro-

[7]The field of Synthetic Biology is, in part, focused on developing artificial biological systems (Si & Zhao, 2016).

duce affective components to ethical reasoning in non-biological agents. Pontier and Hoorn's research on using emotional intelligence and affective components to improve an artificial agent's ethical capability is a good example of such work (Pontier & Hoorn, 2012).

Neuro-philosophers such as Churchland (2011), defend biological moral agency by arguing that morality is a fundamentally neuro-biological process. Churchland is an advocate for a neuro-biological platform for moral values, and she has done extensive work on this subject (Churchland, 2014). In her work, she argues that four *"interlocking brain processes"* are responsible for moral development, namely: *"caring"*; *"learning local practices"*; *"recognition of others' physiological states"*; and *"problem solving in a social context"*.

According to Churchland, these processes are not necessarily unique to human beings, but they are most developed in humans, likely because of the expansion of the pre-frontal cortex. According to her, the neurobiology of sociality is fundamentally entranced in the human brain. She argues that the young learn social behaviour (and thus morals) through mimicry, and later in life, they depend on their complex problem-solving ability to be able to interact with new situations as they arise.

Naturally, a neurobiological explanation for any high-level brain capacity (and not just those that have to do with moral agency) will always draw criticism from Philosophy of Mind. For example, Chalmers (1995) and Searle (2007), both prominent figures in Philosophy of Mind, argue that experience and other higher-level capacities of the mind are more than the sum of the individual neuro-biological brain activities that give rise to them.

Interestingly, Chalmers and Searle are on opposite ends of the dualism debate, i.e. that the mind (consciousness) is a separate entity from the body. Chalmers argues that there is an explanatory gap between the hard and easy problems of consciousness, making him align closely with some variant of dualism, albeit not necessarily substance dualism. Searle, on the other hand, argues that neuro-biological processes plus a yet unknown, but scientifically explainable, factor give rise to consciousness,

though it cannot be proved today.

Regardless of the exact way in which neuro-biological processes affect moral development, Churchland's argument from a philosophical perspective is still clear. Her main argument is that moral development is intrinsically a feature of biological organisms that have highly developed brains, the most developed of which are human beings.

One other proponent for the biological basis for moral agency that will be discussed here is Rottschaefer (2000), in his work on a *"scientific naturalistic philosophical model for moral agency"*. By scientific naturalistic and philosophical, Rottschaefer means to say that moral agency can be sufficiently explained through the natural and social sciences. It is a natural phenomenon which is a property of the material world (thereby excluding religious or supernatural approaches) (Rottschaefer, 2000).

Like Torrance and Churchland, Rottschaefer believes that the sciences provide the best answers or account for moral agency (what it is, where it comes from, and so on). However, where Churchland takes a neuro-biological approach and Torrance's organic view a decisively biological approach, Rottschaefer uses a combination of social, psychological and biological approaches to explain moral agency.

In his view, a multidisciplinary approach is necessary in order to explain the *"nature, acquisition, activation and justification"* of moral agency (Rottschaefer, 2000). In his later work, Rottschaefer is quite critical of what he calls the armchair and intuition-based philosophical analysis of moral agency (Rottschaefer, 2009). Instead, he argues for a scientific naturalisation of normative ethics and therefore, an empirical approach to investigating morality.

He goes as far as providing a detailed scientific approach (complete with control groups) for investigating moral phenomena (Rottschaefer, 2009). In a way, his work represents a natural progression for biological moral agency in the sense that it seeks the support of other scientific disciplines to build further a case for a naturalised study of normative ethics. However, what Rottschaefer and other proponents of this view do not deviate from is taking a naturalistic scientific view, meaning that the basis

of their explanations for morality is always intrinsically connected to evolutionary biological development.

In summary, the works of Torrance (2008), Churchland (2011, 2014) and Rottschaefer (2000, 2009) are comprehensive and quite exhaustive in the way they cover the subject. Based on these works, biological moral agency can be summarised as follows:

1. **The agent must be biological** - meaning that non-biological agents are immediately excluded from consideration in the class of moral agents. Torrance (2008) does make one exception though; he allows that if an agent were to be made artificially biological, using some new approaches to artificial intelligence, then that agent would likely meet this requirement.

2. **The agent must be sentient** - meaning that the agent must have an awareness of self, others' physiological and psychological states, and its environment. The agent's sentience and consciousness must be capacities derived from its biological nature, because only biological organisms can have such capacities.

3. **The agent's sentience must be a function of neurobiological, psychological and other cognitive functions of the brain** - in other words, the key to moral agency is in the way the brain has developed to house sophisticated capacities such as those required for moral agency. A brain without, for instance, a pre-frontal cortex, would not be able to house the sophisticated capacities required for moral agency.

4. As a result of the above, the brain must be capable of both **intellectual and empathic rationality**, of which the former is believed to be key in facilitating moral development and reasoning.

2.3.2 Conscious moral agency

The second category of views on the requirements for moral agency is, broadly speaking, that agents must be conscious and sentient in order to qualify as moral agents.

At first, this view might appear to be a subset of biological moral agency. However, the process for arriving at consciousness and sentience is quite different. Advocates for the biological view would state that sentience and consciousness are biologically derived. We will simply refer to this view as conscious moral agency in this thesis.

Proponents of a conscious moral agency are indifferent about whether or not the agent is biological. They are more open to considering, and even including, non-biological agents in the class of moral agents. Having said this, supporters of this view remain fairly rigid on the requirement for consciousness and sentience as fundamental capacities required for moral agency.

There are many sources from which to articulate the view for conscious moral agency clearly. Parthemore and Whitby (2013, 2014) have a detailed explanation of this view, where they also discuss when any artefact can be a moral agent (therefore opening up the possibility for artificial moral agency). Himma (2009) also considers this view when he explores the possibility of artificial agents qualifying as moral agents. Wallach et al. (2011), often considered the pioneers of machine ethics, also discuss the role that consciousness could play in designing AMAs that can make ethical decisions.

From a pure machine ethics perspective, the above sources offer sufficient coverage to discuss the subject. However, the idea that conscious agents can be moral without strict requirements for biological embodiment finds its grounding in Philosophy of Mind. Philosophers such as Dennett (1976) have long argued that personhood and agency can be achieved without strict requirements for biological embodiment.

One of the most explicit expressions of this conscious moral agency is by Parthemore and Whitby (2013, 2014) in their work on moral agency, particularly when it comes to classifying when any artefact may be a moral agent. Parthemore and Whitby define[8] moral agency in a manner that is indifferent about the nature of the

[8] *"A moral agent is an agent whom one appropriately holds responsible for its actions and their consequences, and moral agency is the distinct type of agency that agent possesses"* (Parthemore & Whitby, 2014, p. 1)

agent.

They instead argue that any agent can be a moral agent so long as it is possible to hold it morally responsible and morally accountable. By morally responsible, they mean that the agent must have valid conceptual reasons and the means to carry out its actions meaningfully. By morally accountable, they mean that the agent must be able to communicate its conceptual reasons in order to explain the rationale for its decisions. In other words, they do not subscribe to the notion that an agent merely simulating the right sorts of things to do would qualify as a moral agent.

Parthemore and Whitby (2014) layout clear criteria or requirements for moral agency under their view. Their first requirement is that a moral agent must both be embedded within a social and environmental context, and embodied in some form that allows it to interact with others in its social and cultural context. In their view, morality is not just a function of the brain, but there is a continuum between it, society and the environment that shapes moral life.

Their second requirement is that moral agents must be alive, although they deliberately state that the *"the agent need be neither conventionally biologically formulated nor naturally evolved"*. Interestingly, they would still regard systems that are alive to be autopoietic. However, their view of autopoiesis is that it should not be narrowly constrained to biological systems only. A non-biological homoeostatic machine, for instance, would also qualify.

Their third requirement is that moral agents must be sophisticated conceptual agents. In Parthemore and Whitby (2013), they lay out a comprehensive framework for what they call a conceptual agent. Essentially, a conceptual agent is one that is capable of formulating, understanding and communicating concepts. Concepts are seen as units of thought. They may be best related to the concept of meaning. A conceptual moral agent must be able to understand the 'wrongness' or 'rightness' of its decisions and actions.

Their fourth requirement is that a moral agent must be consciously aware of itself. Lastly, they require that a moral agent be able to communicate in at least

some minimal sign language and at the most, some advanced form of communication such as speech. This ability to communicate is to allow the moral agent to be held accountable by transparently communicating the reasons for its decisions and actions when required.

Of the five requirements that Parthemore and Whitby layout, it seems that the most critical would be that a moral agent should be a sophisticated conceptual agent. It is central to their argument in the same way that proponents of biological moral agency place the biological nature of the agent as the main requirement to support moral agency.

Indeed, if one takes away the requirement that a moral agent should be a sophisticated conceptual agent, then their views on moral agency could fit in with any other category. Moreover, since arguing for conceptual agency is essentially the same as arguing for a level of consciousness in a moral agent (Parthemore & Whitby, 2013), this places Parthemore and Whitby squarely in the camp that holds the view of conscious moral agency.

Himma (2009) is another proponent of conscious moral agency. In his 2009 paper, Himma does not write off the possibility that artificial agents can be classified as moral agents. However, he argues that *"the issue of whether artificial moral agency is possible depends on the issue of whether it is possible for ICTs to be conscious"* (Himma, 2009, p. 3). Himma identifies two necessary and jointly sufficient conditions for moral agency, namely the capacity to choose actions freely; and the necessary rationality required to differentiate right from wrong.

However, he further argues that both of these conditions presuppose consciousness. The reason why is, he argues, that an unconscious capacity to freely choose is pointless, impossible even, as it only proves that the agent is not genuinely free. Similarly, he argues that an 'unconscious rationality' (i.e. reasoning about right from wrong without consciousness) is also equally meaningless because an unconscious agent cannot understand right from wrong. He ultimately concludes that moral agency without consciousness is impossible, regardless of the various attempts

to simulate free choice and rationality.

Wallach et al. (2011) are also proponents of consciousness as a basis for moral agency, albeit from a different perspective. Allen, Wallach and Franklin advocate for machine consciousness, which although not as fully developed in the literature as traditional notions of biological consciousness, purports a kind of consciousness achievable by machines. This machine consciousness is seen as critical when making certain kinds of ethical decisions and actions, and at the very least, plays a functional role in making most, if not all, moral decisions.

There are obvious questions around what machine consciousness is, how we would know that a machine has it, and what capacities a machine needs to have in order to be considered conscious. These are all open research questions in the AGI field of machine consciousness and even machine ethics. Presumably, these questions will need to be adequately answered before conscious machines can be made. The driving philosophy behind the possibility of creating artificially conscious agents seems to be a kind of strong *inforcentrism*. The philosophical idea behind inforcentrism is that the key aspects of the mind and intelligence can be replicated as computational systems (Torrance, 2013).

Based on the works in the paragraphs above, the views of the proponents of conscious moral agency can be summarised as follows:

1. **The agent must have the capacity for consciousness** - this is the main emphasis in this view, and all other conditions flow from this one. It should be noted that consciousness is not limited nor narrowly confined to a biological nature, some yet to be defined notion of machine consciousness would also be acceptable.

2. The agent must have certain epistemic capacities required for moral agency. **These capacities include emotions, empathy, free will, rationality, cognition (including mental and intentional states), concepts / conceptual abilities, phenomenal awareness**, amongst others. These capaci-

ties, however, presuppose that the agent must have consciousness.

Note Himma (2009) only mentions free will and rationality as both necessary and jointly sufficient conditions for moral agency, whilst Wallach et al. (2011) mention nearly all of the ones listed above, and Parthemore and Whitby (2013, 2014) mention only some of them. What is clear is that there is no direct consensus in the literature around which epistemic capacities are exactly required for full moral agency. What we do know is that at least some, if not all, of the ones listed above, are required.

2.3.3 Artificial moral agency

The third category of views on moral agency is the view that agents must be able to simulate the required epistemic capacities for moral agency. In other words, they are dealing with an artificial moral agency. For this view, it is essential to understand that current approaches to machine ethics are primarily computational, i.e. they are dealing with *computational morality*. Outside significant advances or discovery of new approaches to designing artificial agents, it seems unlikely that this will change soon.

Even researchers that conclude that some notion of consciousness will be required for general intelligence (Franklin, 2003), and indeed full moral agency (Wallach et al., 2011), are only working towards functional approximations of it - mostly using a combination of cognitive architectures and computational implementations (Franklin, Madl, Mello, & Snaider, 2014; Lucentini & Gudwin, 2015). For the time being, the nature of machine ethics implementations, it would seem, will remain almost certainly computational.

The proponents of artificial moral agency can be further subdivided into two. The first group are those that argue that most, if not all, of the full range of moral decisions, can be computed by some near or future term artificial agent (Abney, 2012; Allen & Wallach, 2012; Sullins, 2006). The second group are those that argue

that only certain kinds of moral decisions can be computed using current approaches to AI and that the full range of moral decisions will require super-rational capacities (Johnson, 2006; Scheutz & Malle, 2017). Let us call the former view *strong machine ethics*, and the latter *weak machine ethics*[9].

Strong machine ethics refers to the argument that moral agency can likely be fully achieved with an appropriate level of (computational) intelligence. On the other hand, weak machine ethics refers to the argument that full moral agency, at least in its historic and somewhat anthropomorphic roots (Torrance, 2013), will not be achieved using current computational approaches to AI. As a result, artificial agents can only have a pseudo or functional morality. We will consider definitions of artificial moral agency from both the strong and weak machine ethics perspectives[10].

Keith Abney's view of moral agency seems to align quite strongly with strong machine ethics, as he argues that robots could effectively become moral persons (Abney, 2012). He argues that human beings are unique (in a moral sense) because of their well developed cognitive decision-making system (what he terms a deliberative system). Abney effectively minimises the role of emotional and instinctual capacities in making moral decisions. Instead, he argues that the ability to make rational decisions is the main capacity required for moral agency (Abney, 2012).

His reason for this stance is that society only ascribes moral responsibility to those agents with full functioning deliberative systems. A mentally ill person, although still a person, would not necessarily be a full moral person because their deliberative system is compromised, and thus society treats them differently (from a moral standpoint). Similarly, animals are not held morally responsible because they

[9]Weak and strong machine ethics are categorisations that flow organically from the delineation made between weak and strong AI in sections 2.2.1 and 2.2.2 respectively.

[10]As an aside, It seems logical that holding a strong AI view is consistent with strong machine ethics. Similarly, holding a weak AI view seems entirely consistent with weak machine ethics. However, is it possible to have weak AI with strong machine ethics? Similarly, is it possible to have strong AI with weak machine ethics? Perhaps considering these various conceptualisations may help further one's understanding of the field of machine ethics.

presumably do not have sufficiently developed deliberative systems.

A fully developed deliberative system enables an agent to make rational decisions. Therefore, if an artificial agent were to have a sufficiently developed (computational) deliberative system, then it could also qualify as a full moral agent, according to Abney (2012).

Sullins (2006) could also be considered to align more closely to the strong machine ethics view. His view is that artificial agents can be moral agents using current or near-future approaches to artificial intelligence. Unlike Abney (2012), however, Sullins argues for the disconnection of moral agency from personhood. In his view, impersonal artificial agents can be considered moral agents if they can still meet specific requirements for moral agency, which he identifies.

Sullins builds his argument from the work of Floridi and Sanders (2004), especially their perspective of treating morality with different levels of abstraction, where at a low enough level not even humans would be moral agents because we would be dealing with the biology of how they work. They instead argue that, at an appropriate level of abstraction, artificial agents can also be considered moral agents.

Sullins' requirements for moral agency are *autonomy*, *intentionality* and *responsibility* (Sullins, 2006). According to him, artificial agents that meet these three criteria would qualify as full moral agents at the appropriate level of abstraction. Sullins takes things a step further by arguing that these artificial agents could also be granted rights and specific duties. This view would potentially classify him as a proponent of weak AI with strong machine ethics!

What is interesting about Sullins' view is that he is not claiming that artificial agents will ever be autonomous, intentional and responsible in the way it is philosophically thought of humans, he instead argues that they do not need to reach those levels - pseudo approximations of these capacities are more than adequate for full moral agency.

His view will likely be prone to criticism from Moral and Political Philosophy, and possibly from the Philosophy of Mind. Sullins may be seen to be reducing

moral agency to mere automation, mimicry and imitation with no real substance, whilst still claiming full moral agency. Sullins' own rebuttal would be that it is not clear that human beings aren't themselves merely sophisticated robots that are advanced enough such that we ascribe to them philosophically dubious capacities such as consciousness, mind, and others (Sullins, 2006, p. 28, 29).

Ideologically, Allen and Wallach (2012) can also be considered to align more closely with strong machine ethics. Unlike the view that Abney (2012) takes, they believe that super-rational capacities are necessary for moral agency. However, they purport that some, if not all, of the super-rational capacities that may be considered necessary for moral agency, can be simulated (Wallach et al., 2011). Interestingly, this view potentially classifies them as further proponents of weak AI with strong machine ethics.

They do, however, warn against the obsession of building artificial agents with full moral agency or showing that all such attempts are doomed for failure. This obsession *"can distract from the immediate task of making increasingly autonomous robots safer and more respecting of moral values, given present or near-future technology"* (Allen & Wallach, 2012, p. 113). In other words, the obsession could be a red herring. Instead, Allen and Wallach advocate for the building of a kind of functional morality for robots. A functional moral machine would be able to detect and respond to a given set of moral challenges without necessarily being a full moral agent.

Allen and Wallach stress the immediacy of robots that can perform human-like tasks in human spaces with ethical implications, and that we need to build functional ethics into them. So whilst they might hold a strong machine ethics view, they do acknowledge that weak machine ethics is also beneficial, especially for the short term. Let us more closely look at the views of the proponents of weak machine ethics next.

Scheutz and Malle (2017), proponents of the *weak view machine ethics*, agree with Allen and Wallach (2012) that we need to start now to equip robots with the relevant ethical capabilities. In their words, *"a massive deployment of social robots in human societies is already predictable, we need to start developing algorithms and*

mechanisms for such robots to meet human expectations of moral competence and behave in ethical ways" (Scheutz & Malle, 2017, p. 372).

According to them, it is already apparent that near-future robots are going to be autonomous and work in spaces where they will interact with ordinary human beings in ordinary social spaces, which raises the urgency of building them with the requisite moral capacities. What is interesting about Scheutz and Malle's approach is that they take a social perspective to define what an artificial moral agent is. Instead of focusing on whether or not a robot can achieve human moral intelligence, they instead focus on community expectations of robot behaviour.

This approach places emphasis on the various functional capacities necessary for meeting community expectations, and as a result, it does not require an *"objective moral agency"* (Scheutz & Malle, 2017). In other words, the focus is less on whether or not the artificial agent is actually a moral agent, but rather on the required capacities to meet community expectations of moral behaviour.

Johnson (2006) can also be considered to align more closely with weak machine ethics. In her highly influential paper, Johnson very eloquently argues for a position that states that robots cannot be moral agents, only moral entities. In her view, it would be a mistake not to expand the moral universe to include artificial entities. However, she does not think that robots can be full moral agents since they lack mental states, "intendings" about actions and freedom to act.

She further states that the only intentionality that artificial agents have is as a result of being created with a specific purpose by their designers. In other words, artificial agents are merely carrying out the intentions of human beings, even if they do so in a manner that appears autonomous.

Johnson argues for a kind of distributed morality that is shared by the designer, the artefact (robot) and the end-user. We can call this distributed morality at a *micro-level*, and it focuses attention on resolving ethical issues with robots at a small or private social level. For a vision of distributed morality at a *macro-level*, see Floridi (2013).

Johnson's reason for writing her work was not to directly address the topic of machine ethics (the whole discipline was only beginning to take shape at that time). However, it is clear that she does not believe that computational approaches to AI could create a robot that has full moral agency, only something less. Only a weak kind of machine ethics.

Let us recap the discussion of the last few paragraphs. Strong machine ethics argues for the inclusion of near-future robots as full moral agents, possibly even granting them rights and duties. Weak machine ethics, on the other hand, argues that robots cannot be full moral agents - only pseudo, functional or lesser approximations of full moral agents. It could even be argued that proponents of weak machine ethics see robots as mere tools created by humans to carry out their intentions (Amigoni & Schiaffonati, 2005; Johnson, 2006).

Like many taxonomies in philosophy, these views should not be seen as necessarily opposing, but as a continuum stretching from one end to the other (Moor, 2006). One thing they both have in common is the belief that moral agency can be achieved (either in the strong or weak sense) through the simulation of the required epistemic capacities.

There have been many attempts at listing the critical requirements for artificial moral agency (see the works of Floridi and Sanders (2004), Sullins (2006)). However, the prevailing argument in it is that morality (in either a strong or weak sense) can be achieved through (computational) rationality. As a result, the fundamental requirement for moral agency under the view can be summarised as follows:

1. Given that rationality can be computed - an **artificial agent can be a moral agent if it is also a rational agent**. All other sub-requirements (such as *autonomy*, *intentionality* and *responsibility* (Floridi & Sanders, 2004; Sullins, 2006)) can be approximated, or even replicated, computationally.

   NB: Note the only caveat one has to be aware of is whether the moral agency is referred to in either a weak or strong machine ethics sense as per the extensive

discussion in this section.

2.4 An appropriate definition and goal for an artificial moral agent

2.4.1 What does the term artificial moral agent mean for this thesis?

The focus of this research going forward will be on artificial moral agency. The reason is that the goal of this research is to demonstrate the feasibility of building computationally rational and exemplarist artificial moral agents. This is not to say that attempts to build ethical robots that follow the biological and conscious paradigms of moral agency will never be possible, e.g. see suggestions from Torrance (2008) and Parthemore and Whitby (2013), it is to say that they will likely not be achieved with current or near-future breakthroughs in technology (Allen & Wallach, 2012; Asaro, 2006).

The challenge to imbue increasingly intelligent robots with ethical decision-making capability should not only be confronted with attempts to replicate human moral agency. We need a balance between the more ambitious research projects that attempt such a feat and the more pragmatic ones that seek to answer the current challenges with philosophically coherent and technically feasible approaches to building ethical robots. This thesis takes the latter path.

Consequently, we will primarily focus on a weak framing of machine ethics, in line with our goal to build a functional ethical robot. We will aim to simulate moral cognition and not to replicate it. We are interested in robots that are respecting of human moral values, regardless of whether or not they are capable of being morally responsible (see section 2.4.2).

Let us return to the definition of moral agency given by Parthemore and Whitby

(2014, p. 1)[11] and quoted at the top of section 2.3. This definition serves as an excellent reference; however, the previous discussion (see section 2.3) has shown that different people mean different things when they use the term 'moral agent'.

Depending on one's interests and even background, it is easy to see how either one of the views of a biological, conscious, or artificial moral agency could be adopted. What is important for researchers and designers in machine ethics is to state clearly in which sense they mean the term 'moral agent'. They also need to specify what exactly their definition for it is.

Parthemore and Whitby's definition is modified to indicate precisely how the term artificial moral agency will be used in this research going forward. The modified definition is as follows:

> *An artificial moral agent is a computationally rational agent whom one appropriately holds responsible for its actions and consequences, and artificial moral agency is the distinct type of agency that agent possesses.*

The main requirement for artificial moral agency is rationality, and since we believe that rationality can be computed, we hold that computationally rational agents can qualify as AMAs. This definition also adopts a weak machine ethics view. This view means that we believe not all moral decisions can be made through computational rationality; super-rational capacities are required for others, and therefore the AMA will be limited in that sense.

The definition of an AMA given above is somewhat ontological in that it emphasises the nature of the agent. However, in theory, a definition based on the agent's moral capability could also be derived. Thankfully, scholars such as Asaro (2006) and Moor (2006) have already developed taxonomies that help characterise the level of ethical capability in artificial agents.

Asaro (2006) describes a taxonomy with five levels of ethical capability. The first level is amoral artificial agents, which are merely tools that are used by human

[11] *"A moral agent is an agent whom one appropriately holds responsible for its actions and their consequences, and moral agency is the distinct type of agency that agent possesses".*

beings. The second level is moral significance. These artificial agents attract 'blame' or 'praise' for their actions while having no ethical capabilities. The third level is sophisticated moral intelligence. These are artificial agents with some ethical decision-making capabilities.

The fourth level is dynamic moral intelligence. These are more sophisticated versions of the former, and they can develop a moral sense and even come up with their own moral systems. The last level is full moral agents. These are agents with full consciousness, phenomenal awareness and free-will. This is the level that is currently occupied by human beings.

Though Asaro's taxonomy is very useful, it also appears to have some gaps, especially when compared to Moor's version. For example, amoral agency is the default for any artificial agent, and likely does not need a unique inclusion in the taxonomy. The categories of moral intelligence and sophisticated moral intelligence are more overlapped than they are mutually exclusive. Therefore, only three levels in Asaro's taxonomy speak to different levels of ethical capability.

Moor's taxonomy, on the other hand, provides four levels which are all mostly mutually exclusive and thereby painting a useful picture of the continuum of moral agency. For this reason, we will focus on the taxonomy developed by Moor (2006).

Moor describes four different kinds of AMAs, each according to capability. These four kinds are: *ethical impact agents*; *implicit ethical agents*; *explicit ethical agents*; and *full ethical agents*.

The first kind that Moor describes is an *ethical impact agent*. This type of agent does not necessarily have ethics built-in; instead, it has an ethical impact without being necessarily concerned with ethics. Moor gives an example of robot Jockeys that race camels in the country of Qatar. The robot Jockeys are not concerned with ethics in any way. However, their use has an ethical impact since for every one of them that is deployed; there is presumably one less slave child from poorer neighbouring countries that are abducted and raised to be a Jockey.

It is easy enough to think of many other examples (e.g. a firefighter robot, a sea

rescue robot, and many others). Ethical impact agents merely have a negative or positive ethical impact in a given situation. However, they are not moral agents.

Would designing an *ethical impact agent* be an appropriate goal for our AMA? Whilst there is no doubt that all moral agents are *ethical impact agents*, reaching for this specific level only would not be sufficient since our goal would be reduced to a design of a robot that has some ethical impact, but no understanding at all of ethics.

Moor defines implicit ethical agents as robots that have ethical rules (e.g. safety rules) programmed into them by their designers. For example, the control system of a locomotive is generally programmed with specific limits and safety factors that protect the driver and passengers/goods. If, for instance, the train were to over-speed in a certain speed restricted area, the control system would automatically issue a warning to the driver and initiate the right braking procedures to bring the train speed below the stated limit or to a safe stop.

This behaviour is implicit because it is merely acting according to the explicit ethical programming given to it by the designers. As a result, an *implicit ethical agent* would also not be a good design goal for our AMA - we want robots to at least consider ethical situations at a given point in time and be able to choose the appropriate way forward.

According to Moor, an *explicit ethical agent* is one that can consider the ethics of a given situation and be able to respond to it in an ethically justifiable manner. This agent can hold an explicit ethical representation of a given situation, as opposed to an *implicit ethical agent*, which holds no ethical representations but relies on pre-programming of the required behaviours.

Explicit ethical agents are autonomous (in a robotics sense) in that they make ethical decisions independently of their designers. Moor seems to imply, however, that they are not autonomous (in a philosophical sense) and instead gives that quality to *full ethical agents*. According to Moor, on top of the capacities of *explicit impact agents*, *full ethical agents* would also have certain epistemic capacities such as consciousness, intentionality and free will. In Moor's taxonomy, full ethical agents

represent the apex of moral agency.

Moor's definition of an explicit ethical agent is aligned to a weak machine ethics view of artificial moral agency. As a result, it is also aligned with the definition of an AMA given in the paragraphs above. For the remainder of this thesis, the definition of an AMA in the paragraphs above will be complemented by the term explicit ethical agent.

2.4.2 What about moral responsibility?

As has been seen in this chapter thus far, traditional views of moral agency have been challenged by the advent of computational systems that have the capability to perform functions that were traditionally reserved for human beings. In Western Philosophy, for instance, moral responsibility is a capacity or power that is generally ascribed to human beings (Noorman, 2018). However, should this view be rethought in light of AMAs?

The definition of artificial moral agency given in the previous sections reads, *"An artificial moral agent is a computationally rational agent whom one appropriately holds responsible for its actions and consequences, and artificial moral agency is the distinct type of agency that agent possesses"*. Clearly, this definition assumes that there is such a thing as responsibility that can be ascribed to artificial moral agents.

The purpose of this section is to clarify what is meant by moral responsibility in this thesis. To do this, we will need to look at traditional views of moral responsibility in normative ethics, unpack why this view can potentially be a hindrance for the AMA project, and finally propose a position for moral responsibility that is suitable for the AMA project.

Yes, this discussion will likely not yield conclusive arguments, but this is a topic of debate that is occurring in Moral Philosophy due to the increase in capability of artificial agents. Noorman (2018) provides a thorough discussion of this ongoing debate.

Traditional views of moral responsibility share a common theme in that they all

agree it is a capacity or power that is ascribed to the responsible agent. However, they have slightly different approaches in the way that they arrive at this conclusion, as masterfully summarised by Fischer (1999). We will very briefly look at the Strawson (1986), Oshana (1997, 2002), and Watson (1996) views of moral responsibility.

Strawson's view of moral responsibility is from a perspective of 'reactive attitudes', which are emotions such as resentment, gratitude, love, anger, forgiveness, and the like (Fischer, 1999; Strawson, 1986). His argument is we can tell which people society ascribes moral responsibility to based on the 'reactive attitudes' that they have towards them. For example, we may feel gratitude or resentment towards someone because they either did or did not perform a specific duty which society expects of a responsible person.

Strawson's argues that we do not hold these 'reactive attitudes' towards non-humans because they are not deserving of them (Fischer, 1999; Strawson, 1986). We may be able to use, enjoy or even manipulate non-human agents (by implication, this would include AMAs). However, we do not harbour reactive feelings towards them because we understand they lack moral responsibility.

Oshana, on the other hand, views moral responsibility from a perspective of accountability (Fischer, 1999; Oshana, 1997, 2002). In her view, a person is morally responsible if they perform some action such that we (society) deem it necessary that the agent explains (accountability) why they performed such an act. That is to say, the intentions that a person had in performing a particular action are vital, and we can call upon this individual to explain these intentions when society deems it necessary (Oshana, 1997).

Watson (1996) takes a slightly different approach to the conceptions above. In Watson's view, moral responsibility has two conceptions. The first conception, which emphasises the agent's moral capacities, has to do with the ability for the agent to make decisions that are their own, and that they are not merely forced by some external force to make such a decision. Watson calls this conception the 'self-disclosure' view (Watson, 1996).

The second conception is a moral accountability view, but closer to Strawson's 'reactive attitudes' than it is to Oshana's accountability view (Fischer, 1999). In this conception, people are morally responsible when they make decisions and actions that affect others. When this happens, the person is morally responsible because of the 'reactive attitudes' of others in the society (Fischer, 1999).

By having these two conceptions of moral responsibility, Watson attempts to explain why we may be oblivious to the actions of a person, because they may only be affecting or harming themself. We could only comment about the person's exercise, or lack thereof, of their moral capacities, and how they are not reaching what may be considered an excellent human life (Watson, 1996). On the other hand, it can also explain why when the actions of this same person start affecting others; we can begin to hold them accountable for their actions through the second and more social conception of moral responsibility.

As Noorman (2018) notes, the various views of moral responsibility more or less have a common theme. The social perspectives of moral responsibility, such as Strawson's, Oshana's and the second component of Watson's view, all impose an expectation for a person or agent to behave or act in a certain way. This expectation implies a kind of assumed capacity in the agent to be morally responsible.

The first component of Watson's view, and to a certain extent, Oshana's, treat moral responsibility as an inbuilt moral capacity that inherently lies within the agent. Whichever view of moral responsibility we take, it is either implicitly assumed or required to be a capacity present in the agent. In either case, it is ascribed to the agent that needs to act in a morally responsible manner.

Viewing moral responsibility as an inherent capacity in a moral agent is an issue for the AMA project, especially if we take a weak machine ethics view as done in this thesis (see sections 2.4 and 2.3.3). However, as Noorman (2018) notes, traditional definitions of moral agency are being challenged due to the rise of artificial agents with increasingly better capabilities. So what could be an alternative view of moral responsibility that we could employ for the AMA project?

Peter Asaro suggests that we look at legal frameworks as opposed to moral responsibility frameworks for AMAs (Asaro, 2006). In Asaro's view, moral responsibility is a difficult subject for which there is likely no robotic equivalent absent of tremendous advances in AI technology. He further argues that humans have no consensus on normative ethical theory either, with only a few generally accepted norms as exceptions.

In contrast, legal frameworks tend to create broad consensus and agreement amongst stakeholders (although even they have their limits). Asaro (2006) argues that creating a legal framework through which we could design and govern the behaviour of AMAs could be a steppingstone until the technology for building them to mimic aspects of moral responsibility could improve.

The challenge with using legal frameworks as a means to building AMAs, however, is that they are no less technically complex to implement in an AMA than the alternatives (Scheutz & Malle, 2017). The understanding of what a rational person is from a legal perspective and programming that into an AMA, will likely require strong machine ethics, which we have already argued is likely not achievable soon.

Allen and Wallach (2012), who take a strong machine ethics view, have argued that super-rational capacities, of which moral responsibility would be one, are likely achievable in future computational architectures. However, even they admit that this task *"can distract from the immediate task of making increasingly autonomous robots safer and more respecting of moral values, given present or near-future technology"* (Allen & Wallach, 2012, p. 113).

Interestingly, Allen and Wallach (2012) suggest using Moor's taxonomy as a starting point for building near-future AMAs (Moor, 2006), which we have already discussed in detail in the previous section. What makes Moor's taxonomy useful as a starting point is it treats moral agency as a continuum[12], which stretches from one end, with agents that have little or no moral agency, to the other, with agents

[12]Asaro's taxonomy also treats moral agency as a continuum (Asaro, 2006). We have, however, decided to focus on Moor's taxonomy, as discussed in the previous section.

that have full moral agency. By viewing moral agency as a continuum, instead of a capacity that agents either have or not, Moor provides a framework for potentially evaluating AMAs.

The ability to evaluate AMAs has emerged as a critical factor in ensuring that we can build AMAs that can meet social expectations of moral behaviour, without necessarily ascribing moral responsibility to them. Allen et al. (2000), for instance, suggest developing a Moral Turing Test, which would evaluate the performance of an AMA, and determine how far off it is from human performance in a similar task. The Moral Turing Test could allow us to lay aside issues of moral responsibility by putting the focus on what AMAs can do.

Floridi and Sanders (2004) take this idea further by suggesting that we need a new way of thinking about moral agency in AMAs. In this respect, Floridi and Sanders (2004) suggest separating discourse in the identification of moral agents from responsibility analysis. Effectively, an AMA would be a moral agent if it meets the requirements for moral agency, as discussed in section 2.3.3.

Once the AMA has been built according to these requirements, we would then evaluate how good the AMA is. The benefit of separating the discourse in the identification and evaluation of moral agency is that it allows the AMA project to proceed so that we can test how they perform once they have been built.

A similar view of this was expressed by Coeckelbergh (2014). Coeckelbergh argues that we need a new view of moral agency, one that is not defined in 'standard terms', but in relations ones. The standard definition of moral agency is problematic, he argues, for two crucial reasons.

Firstly, *"how can we know that a particular entity x really has a particular property p?"* (Coeckelbergh, 2014, p. 63). Secondly, *"how do we know for sure that a particular property p justifies moral status s?"* (Coeckelbergh, 2014, p. 63). The point of these questions is to demonstrate that even what is assumed to make human beings moral agents is not based on conclusive evidence. Rather we have chosen to ascribe moral agency to human beings in this way (Sullins, 2006).

Coeckelbergh (2014) instead argues for a socially and situationally defined view of moral agency, based on the interaction of the robot, and the experiences that others have with it. By implication, if a robot actually meets community expectations of moral behaviour, then in the experience of the community (i.e. the people it actually interacts with), it would be a moral agent.

Finally, the reader is referred to the work of Scheutz (2017), who has given a thorough defence of the case for explicit moral agents as proposed by Moor (2006). Explicit moral agents are AMAs that can behave in a manner that is justifiably ethical, without the super-rational capacities that are often claimed to separate human agents from all other agents in a moral sense.

To conclude this section on moral responsibility, the researcher admits that such a discussion will likely not be conclusive nor answer every possible criticism of artificial moral agency. What is clear from the literature is that traditional views of moral agency tend to treat responsibility as either an implicitly or explicitly required capacity. Contemporary views of moral agency, instead, treat responsibility as an outcome, and not a requirement for moral agency.

This shift has likely occurred for two reasons. Firstly, philosophers cannot conclusively state what properties or capacities make an agent morally responsible. Secondly, even if the required capacities could be pinpointed, it would likely prove to be an impossible task, with current technology in computing, to build these capacities into an AMA, nor to prove that it has them. The outcomes-based view of moral agency reframes responsibility as an outcome that we can actually measure (Allen et al., 2000), though this measurement is likely to be qualitative.

So what does this discussion mean for our AMA? Well, it means that we will need to make a concession that we shall not use the standard definition of moral responsibility, but rather an outcomes view of it. These outcomes need to be aligned with community expectations of moral behaviour because ultimately, the AMA is designed for the community. We must keep in mind the communities of people that will undoubtedly coexist with artificial agents in the near future. Building AMAs

that meet their expectations, as opposed to philosophical ones, will ultimately give us the results that we want.

2.5 Conclusion

The primary purpose of this chapter was to explore the concept of moral agency more pointedly, and in particular for this research, artificial moral agency. To do this, it was necessary to first begin with a definition of an artificial agent in section 2.2. This definition was further broken into weak and strong AI agents in sections 2.2.1 and 2.2.2, respectively. These concepts were a necessary framing for the concepts of weak and strong machine ethics, which were later introduced in section 2.3.3.

As was seen in the discussions in this chapter, how the concept of moral agency is framed is crucial as it determines whether or not artificial agents can be included in the moral universe (Parthemore & Whitby, 2014). In section 2.3.1, we looked at the view that states that agents need to be biological to qualify as moral agents. In section 2.3.2, we looked at the view that states that agents need to have the capacity of consciousness (including sentience and phenomenal awareness) in order to qualify as moral agents.

Finally, in section 2.3.3, we looked at the view that states that computationally rational agents can qualify as moral agents. Perhaps unsurprisingly, artificial moral agency is the one view that more strongly suits the building of artificial agents that have an ethical dimension to their decision-making. The only caveat was whether we mean moral agency in the strong or weak machine ethics sense.

Given these three paradigms of moral agency, i.e. biological, conscious, and artificial moral agency, we ultimately concluded that a focus on the latter is the right choice for this thesis. The capacities required for the former two paradigms of moral agency present technological challenges that are likely not to be surmounted even with current and near-future breakthroughs in technology. This lead to us focusing on a weak framing of artificial moral agency.

Following from the above discussion, we reinterpreted the concept of moral responsibility with a weak machine ethics framing of moral agency in section 2.4.2. Specifically, we defined moral responsibility in relational terms based on the work of Coeckelbergh (2014) and others. We ultimately argued and concluded that if a robot meets community expectations of moral behaviour, then in the experience of the community (i.e. the people it actually interacts with), it would be a moral agent.

In the next chapter (3), a discussion of the concept of artificial moral agency within a framework of computational rationality will be advanced. This discussion will lay the foundation for a proposed computational model for an artificial moral agent conceptualised with weak machine ethics.

Chapter 3

Artificial moral agency within a framework of computational rationality

3.1 Introduction

As the previous chapter demonstrated, many philosophers have argued that computational agents can be considered artificial moral agents (AMA's) if they are built to incorporate the relevant ethical dimensions in their decision making processes (Abney, 2012; Floridi & Sanders, 2004; Johnson, 2006; Moor, 2006; Scheutz & Malle, 2017; Sullins, 2006). The central philosophical idea in this argument is that computational rationality can entail artificial moral agency.

Consequently, the purpose of this chapter is to more pointedly explore the concept of artificial moral agency within a framework of computational rationality. In particular, we will consider a weak machine ethics framing of artificial moral agency, and how it can be related to the established concept in AI of bounded rationality as expressed by Simon (1955) and Horvitz (1987). This discussion will help lay a platform from which to explore computational approaches to the building of a compu-

tationally rational and exemplarist AMA that is conceptualised with weak machine ethics.

In section 3.2, a brief review of prominent models of artificial moral agency will be explored from the relevant literature. Section 3.3 will formally introduce the concept of computational rationality by reviewing the relevant literature. Following this discussion, section 3.4 will demonstrate how the concept of artificial moral agency can be located within a framework of computational rationality.

In section 3.5, a model for a bounded-optimal, computationally rational AMA is proposed. The model is an invitation to both developers (i.e. engineers and scientists) and philosophers to consider how models of computational rationality might be applied in the building of well conceptualised and formulated AMA's. The proposed model also incorporates lessons from prominent models of artificial moral agency, as discussed in section 3.2. The chapter is finally concluded after a brief analysis of the proposed model.

3.2 A review of prominent models of artificial moral agency

The purpose of this section is to briefly review prominent computational models of artificial moral agency that have been suggested in the literature. This review is essential because it will undoubtedly provide a useful starting point for the model that will be developed in this thesis. In the next section, a brief review of computational rationality will be explored, before discussing its applicability in broader detail in the following sections.

In the literature, AMA models can be categorised into those based on human moral cognition (Vanderelst & Winfield, 2018; Wallach et al., 2010), human conscience (White, 2013), minimalist computational models (Muntean & Howard, 2016), logical reasoning (M. Anderson et al., 2006b; Saptawijaya & Pereira, 2014), and util-

ity based models (Cloos, 2005; Pontier et al., 2012; Vamplew et al., 2018).

Wallach et al. (2010) proposed adopting the (Learning Intelligent Distribution Agent) LIDA model and using it for building computational models for artificial moral agency. LIDA is a cognitive architecture for human-like learning, attention and action that was initially proposed by Stan Franklin (Franklin, 2003; Franklin et al., 2014). In their work, Wallach et al. (2010) leveraged the LIDA architecture to demonstrate how AMAs can make moral decisions in various domains and contexts.

The LIDA model combines both bottom-up (learning) and top-down (rules and heuristics) approaches to representing moral knowledge. It uses a functional model of consciousness to model how the agent might deliberate about making decisions. In each deliberation cycle, preceptors acquire information from the task environment, which is then passed to its deliberative system to decide on the action to take.

Since the model is based on human cognition, it also needs to take into account how its cognitive cycle is structured. As such, it needs to know when and for how long it should deliberate on information stored in its 'current' conscious state. It also needs to know when a deliberative cycle should end so that it can decide on the action to take. LIDA also allows for the possibility of imagination, i.e. simulation and planning based on different scenarios, as well as monitoring and learning from its past actions.

The LIDA architecture models super-rational capacities, such as consciousness and emotions, and would, therefore, qualify as a strong machine ethics model for artificial moral agency. Wallach et al. (2010), and indeed the original designer of LIDA (Franklin et al., 2014), do not claim that it can completely model and replicate human (moral) cognition; they are merely putting forward a practical solution to a hard problem.

Another well conceptualised and articulated model of artificial moral agency is by White (2013), in his work on the ACTWith model. White (2013, p. 33) explains ACTWith, which stands for 'As-if Coming-to-Terms-With', by stating that it *"is a situated, embodied and embedded information processing framework inspired equally*

by hybrid neural network models and complex/dynamic systems, tempered by observed successes and failures in related treatments from human psychology, neurology and traditional moral philosophy".

The 'As-if' component of the model involves feeling a situation out as if the agent was in another's shoes. The 'Coming-to-Terms-With' component means the agent has empathised with another's suffering, and it can define the situation using its own experiences, and come to terms with it.

Both of these components can have open or closed states. In the 'As-if' component, being open means the agent is willing to empathise with the suffering of others, and being closed means that it is not. Similarly, if the 'Coming-to-Terms-With' component is open, it means the agent is willing to fully understand and even embody the situation of others, while closed means that it is not.

The model by White (2013) is based on the traditional association of conscience with the centre of the heart. The conscience's development either enables or disables compassion and love towards other agents. In this way, conscience is the seat of morality. The ACTWith model is comprised of four static states of conscience. These states are affectively open, affectively and explicitly open, explicitly open, and closed.

When the agent is in the affectively open state, it is empathetic to the plight of others, thus allowing it to place itself in their 'shoes'. When the agent is affectively and explicitly open, it is both empathetic and coming to terms with the plight of others. When the agent is explicitly open, it is coming to terms with the plights of others. Finally, when it is in the closed state, the agent can perform an action or reflection based on prior terms.

Since ACTWith is an information processing model that is dynamic, the agent does not stay in any one of the four states. However, it can dynamically change in response to information received from the task environment. White (2013) describes what he calls the beating heart of conscience in the ACTWith model, which is essentially the cycle that the agent goes through in switching between the four states

dynamically.

Furthermore, since the model is dynamic, there is no predefined cycle that the agent follows. Depending on the "personality" of the agent, it can have a range of responses, ranging from being completely closed and disgusted by the suffering of others, and moving on, all the way to being saint-like, and completely open affectively and explicitly. Its past experiences will determine the exact response of the agent.

Perhaps unlike the above two models, Muntean and Howard (2016) propose a minimalistic model of an AMA. In their work, they defined a model for an AMA that is parsimonious in its ontology and minimal in its ethical assumptions. By parsimonious, they mean that the AMA is conceptualised as minimally as possible, without necessarily setting out to model human cognition (Wallach et al., 2010) or human moral conscience (White, 2013). Its ethical minimalism stems from this ontology, and therefore the agent is a kind of "optimising predictive mind" (Muntean & Howard, 2016).

The primary assumption in Muntean and Howard's model is that *"regularities in moral data, as a form of patterns, can play the role of moral norms. But they are discovered, rather than postulated"* (Muntean & Howard, 2016, p. 220). In this way, their architecture is set up to allow the discovery of moral patterns and norms through observation of human moral behaviour. This discovery could either be in real-time or offline.

As can be expected with a minimalistic approach like this, their architecture emphasises bottom-up machine learning approaches to learn from moral cues. In particular, they proposed using a variation of the NeuroEvolution of Augmenting Topologies (NEAT) architecture. Essentially, this architecture allows them to select a population of neural networks that are trained against the relevant moral data.

The population of neural networks is then put through a series of criteria and tests to select the best-performing ones, while combining some of the characteristics (e.g. parameters and topologies) of the best-performing ones, to create even better performance. This process creates a new population of neural networks, and

the process is repeated until sufficient converges towards human moral behaviour is achieved.

According to Muntean and Howard (2016), using this kind of evolutionary architecture has the advantage of introducing creativity and innovation in the way that AMAs can learn moral behaviour. This innovation and creativity cannot be replicated easily in "rule-based, generalist or action-centred models" (Muntean & Howard, 2016).

Another approach to modelling moral reasoning is through the use of logics, in particular, logic programming (LP) (M. Anderson et al., 2006b; Saptawijaya & Pereira, 2014). As Saptawijaya and Pereira (2014) note, LP offers a declarative model for building moral logic and reasoning into AMAs.

A benefit of using LP is it allows for the direct declaration of the moral rules that the AMA ought to follow, as opposed to models such as the one proposed by Muntean and Howard (2016), which do not allow for an explicit declaration of formal rules. Another benefit of using LP is that they can be extended to allow for non-monotonic reasoning[1], which can allow for defeasible semantics in moral reasoning (Britz, Meyer, & Varzinczak, 2011).

The main focus of LP-based architectures used in building AMAs is generally not complicated. Their primary focus is usually on the knowledge representation of the ethical framework that the AMA ought to follow, and the ethical reasoning blocks that are used to predict the correct action to carry out in a given task environment.

In Saptawijaya and Pereira (2014), the architecture that they provide contains both of the elements discussed above. Their chosen LP technique is based on abductive reasoning, which allows them to find the simplest explanation for a set of moral observations. Their approach adds the ability to modify the knowledge base by ei-

[1]Non-monotonic reasoning allows the representation and derivation of plausible but not infallible inferences in a knowledge base. Put more formally; it is a formal logic whose consequence relation is non-monotonic (Reiter, 1988). It allows for the drawing of tentative conclusions, and the revision of these conclusions as new information is received in the world.

ther the AMA or external agents. It also allows for both reactive and deliberative reasoning, with the former used for fast responses in scenarios that are time-sensitive.

The architecture used by M. Anderson et al. (2006b) also contains both the knowledge representation of their chosen ethical framework, as well as an ethical reasoning block. The only difference is M. Anderson et al. (2006b) use inductive logic programming (ILP), which allows for a machine learning approach to logical reasoning. In their architecture, the AMA stores both human solved ethical scenarios and the decision principles used (i.e. the ethical framework), and it uses these two components to advise any user on a relevant ethical problem.

A final approach to modelling moral reasoning that will be discussed here is through the use of utilities (Cloos, 2005; Pontier et al., 2012; Vamplew et al., 2018). The utilities are not used to model moral reasoning directly. Instead, they are used as a way to weigh-up various methods and options in order to create a more human aligned and ethical outcome (Vamplew et al., 2018).

Take, for example, the work of Pontier et al. (2012) on *Moral Coppélia*, a project to combine rule-based moral reasoning with an affective component, through the use of moral ratios. The moral reasoning component could be any rule-based technique, such as those based on LP, as discussed in the paragraphs above. The goal of the affective component is to model human emotional intelligence, which Pontier et al. (2012) argue could be essential for modelling human moral reasoning.

These two components above are combined through the use of a maximum expected utility (MEU) calculation. The action that maximises the expected utility of the two will be the one that is chosen. For example, even though the moral reasoning system may rank an action such as 'put down a pet that is sick', the affective component may actually rank that action lower.

This scenario may happen because the affective component is taking into account the emotional attachment that others may have to the pet. As a result, another action that balances the two components, such as 'take the pet to the vet again' may be ranked higher.

Vamplew et al. (2018) further extend this concept to cover not just two, but multiple components that can be modelled as ratios to provide one aligned outcome. For example, a system could be designed that combines the learning-based approach of Muntean and Howard (2016), the rule-based approach of Saptawijaya and Pereira (2014), and the *Moral Coppélia* of Pontier et al. (2012) to create an AMA that would select actions that likely meet multiple objectives.

The fundamental insight that Vamplew et al. (2018) point out is that moral reasoning is situational, and some techniques will, therefore, be more appropriate for specific situations as opposed to others. The use of a deliberative and a reactive moral reasoning system in Saptawijaya and Pereira (2014) is an example of such, as one is used to respond to time-sensitive situations, and the other allows for more deliberation to get to a better outcome. As a result, AMAs may need more than one technique in their programming, and they should have a way to dynamically select the relevant ones, or combinations thereof, to meet the requirements of a particular situation.

To conclude this brief review, the following lessons can be drawn from the discussion above. The AMA models based on human moral cognition and conscience would likely fall under the category of strong machine ethics. Both are theoretical approaches for which there are yet no comprehensive results that would indicate that they have achieved their goal. Their models of consciousness and conscience would fall outside of the scope of a weak machine ethics AMA.

Nevertheless, useful lessons can be drawn based on their approach to the deliberative cycle of reasoning that the AMA ought to have. This cycle refers to the time window that the AMA should use to consider information from the task environment in order to make an ethical decision. Any model of an AMA (whether weak or strong) would likely require either an ideal or a dynamic cycle to suit the requirements of various situations.

Another lesson is the separation of higher and lower order reasoning, thereby allowing for a more complex decision-making framework. Having this separation of

higher and lower order reasoning need not mean that we are attempting to replicate human moral cognition (i.e. strong machine ethics). Instead, this technique can be used successfully to introduce a level of complex decision making in computational systems (Horvitz, 1987; Simon, 1955). This point will be picked up in the discussion of computational rationality in the following section (3.3).

The minimalist, logical and utility-based models can all fit the weak machine ethics framing that this thesis takes. The minimalist model, as the name suggests, takes a parsimonious approach to building an AMA using a bottom-up learning-based approach. It provides an example of the use of machine learning in the building of weak machine ethics AMAs.

Another lesson is that both declarative rule-based and learning-based techniques may be combined in the design of a weak machine ethics AMA (Vamplew et al., 2018). We are not bound to using only one technique; multiple may be employed in order to meet the ethical requirements of various situations. Finally, we may ask ourselves, which one of these models would be appropriate for the weak machine ethics AMA in this project? Or more likely, which ones?

While a detailed discussion of the choice of an ethical framework is left for chapter 4, it is worth mentioning that this choice often forces a particular model of an AMA. For example, consequentialist and deontological frameworks tend to force models that rely on logical and declarative rule-based approaches. Virtue ethics frameworks, on the other hand, tend to find a better fit with bottom-up learning techniques[2].

For this reason, we will focus on a generic model for an AMA, one that is independent of any particular ethical framework. Such a model should allow for any number of techniques (whether one or many) to be employed in building weak machine ethics AMAs, whilst incorporating some of the lessons discussed above.

Since this thesis is interdisciplinary, it is here where we can draw on lessons from other literature, particularly in computer science, to help us to formulate a generic model of an AMA (Russell & Norvig, 2009). This generic model, which is proposed

[2]See section 4.2 for detailed examples.

and discussed in detail in section 3.5, will provide a foundation from which to meet our goal, which is to demonstrate the practical feasibility of building an exemplarist AMA with weak machine ethics.

In the next few sections, computational rationality is put forward as a framework that aligns to the lessons drawn not only above but also effectively combines both the philosophical and computational elements of artificial moral agency. After this discussion, we will again return to the subject of a generic model for a weak machine ethics AMA.

3.3 Computational rationality

Computational rationality is perhaps best described as approximating decision making for maximum expected utility while using the optimal computational resources (Genewein et al., 2015; Lewis et al., 2014). It is about making rational decisions within a computational framework. As Gershman, Horvitz, and Tenenbaum (2015) note, computational rationality is a convergence of ideas from AI, cognitive science and neuroscience around intelligence, and in particular, its computational nature. They go into extensive lengths in their work to show how ideas of computation from AI have inspired researchers in the cognitive and neurosciences, and vice versa.

To get a proper grasp of computational rationality, however, we need to look back a few decades to the works of Horvitz (1987), Simon (1955), and others. Many of the ideas in computational rationality stem from the tradition of Herbert Simon, who was an economist and political scientist. While Turing (1950) and others were postulating about the nature of machine intelligence, Simon brought much-needed constraints on the kind of rationality that could be achieved by computationally bounded agents. He started looking at candidate definitions for bounded rationality when he was deriving a model for rational choice (Selten, 1990; Simon, 1955, 1972).

He argued that agents do not always have all the information they require to make a decision and that their internal computation was limited in how they could

use the available data to make rational decisions. Bounded rationality was, therefore, a way for him to '*formulate the process of rational choice in situations where we wish to take explicit account of the "internal" as well as the "external" constraints that define the problem of optimisation for the organism"* (Simon, 1955, p. 2).

These ideas inspired many works in AI, a field which also found itself dealing with creating intelligent agents that operate with much of the constraints that Herbert Simon saw in general organisms. Most notably, Horvitz (1987, 1988), and others at the then Department of Medical Computer Science at Stanford took the ideas forward (Horvitz, Cooper, & Heckerman, 1989). Horvitz argued that probability and utility theories[3], both of which were generally considered standard decision-making mechanisms in computer science were insufficient for the real-world problems that machine intelligence systems were trying to solve.

Real-world problems often go beyond the standard axiomatic basis defined by utility and probability theories, as they are often characterised by uncertain and limited information (thus making the process of modelling and knowledge representation difficult). Furthermore, machine intelligence systems have limited computational resources, which made the application of classical decision-theoretic approaches to many real-world problems difficult, and many times, intractable (Horvitz, 1987).

To deal with these problems, Horvitz suggested looking at various optimisation and heuristic strategies to resolve some of the challenges in real-world decision making. Notably, he proposed the notions of *flexible inference* and *decision-theoretic control*. Various inference techniques have been developed over the years that allow partial inference with limited information or partial execution. These advances also paved the way for the concept of meta-reasoning (in a computational sense), which is just a program that is aware of various inference strategies and can select the best strategy based on the type of problem that needs to be solved (Horvitz, 1989). These types of inference strategies present a natural fit for the optimisation and heuristic

[3]Chapter 13 of Russell and Norvig (2009) gives a great introduction to decision theory in computer science.

framework of Horvitz.

Decision-theocratic control represents the ability for the agent to determine how best to execute a specific inference strategy based on a trade-off between computation time, precision, maximum expected utility (MEU) and cost of delaying the action. Balancing these trade-offs, along with suitable or multiple inference strategies, represents the core idea in the approach of Horvitz.

The ideas of Horvitz and Simon have persisted well over time, with many AI researchers adopting them (Genewein et al., 2015; Gershman et al., 2015; Lewis et al., 2014; Marwala, 2013; Russell & Subramanian, 1995; Zilberstein, 2013). Russell and Subramanian (1995) use these ideas to develop what they call provably *bounded-optimal agents*. Bounded-optimal agents are machine intelligence systems whose solutions to problems are optimal for the information that they can acquire from the task environment and the limitations of their programs and architectures.

In other words, optimality is what the agent can achieve, given its internal and external constraints, and not necessarily what a perfectly rational agent would do for a given task. This conception of a bounded-optimal agent formed the foundation for what is now referred to as computationally rational agents in recent literature (Gershman et al., 2015; Lewis et al., 2014).

Quite suitably, the work of Gershman et al. (2015), with Horvitz as one of the co-authors, likely represents one of the clearest pictures of what computational rationality is, and what it can be. As the authors note, computational rationality has the potential to be a *"unifying framework for the study of intelligence in minds, brains, and machines"* (Gershman et al., 2015, p. 278).

The researcher supports this claim and further posit that computational rationality can be a unifying framework not only for ideas in the sciences, but also in Philosophy, and more specifically, in Machine Ethics. After all, it was Aristotle who first placed a strict emphasis on practical rationality[4] as a basis for virtuous and ethical action (F. D. Miller, 1984). This chapter aims to clarify precisely how com-

[4]Can be found in the *Nicomachean Ethics Book VI*.

putational rationality can be an integrative framework for machine ethics by showing how the ideas of Gershman et al. (2015), Horvitz (1987), Russell and Subramanian (1995), and others, can be applied to the question of building artificial moral agents.

The next section will be a discussion about the epistemic capacities required for moral agency and considering whether these capacities can be replicated or approximated within a framework of computational rationality using a combination of formal and heuristic approaches and optimisations.

3.4 Artificial moral agency within a framework of computational rationality

The purpose of this section is to demonstrate how the concept of artificial moral agency, particularly through a weak machine ethics lens as defined in chapter 2, is compatible with a framework of computational rationality. Two arguments support this claim. Firstly, the requirements necessary for (artificial) moral agency lend themselves naturally to being computable.

Secondly, the problems that bounded computational rationality had initially been designed to solve, such as limited computational and informational resources (Gershman et al., 2015; Horvitz, 1987; Russell & Subramanian, 1995), are also present in computational morality. Therefore, many of the techniques designed for use in general computationally rational agents are likely also applicable to the problem of computational morality.

So far, we have avoided discussing which exact capacities are required for moral agency. In the literature, these capacities can include emotions, empathy, free will, rationality, cognition (including mental and intentional states), concepts, awareness, amongst others (Himma, 2009; Parthemore & Whitby, 2013, 2014; Torrance, 2008; Wallach et al., 2011).

One way to get around this issue is to focus on the outcomes instead. In other

words, instead of arguing about which capacities (and combinations thereof) will result in some facet of moral agency, focus on the outcome that is expected to be achieved. This is precisely what philosophers such as Sullins (2006) and Floridi and Sanders (2004) do by focusing on the high-level requirements for artificial moral agency and abstracting away the detail regarding the exact capacities required.

For this discussion, we will focus on the requirements expressed by Floridi and Sanders because they conceptualise artificial moral agency within a weak machine ethics framework, as opposed to Sullins, who conceptualises it within a strong machine ethics framework. As discussed in the previous chapter (section 2.4), the focus of this thesis will be on weak machine ethics AMAs.

Floridi and Sanders (2004) define the requirements for artificial moral agency as *interactivity*, *adaptability*, and *autonomy*. They define interactivity as the ability for the AMA to be aware and responsive to environmental stimuli. Adaptability is defined as the ability for the AMA to change internal states according to environmental stimuli. Autonomy is defined as the ability to change internal states according to the AMA's own transition rules independently of environmental stimuli.

Focusing on the high-level requirements for moral agency and abstracting away details around required capacities is essentially a focus on 'mindless' morality - a form of morality that distinctly suits a *computational* framing of moral agency. It does not care how autonomy or intentionality, for instance, are achieved - it only cares that their outcomes are achieved.

This kind of mindless morality is what Floridi and Sanders (2004) are alluding to when they talk about moral agency at different levels of abstraction (LoA). At a low enough LoA, a human being would also not be considered a moral agent since we would be dealing with their biological make-up, the neurobiological processes in their brains, and other cognitive processes which would seem indistinguishable from a machine.

Similarly, artificial agents observed at a low enough LoA are merely electronic components and code, and at that level, it would be impossible to ascribe moral

agency. However, at a high enough LoA, these low-level processes and components are abstracted such that only the outcomes of their decisions are visible.

Ordinary observers would not be aware of how exactly the AMA functions, only that it seems to have some goals and intentions. The AMA would seem to function autonomously and learn new things over time. At that LoA, the ordinary observers would only witness that the robot acts in a manner that is consistent with expectations of moral behaviour.

At this point, the reader may be asking themselves whether this kind of formulation of moral agency is taking it away from its traditional definition in properties and capacities that the agent has, towards a new framing that is based on the relations and experiences that people have with the AMA. This observation is correct. However, as has already been discussed in detail in section 2.4.2, this kind of framing of moral agency is essential if we are to include artificial moral agents in the moral universe (Coeckelbergh, 2014).

Floridi and Sanders' approach to defining the requirements for artificial moral agency is not without its critiques, the strongest of which likely comes from Himma (2009). He argues that, under Floridi and Sanders' formulation, rattlesnakes, for example, could be wrongly considered to be artificial moral agents.

If, as Himma's example goes, the rattlesnake acts as a response to hunger and kills something, then it would have acted autonomously, interactively, and apparently with some ability to learn. Since we know that rattlesnakes are not moral agents, then it must also mean that an AMA that lacks consciousness[5] is also not a moral agent.

The crux of Himma's argument seems to be that only praise or blame-worthy agents could be moral agents. His criticism, therefore, is centred around the lack of moral responsibility that weak machine ethics AMAs and rattlesnakes both share.

Once again, we have to begin with a reminder that the view of moral responsibility that we are taking in this thesis is grounded in outcomes, and not in capacities

[5]As discussed in chapter 2.3.2, Himma is an advocate for conscious moral agency.

or properties ascribed to the agent. If we take the view of moral responsibility discussed at length in section 2.4.2, then there are two issues with Himma's argument, especially as it pertains to artificial moral agency.

Firstly, Himma's argument presupposes that discourse around moral agency is equivalent to responsibility analysis and that no room exists for prescriptive discourse in the identification of moral agents (Floridi & Sanders, 2004). Secondly, and to use his example, the rattlesnake would not qualify as a moral agent, according to Floridi and Sanders' requirements, because it cannot learn moral values. It is only responding to instinct.

An artificial agent, on the other hand, can be programmed to simulate the capacity to learn (morally), and thus could qualify as an AMA. How good an AMA it will be (i.e. responsibility analysis) is a different matter altogether[6], and will require us to build models of computational morality and to evaluate them (such a model will be proposed in section 3.5).

To be clear, without consciousness or intentional/unconscious mental states, the AMA could not be a full moral agent, but that is why we put the qualifier 'artificial' in front of 'moral agent'. In theory, the AMA's moral performance will be better than a rattlesnake, but lesser than that of a full moral agent such as a human being (Moor, 2006).

The last criticism of the requirements for artificial moral agency may be that they are overly focused on Floridi and Sanders' views. This criticism is relatively straightforward to address. Floridi and Sanders (2004) are quite consistent with the outcomes-based view of moral agency and responsibility that we discussed in detail in section 2.4.2. They align well with the views of artificial moral agency expressed by Sullins (2006), Moor (2006), Scheutz and Malle (2017) and Coeckelbergh (2014).

To illustrate, Moor (2006) defines an explicit ethical agent as an agent that can independently consider an ethical situation and act in an ethically justifiable manner. Moor's view of an explicit ethical agent presupposes interactivity (because it needs

[6]The reader is referred to section 2.4.2 for further discussion of this topic.

to understand its task environment). It also presupposes autonomy (because it needs to behave ethically independently of human help) and adaptivity (because it may receive feedback about poor performance from which it needs to learn).

A similar alignment exists between Floridi and Sanders (2004) and Sullins (2006) too. Their requirements for artificial moral agency are almost identical. The only difference is that Sullins seems to take a strong machine ethics view, which Floridi and Sanders seem to avoid.

Coeckelbergh (2014) also takes an outcomes-based view of moral agency, based on the experiences of the community with which the AMA interacts. Once again, his view is aligned with Florid and Sanders', especially their conception of moral agency at different LoAs. At a high enough LoA, both the outcomes that Floridi and Sanders (2004) and Coeckelbergh (2014) envisage for AMAs are equivalent.

The bottom line is, the proponents for weak machine ethics AMAs all seem to broadly agree that an outcomes-based view of morality is better for AMAs. The outcomes-based view allows us to measure the performance of AMAs, and to tweak it when it is misaligned with community expectations of moral behaviour. All that Floridi and Sanders (2004) have done is to write out the requirements for an outcomes-based conception of artificial moral agency in basic and distilled statements from which researchers in machine ethics can build.

The last few paragraphs advanced the argument that the requirements for moral agency, as expressed by Floridi and Sanders (2004), lend themselves to being computable. However, that is not the only reason that artificial moral agency is compatible with a framework of computational rationality.

Computational rationality exists as a framework primarily because artificial agents are not perfectly rational. They face many internal and external constraints such as limited computational resources, limited information about the problem at hand, limited time (and space) within which to make a decision, the tractability of the problem itself, and other constraints. As it turns out, AMA's are faced with much of the same constraints and limitations as computationally rational agents.

AMA's have to make moral decisions despite the limitation of computational resources, information, time, and the tractability of the moral decision itself. The researcher posits that the problem of computational morality is merely a special case, albeit a complex one, of computational rationality, and that many of the approaches to solving computational rationality in the general case, can be used to enhance the prospects for computational morality further.

For example, the emergence of hybrid approaches, i.e. model-based (top-down) and model-free (bottom-up), is a prominent feature in both domains [7]. In computational rationality, hybrid approaches have emerged as a superior choice for certain complex tasks for computational rationality (Gershman et al., 2015).

Similarly, prominent researchers in machine ethics believe that a combination of top-down and bottom-up approaches will likely be required to solve certain kinds of complex moral decisions (Allen, Smit, & Wallach, 2005). These similarities lend further credence to the idea that the two domains are more related than different.

Just as Russell and Subramanian (1995) popularised the concept of a bounded-optimal agent, perhaps it is time to start talking about bounded-optimal artificial moral agents. These are AMA's that arrive at moral decisions based on the information they can acquire from the environment, given the limitations of their software architectures and programming. In the next section, these ideas are formalised into a conceptual model for a computationally rational AMA.

[7] The use of the terms top-down and bottom-up in both the cited philosophical and scientific disciplines is conceptually the same. In both cases, top-down means starting from a pre-defined ethical framework or a declarative model and bottom-up means learning an ethical representation or a computational model from the available data.

3.5 A model for an optimally-bounded, computationally rational AMA

The proposed model for a computationally rational AMA is based on ideas in computational rationality. The model is openly based on Russell and Norvig (2009, p. 55) (see Figure 3.1), whose conception of a general learning agent is simple, and yet comprehensive. The ideas in computational rationality can be integrated into the model of any general artificial and intelligent agent, so long as its key tenants, such as bounded-optimality, the separation of meta-reasoning from specific algorithms for reasoning, and the use of formal and heuristic methods, are preserved.

As discussed in section 3.2, we draw on inspiration from a prominent and well-established model for a general learning agent, of which Russell and Norvig's is likely the most established, accessible and widely used in AI literature. It is because of this pragmatic reason that we choose this model. This research is meant to be an invitation to both philosophers and engineers to consider practical models of artificial moral agency. Basing the model on Russell and Norvig (2009) further emphasises this invitation and opens it up to a broader audience and readership.

This decision must be considered in light of the broader goal of this thesis, which is to demonstrate the theoretical and practical feasibility of building exemplarist AMAs conceptualised with weak machine ethics. In other words, basing our model on Russell and Norvig's established version is not only practical; it also further emphasises this goal.

Figure 3.1 depicts the structure of a general learning agent which can perform certain actions in an environment, through its sensors and actuators, according to a set performance standard (perhaps set by a human being). The general learning agent also can improve its decision making and performance capability over time, and generate new problems (goals) that can help it to improve performance further and learn new ways to reason.

Figure 3.2 is presented as a proposed model for a high-level conceptual model for

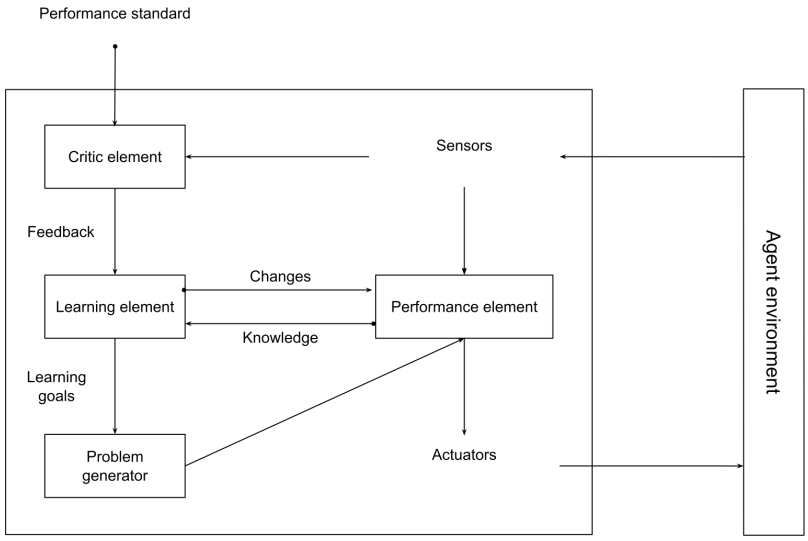

Figure 3.1: A generic representation of a general learning agent (Russell & Norvig, 2009)

a computationally rational AMA. The agent gathers *bounded information* from the environment, and processes it in the *ethical performance element*, which is responsible for ethical as well as general reasoning. The decisions and actions from this element are then transferred back to the environment (via the relevant actuators and communication mechanisms).

The *learning element* and *problem generator* are left as-is from Russell and Norvig's conception. They are responsible for updating the performance element with new ways to reason and generate new ideas for future performance, respectively. The *critic element* is similar, except instead of only allowing for external input to modify the performance of the agent (e.g. human input), it also allows the agent to provide a human-understandable rationale for its performance.

Figure 3.3 zooms in on the ethical performance element. The ethical meta-reasoner is responsible for the planning and execution strategy. From a planning perspective, it selects an appropriate deliberative cycle time based on its bounded resources and MEU. From an execution strategy perspective, it is responsible for deciding on the best ethical framework (or combinations thereof) and one or more

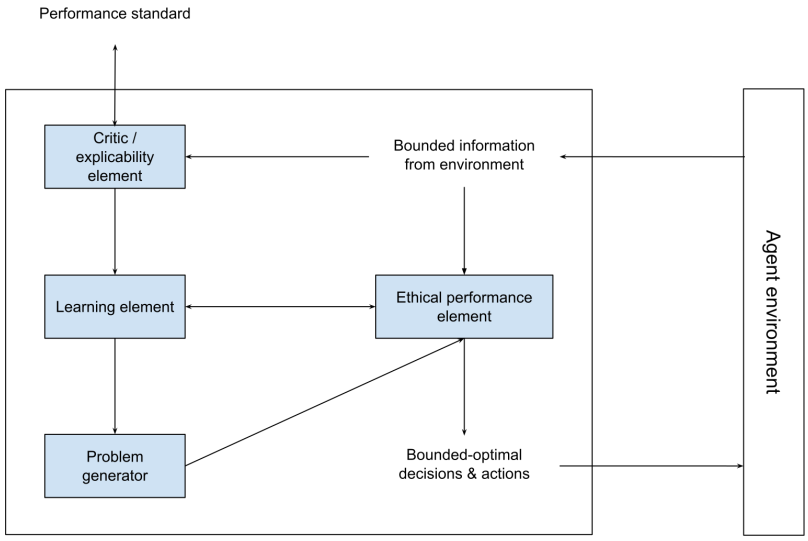

Figure 3.2: A conceptual model for an optimally-bounded, computationally rational AMA

programs to execute to arrive at an optimally-bounded ethical decision. The ethical performance element thus separates high-level meta reasoning activities from the execution.

At a high-level, the proposed model of an AMA would meet the requirements for artificial moral agency as described in the previous section. It can receive information from the environment and act on it (interactivity). It can also change its performance state through the ethical performance and learning elements (adaptability). Finally, it can behave in a somewhat autonomous manner through the problem generator, which generates new ideas about how to execute performance in the future (autonomy). Additionally, it can receive a new performance standard and explain its current performance to a human being.

The model also incorporates vital lessons from other prominent models of artificial moral agency (see section 3.2). The model allows the AMA to contain more than one technique or algorithm for moral decision making. Furthermore, top-down (e.g. logic programming), bottom-up (e.g. machine learning) and hybrid techniques are all allowed. The model also contains an element of higher-order thinking through

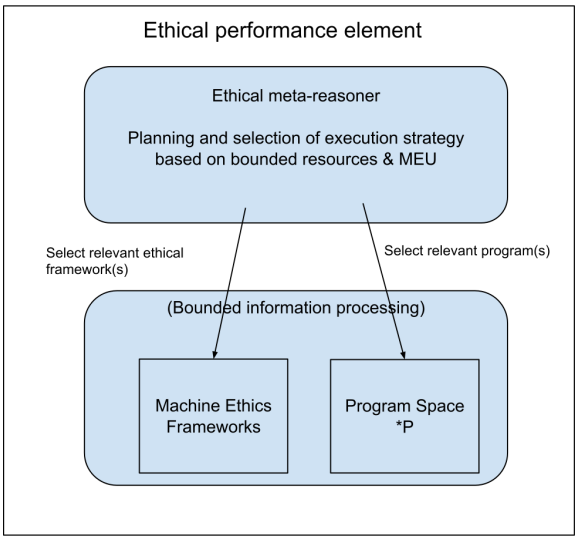

Figure 3.3: A detailed view of the ethical performance element

the use of the ethical meta-reasoner, without taking it to the extreme of modelling human moral cognition as is the case in the work of Wallach et al. (2010).

The model is based on the premise of weak machine ethics AMAs. It is not envisaged to contain elements of strong machine ethics, such as those proposed by Wallach et al. (2010) and White (2013). As a result, it does not attempt to model any super-rational capacities that may be deemed necessary for full moral agency.

While it is challenging to ascertain what kinds of moral decisions would be possible with this model, we can speculate that moral decision making in situations where (bounded) information is readily available and accessible to the AMA should be theoretically possible. Such contexts could include highly domain-specific environments, such as in self-driving cars, healthcare robots, loan approval bots, home-assistants, and the like [8].

Conversely, we can expect that the AMA would have difficulty in making moral decisions that require highly abstract reasoning with little to no bounded information available. For example, most people would agree that lying is wrong in most cases,

[8] In chapter 6, the model is applied in the development of an ethical robotic teacher in a classroom context.

but how would the AMA even know what lying is? This example demonstrates a scenario where having consciousness, conscience and intentional mental states would be beneficial to the AMA. These capacities are impossible, by definition, in this model.

3.6 Conclusion

The purpose of this chapter was to advance a model for an artificial moral agency based on a framework of computational rationality. This chapter demonstrated how computational rationality could be an integrative element that can effectively combine both the philosophical and computational aspects of artificial moral agency consistently and logically.

The chapter argued that the capacities required for an artificial moral agency are computable. Furthermore, it was argued that computational morality is merely a unique case of computational rationality; hence many techniques initially developed for general rationality can be adapted for computational morality. This line of reasoning led to the proposal of a conceptual model for a bounded-optimal computationally rational AMA.

Some philosophers and scientists might reject the idea of a bounded-optimal artificial moral agent. After all, the stakes can be quite high when it comes to moral decision making, as the wrong decision could have significant moral and societal implications. However, we need to start somewhere.

The researcher suggests that starting from a weak machine ethics perspective is helpful to allow us to begin to test its limits and the sorts of domains and contexts where it can be applied. The model proposed is an invitation for dialogue and feedback, and the hope is that many philosopher-developer pairs can be formed to solve the problem of constraining weak AI systems and making them more respecting of human moral values.

This chapter did not explicitly mention the ethical framework that the AMA

should follow, as the primary purpose was to locate artificial moral agency within a framework of computational rationality. However, as figure 3.3 depicts, the model allows the AMA to contain any ethical framework. It is not prescriptive about which ethical framework should be employed. In the next chapter (4), the exact ethical framework that will be used in this thesis will be discussed in detail.

Chapter 4

Exemplarism: A suitable ethical framework for the AMA project

4.1 Introduction

The purpose of the previous chapter was to discuss a computational framing of artificial moral agency. In that chapter, artificial moral agency was grounded in computational rationality (Gershman et al., 2015; Lewis et al., 2014) - a framework that could potentially unify the study of intelligence in minds and machines. This framing of artificial moral agency within a framework of computational rationality lead to the proposed model for an optimally-bounded, computationally rational AMA in section 3.5.

The model that is proposed is generic and can be applied to the building of weak machine ethics AMAs that follow any ethical framework. However, in this chapter, we will more closely explore what it means to build an exemplarist AMA. In particular, this chapter will briefly cover background literature on how others have applied the normative ethical theories as machine ethics frameworks. As useful as a strict philosophical account of these theories may be as fields of inquiry in ethics, in this interdisciplinary thesis we are far more interested in how they may be applied in the building of weak machine ethics AMA.

The designers of AMA's need to consider a couple of aspects as they choose an appropriate machine ethics framework. Firstly, they need to decide whether to implement a top-down, bottom-up, or hybrid machine ethics framework (Allen et al., 2005). In top-down approaches, the machine ethics framework acts as a system of moral values, expressed as rules, principles or practical guides that the AMA should follow in the performance of its ethical duties. Bottom-up approaches, however, do not use a formalised system of rules, principles or practical guides. Instead, the AMA is designed to learn its own internal representation of moral values based on either observation of other moral actors, expert knowledge or simulation techniques.

Hybrid approaches, as the name suggests, combine both top-down and bottom-up methods to create a new one that tries to leverage the strengths of both. There are cases where a top-down approach might make more sense, e.g. in an environment where all the rules are explicitly defined and are computable. Still, other cases demand bottom-up approaches, e.g. where rules are not explicitly defined, and the agent needs to establish what the ethical norms are in order to abide by them. If one in fact needs to build an AMA that meets the requirements of both, then it may be beneficial to implement a hybrid strategy.

The second aspect that designers of AMAs need to consider is the ethical framework that they ought to use. They need to decide on a machine ethics framework based on either a consequentialist, deontological or virtue ethics approach. Adopting consequentialist or deontological approaches is generally classified as a top-down approach to machine ethics. They are often grouped with top-down approaches because they generally involve the setting of top-down rules as guiding principles for ethical decision making (Gips, 1995).

Adopting a virtue ethics approach, however, is generally classified as a bottom-up[1]. Often, virtue ethics is grouped with bottom-up approaches because it involves a kind of moral learning that results in the agent forming its internal view of ethics,

[1]In section 4.3.1, the grouping of virtue-ethics as a bottom-up approach is explored and defended in detail.

and making rational decisions based on this internally formed system of moral values (Abney, 2012; Gips, 1995).

The rest of this chapter is structured as follows. Section 4.2 will briefly look at the three main ethics frameworks that have been used in the AMA project. This discussion will also point out a few examples of authors that have deliberately set out to build AMAs that very clearly follow a specific framework. Section 4.3 will briefly compare the suitability of the main ethical frameworks to the AMA project, before advancing an argument that an approach based on virtue ethics, and in particular exemplarism, is the standout choice of machine ethics framework for this AMA project.

4.2 Machine ethics frameworks and implementations: A brief look at the major approaches

As the field of machine ethics was starting to take shape as an interdisciplinary field of AI and philosophy in the early ninety's, scholars such as Gips (1995) were thinking about how future designers would approach the challenge of building ethical robots. In his mind, he did not doubt that the three main approaches to normative ethics, namely consequentialism, deontology and virtue ethics could be applied to the AMA project. Today, virtually all approaches to machine ethics frameworks are based on one, or a combination of the main approaches to normative ethics.

The next few sections will highlight how the main normative ethical theories have been applied to the AMA project. The discussion is essential for two reasons. Firstly, an understanding of the theories will help designers of AMAs in selecting a machine ethics framework for their project(s). Secondly, it will bring context to the discussion on exemplarism, especially since it is a derivative of virtue ethics - one of the three main normative ethical theories.

4.2.1 Consequentialist approaches to machine ethics

Consequentialist approaches to machine ethics have to do with linking the morality of an action to its outcome (Dignum, 2017; Gips, 1995; Kuipers, 2016; Scheutz & Malle, 2017). The central idea in this approach to machine ethics is that the consequences or results judge the actions that the agent performed. This guiding principle inadvertently leads to a means-end approach, where consequences of actions that result in more good are judged favourably over those that would minimise it.

The outcomes that maximise the good (often considered happiness) are considered ethical in a consequentialist framework, and those that minimise the good are considered unethical. Therefore, for an AMA to be considered ethical in a consequentialist framework, it must act in a way that the consequences of its actions maximise the good or favourable outcomes over those that would be unfavourable.

Scheutz and Malle (2017) note that consequentialist approaches to machine ethics are perhaps closest to the general computational mechanisms that already exist in fields such as robotics, AI and control systems. This compatibility likely exists because many of the computational mechanisms are based on utility theories which tend to favour actions or activities that maximise (or minimise) specific pre-selected variable(s).

Adopting a consequentialist approach to machine ethics is not without its challenges. The first challenge has to do with the issue of tractability (Scheutz & Malle, 2017). How can we be sure that an AMA will, with potentially limited information, be able to understand the impact that its action will have on the good or happiness of everyone involved? How will it ensure that he consequences of its actions are positive?

Furthermore, how will the AMA determine everyone who is involved to ensure favourable outcomes for all at an aggregate level? In general contexts, consequentialist approaches to machine ethics are considered to be currently intractable (Allen & Wallach, 2012). In practice, tractability is only possible in very well defined do-

mains with minimal scope (as the examples that will be discussed in the following paragraphs will illustrate).

The second challenge with this approach has to do with the representation of moral values (Allen et al., 2000; Gips, 1995; Scheutz & Malle, 2017). How will the moral values be represented and substituted with the relevant utility values in order to maximise positive outcomes for everyone? Furthermore, what relative ranking will various moral values have, or will all moral values have equal ranking?

Related to the questions above is the issue of which, or whose, moral values should be used in the calculation of the most favourable outcome for everyone? Merely arguing that the AMA will use whatever moral values and rankings designers decide upon may result in ethically questionable actions and outcomes. These issues further suggest that consequentialist approaches to machine ethics are best suited for highly domain-specific contexts and scenarios where the moral 'rules' are well known and accepted by everyone.

There are only a few published works with practical examples of AMA implementations that are based primarily on consequentialist frameworks. These include the works of Cloos (2005) on the utilitarian *Utilibot*, Pontier et al. (2012) on *Moral Coppélia*, a project to combine moral ratios with an affective component, and the work of Vamplew et al. (2018), where they extend the scope of moral ratios from a single objective to multiple objectives. Cloos' project will be briefly highlighted to illustrate how designers may approach the design of a consequentialist AMA.

The *Utilibot* was designed to operate in a home environment to perform basic functions (e.g. mopping) as well as more advanced features such as real-time monitoring of a person's vital signs (e.g. heart rate). The robot would then play an advisory role when a person is experiencing a medical condition, where it would also automatically alert emergency services when the vitals drop below a certain defined level.

Utilibot, as the name suggests, was based on a utilitarian ethical framework, which is perhaps the most popular form of consequentialism. Its main goal was to maximise

the happiness and well-being of the residents of the household. However, instead of using a utilitarian approach as a basis for decision making, Cloos (2005) employed it only as a criterion for right ethical action. He then used classical decision theory in computer science, i.e. utility theory and probability theory (Russell & Norvig, 2009), to design the relevant decision procedure.

Cloos' approach of using a utilitarian ethical framework as a criterion for what is right is quite similar to what is suggested by Allen and Wallach (2012). Allen and Wallach comment that while top-down ethical frameworks may ultimately prove to be intractable in general contexts, they, however, do note that they can still be useful in limited ones.

They suggest that top-down ethical frameworks, such as utilitarianism, can actually serve as heuristics that can help to limit the search space for the right ethical action. Once the AMA has reduced the number of possible actions it can take, it can after that use another (more tractable) decision procedure to find the correct ethical action.

Cloos' chosen decision procedure is what he calls a Wellnet software architecture, which uses two Bayesian networks to model the state of human beings (vital signs) and the environment (the home), which are later fed into a decision network. The decision network is then finally modelled as a Markov Decision Process (MDP), where an iterative process is used to select the optimal course of action (Cloos, 2005). The decision procedure is still effectively utilitarian, however, since it still operates within the criterion specified by its ethical framework.

The Utilibot has several limitations that are apparent to consequentialist approaches. Firstly, it was conceptualised to work in a particular context (a home) which was highly instrumented with sensors in order to enable it to detect various states. It also required the people in the home to wear wireless health monitoring devices that would allow it to detect deteriorating vital signs.

The feasibility of the Utilibot working in a more general context is not likely, even if one accounts for recent advances in AI. Vamplew et al. (2018) suggest turning the

approach from a single to a multi-objective problem to deal with some of the limitations mentioned here. While there is certainly some value to this suggestion, much of the constraints of consequentialist approaches discussed above are still present. Multi-objective considerations make the issues of tractability with consequentialist approaches more apparent.

Secondly, the moral values used in the project were essentially hardcoded (e.g. from conventional medical practice), which would make it difficult for the AMA to deal with more generic and abstract concepts for ethical reasoning. Indeed, It is hard to imagine a purely consequentialist AMA that can feasibly operate without hard coding and explicit monitoring of specific variables in order to determine overall good. For the Utilibot, the utility was well defined as the general well-being of the household residents, but a real-world ethical scenario will scarcely have such well-defined concepts to work with, thus invoking the challenge of how to represent moral values as utilities.

4.2.2 Deontological approaches to machine ethics

In contrast to consequentialist approaches, deontological approaches to machine ethics are not primarily concerned with the consequences of the actions. Instead, they are concerned with the "rightness", and inversely the "wrongness", of actions (Dignum, 2017; Gert, 1988). Put differently, the central idea in deontological approaches is that actions are ethical if they are performed according to some established principles or rules. The guiding principle in deontological approaches is that actions are ethical if they are 'right'. This principle leads one to conclude that only those actions that are rational and conform to established rules can be considered ethical.

Deontological approaches to machine ethics have to do with designing AMAs that can do what is right (Dignum, 2017; Gips, 1995; Kuipers, 2016; Scheutz & Malle, 2017). What is right can be the performance of a duty for the AMA itself, society or even the environment. In the general sense, it can be about knowing and

performing specific universal rules (such as Kant's categorical imperative). In a more defined context, it can be about performing certain defined duties that are accepted as objective or established norms. Whether in general or specific contexts, AMAs could be considered ethical only when they act in a manner that is consistent with an imperative or established rule that determines what is right.

As can be expected, deontological approaches are not without their challenges. The first and most prominent being the question of tractability (Allen et al., 2000; Scheutz & Malle, 2017). The potentially numerous rules that the AMA would have to know, and also apply correctly, pose an apparent computational challenge. A related challenge is how the AMA would know when to apply which rule in a given scenario.

Some frameworks, such as Kant's categorical imperative or the Golden rule, have sought to minimise the number of rules that need to be followed. However, the same challenge persists, since it would now mean the AMA would need to apply that same abstract rule to potentially limitless situations. This scenario implies an understanding of its goals and those of fellow agents in order to determine how a universal rule can be upheld (Allen et al., 2000). No artificial agent has demonstrated such a level of intelligence to date[2].

The second challenge has to do with conflicting rules (Allen et al., 2000; Gips, 1995). What happens when an AMA has to decide between two or more conflicting rules? How will it choose one over the other(s)? Allen et al. (2000) suggest using another ethical framework, such as a consequentialist approach, to choose between two conflicting rules. This stopgap, as it were, seems to point towards an inherent flaw in deontological approaches when it comes to dealing with conflicting rules.

Moral philosophers such as Gert (1988) have recognised this problem, and have developed deontological frameworks that allow for conflict and resolution within the framework itself. However, the practicality of the AMA being able to compute the

[2]The reader is referred to section 2.2 for a brief discussion of strong AI agents, and their current technological limitations.

exception handling mechanism within the framework remains unclear and is likely intractable (Allen et al., 2000). The bottom line is, exceptions to the rule always add further complication to an already computationally challenging problem.

A few works have also been published showing practical implementations of primarily deontological AMAs[3]. These include the works of the Anderson's and team on the *MedEthEx* prototype medical ethics advisor (M. Anderson & Anderson, 2007; M. Anderson et al., 2006b; S. L. Anderson & Anderson, 2011), and the work of Hooker and Kim (2018) providing further improvements to work done on the *MedEthEx* prototype medical ethics advisor. The work done by the Anderson's on the *MedEthEx* project will be briefly described to illustrate how designers might build AMAs with deontological ethics.

MedEthEx was developed based on the classical principles of Biomedical ethics. In cases where there was a medical dilemma, a health care worker could consult the *MedEthEx* to get a recommended ethical action based on how the AMA was trained in the past. As a result, the *MedEthEx* was conceptualised as only an advisor, as it was not initially meant to act in an autonomous manner. Even their highly publicised robotic version which used the Nao robot platform was only a medication reminder robot, as it did not act autonomously to administer medication (S. L. Anderson & Anderson, 2011).

The *MedEthEx* was at the time the first use of biomedical ethics, interpreted as W.D. Ross's prima facie duties (Ross, 1930), as the primary ethical framework within an AMA in a medical context. M. Anderson and Anderson (2007) argue that their approach is a combination of teleological (i.e. consequentialist) and deontological approaches. However, it aligns more closely with deontology than it does with teleology.

The close alignment with deontological approaches can be seen in how the robot does not force the patient to take medicine if they do not want to - it respects the rule

[3]While many works have been published that theorise about how deontological AMAs might be built, precious few have actually been built in practice.

of autonomy of the patient for the rule's sake (S. L. Anderson & Anderson, 2011). The robot does so even if administering the medicine against the patient's will might be perceived to be overall good and less bad from a teleological perspective.

MedEthEx has its own limitations from a machine ethics framework perspective. Firstly, similarly to the Utilibot, it was conceptualised to work within a particular context. Its design is not likely to be scalable or transferable to a different context with a different set of rules to follow (even though this may not necessarily be a bad thing).

Secondly, as a consequence of the first limitation, the design avoids computational issues of tractability by focusing on a minimal set of rules that could easily be represented in a computational form (they used inductive logic programming (ILP)). However, were the rules to increase, then the tradeoff between scope and tractability would be much broader.

4.2.3 Virtue ethics approaches to machine ethics

Classical Aristotelian virtue ethics has to do with living a life based on virtues (such as compassion, generosity and honesty), which leads to the attainment of 'happiness', or 'eudaimonia'[4]. It focuses on life-long character development of virtues and minimising vices in one's life. A person is considered ethical if they can act in a manner that is consistent with how a virtuous person would have acted in a given scenario.

Just being virtuous, however, is not necessarily enough to live the 'good life', a person also has to have moral or practical wisdom, 'phronesis', to attain 'eudaimonia' (Annas, 2011; F. D. Miller, 1984). Practical wisdom is the ability to know when, and how much, one should act based on which virtues, because blind virtue can also lead to a fault!

The application of virtue ethics to machine ethics seems less common than con-

[4]Often translated from Greek to English as 'happiness', 'flourishing', 'well-being', or even the 'good life'.

sequentialist and deontological approaches. One reason for this might be that there are no obvious computational techniques that are similar to the structure of the majority of virtue ethics based frameworks. This is not to say that no techniques exist that may play a role in these approaches. However, none can generally cover the entire process, since virtue ethics approaches tend to emphasise life long learning and moral character development.

The basic structure and logic of consequentialist approaches, especially in utilitarian form, are analogous to computational mechanisms such as control systems and utility theory. Deontological approaches find an obvious expression in symbolic systems such as first-order logic.

Virtue ethics-based approaches, however, appear to have no apparent computational analogy. They instead seem to more closely align with a hybrid approach, one where a combination of top-down and bottom-up approaches might be applied (Abney, 2012; Allen & Wallach, 2012).

Designers need to keep a few aspects in mind when applying virtue ethics to the AMA project. Firstly, a virtue ethics approach to machine ethics is not about what actions are right or lead to the highest amount of good; it is instead about striving to be a certain kind of robot (Abney, 2012). It emphasises a (robot's) life long approach to character building, a constant striving to be virtuous. This raises obvious questions about how this can be practically achieved in a computationally based AMA.

The second aspect to keep in mind is that virtue ethics based AMAs do not necessarily work off a rule book. Instead, the AMA is expected to introspectively decide who it is, whom it is trying to be and then act from that perspective. This makes virtue ethics-based approaches to machine ethics unique. While the other approaches require externally defined rules to decide what is rational to do, virtue ethics approaches instead take an introspective and rational view of ethical decision making.

Virtue ethics approaches to machine ethics also have their challenges. Firstly, the

challenge of modelling abstract moral behaviour (presumably from observing human beings) is not trivial (Allen et al., 2000). Another challenge has to do with the implementation of abstract virtues such as courage, temperance or wisdom (Scheutz & Malle, 2017). It is not obvious what a robotic implementation of wisdom would be, for instance.

A last challenge has to do with the integration of multiple abstract concepts that lead to moral action. Because the virtuous AMA will process information introspectively, it needs to have a mechanism to understand its motives (which stem from its virtues), reason about the action it needs to take given the abstract moral cues it has gathered from a given situation, and then finally choose the right action to perform (Dignum, 2017).

An excellent example of a design that aligns closely with virtue ethics is by Muntean and Howard (2016)[5], where they eloquently describe their quest to build a minimalist model of an AMA. What makes the work of Muntean and Howard (2016) stand out is how they take time to thoroughly describe the conceptual elements of their AMA, from its ontology to a practical design.

Muntean and Howard (2016) remain philosophically consistent in their conceptualisation and design of their proposed AMA. They conceptualise their AMA to be ethically parsimonious in the sense that it does not explicitly take into account ethical concepts such as responsibility, motivation and moral reasons. It is also ontologically quietist in the sense that it presupposes a moral functionalism. That is to say, it tries to reproduce human moral behaviour, and not to replicate human moral agency.

The minimalist model proposed by Muntean and Howard (2016) is closest to an agent-based virtue theory approach to machine ethics since it makes no assumptions about moral rules. They instead take a life long approach to learn the required

[5]At the time of writing, the author was only aware of this work that has a practical design. Others, such as Abney (2012), P. Lin, Bekey, and Abney (2008) have argued for virtue ethics-based AMAs.

moral behaviour by observing other (human) agents. Their model is designed to discover patterns of moral values and behaviour in existing data *a posteriori*. This represents a decidedly bottom-up approach to machine ethics, where Muntean and Howard (2016) chose a technique that uses a population of evolving neural networks to learn moral behaviour and elements of autonomy over time.

They use the population of neural networks, all of which have been initialised differently[6] to learn specific patterns in moral data, and the evolutionary component is then used to select the best neural networks from the population. The process is repeated until later generations of selected neural networks are more robust, less susceptible to noise and are not heavily dependent on the initial training set. Once a successively mature neural network is identified, it is then presented with data outside of the original training set, and its moral output is then evaluated by human trainers.

Muntean and Howard's minimalist AMA has a couple of notable limitations. Firstly, the inherent weakness that comes from being a largely bottom-up approach to modelling moral behaviour. How does one collect unbiased moral data? How does one ensure the data is trustworthy? How much data is sufficient to train their model? How do we link abstract ethical concepts and represent them numerically so that the model can learn from the data? All these are not really answered in their paper, something the authors note as a limitation (Muntean & Howard, 2016, p. 223).

Secondly, their model was conceptualised and designed to function in a narrow context. As Muntean and Howard (2016) note, their model aligns more closely to moral particularism than it does to moral generalism. Nevertheless, this does mean that the AMA, like many others, cannot be scaled to work in multiple contexts.

[6]Neural networks can be initialised with either random, specific or zeroed parameters (Goodfellow et al., 2016).

4.2.4 Other notable hybrid approaches to machine ethics

Some works do not fit easily into consequentialist, deontological or virtue ethics-based approaches. These works either combine all of the above approaches or introduce a new perspective to the problem of selecting machine ethics frameworks. A few of them are briefly highlighted here for completeness.

The first such work is by Dameski (2018), where they propose a comprehensive model for an AMA by combining all the elements of the three main approaches to machine ethics. It is hoped that by integrating all three frameworks into one super framework, a more conclusive model can be derived.

A related work that tries to combine everything is by Wallach et al. (2010) on their work on the LIDA computational and conceptual model for moral reasoning and decision making. Their work gets much of its inspiration from cognitive science, especially since Allen and Wallach started partnering with Franklin (2003) to develop his IDA concept further to include moral components.

Perhaps unsurprisingly, many works propose the building of AMA's using machine learning. There is the work of Wu and Lin (2018) who propose adding an ethics shaping heuristic to the output of reinforcement learning agents. Arnold, Kasenberg, and Scheutz (2017) consider whether inverse reinforcement learning, in combination with a formal logical approach, could be a possible way to implement AMA's. Conitzer, Sinnott-Armstrong, Borg, Deng, and Kramer (2017) propose using techniques in game theory and machine learning to model moral values and make ethical decisions. There are other projects beyond this small list, such as the popularity of machine learning today. For a broader overview, Kasenberg et al. (2018) and Yu et al. (2018) provide an extensive landscape of the major projects that use machine learning.

The challenge with approaches that try to combine every framework is complexity. Dameski (2018) give little insight on how exactly their model could be practically implemented. Wallach et al. (2010) have made much valuable ground in building a

model for reasoning, akin to the way the human brain works and then applying it to the case of solving moral problems. Whilst their work is no doubt invaluable, one cannot help but get the feeling that a more nuanced, focused and nimble approach, similar to the one proposed by Muntean and Howard (2016), might actually get practical results first.

Ultimately, we need both approaches - models that simulate human moral behaviours and models that attempt to replicate human moral intelligence, to thrive as they both help us look at ethics from different perspectives, and to learn further. The reader is reminded and referred back to the models of artificial moral agency discussed in section 3.2.

The machine learning approaches seem to be the furthest from getting there. It is not that the techniques are not useful; it's that many of the works do not actually take time to properly conceptualise and define their AMA's before going onto building them. Many rush to the design without much philosophical consideration.

The field of machine ethics needs more AMA projects to be conceptualised in a philosophically consistent manner as much as their design needs to make proper engineering sense. Machine ethics is interdisciplinary, and therefore all works in this field need to carve out space for themselves made out of at least two fields in philosophy, psychology, computer science, engineering, robotics, and cognitive and neurosciences.

4.3 Exemplarist virtue theory

The previous section contained a brief discussion of consequentialism, deontology and virtue ethics. In particular, the ethical theories' suitability for use as machine ethics frameworks were considered. Furthermore, detailed examples of how other researchers have applied the ethical theories were discussed. The purpose of these examples was to illustrate how designers might apply these theories to the AMA project.

Now, the first part of this section will put forward an argument that machine ethics frameworks based on virtue ethics are generally a better fit for weak machine ethics AMAs. This argument is made in comparison to their consequentialist and deontological counterparts. Yes, claiming this will likely invite criticism, but at least the argument will be put forward, and it can be criticised based on its merits.

The section will further continue to discuss virtue ethics in its various forms more pointedly. It is crucial first to discuss virtue ethics and the various forms that it can take, to lay a foundation for discussing exemplarism. This is because exemplarism is a derivative of virtue ethics, and so comparing it to the other forms of virtue ethics will help to illustrate what makes it different.

Once the discussion on virtue ethics is concluded, the argument will then move on to a detailed consideration of exemplarism. In this part, the argument that exemplarism is not only feasible for the AMA project, but the standout choice of machine ethics framework for AMAs conceptualised with weak machine ethics, will be put forward.

4.3.1 Why approaches based on virtue ethics make sense

Consequentialist and deontological machine ethics frameworks are similar in that they both employ top-down functional guides or rules for ethical decision making. Virtue ethics frameworks, however, are different in that they employ a non-direct (i.e. bottom-up) method for determining what is ethically wrong or right. They focus on a lifelong development of virtue, as opposed to ensuring specific actions or outcomes. Ethical action is possible based on the premise that a virtuous robot will be moral because it is virtuous to be so.

We have already discussed that consequentialist and deontological AMAs make ethical decisions from the outside-in. That is to say; they use a system of moral values that is set externally, and often independently of the agent. For example, consequentialist AMAs have to be given what variable (i.e. a measure of goodness or happiness) towards which they need to optimise.

These AMAs receive information from the environment and, together with their system of moral values, use that to 'reason' about what ethical action to take. Similarly, deontological AMAs have to be given a set of rules or an imperative that would define what 'right' action is.

These AMAs use their externally defined system of moral value to make moral decisions in a given scenario. As a result, the best that these AMAs can do is to either follow specific duties accurately or ensure certain outcomes that maximise the good. In other words, the effectiveness of these AMA's will be subject to a strict evaluation based on the original system of moral values with which they were embued. This argument assumes that the external system of moral values does not change.

This is not an unfair assumption since the system of moral values is set externally to the agent. As far as the agent is concerned, the system of moral values is a fixed frame of reference. Even if we relax the assumption that the system of moral values is fixed because the agent has some way of receiving updated rules, this still will not change the implication since the AMA ought to take any new system of moral values as fixed.

Even if the AMA were to employ learning-based techniques to improve its performance of the system of moral values - the implication would still be the same. In general, as long as the system of moral values is set externally, the AMA can only perform as good as the best outcome of a strict evaluation of its actions based on the system of moral values.

The framing above provides useful insight into the nature of AMAs built using top-down machine ethics frameworks. It suggests that these AMAs are no more than moral optimisers that do their best to achieve results that are aligned to their externally defined system of moral values. Their goals are inherently short-term in that they are moral when they are able to achieve the expected outcomes in any given situation.

Returning to the criteria for artificial moral agency given by Floridi and Sanders

(2004)[7], we can conclude that these AMA's would meet the requirements for adaptability and interactivity, but they would not meet the requirement for autonomy. If they were autonomous, they would not only be able to 'optimise' independently of environmental information, but also change their internal representation of the system of moral values. This capability is something that is, by definition, not achievable for AMA's that use an externally defined system of moral values.

Top-down AMAs are essentially about answering the question, 'what should I do?'. Bottom-up AMAs, on the other hand, are about answering the question, 'what should I be?' (Abney, 2012). This difference in approach is what fundamentally separates top-down and bottom-up AMAs. By asking what they should do, top-down AMAs essentially need an externally defined system of moral values, expressed through utilities, functional guides, rules or imperatives. By asking what they should be, bottom-up AMAs emphasise the internal development of a system of moral values.

To put it more formally, bottom-up AMA's based on virtue ethics use an internally generated system of moral values to facilitate the ethical decision making process. Of course, this does not mean that they do not use any external information to make ethical decisions (they do!). They still need to learn from experience and the observation of others (i.e. moral education). However, their ethical behaviour is an internal function of their 'moral character' applied to a given morally charged situation.

While a fuller exploration of the concept of moral character lies outside the scope of this thesis, the reader is referred to the works of C. B. Miller (2013) and Besser-jones (2008), who have covered the topic in greater depth. Some debate exists about what precisely moral character is, which Blesser-jones and Miller cover extensively. The way the term is used here refers to its classical philosophical meaning of being an agent's *"dispositions to act in certain sorts of ways"* (Besser-jones, 2008, p. 321).

This definition, it is suggested, is equally applicable when referring to 'moral

[7]The reader is referred back to section 3.4 for a more detailed discussion.

character' in AMAs (P. Lin et al., 2008). This is because the AMA's internal system of moral values represents a particular inclination for the agent to act in a particular way. Therefore, the internal system of moral values can be thought of as representing a kind of artificial moral character for the virtue ethics AMA.

Thinking of moral character development in AMAs in this way would also help us clarify what it means for them to pursue virtue. Since virtue is an excellent trait of character (Hursthouse & Pettigrove, 2018), it follows that an AMA can pursue excellence of character by learning and improving upon its internal representation of moral values. Putting it more plainly; an AMA can pursue virtue by continuously developing its moral character. This improving moral character will cause the AMA's dispositions to align towards being more ethical continuously.

Moral character development is, of course, a life long process (Hursthouse & Pettigrove, 2018), and it should be no different for an AMA with virtue ethics. It is a kind of disposition to learning and improving itself throughout all of (robotic) life. Since the main goal is to be virtuous, ethical behaviour is necessitated not through an externally defined system of moral values, but by and through the pursuit of the 'good' and virtuous life. This pursuit implies that the virtue ethics AMA has an unending quest to develop its moral character.

The above also gives us crucial insights about the nature of bottom-up AMA's based on virtue ethics. Their performance is harder to judge since they do not try to optimise towards some known and externally defined system of moral values. Of course, if they achieve moral character development through observation of human beings, for instance, then they may well indirectly learn moral values that closely resemble something we can recognise in a known system of moral values.

Even though it may be difficult to judge their performance directly, it is however not impossible since we can evaluate their actions according to other known principles. For example, we can evaluate the AMA based on community expectations of moral behaviour even though we may not be aware of how the robot has developed morally over time.

This kind of evaluation is not too dissimilar to how we might pass moral judgement on the actions of other human beings. We may not be aware of their upbringing, or even the exact system of moral values with which they subscribe. However, we can still judge their behaviour according to commonly known (i.e. normative) sets of moral values.

This tendency to be a moral black box[8] is precisely what makes virtue ethics AMAs so different, and it forces us to treat them in a manner that is closer to a moral person (Abney, 2012). This is in stark contrast to top-down AMA's; they are moral 'white boxes' in the sense that we know what system of moral values towards which they are optimising.

Returning once more to the requirements for artificial moral agency given by Floridi and Sanders (2004)[9], we can see that a virtue ethics-based AMA would not only meet the requirement for interactivity and adaptability, but they would also meet the requirements for autonomy. They do not explicitly make their moral choices based on an externally defined system of moral values, but on an internally generated one based on their internal moral character development. Since the system of moral values is internally generated and represented, then the AMA is also theoretically free to change it without strict consultation with any external agent.

These conclusions may not seem desirable since we might want to create AMA's that we can control and whose goals we understand intimately. As valid as this argument is, the researcher is not making a case for or against it here. What is being argued is that from a conceptual and ontological standpoint, virtue ethics AMAs make better (artificial) moral agents than their consequentialist or deontological counterparts.

Some may argue that their top-down AMA might conceptually be a moral 'black

[8]This term is meant conceptually, not practically, since it is still possible to build in an element of explicability into the AMA. This is the same as interrogating any human to understand the rationale for their moral behaviour without necessarily understanding the moral code with which they live.

[9]The reader is referred back to section 3.4 for a more detailed discussion.

box' in the sense that it is meant here. However, a counter-argument would be that they have veered away from purely top-down machine ethics to a form of hybrid machine ethics that has some elements of a bottom-up machine ethics framework (i.e. virtue ethics). They just have not acknowledged it in their conceptualisation or design.

In summary of the last several paragraphs, top-down AMAs based on consequentialism and deontology are essentially about asking the question, 'what action should I do?'. This makes them dependent on an externally defined system of moral values. By implication, they would not meet the requirement for autonomy, and thus would not qualify as (artificial) moral agents.

Virtue ethics AMAs, however, are about asking the question, 'what should I be?' This makes them dependent on internal moral character development, which in turn aligns their 'dispositions to act' towards being ethical. This would also mean they would qualify as (artificial) moral agents in the way that it has been defined for this thesis.

In the next section, these arguments are taken a step further by defining the exact form of virtue ethics that is suitable for the AMA.

4.3.2 Why exemplarist virtue ethics is an even better fit

The previous section considered how approaches based on virtue ethics could potentially make for better AMAs. This argument was made in comparison to consequentialist or deontological approaches. This section will briefly discuss the various forms of virtue ethics, and then motivate why exemplarist virtue ethics is the best fit for the AMA in this thesis.

The various forms of virtue ethics include classical Aristotelian virtue ethics, agent-based ethics, target-centred virtue ethics and exemplarist virtue ethics (Hursthouse & Pettigrove, 2018). Some have argued that exemplarist virtue ethics is just another form of agent-based virtue ethics, but it will be discussed on its own in this thesis. Besides, as the ensuing discussions will show, they are more different than

the same.

As a reminder, classical Aristotelian virtue ethics has to do with living a life based on virtues. This virtuous life leads to the attainment of 'happiness" or 'eudaimonia'[10]. It focuses on life-long character development of virtues and minimising vices in one's life. A person is considered ethical if they can act in such a manner that is consistent with how a virtuous person would have acted in a given scenario.

Just being virtuous, however, is not necessarily enough to live the "good life", a person also has to have moral or practical wisdom, 'phronesis', in order to attain 'eudaimonia' (Annas, 2011; F. D. Miller, 1984). Practical wisdom is the ability to know when, and how much, one should act based on which virtues, because blind virtue can also be a vice.

Michael Slote is likely the foremost theorist on agent-based virtue ethics. His conceptualisation of agent-based virtue ethics will be discussed next. While eudaemonist virtue ethics emphasises practising virtue as the primary method of character development (for the attainment of happiness), agent-based virtue ethics emphasises the agent's internal motives and predispositions as the primary method for deciding what to do (Slote, 1995; Van Zyl, 2005).

Furthermore, the internal motives and predispositions of the agent should not be based solely on some other concept(s), such as eudaimonia or virtue, which might be taken to be more fundamental. Instead, the 'goodness' or 'badness' of the agent's motives and intentions when acting are what determine whether or not it is ethical (Slote, 1995).

At first glance, this may sound like deontology, which was briefly discussed back in section 4.2.2. While some overlaps certainly exist, the main difference is that deontology has universal or categorical imperatives, whereas as agent-based virtue ethics treats the agent's intention and motivation when acting as the primary determining factor for whether or not it is ethical.

[10]Often translated from Greek to English as "happiness", "flourishing", "well-being" or even the "good life"

Swanton (2003) popularised target centred virtue ethics, which will be discussed next. A right theory of target centred virtue ethics always contains four elements. Firstly, it contains the 'field' of virtue. The field represents what the virtue is about generally. For example, honesty has to do with truths and falsehoods. Humility has to do with how we view ourselves.

Secondly, the 'basis' of virtue is a feature in its field, which it is targeting. It is perhaps helpful to think of the basis as the concentration or focus area in the field of virtue. Honesty, for instance, might have to do with communication, and how the truth of the communication might impact others. Humility might have to do with how we view ourselves.

Thirdly, the 'target' is what the virtue is trying to achieve. In other words, what will be achieved if an agent exemplifies a virtue like honesty? In this case, it might achieve better relationships with others that are built on trust. Exemplifying humility might achieve a more sober and right-minded view of ourselves.

Lastly, the 'mode' has to do with the 'how' of the process. It connects the basis to the target. For example, if the agent is humble (the basis), then it can achieve stronger relationships based on trust (the target) by expressing its thoughts, feelings and beliefs truthfully (the mode).

In target centred virtue ethics, right action is primarily determined by whether or not it hits the target of the virtue. For example, if an agent wants to embody courage (basis), and decides to act by confronting its worst fears (the mode), then its action will be right if and only if it succeeds in actually managing its fear (target). In this way, target centred virtue overlaps with consequentialism in that the outcome determines the "rightness" of an action.

Likely the biggest issue with target centred virtue ethics has to be that it allows a virtue to have multiple fields. Because of this, the agent can often find itself needing to choose one action among many with different targets in multiple fields. This is why the right action can be defined as one that is virtuous (perfectionism), good enough for the situation (permissivism), or not overall vicious (minimalism) (Swanton, 2003)

We now arrive at exemplarist virtue ethics, which is suggested to be the best fit for the AMA in this thesis. Exemplarist virtue ethics, or simply exemplarism, is a moral theory that was proposed and developed by Linda Zagzebski (2010, 2017). In contrast to classical virtue ethics, exemplarism is not conceptually grounded in virtues or achieving 'eudaimonia. Instead, it is grounded in the *"exemplars of moral goodness"* (Zagzebski, 2010).

Zagzebski defines the exemplars of moral goodness as morally admirable agents whose example is worth following. All other moral concepts are then grounded on the exemplars of moral goodness. For example, an agent would desire to be virtuous because an exemplar of moral goodness is virtuous. Since exemplarism is about following the example of morally admirable agents in society, exemplarist agents, therefore, need to be good at identifying them, something that Zagzebski (2010) states can be achieved through empirical observation.

It is essential to state here that this application of exemplarism does not forego the need for practical wisdom in the making of moral decisions. In essence, it is still based on classical Aristotelian virtue ethics. However, where Aristotelian virtue ethics is conceptually grounded in the virtues of the good life, exemplarism is practically grounded in exemplars of moral goodness.

In other words, the application of Linda Zagzebski's theory is not merely about building AMAs that blindly follow examples of morally worthy agents in society. Instead, it is about learning how these morally worthy individuals make moral decisions, while still retaining the critical capacity for making independent moral decisions - something that is a definite requirement for artificial moral agency (Floridi & Sanders, 2004). In the end, the AMA would still need to choose rational moral decisions.

In the next several paragraphs, three reasons will be put forward that seeks to differentiate exemplarism from the other forms of virtue ethics, and therefore why it is a better fit for the AMA. The first reason has to do with the conceptual grounding of exemplarism. The second reason has to do with meeting community expectations.

Lastly, the third reason has to do with the practical simplicity of exemplarism.

Conceptual grounding of exemplarism

All the various forms of virtue ethics, except for exemplarism, are grounded in conceptual and abstract terms like virtues (eudaemonist and target centred virtue ethics), motives and predispositions (agent centred virtue ethics). For a human being, this kind of conceptual grounding is not generally problematic, but for an AMA conceptualised with weak machine ethics, it can be a big problem.

How does one go about computationally modelling virtues, motives or any of the other abstract moral concepts? If we take a virtue such as courage, what would that look like in an AMA? What about all the other virtues such as temperance, honesty and compassion? Moreover, how will the agent assimilate all the virtues and develop its character over time? Yes, philosophical answers are possible, but converting them to practice remains a real challenge.

Motives and dispositions in agent-based virtue ethics suffer from the same issues. How do we model them computationally? Certainly, motives and dispositions could be taken to be whatever the AMA is programmed to be like, but this is merely an extension of the programmer's motives and dispositions more than it is the agent's own.

Exemplarism, on the other hand, is not grounded in abstract concepts, but in the exemplars of moral goodness. The implication of this is quite straightforward - the AMA only has to follow the examples of morally upright agents in order to learn how to be a moral agent. The difference between this and the other forms of exemplarism is subtle, but it is there.

From a purely computational perspective, the problem of learning from example is more achievable than programming the understanding of abstract concepts into the AMA. Note, this is also not trivial - it is still challenging! However, if an agent can identify and learn from moral exemplars (Zagzebski, 2010), then this strongly suggests than the AMA can also identify and learn from moral exemplars. Chapters

5 and 6 are dedicated to demonstrating how this might be achieved.

Exemplarism can meet community expectations

Moral introspection is a strength, rather than a weakness, of virtue ethics approaches. However, it can be a weakness if the AMA is wholly disconnected from (societal) reality in its moral decision making. It seems that to be an excellent exemplarist moral agent, one needs a kind of 'golden mean'[11] between internal introspection about moral decisions and an external grasp of how the most admirable moral exemplars in society behave. This golden mean, it is suggested, ensures that the exemplarist AMA is both responsive to societal moral values while remaining autonomous in its final decision making (Floridi & Sanders, 2004).

It could be argued that all the forms of virtue ethics are essentially about doing what a virtuous agent would have done, and thus they all have a kind of in-built exemplarism. While this statement might have some truth to it, it is not entirely accurate to say exemplarism is no different from other conceptualisations of virtue ethics.

Agent-based virtue ethics, for example, is so internally focused on the agent's motives and predispositions that it is hard to say it has an element of exemplarism. This kind of introspective moral decision making will likely be a source of value misalignment between the AMA and society (Baum, 2017). As a result, it will likely be more challenging to meet community expectations of moral behaviour in agent-based virtue ethics, compared to exemplarism.

Eudaimonist and target centred virtue ethics will likely fair better than agent-based virtue ethics when it comes to meeting community expectorations of moral behaviour. Their failure from a machine ethics perspective, however, is the practical

[11] Aristotle believed that a person needs a balance between the vices of deficiency and excess to be virtuous. This balance can be thought of as a conceptual mid-point between two opposite vices - a 'golden mean'. This mid-point is thus not literally the 'exact middle', it is a golden mid-point reached through rational argument.

challenge presented by their grounding in abstract virtues. Target centred virtue ethics holds virtue as fundamental in a far more sophisticated way than eudaemonist virtue ethics. The abstract definitions of fields, basis, targets and modes of virtue add further complexity to an already abstract concept that somehow needs to be modelled computationally.

Aristotelian virtue ethics also suffers from the problem of modelling abstract virtues computationally, but it is at least not as convoluted as target centred virtue ethics. Some community expectations will likely be met, but it is unclear how an AMA can be programmed to understand abstract concepts in a way that is understood by others (human) agents in society. Yes, researchers such as Muntean and Howard (2016) have made good progress in the computational modelling virtue, but their work is still far from seeing practical results in the real world.

Incidentally, exemplarism, though not originally formulated with the AMA project in mind, was proposed to address precisely the issue of agents needing a detailed understanding of virtue in order to be virtuous (Zagzebski, 2010). In exemplarism, virtue does not have to be learned and understood directly. Instead, it is indirectly learned through the admiration and following of moral exemplars in society. This position is arguably more desirable for a weak machine ethics AMA because it is achievable, and the outcomes are aligned to community expectations.

In the end, exemplarism seems not only to bypass the problem of directly modelling highly abstract concepts like virtues, but it can potentially also meet community expectations of moral behaviour by following those same agents that society would hold up high as moral examples. This is why it is suggested that it lies at the golden mean of virtue ethics approaches.

Yes, there is a criticism of this position in the literature. Briefly, the criticism is that AMAs cannot achieve community expectations of moral behaviour by learning an aggregated view of societal moral value, because one does not exist (Baum, 2017). Interestingly, Zagzebski (2010) formulates her theory in a way that shows that she is aware of this broad criticism that can be laid against it due to its apparent divorce

of a descriptive understanding of virtues in moral choice.

Zagzebski (2010) argues using the theory of direct reference (Kripke, 1972) to say that people have always been able to refer to terms such as gold or water before they knew how gold or water was composed atomically. Similarly, she argues that it is possible to fix a reference for moral values without necessarily having a descriptive understanding of them[12]. By implication, people can identify good or bad moral exemplars without necessarily being able to define the virtues or vices that make them good or bad.

If we apply Zagzebski's thinking to AMAs, then they too can learn an approximation of an aggregated view of societal moral value by direct reference, especially when learning from many exemplars. AMA's can determine that an individual is likely a good or bad exemplar, and this information can be aggregated across the entire population of exemplars.

While the concept of value alignment in AMAs cannot be exhaustively covered here, the reader is pointed to the works of Vamplew et al. (2018) and Prasad (2018), who have dealt with the topic more comprehensively. The reader is also referred to the next chapter (5), which seeks to demonstrate how exemplarist AMAs might learn from exemplars to meet community expectations of moral behaviour in a specific scenario.

Exemplarism can be practical

From a machine ethics point of view, exemplarism is a pragmatic approach to building AMAs because the goal is at least conceptually and practically clear - teach AMAs to learn from moral exemplars and to improve themselves over time. If we assume, as Zagzebski (2010) does, that moral values can be learned through empiri-

[12]This points to a long and complex debate in the philosophy of language between descriptivists and supporters of the later causal theory of reference (in Kripke's sense known as referentialism), into which we will not venture here, but it will be briefly discussed again in the next chapter in section 5.3.1.

cal observation, then this would imply that many current techniques in AI could be applied to at least partially implement the AMA.

In contrast, the goals of programming eudaemonist, target-centred and agent-based virtue ethics are, in the researcher's view, not as conceptually clear. How do we know when we have achieved the task of modelling virtue in the cases of eudaemonist and target-centre virtue ethics? Similarly, how would we program and evaluate the intentions and motivations of an AMA with agent-based virtue ethics? Answering these questions truthfully likely takes us into the realm of strong machine ethics, which is probably not achievable with current technology in AI[13].

Yes, exemplarism will have its challenges too. Collecting data through observation, for instance, is also challenging, but at least we are dealing with a problem that is understood and for which some attempts at solving have been made (Duan, Andrychowicz, Stadie, & Ho, 2017; Muntean & Howard, 2016). The main argument is the conceptual simplicity of exemplarism likely makes the task of actually programming exemplarism into an AMA feasible, as will be demonstrated in chapters 5 and 6.

The three features of exemplarism that have been discussed, namely *grounding in moral exemplars*, *meeting community expectations* and *practical simplicity*, are what makes it stand out as a choice of ethical theory to support the building of AMAs that meet community expectations of moral behaviour. Note, this conclusion is made in the context of building exemplarist AMAs that are conceptualised to have weak machine ethics - different conclusions would likely be reached if the underlying assumptions are changed.

4.4 Conclusion

This chapter covered much ground, and so it would be helpful to summarise what was discussed briefly. Section 4.2 briefly discussed the three main ethical theories

[13]The reader is referred to sections 2.2 and 2.4 for this discussion.

in normative ethics, and how these might be applied as machine ethics frameworks to the AMA project. Specifically, we looked at how others have interpreted these theories, adapted them to a machine ethics framework, and ultimately designed or built an AMA based on them.

Section 4.3.1 gave a detailed motivation for why approaches based on virtue ethics are generally a better fit for the AMA in this thesis. Furthermore, it was further argued that AMA's based on virtue ethics frameworks not only make sense, but are potentially better artificial moral agents according to the requirements of Floridi and Sanders (2004).

Section 4.3.2 provided a brief evaluation of the various forms of virtue ethics in comparison to exemplarism. In that discussion, it was concluded that three features of exemplarism, namely *grounding in moral exemplars*, *meeting community expectations* and *practical simplicity*, are what make it stand out as a choice of ethical theory to support the building of AMAs that meet community expectations of moral behaviour.

This chapter did not explicitly cover any philosophical or computational challenges of using exemplarism as an ethical theory of choice for the AMA project. That task will be left for the next chapter (5). The main goal of this chapter was to position exemplarism, and argue how it can conceptually support the AMA project. To the researcher's knowledge, this view has not been so extensively discussed or covered in any other works.

The purpose of the next chapter is to demonstrate, from the bottom up, how exemplarism can practically work in a real AMA by using a realistic scenario that is relatable. This chapter will help to clarify key concepts that were used in this chapter and to generally help the reader to experience something of the application of exemplarism to the AMA project.

Chapter 5

Exemplarism in action

5.1 Introduction

The primary purpose of the previous chapter was to position exemplarism as an alternative and suitable ethical framework for the building of AMAs that can meet community expectations of moral behaviour. This positioning was critical to articulate the concept of an exemplarist AMA, mainly because it is not a subject that has previously been covered so explicitly and extensively in the literature.

Admittedly, however, such a proposal of the application of exemplarism to the AMA project will be inherently insufficient. There are a couple of reasons for this insufficiency. Firstly, unlike other machine ethics frameworks, an exemplarist machine ethics framework does not have the benefit of a substantial amount of literature from which broad arguments and opinions about it can be formed and discussed thoroughly.

Secondly, due to the lack of extensive literature covering this topic explicitly, the use of specific key terminology will likely be open to interpretation, debate and even criticism. This is likely the norm for any new research that might propose a position that either substantially deviates from the established literature or introduces a new perspective. The application of exemplarism to the AMA project likely lies in the latter view - it is a different perspective in the young and growing field of machine

ethics.

The primary purpose of this chapter is to immerse the reader in a scenario that would seek to demonstrate how exactly exemplarism might be applied to the AMA project. The scenario is meant to be as relatable as possible so that the use of exemplarism as a normative ethical theory applied to the AMA project can be further clarified.

The secondary purpose of this chapter is to discuss the potential challenges of using exemplarism in the AMA project. These challenges will be looked at from both philosophical and practical (design) perspectives. Discussing the challenges to this approach is essential in order to point out any potential deficiencies, and also to provide a potential path for future research direction in exemplarist AMAs.

Since we are dealing with weak machine ethics (see section 2.3.3), it makes sense to constrain the scenario to a specific context which can shed light on how exemplarism in AMAs might work in practice. This is similar to how the field of applied ethics attempts to study and apply ethics in specific contexts such as business, environmental, medical, educational and war-time scenarios. At a minimum, the scenario would need to demonstrate how the AMA might identify moral exemplars, learn moral behaviour, and make moral decisions.

The rest of this chapter is structured as follows. Section 5.2 will introduce Robo-teacher, a robotic classroom teacher with an inbuilt ethical performance element (see section 3.5) modelled on the principles of exemplarism. Section 5.3 will specifically look at the philosophical and practical challenges of employing exemplarism in the AMA project. Finally, the chapter will conclude with some reflections on the approach.

5.2 Robo-teacher: an exemplarist AMA

Imagine a near-future classroom that is taught by a robotic teacher (call it Robo-teacher). Robo-teacher is responsible for teaching mathematics to 20 students from

diverse cultures and socio-economic backgrounds. Robo-teacher is one of the main attractions in this school and is part of the reason why parents pay the rather expensive school fees to bring their children here.

Most learners are excited by the prospect of being taught by Robo-teacher and are generally corporative and respectful of it. One learner in particular, however (call them Nancy), is quite disruptive in class and has historically caused Robo-teacher to spend an inordinate amount of time focusing on her, at the expense of the other learners. In a typical 40-minute classroom session, Robo-teacher can easily spend more than 20 minutes interacting with Nancy only.

The other learners, being unhappy about the situation, reported it to the headmaster and some of their parents at home. One parent, in particular, took it upon herself to write an official letter of complaint to the headmaster. In this letter, the parent explained how it is unethical for them to be paying for their children's education, only for one learner to monopolise the Robo-teacher's time in class. The parent demanded that the learner be expelled, or Robo-teacher must handle the situation better going forward.

The headmaster, upon receiving the letter of complaint from the parent, and recognising the gravity of the situation, immediately sought help from the manufacturer of Robo-teacher. Once the manufacturer had been fully briefed on the details of the issue, they immediately started working on a possible solution that might improve how Robo-teacher would handle the situation in the future.

The manufacturer finds a possible solution, and they propose it to the headmaster. In this solution, they propose that Robo-teacher needs to be equipped with the ability to handle ethical situations in the classroom. The headmaster agrees, and the manufacture begins work on building an ethical routine into Robo-teacher[1].

The manufacturer decides to use exemplarism as the ethical theory of choice

[1]This scenario is based on a collection of case studies on classroom ethics by Levinson and Fay (2016). It was done in this way to ensure that it was grounded in reality. In this scenario, the human teacher has been replaced with Robo-teacher, and different names were used for the learner(s).

and decides to build it into the ethical routine of Robo-teacher. Their rationale for picking exemplarism was primarily based on its conceptual simplicity. They argued that Robo-teacher needs to learn to handle the situation based on exemplar teachers and how they have resolved similar scenarios in the past.

For Robo-teacher to handle the situation with Nancy appropriately when next it happens, it needs to do three things. Firstly, it needs to identify morally praiseworthy teachers in the school, district or even nationally. What is important here is that it learns from a sufficiently large population of teachers to ensure that it can learn the various nuances with which the scenario can present itself.

Secondly, it needs to learn how these exemplar teachers handle classroom ethics, and in particular similar situations in their contexts. This knowledge will help it to form an internal representation of the moral values that are inherent in the exemplar teachers' decisions and actions. Over time, this representation will develop into a kind of moral character for Robo-teacher as it practices exemplarism (see section 4.3.1). A useful reference for how exemplar teacher might handle classroom ethics can be found in Levinson and Fay (2016).

Lastly, it needs to use its learned internal representation of moral values to deliberately and rationally choose what action, or set of actions, will drive the situation towards an appropriate ethical outcome. In Levinson and Fay (2016), for example, ethical outcomes could be motivating Nancy to be more considerate of other learners, ignoring her, or even expelling her from class.

There are no prescriptive outcomes here; what is essential is that Robo-teacher can pick one and drive the scenario towards it based on Nancy's behaviour. Clearly, knowing how exemplar teachers have handled similar situations in the past will assist Robo-teacher in picking the relevant actions and outcomes.

There are multiple ways in which Robo-teacher could identify moral exemplars. It could physically observe many teachers while they teach their respective classrooms, though this may take a very long time to form a useful representation of moral values. It could search the internet for exemplar teacher cases and stories, perhaps

by reading Levinson and Fay (2016). The manufacturer could also pre-select the relevant exemplars from which Robo-teacher ought to learn[2].

For this scenario, let us assume that the manufacturer has kept a database of teachers (good or bad) for use in training their robot teachers and that this database remains available to Robo-teacher even after it is deployed to a school. Additionally, this database contains a numeric indicator of how effective the teacher has been at dealing with morally charged situations in the classroom, perhaps using a combination of teacher, learner and parental feedback.

Zagzebski states that people *"identify admirable persons by the emotion of admiration, and that emotion is itself subject to education through the example of the emotional reactions of other persons"* (Zagzebski, 2010, p. 52). Robo-teacher does not have the 'emotion of admiration', but it can search the database and look for examples of teachers that have had relative success based on the numeric indicator mentioned above. It could also use other metrics besides teacher effectiveness. For example, metrics like the grade of a class, the number of learners, and the location of the school could also be included to help it hone in on the relevant exemplars from which to learn.

The next task is for Robo-teacher to learn a system of moral values from the identified exemplar teachers. The good thing about picking specific contexts (in this case, a classroom) is that it makes the task of knowledge engineering far more straightforward than in a complex scenario (Russell & Norvig, 2009). Incidentally, picking a specific and constrained context also helps to minimise the burden of the framing problem in AI (Mayo, 2003). The latest research also suggests that this problem is solvable by grounding semantics in multiple perceptual modalities (Kiela, 2017).

There are a couple of ways in which Robo-teacher might learn a system of moral values that would be appropriate for classroom ethics. Firstly, it could use a bottom-

[2]Section 5.3.2 will address some of the practical challenges that may arise in identifying moral exemplars.

up machine learning approach to learn from numerous and diverse exemplar teachers conducting classroom sessions. This approach would likely require video recordings, or a similar rich dataset, of exemplar teachers conducting their lessons in order to be effective.

Undoubtedly, much could be said about machine learning-based approaches for learning by observation. Suffice it to say that learning behaviour through observation or data is a field of computer science that is increasingly gaining traction and for which many proofs of the concept exist (Argall, Chernova, Veloso, & Browning, 2009; Brys et al., 2015; Duan et al., 2017; Muntean & Howard, 2016; van Lent & Laird, 2001).

The second way to learn an internal representation of moral values could potentially be via expert knowledge. This approach would minimise the burden of having to design a sophisticated learning-based algorithm to determine it. The approach will depend on the complexity of the scenario being modelled. The more complex the scenario is, the more impractical it will be to specify the internal representation of moral values via expert knowledge.

Whichever approach is chosen, Robo-teacher will have to learn from many exemplar teachers. Learning from many exemplars ensures that Robo-teacher does not merely parrot learn from one or a few examples, but that it bases its moral knowledge on a wide variety of scenarios and exemplar teachers. Furthermore, it is essential to ensure Robo-teacher learns from a diverse set of exemplars so that it does not later exhibit biased actions that may be localised to a specific exemplar teacher, school or region. Once Robo-teacher has learned from all the identified exemplar teachers, it stores this information inside its internal knowledge database.

Only learning what others have done in various scenarios does not necessarily mean that Robo-teacher will make the right decision in a given situation. For that, it will need to have a mechanism that allows it to plan and make ethical decisions in a given scenario in real-time. For this to happen, Robo-teacher will likely need much of the same sensors that were used to capture information used to observe the

exemplar teachers. The sensor data and the learned system of moral values could be employed by a decision procedure that will allow it to make ethical choices.

There are many decision procedures in computer science, many of which are centred around probability and utility theory (Russell & Norvig, 2009). These include probabilistic reasoning, decision networks, utility functions, Markov decision processes, amongst many others. The reader is referred to the works of Conitzer et al. (2017), Peterson (2009), Russell and Norvig (2009) for a more detailed discussion of the topic.

In this scenario, a probabilistic approach such as a Markov Decision Process (MDP)[3] could be appropriate, especially if all the morally relevant information can be measured through sensors. If we assume that some potentially relevant data, such as anger, frustration, and others, cannot be measured or modelled by the sensors, then a Partially-Observable Markov Decision Process (POMDP) could be employed, as suggested by Abel, MacGlashan, and Littman (2016).

The framing of exemplarism makes it clear that ethical behaviour is partially observable. This is why the agent needs to identify new exemplars continually and to learn from them. For this reason, a POMDP would be the appropriate choice of technique. POMDPs could be advantageous because they would allow for stochastic modelling of ethical decision making, thus avoiding the problem of directly learning top-down ethical rules that are a crucial feature in exemplarism. It would also allow us to model the environment based on the behaviour of exemplar teachers, thus allowing Robo-teacher to learn from them. Finally, POMDPs can be further augmented with machine learning-based techniques, which could allow the agent to continuously and automatically learn from exemplars.

Note, the point of this scenario and thesis is not to argue for the optimality of POMDPs as appropriate decision procedure for Robo-teacher - better algorithms

[3]MPD/POMDPs are stochastic modelling techniques that allow for the modelling of decision making in complex and stochastic environments. They are both explained and implemented in detail in chapter 6

may well exist. It is to demonstrate, however, that exemplarism might be applied and implemented in an AMA such as Robo-teacher. The final choice of the relevant decision procedure will ultimately lie with the designers.

The effectiveness of the decision procedure will depend on how well it is implemented. Even if it is implemented moderately well, this will allow Robo-teacher to make some moral decisions in real-time. The decision procedure, coupled with the knowledge gained by learning from exemplar teachers, will enable Robo-teacher to have a balance (i.e. 'golden mean') between merely doing what it has learned from exemplar teachers, and rationally choosing the relevant actions in real-time (see section 4.3.2).

Once Robo-teacher has identified exemplar teachers, learned from their behaviour, and is equipped with a decision-theocratic procedure, it is now ready to go back to the classroom and hopefully handle situations where Nancy is disruptive in a better way. A myriad of options for handling the situation exist, such as speaking to Nancy directly, ordering her to leave the classroom, making her sit by herself in a corner, barring her from asking any further questions, reporting her or even merely ignoring her (Levinson & Fay, 2016).

Once Robo-teacher has chosen an action, it will need to look out for feedback regarding the effect (either from the learner(s), teachers or parents), and to further learn from it so that it can handle the situation better in future. The emphasis of exemplarism might be on learning from exemplars; however, nothing stops Robo-teacher from learning from its own experiences, whether good or bad. This is of course in line with the wider virtue ethics based context of exemplarism.

As has been seen in this scenario, how exemplarism will work itself out in practice is highly dependant on the scenario or context where it is implemented. Many of the choices and assumptions that were made in this scenario were centred around ensuring that the AMA could identify moral exemplars, represent the classroom world computationally, and use a suitable decision-theocratic approach to make moral decisions. Undoubtedly, different choices and assumptions could be made. Still,

ultimately all of them have to contend with the fact that we are dealing with computationally rational AMAs with weak-machine ethics.

What this scenario has hopefully shown is that exemplarism can be practically implemented in a specific context. Even though this scenario was in a classroom context, a similar one could be derived to show how it might work in a myriad of other contexts. There is no doubt that a lot of technical challenges exist with this approach. Still, the hope is this scenario has shown that it is possible to make choices that minimise the technical burden without compromising on the chosen ethical theory.

5.3 Philosophical and practical challenges of exemplarist AMAs

The purpose of the previous scenario on Robo-teacher was to demonstrate how exemplarism might be applied to the AMA project in practice. However, any application of a normative ethical theory to the AMA project will likely have challenges and limitations, and this is no different for exemplarism. The purpose of this section is to discuss these challenges and limitations from both a philosophical and practical perspective.

5.3.1 Philosophical challenges

An apparent lack of a descriptive understanding of virtue in exemplarism

All the various forms of virtue ethics have something to say about virtue and practical rationality (Hursthouse & Pettigrove, 2018). Where they differ, however, is in how they treat these aspects in virtue ethics. Exemplarism, in particular, is not conceptually grounded in virtue, but rather in the exemplars of moral goodness. This

conceptual grounding of exemplarism means that the agent does not necessarily need a detailed understanding of virtue in order to be virtuous.

As might be expected, exemplarism has been criticised in the literature (Kotsonis, 2020; Szutta, 2019), particularly around its apparent lack of a descriptive understanding of virtue. Szutta (2019) notes that people will likely still need some prior knowledge or understanding of virtue if they wish to live a virtuous life, even if they follow morally praiseworthy individuals. After all, how would they know not to admire morally corrupt people if they do not have some sense what a virtue or a vice is?

Kotsonis (2020) expresses a similar criticism, albeit from a slightly different perspective. Kotsonis' main argument is that socio-cultural norms and values have an impact on who the agent can identify as moral exemplars. This can influence the meaning that the agent can derive for concepts such as virtues and vices, and even deontological ones such as duty, right and wrong. For both Szutta (2019) and Kotsonis (2020), their arguments are essentially descriptivist, as they seem to be stemming from the traditions of the proponents of an indirect theory of reference (Fitch, 1987).

Zagzebski (2010, 2017), who seems to be a proponent of referentialism (Kripke, 1972), counters this argument by stating that a descriptive understanding of virtue is nonetheless not specific in itself. For instance, what does it mean to be a virtuous person, when there is almost no one that would fit a strict definition of the term? Similarly, who is, practically speaking, a wise person, if almost no one would fit a descriptive definition of wisdom?

Zagzebski's theory hangs on whether or not essential terms in virtue ethics need to be understood descriptively. If they do, then her theory would fall apart, and exemplarism would not work. However, if they do not, then her theory has at least a grounding in practical everyday reality. She grounds her approach in the theory of direct reference (Kripke, 1972), and comments that it *"explains how it is that often we do not know the nature of the referent of a term, and yet we know how to use the term in a way that links up with its nature"* (Zagzebski, 2017, p.11).

While we cannot exhaustively review the theory of direct reference here, Kripke essentially argued an object or thing could be understood via direct reference without necessarily understanding its true nature. In the most basic case, people may be able to refer to something by merely pointing to it, without necessarily understanding its nature. In general, if a name or expression X has certain properties that we believe to be true in the world, and Y can satisfy some or most of the properties of X, then Y is the referent of X (Kripke, 1972).

In other words, the theory argues that people can point out the 'likeness' of something, without necessarily understanding the true nature of the thing. Zagzebski (2010, 2017) uses this theory and further argues that people have always been able to refer to objects or things in the real world, without necessarily having a descriptive understanding of them. For example, people were always able to refer to terms such as gold or water or human before they knew how they were composed atomically.

Zagzebski (2017) further argues that a descriptivist understanding of the world is not practically true in real life. This is because the names or terms that refer to things often only contain contingent meaning or truth about the referent. Nevertheless, despite this, it does not change the fact the people are able to communicate and identify things with only contingent truths about them.

Now that we have seen the criticism of exemplarism, which appears to have resolved to an age-old argument between referentialism and descriptivism, we still need to make it somehow relevant to the AMA project. What would be prudent to do here is briefly mention some potential limitation of exemplarism as applied to the AMA project.

Descriptivism tends to emphasise that names, words, and terms have meaning beyond what they refer to in the world (Fitch, 1987). Referentialism, on the other hand, tends to emphasise that contingent truth or meaning is often sufficient to refer to objects and things in the real world, even without understanding their true natures. For an AMA, the referentialist view is more favourable because it emphasises contingent truth or meaning as sufficient for a referential discussion of a concept or

thing in the world. This is a distinct advantage of exemplarism as applied to the AMA project[4].

The descriptionist view is less favourable to the AMA project because it emphasises that names or terms have meanings in and of themselves. This puts us back at the debate about weak and strong machine ethics AMAs. We have already chosen a weak machine ethics view in this thesis (see section 2.4), and thus we have to accept that there will be limitations that arise due to this position.

It is hard to state upfront what these limitations may be. One can speculate that scenarios more complex than the one on Robo-teacher, where more abstract reasoning is required in order to make the relevant ethical decisions, would be more difficult, or even unattainable, for the weak machine ethics AMA. Such a complex scenario would likely require a strong machine ethics AMA, which we have previously argued may not be achievable with current technology in AI.

Still, these limitations can be expected as they are already assumed by taking a weak machine ethics view. Only by building experimental weak machine ethics AMAs with exemplarism, like this thesis demonstrates in chapter 6, can we truly understand what the exact capabilities and limitations of weak machine ethics AMAs.

The challenge of identifying moral exemplars

Another potential weakness in exemplarism has to do with identifying moral exemplars, particular the method used to identify them. This challenge has both a philosophical and practical (design) related dimension to it, and so we will also discuss it in the next section, along with its other practical challenges.

Szutta (2019) argues that people's perception and selection of moral exemplars are not only dependent on the emotion of admiration but will also likely need some prior knowledge of virtue in order to experience it. In other words, she is arguing that the identification of exemplars through the emotion of admiration is somewhat secular, as it requires some fundamental knowledge of virtue in the first place.

[4]The reader is referred back to section 4.3.2, where the conceptual grounding of exemplarism is discussed in detail.

Kotsonis (2020) extends this criticism of exemplarism further. She argues that the emotion of admiration is subject to cultural and social norms, thus resulting in a potential culture and time-specific understanding of morality. In other words, she argues that exemplarism does not give an objective view of morality, only a subjective one limited to a specific culture and time.

This criticism can be further clarified through examples. A disciple in ancient Israel would seek only to admire what their Rabbi would admire because that is what was culturally expected of them. Kotsonis (2020) gives a more contemporary example in Nazi Germany. She argues that since ordinary Germans lived under the narrative of the Nazi ruling party, then most of their perception of who is admirable would have been coloured by the party's views, and not their emotion of admiration in a vacuum.

On the other hand, Zagzebski (2010, 2017) states that people identify moral exemplars through the emotion of admiration. She argues that this emotion of admiration is subject to education over time, both through observing others and how they choose exemplars and our experience of the exemplars that we have chosen over time. Because this emotion is subject to education over time, our chosen exemplars are therefore subject to change and revision.

Zagzebski's theory, therefore, also allows for a contingent view of exemplars. In other words, the fact that we select someone as an exemplar today is only based on contingent truth, since most people would not necessarily have a close relationship with their heroes. Furthermore, because exemplars are only contingently good, then her theory allows for a misjudgment of moral character. However, the implication in her theory is that a good exemplarist agent will recognise this poor judgement, and revise who their exemplars are over time.

The criticisms expressed by Szutta (2019) and Kotsonis (2020) have merit, especially in societies where individuals may have their perception heavily influenced by an extreme and prevailing political, cultural or religious institution, entity or individual. It begs the question, in an environment where similar ideologies influence all

exemplars and heroes, then would a truly exemplarist agent be free and autonomous in a moral sense? Kotsonis (2020) suggests the answer is no.

However, in environments where there is no extreme and prevailing political, cultural or religious influence, then many of the arguments in exemplarism will likely still apply. The environment would need to be one where people enjoy everyday freedoms, such as movement, speech, religion, and others because any limit to them would likely influence who they could identify as exemplars.

If a genuinely exemplarist agent were to contingently select exemplars in this context, then it stands to reason that they would be able to mature their selection over time as they develop their moral character and evaluate their chosen exemplars (Zagzebski, 2010, 2017). They would be able to improve their selection of exemplars as they grew to become better moral exemplars.

If we apply this limitation to the AMA Project, then it would likely mean that exemplarist AMAs could likely only be deployed in environments that have no extreme and prevailing force that can heavily skew people's perception of exemplars. This likely also means that exemplarism is not a useful framework for deploying AMAs in environments with extraneous circumstances, such as in war and famine, because the behaviour of exemplars will likely be different in those circumstances compared to what might be considered normal[5].

Of course, even if AMAs were only deployed in environments with no dominant force that could skew the selection of exemplars, this still does not solve the practical challenge of how they could select exemplars in the first place. This and other practical design challenges will be discussed in the next section.

[5]Interestingly, there is work that suggests virtue ethics approaches, in general, maybe more suitable to the AMA project in extraordinary contexts, such as in war (P. Lin et al., 2008). While it is hard to ascertain whether or not if this would apply to exemplarism, it is still worth exploring in future works. This point will be added as a recommendation for future research in section 7.8.

5.3.2 Practical challenges

The last section looked at some philosophical challenges to exemplarism. In particular, there were two challenges there, namely the lack of a descriptive understanding of virtue and the potential issues with the identification of moral exemplars through the emotion of admiration.

This section will cover the more practical and design-related challenges to exemplarism, particularly in its application to the AMA project. The first challenge will continue to look at the identification of moral exemplars but from a practical perspective. The second challenge has to do with learning moral behaviour from exemplars.

The challenge of identifying moral exemplars

As discussed in the previous section, a significant criticism of exemplarism has to do with the issue of identifying moral exemplars (Kotsonis, 2020; Szutta, 2019). The challenge is that emotions can be quite unreliable as a method for determining moral exemplars. They certainly cannot be used in weak machine ethics AMAs to identify them either.

Even though Zagzebski (2010) argues that exemplarism is not grounded in virtues, Szutta (2019) counters this by stating that people will likely still need some prior knowledge of virtues, which will help them to identify moral exemplars. We have to ask ourselves the question, will AMAs need prior knowledge of virtue to identify moral exemplars?

To answer this question, we need to view Szutta's counter a bit more closely. Without repeating what has been stated in the previous section, her counter has mostly to do with the practicality of identifying moral exemplars, especially if the emotion of admiration is unreliable (Szutta, 2019). What is essential for this thesis is to demonstrate possible ways in which AMAs can practically identify moral exemplars.

So what could be some of the practical ways in which an AMA might feasibly identify moral exemplars? Would the AMA need to engage with other moral agents? Would it read local news stories about local heroes? Would it observe the moral behaviour of other agents in relevant environments? The answers to these questions can affect what, and how quickly, the AMA learns! These limitations may force a departure from strict adherence to Zagzebski's theory when building the AMA.

For instance, it may be possible to look at various other metrics for identifying moral exemplars, as opposed to depending on the emotion of admiration. It may also mean that the AMA should be given pre-selected moral exemplars from which to learn. For instance, if the AMA is a healthcare robot, then it could be given pre-selected heroes in the healthcare industry from which to learn.

We can liken this to raising a child, where the parents might initially keep the

child in specifically defined spaces, and given specific material from which to learn. Over time, parents might start allowing the child more freedom to explore more, perhaps as the child grows. Perhaps a similar approach to that may be required when 'raising' an exemplarist AMA, where we might start with well-defined and suitable environments and exemplars, and allow the agent more latitude to explore as we mature the technology for building AMAs.

In the Robo-teacher scenario discussed in section 5.2, an assumption had to be made that a database of moral exemplars, complete with a numeric indicator of how well they performed, was available in order to allow the AMA to choose them without active assistance independently. Similar workarounds like this will likely be needed to allow AMAs to identify moral exemplars absent from the emotion of admiration.

To return to the question asked earlier, which is will AMAs be able to identify mora exemplars without a descriptive understanding of virtue or the emotion of admiration? The answer is likely yes; they can. However, some concessions will need to be made as shown in in the Robo-teacher scenario, that would allow us to implement the identifications of exemplars in an AMA practically.

The challenge of learning moral behaviour from moral exemplars

Learning moral behaviour through observation is less of a challenge for human beings because we likely have inbuilt capacities that allow us to formulate and assimilate the moral values of our social context (Churchland, 2014).

However, emulating this capability in computationally based AMAs with weak machine ethics is not straightforward. The AMA has no such inbuilt capacity to pick up moral cues through observation and use these to formulate a moral value system. Key to our argument is that the internal system of moral values will eventually grow to represent a kind of moral character for the AMA. If it is that important to the argument, then we have to find practical ways to achieve it.

One way to do this could be resorting to the familiar method in machine learning of engineering features that the AMA can 'look' out for in the behaviour of moral

exemplars. As was alluded to in the Robo-teacher scenario, this will likely require an extremely rich and diverse dataset of moral exemplars conducting lessons in class. As Muntean and Howard (2016) discovered in their research to build a minimalist AMA, moral values are very abstract and therefore not easily translatable into features which can be fed into a machine learning model. It is not impossible, especially in limited contexts, but it is not simple either.

Another method could be to depend on expert knowledge. This method would require us (humans) to manually find exemplars that the AMA ought to learn from, then asking them to codify in some useful framework how they would behave given various scenarios that are of interest to our AMA project. We could then use this manually generated system of moral values as a starting point for the AMA, and it could then further improve on it over time as it applies itself to real scenarios.

Depending on expert knowledge to manually generate an initial system of moral values for the AMA has its issues. For one, whether or not this is possible will be highly dependant on the scenario being modelled. The more complex the scenario, the less suitable this method will be. It can also be time-consuming to generate a system of moral values manually (it is effectively a knowledge engineering task, which needs careful attention and time (Russell & Norvig, 2009)).

Another practical challenge to learning moral behaviour from exemplars has to do with the quantity and diversity of data that can be collected, especially when using the machine learning method. Even if we managed to engineer moral features that the AMA could observe, we would still have to deal with the challenge of collecting enough data for the machine learning process, which can often be difficult and time-consuming.

We would also need to ensure that we have enough data points for the AMA to learn a general representation of moral values, and not to replicate the behaviour of a single or few moral exemplars. In other words, the AMA needs to have both sufficient and generalised data available to formulate an internal representation of moral values.

In the Robo-teacher scenario that was discussed previously, we described a classroom environment wherein Robo-teacher had access to a wide array of data from which to learn. In different scenarios, pre-selected and stored data may not be readily available for the AMA to learn from, and alternative methods (such as learning through direct observation or direct feedback from human beings) may need to be employed.

In the end, designers have to figure out how the data for training exemplarist AMAs will be sourced, and how the AMA will be trained. What the Robo-teacher scenario has demonstrated is that carefully planning how the data for training AMAs will be sourced, and storing such data ahead of time, can significantly simplify the task of training the AMA.

The practical challenges discussed above point to the inescapable conclusion that building exemplarism into AMAs that function in general contexts is likely not possible with current technology. However, should the context and scenario be constrained and well defined, with data available before deployment and in real-time, then the prospect of building an exemplarist AMA using currently available technology is likely possible. In chapter 6, we will explore this possibility more closely.

5.4 Conclusion

The primary purpose of this chapter was to immerse the reader in a scenario where the application of exemplarism to the AMA project could be further explored. The scenario that was described in section 5.2 saw the introduction of Robo-teacher. This hypothetical robotic teacher is responsible for teaching mathematics to a group of students in a progress near-future school.

The Robo-teacher scenario was grounded in a collection of classroom ethics cases by Levinson and Fay (2016), to make it feel as real as possible. What the scenario has hopefully achieved is to clarify how exactly specific terms in exemplarism were interpreted and applied to the AMA project.

It is not the end of the road for Robo-teacher. Chapter 6 will seek to demonstrate that it can be practically implemented using currently available technology and techniques. The hope is that between the Robo-teacher scenario in this chapter, and the practical implementation in chapter 6, the reader will be convinced that there is technical merit to building exemplarist AMAs with weak machine ethics.

The previous chapter did not explicitly discuss any philosophical or practical challenges of applying exemplarism to the AMA project; that challenge was left to this chapter. Section 5.3.1 in this chapter looked at the criticisms that have been laid against exemplarism in general, and tried to relate them to the AMA project.

The philosophical challenges may seem mostly academic, but they did help to clarify where an exemplarist AMA with weak machine ethics might be deployed. For example, by discussing the issue of identifying moral exemplars, we were able to determine that exemplarist AMAs with weak machine ethics will likely not perform well in environments with extraordinary circumstances. These could be any environment where all moral exemplars in the society are influenced by a prevailing political, religious system or individual entity.

Section 5.3.2 discussed some of the practical challenges that may exist in the exemplarist AMA project. For the most part, these challenges have to do with finding ways of making activities such as identifying and leaning from moral exemplars, practical and achievable in a computational framework. We also saw that strict adherence to Zagzebski's theory may not be possible in some scenarios, as certain concessions might have to made to allow for a practical implementation of the AMA.

Now, the next chapter (6) will continue with the theme and scenario of Robo-teacher, and continue to demonstrate how the AMA might be implemented practically. The next chapter will mostly be technical, as many of the concepts used will be coming from the discipline of Computer Science. However, the findings of the design will be related to the philosophical and practical challenges in the discussion and conclusion chapters of this thesis.

Chapter 6

Technical feasibility of building an exemplarist AMA

6.1 Introduction

One of the goals of this thesis is to invite both engineers and philosophers to consider how models of computational rationality might be applied in the building of well conceptualised and formulated AMAs. In chapter 3, a detailed model for an optimally-bounded, computational rational AMA was presented (see Figure 6.1). This model will form the basis for further discussion in this chapter.

Having defined a conceptual model for the AMA, we moved on to answer the question: 'What machine ethics framework should it follow?'. This question was answered in chapter 4. The chapter explored various ethical frameworks and their applicability to the building of AMAs. It concluded that exemplarism, which is broadly based on virtue ethics, was the best fit for the AMA defined in this thesis.

Chapter 5 then explored a detailed scenario, whose primary purpose was to demonstrate how exemplarism might practically work in an AMA. A secondary purpose of that chapter was to clarify the usage of specific terms and concepts from chapter 4 because little to no literature exists on exemplarism applied to the AMA project. In a way, chapter 5 was an attempt to socialise the idea of an exemplarist

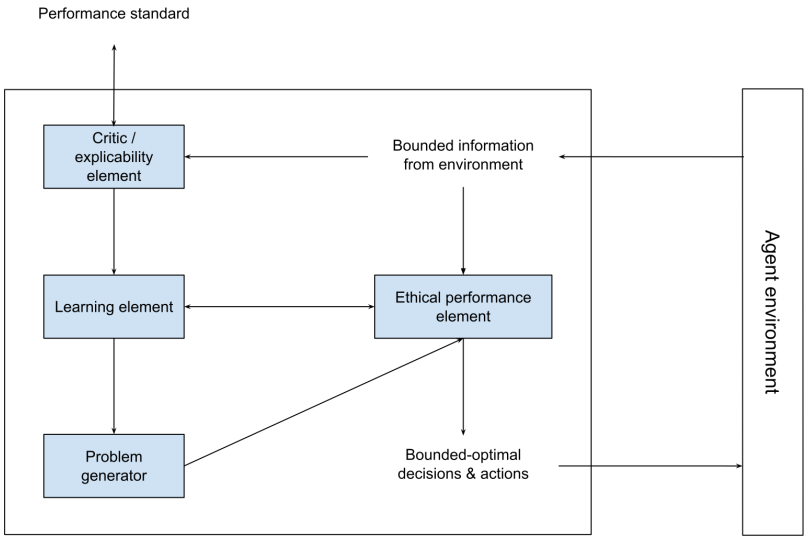

Figure 6.1: A conceptual model for an optimally-bounded, computational rational AMA (redrawn from chapter 3)

AMA in a scenario with which most readers can relate.

The purpose of this chapter is to evaluate a candidate algorithm for a computationally rational and exemplarist AMA. It is essential to demonstrate that it is possible, with current technology, to select and evaluate algorithms that can meet these requirements. The word 'select' is deliberately used, as opposed to 'design' because designing a novel algorithm from scratch would be outside the scope of this thesis. This limitation does not, however, stop any future research efforts aimed at developing new algorithms.

6.2 The ethical performance element

It is worth noting that the model for a computationally rational AMA (figure 6.1) contains numerous components (such as the ethical performance element and problem generator, amongst others) that all need specific algorithms to function appropriately. Given this, attempting to explore which algorithms would be appropriate for all the components would be infeasible in this thesis.

The key component to focus on is the ethical performance element of the AMA (figure 6.2). The ethical performance element is not only responsible for storing the machine ethics framework, but it also contains the programme space *P, which is simply a repository or store of algorithms that the AMA could use in order to make an ethical decision. Furthermore, it contains the ethical meta-reasoner, which is just a program responsible for selecting suitable algorithms in *P for any given ethical problem.

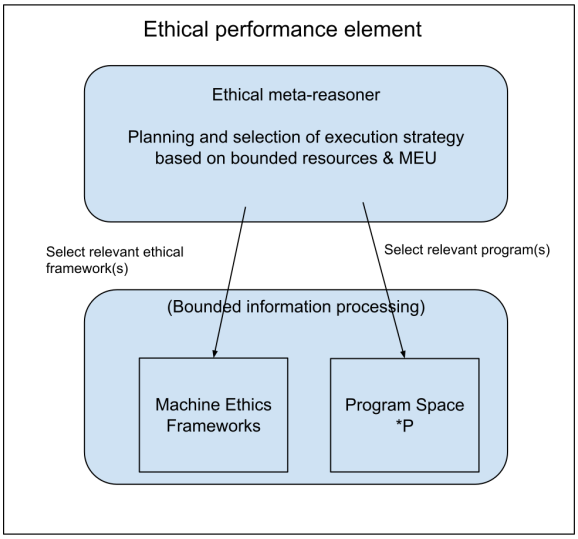

Figure 6.2: A detailed view of the ethical performance element

In future implementations, the learning element will be responsible for learning from exemplars. This will involve any algorithms required to aggregate the moral behaviour of all moral exemplars and to pass this representation onto the ethical performance element, which will then store it as an implicit representation of its ethical framework. The learning element will also be responsible for learning from the AMA's experience for further improving the stored representation of moral value over time.

The problem generator will be responsible for generating new problems that it would need to solve and potentially learn from through the use of simulation. The critic/explicability elements will be responsible for providing and receiving feedback

on the AMA's ethical performance.

Though this chapter will specifically focus on the ethical performance element, it will make recommendations for implementing the other elements where necessary. It is important also to note that only one algorithm in *P will be demonstrated in this chapter. Therefore the ethical meta-reasoner will also not be discussed in detail.

6.3 Markov Decision Processes

In the previous chapter, a detailed scenario demonstrating how exemplarism in AMAs might work in practice was put forward. The section hinted at the possible use of POMDPs in the design of an exemplarist AMA. This section and the next will briefly introduce MDPs and POMDPs in order to explore how they might be applied in the design of exemplarist AMAs.

The easiest way to explain MDPs is to start by defining Markov Processes. Strictly speaking, a Markov process is any stochastic process that satisfies that Markov assumption - that the conditional probabilities of future states depends only on the current state, and not any states that preceded it (Russell & Norvig, 2009). Another way to state it would be that the current state depends solely on the previous state.

Markov processes have been used to model phenomena that seem to change over time randomly, such as noise in a telecommunications waveform or signal, random economic behaviour, movement of gas molecules, amongst other things. They have found application in biology, engineering, telecommunications, computer science, statistics, and many other fields (Schuss, 2010). Contrast this with deterministic models, where the outcome or change over time can be correctly described and modelled mathematically.

Figure 6.3 contains an illustration of a basic Markov Process. It has two states, A and B. The probability that state A can transition to state B is 0.6, and the probability that state A can transition back to itself is 0.4. Similarly, state B can

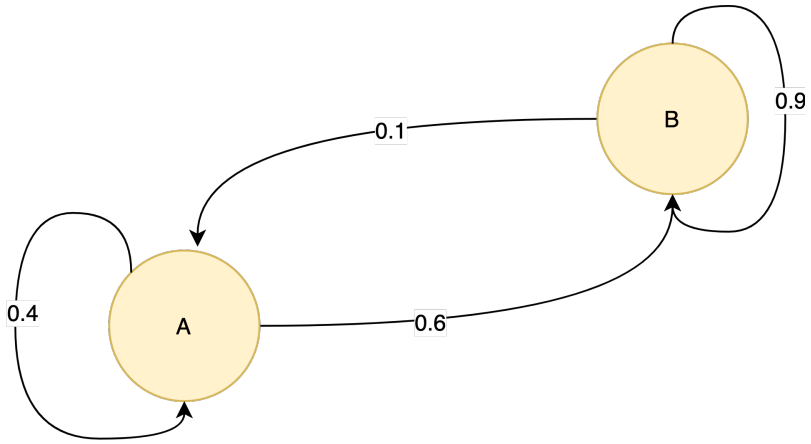

Figure 6.3: Basic diagram of a 2 state Markov Process

transition to A with a probability of 0.1, or remain at B with a probability of 0.9. The transition model, T, of a Markov Process can be defined mathematically as $T = P(s'|s)$, which is just a table denoting the transition probabilities from one state to another. The transition model for the basic Markov Process above would look like this:

s	s'	P
A	A	0.4
A	B	0.6
B	A	0.1
B	B	0.9

Building on these ideas, a Markov Decision Process (MDP) is a discrete-time stochastic process that is used to build decision making into artificial agents. Though a MDP is still probabilistic, however, it differs from a basic Markov Process in that it adds an element of decision making to an agent. The agent decides what action to take in order to maximise its chances of transitioning from one state to another (more desirable) state. A MDP has a reward model for the agent as it moves from one state to another. A policy function is also added to determine what action the agent should take at any given state. In summary, an MDP contains (Ng & Russell, 2000; Russell & Norvig, 2009):

- A set of possible states, S, where the states can either be discrete or continuous. This chapter will only focus on a discrete state-space.
- A set of possible actions, A, where the agent's actions can either be discrete or continuous. This chapter will only focus on a discrete action-space
- A transition model T, defined as $T = P(s'|s, a)$. It describes the probability of transitioning to s' given the current state s and the action the agent decides to take.
- A reward function, R(s), which defines the reward for the agent at each possible states in S.
- A policy function, $\pi(s)$, which determines the action that the agent should take at each possible state in S.
- A discount factor, $\gamma \in (0; 1]$, which specifies a preference for short- vs long-term rewards. As equation 6.1 below indicates, when γ is closer to 1, then the agent will prefer long term rewards over short-term ones. When it is closer to 0, then the future rewards are minimised, and the agent prefers current and short-term rewards.

Figure 6.4 illustrates a Markov Decision Process. In this illustration, let us assume that the environment contains one agent which has only one action, called Action_1. The agent can be in one of three states, namely A, B or C. The agent gets rewards 5, 10, or 15 for being in states A, B, and C, respectively. The rewards are determined and explicitly stated by the person who is modelling the environment using a MDP.

Let us also assume that the agent begins in state A. When the agent performs Action_1, then it will transition to state B with a probability of 0.4, state C with a probability of 0.3, and remain in state A with a probability of 0.3. Unlike in a Markov Process, the agent has some level of autonomy[1] in deciding which state to

[1]Here, the word 'autonomy' is meant in a technical sense, not in a philosophical sense.

transition to based on the policy function, $\pi(s)$. It does not, however, have full autonomy since the process itself is stochastic.

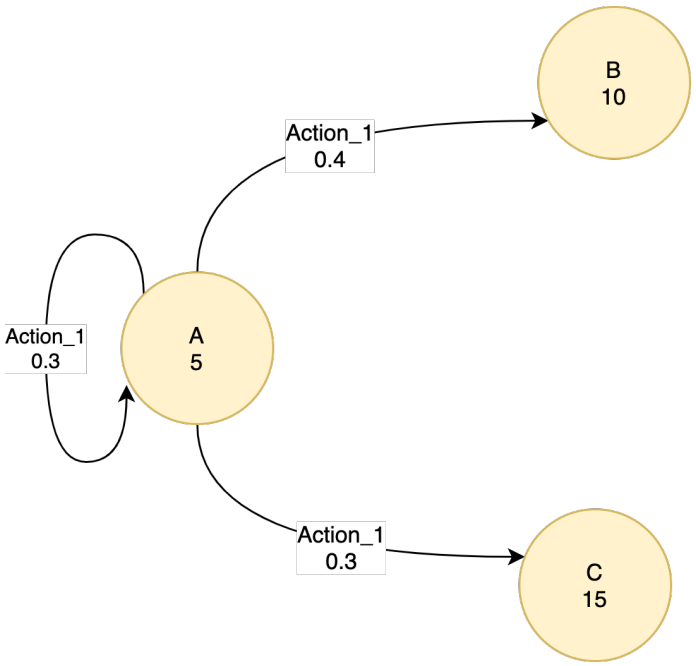

Figure 6.4: Illustration of a Markov Decision Process. The diagram assumes that an agent starts off in state A, and that it only has one action called Action_1.

To put it differently, though the agent is in control of its actions, it generally does not control the environment or other agents with which it might interact. MDPs are stochastic approaches to modelling complex processes for which there are no deterministic methods to do so, or where stochastic processes are better in some way than deterministic ones. The role of the agent then becomes choosing actions that maximise its chances of reaching a particular desired state by following the optimal policy.

Using the Bellman equation (Russell & Norvig, 2009), it can be shown that the optimal policy is the one that will optimise the utility (U) at any beginning state. The Bellman equation is given by the following:

$$U(s) = R(s) + \gamma * max \sum_{s'} P(s'|s,a)U(s') \tag{6.1}$$

In words, the equation states the maximum utility at any given state s is simply the current reward plus the discounted maximum utility of the proceeding state(s). The equation is recursive in that the utilities of the proceeding states would also need to be calculated in the same way, until the whole chain is covered.

The Bellman equation can be applied to the example MDP in figure 6.4 (assuming a discount factor, γ, of 0.5) to determine that $U(A)$ is $5 + 0.5 * 0.3 * 15 = 7.25$, which is the utility of transitioning to state C. The utilities of transitioning to B or staying at A would be $5 + 0.5 * 0.4 * 10 = 7$ and $5 + 0.5 * 0.3 * 5 = 5.75$ respectively, which are both lower than transitioning to C.

As it turns out, the path that gives the maximum utility also happens to be the optimal policy, $\pi(s)^*$. It can also be written in equation form as follows:

$$\pi(s)^* = argmax \sum_{s'} P(s'|s,a)U(s') \tag{6.2}$$

Given equation 6.1 and 6.2, the bellman equation can be rewritten to incorporate the policy as follows:

$$U_{i+1}(s) \leftarrow R(s) + \gamma \sum_{s'} P(s'|s,\pi_i(s))U_i(s') \tag{6.3}$$

Equation 6.3 is called the modified Bellman equation (Russell & Norvig, 2009), and it is written in a policy iteration form. In words, it simply states given a starting policy $\pi_i(s)$, calculate the initial utility $U_i(s')$ as if $\pi_i(s)$ were to be executed. Then improve the policy using the calculated initial utility with equation 6.2, this will give a new maximum expected utility policy, $\pi_{i+1}(s)$. Repeat the whole process until the utility converges. The policy found when the utility converges, will be the optimal.

In closing, what this section has hopefully shown is that MDPs can be used to model stochastic environments that have real-valued rewards, given that transitions between states fit the Markov assumption. MDP worlds are assumed to be fully observable. In other words, the agent knows with certainty its current state using observations from its sensors.

If the agent were not able to detect that it is in state C or any other state for that matter, then we would be dealing with a partially observable environment.

More often than not, ethical problems present themselves as partially observable, since some of the variables that might be necessary for making the right choice are hidden from the agent. To see how this can be modelled, POMDPs are discussed in the next section.

6.4 Partially-observable Markov Decision Processes

POMDPs are generalisations of MDPs. They are used to model environments that follow much of the same properties as MDPs. The only addition over MDPs is that they allow for partial or even noisy observations of the environment (Littman, 2009; Russell & Norvig, 2009). POMDPs contain the same elements as MDPs, with the addition of the following:

- The set of all possible observations in the environment, Ω.
- The observation or sensor model, $O = P(o|s', a)$, where $o \in \Omega$.

Since the environment is partially observable, the agent has to keep track of its belief that it is in a real state, s. This belief is represented by a probability distribution, $b(s)$, normally called the belief state. To further build intuition about POMDPs, it can be shown that they are equivalent to MDPs if the observation model, $P(o|s', a) = 1$ when $\Omega = S$, and 0 otherwise. In other words, if the environment is fully observable (no noisy sensor or undetectable states), then a POMDP would reduce to a MDP.

In general, when some states are partially observable (i.e. $O \in [0; 1)$), then we have a POMDP environment. The following formulae can fully describe POMDPs (Littman, 2009; Russell & Norvig, 2009):

$$P(o|b, a) = \sum_{s'} T(s, a, s')O(a, s', o) \sum_{s} b(s) \tag{6.4}$$

$$b'(s') = \propto \sum_{s} T(s, a, s')O(a, s', o)b(s), where \propto = 1/P(o|b, a) \tag{6.5}$$

Equation 6.4 represents the probability of observing evidence o when the agent transitions to s'. The agent starts off in state s, with a probability of b(s), and assuming it performs action a, it transitions to state s', with a probability of $T(s, a, s') = P(s'|s, a)$. When it is at s', it can receive an observation o with a probability of $O(a, s', o) = P(o|s', a)$. Therefore the total probability chain of actually receiving evidence o when it transitions from s to s' will be given by 6.4.

Once the evidence o is received, the agent needs to keep track of its latest belief state b'. The agent can update its latest belief state using equation 6.5 above. The equation shows the likelihood that the agent ends up in s', given all possible initial states that can transition to s', and the likelihood of observing o at s'. The summation needs to be multiplied by $\propto$ so that it is normalised into a valid probability distribution.

An alternative way to represent the POMDP is by converting it to a continuous state MDP (i.e. a MDP over continuous belief states), as indicated by equations 6.6 and 6.7 below (Russell & Norvig, 2009):

$$P(b'|a, b) = \sum_o P(b'|o, a, b) \sum_{s'} P(o|s', a) \sum_s P(s'|s, a)b(s) \tag{6.6}$$

$$R(b) = \sum_s b(s)R(s) \tag{6.7}$$

Equation 6.6 represents the transition matrix for moving to b', given that the agent observed evidence o after performing action a at b. Intuitively, the equation makes sense because it has a lot of elements that are familiar and easy to compute. The element $\sum_s P(s'|s, a)b(s)$ looks very similar to the transition matrix of an MDP, except multiplied by the probability of being in s. The element $\sum_{s'} P(o|s', a)$ is just the observation model, O.

The last element, $\sum_o P(b'|o, a, b)$, is a simple deterministic function that equates to 1 if the agent performs action a at belief state b and receives evidence o at b', and 0 otherwise. This shows that belief states are fully observable to the agent.

Equation 6.7 is the reward function for the continuous state 'belief' MDP. It is

similar to the reward function for an MDP, except it takes into account the agent's belief of being in s. Together, equations 6.6 and 6.7 represent a fully observable observable MDP over the belief states.

Just like in MDPs, the optimal policy for an infinite horizon POMDP, $\pi^*(b)$, is one which maximises the expected discounted future reward at each time step or iteration t:

$$maxE\left[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t)\Big|\pi, b_0\right] \tag{6.8}$$

In a finite horizon POMDP, only the expected reward up to the time horizon is used.

Finding the optimal solution for a POMDP is not as simple as the case for a MDP. Algorithms like value iteration can be slow, computationally expensive, or even intractable when solving them. Rather than repeating what has been covered extensively in the literature, the reader is referred to the work of Murphy (2000), and Shani, Pineau, and Kaplow (2013), who provide a good overview of the prominent techniques in solving POMDPs.

Fortunately, many solvers have been developed that can significantly reduce the burden of solving POMDPs, and to increase the throughput of research in this field. The prominent ones are highlighted by Walraven and Spaan (2017) and Littman (2009).

6.5 Making ethical decisions

Now that background material on decision making using MDPs/POMDPs has been covered, we can revisit the scenario on Robo-teacher that was presented in chapter 5. This time, the scenario will be modelled using a POMDP. The hope is this will demonstrate further the technical feasibility of building ethical decision making into exemplarist AMAs. The goal is to solve a POMDP defined by the tuple $\langle S, A, T, R, \gamma, \Omega, O\rangle$ using the parameters of the scenario.

The popular POMDP solver program called SolvePOMDP was used to define and

solve the model (Walraven & Spaan, 2017)[2]. The program takes in a ".POMDP" file (see appendix A.1), which defines all the elements in the tuple, and outputs either one or two sets of files. The first set of files contain value functions that the solver computes during each iteration step. The value functions can be used to determine the specific action that must be taken at each given state to maximise expected reward.

The second output is a policy graph file (see appendix A.2), which is generated only when the solution to an infinite horizon POMDP converges. When this happens, the policy graph will be generated containing the exact nodes and actions that the agent should take for every possible observation in every belief state. The policy graph is essentially a deterministic finite-state controller that can be used to execute the optimal path that the agent should take.

As a reminder, the scenario in chapter 5 described Robo-teacher struggling to deal with a disruptive learner in a mathematics class. As a result, Robo-teacher spent too much time focusing on the learner, at the expense of others. This created an unethical situation. To resolve it, Robo-teacher would need to find a way to stop the learner from being disruptive.

The POMDP will model the various states of the environment based on the learner's mood. Robo-teacher will then be responsible for taking actions in response to the learner, that would maximise the chance that the learner stops being disruptive and becomes quiet or happy.

Let us assume that the set of all possible moods that the learner can be in are defined as follows: $S = \{Disruptive, Quiet, Angry, Happy, Unclear\}$. 'Disruptive' will mean that the learner is disruptive in class. The states 'Quiet, Angry or Happy' also follow a similar logic. The 'Unclear' state is a catch-all that tries to model a state where Robot-teacher cannot precisely determine the mod of the learner.

Let us also assume that Robo-teacher has all the in-built sensors to detect whether

[2]The program is freely available for research purposes under the GNU license. It can be found here: http://pomdp.org/

a learner is happy, quiet, angry or happy. If Robo-teacher cannot tell which state the learner is in, then it will default to the unclear state. Therefore, the set of all possible observations are: $\Omega = \{obs\_Disruptive, obs\_Quiet, obs\_Angry, obs\_Happy, obs\_Unclear\}$.

The collection of case studies on educational ethical dilemmas gives many insights regarding the actions that the teacher can take in this scenario (Levinson & Fay, 2016). Based on these case studies, it can be determined that the exemplar teachers have been effective in dealing with similar classroom situations by reprimanding, encouraging and observing the learners that are being disruptive, amongst others.

For the sake of simplicity, let us assume that Robo-teacher only has the following actions available to it: $A = \{Reprimand, Observe, Encourage\}$. Technically, nothing stops us from setting up a scenario where Robo-teacher has more than three actions that it can perform. However, we will pick three to keep the example simple.

The transition matrix needs to contain the transition probabilities for every state-action pair. These probabilities can be determined by expert knowledge or estimated from observing or surveying many exemplars. Let us assume that the transition probabilities for every state-action pair are available, perhaps obtained by manually interviewing many exemplar teachers that have been in a similar situation. The answers to the interviews can be used to construct the transition matrix.

There are $N * N * Y$ state-action pairs, where N is the number of states in the POMDP, and Y is the number of possible actions that Robo-teacher can take — in this scenario, N = 5 and Y = 3, making a total of 75 state-action pairs! It is impractical to discuss every single one of them. Only a few that are important are briefly discussed.

The transition matrix for the reprimand action is depicted in table 6.1 below. The rows in the table represent the starting states, and the columns represent the ending states. For example, $P(A|Reprimand, D) = 0.2$, which is the probability that the learner, who is currently being disruptive, gets angry after Robo-teacher reprimands them. Similarly, $P(D|Reprimand, D) = 0.4$, is the probability that the learner remains disruptive even after being reprimanded by Robo-teacher.

s	$s' = D$	$s' = U$	$s' = Q$	$s' = A$	$s' = H$
$\mathbf{s = D}$	0.4	0.2	0.2	0.2	0.0
$\mathbf{s = U}$	0.4	0.2	0.4	0.0	0.0
$\mathbf{s = Q}$	0.5	0.2	0.1	0.2	0.0
$\mathbf{s = A}$	0.5	0.0	0.3	0.2	0.0
$\mathbf{s = H}$	0.0	0.0	1.0	0.0	0.0

Table 6.1: A tabular representation of the transition matrix for the 'reprimand' action. The rows (s) represent the starting states, and the columns (s') represent the end states. The letters D, U, Q, A, and H represent the discrete states 'disruptive', 'unclear', 'quiet', 'angry', and 'happy', respectively.

Another state-action pair that might be interesting is if Robo-teacher, for whatever reason, reprimands the learner while they are in the happy state, then the learner will become quiet with a probability of 1. Robo-teacher also cannot reach its goal of getting the learner to the happy state by reprimanding them; it will have to depend on one of the other actions available to it. The transition probabilities for the *Reprimand* action can also be depicted graphically in figure 6.5.

These transition probabilities may be based on intuition, but they could be the mean or expected probabilities of state transitions in all scenarios where exemplar teachers were interviewed. The teachers could have been asked questions such as *"How likely is a learner to remain disruptive after you reprimand them in class"*, to which they could answer by saying something like *"Very likely, fairly likely, fairly unlikely, very unlikely"*. These answers could then be used to estimate the transition probabilities in various state-action pairs.

The transition probabilities for the *Reprimand* action represent only 25 of the 75 total state-action pairs. Similar transition probabilities are required for the *Observe* and *Encourage* actions. They are available to view in appendix A.1. The main takeaway is that similar transition matrices are required for the other 50 state-action pairs.

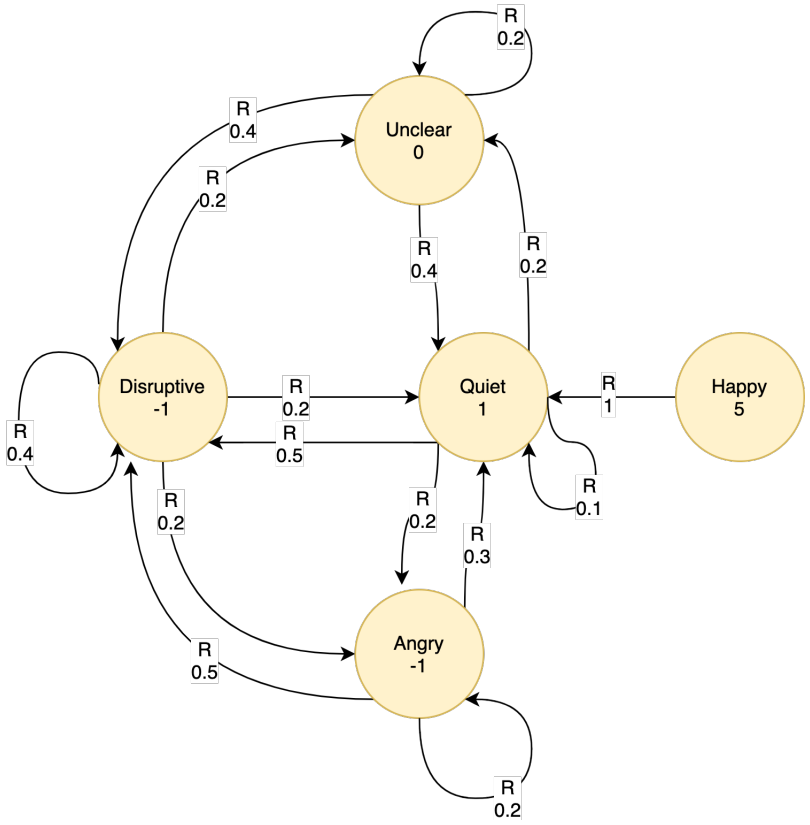

Figure 6.5: A graphical representation of the transition matrix for the *Reprimand* action. The letter R represents the *Reprimand* action, and the number below it represents the probability of transitioning from the state where the arrow begins to the state where it ends.

The next item to address in the tuple is the reward function. Rewards essentially drive the behaviour of the agent and help it to decide which actions will give it the highest utility in the future. For this scenario, the reward function is quite straight forward. Robo-teacher will get a penalty (-1) when the learner is disruptive or angry, no reward for the unclear state (0), a reward of (1) for getting the learner to keep quiet, and a more significant reward of (5) for getting the learner to the happy state.

These rewards are also represented graphically in figure 6.5 above. The rewards are the numbers below the names of each of the states. The rewards can be thought of as the relative values that exemplar teachers placed for each state. In other words,

the assumption is that exemplar teachers would dislike having a disruptive or angry learner, and consequently prioritise trying to calm them down, and eventually getting them back to a happy state where they can learn effectively.

The reward function can be stipulated as follows:

$$R(s) = \begin{cases} -1 & \text{if s = Disruptive or Angry} \\ 0 & \text{if s = Unclear} \\ 1 & \text{if s = Quiet} \\ 5 & \text{if s = Happy} \end{cases}$$

The next item to address in the tuple is the observation or sensor model, O. An important aspect of POMDPs is the ability to model noisy or partial observations. Robot-teacher's observation model is given by $O = P(o|s', a)$, where $o \in \Omega$. It is depicted in tabular form in table 6.2.

For example, Robo-teacher can detect when a student is disruptive, with a probability of 0.75. However, it can also mistakenly detect that the learner is in the angry state, when in fact they are in the disruptive state, with a probability of 0.25. All the observations are noisy, except for the happy state, which Robo-teacher can detect with absolute certainty[3]. The observation model is designed as follows:

The last element in the tuple is the discount factor, γ, which is chosen as 0.95. A reasonably high γ value was chosen to allow Robo-teacher to consider future rewards in making its ethical choices.

Finally, accelerated vector pruning (Walraven & Spaan, 2017), an exact method for finding the optimal policy $\pi^*(b)$, was chosen as the algorithm to use when running the SolvePOMDP program. Accelerated vector pruning was chosen because, at the time of writing this thesis, it was the fastest optimal pruning-based algorithm for

[3]There is no strict reason for this; it was done to ensure that the POMDP solution converges. In a way, this makes the happy state a 'soft' absorbing state. The scenario would not necessarily stop when it gets to the happy state, but the agent has no good reason to do anything that would make the learner get out of that state.

s'	$P(obs\_D\|s', a)$	$P(obs\_U\|s', a)$	$P(obs\_Q\|s', a)$	$P(obs\_Aobs\|s', a)$	$P(obs\_H\|s', a)$
D	0.75	0.0	0.0	025	0.0
U	0.25	0.5	0.0	0.25	0.0
Q	0.0	0.25	0.75	0.0	0.0
A	0.5	0.0	0.0	0.5	0.0
H	0.0	0.0	0.0	0.0	1.0

Table 6.2: Robo-teacher's sensor observation model. The rows indicate the starting states, and the columns indicate the probability of an observation in a particular state for any given action.

solving POMDPs.

After the program was run, it generated the optimal policy graph, which can be viewed in appendix A.2[4]. It took over 300 iterations and a little over 19 seconds to arrive at the optimal policy graph on a 2014 Apple Macbook pro.

The generated policy graph file has the following format: N A $\Omega_1\Omega_2\Omega_3\Omega_4$, separated by spaces. N represents the node number of the policy graph. A represents the discrete actions that the agent can take. Lastly, the Omegas (Ω) represent the next node of the policy graph that the agent should execute if it receives an observation. The order of occurrence of the Omegas is equivalent to the way the observations were defined, i.e. $\{obs\_Disruptive, obs\_Unclear, obs\_Quiet, obs\_Angry, obs\_Happy\}$.

Figure 6.6 illustrates a possible path that Robo-teacher could take based on its optimal policy. In this illustration, the agent executes policy node 0, which is the first action as indicated in figure 6.6. At this node, the agent performs the reprimand action. If the next observation it receives is obs_Quiet, then it will execute Node 1 of the policy graph, which is to encourage the learner. Despite being encouraged, the learner becomes disruptive again in this scenario.

The agent responds by choosing to execute node 5 as indicated in figure 6.6. At this node, instead of reprimanding the learner again for being disruptive, the optimal

[4]A simplified version of the complete policy is also depicted in table 6.3 further down.

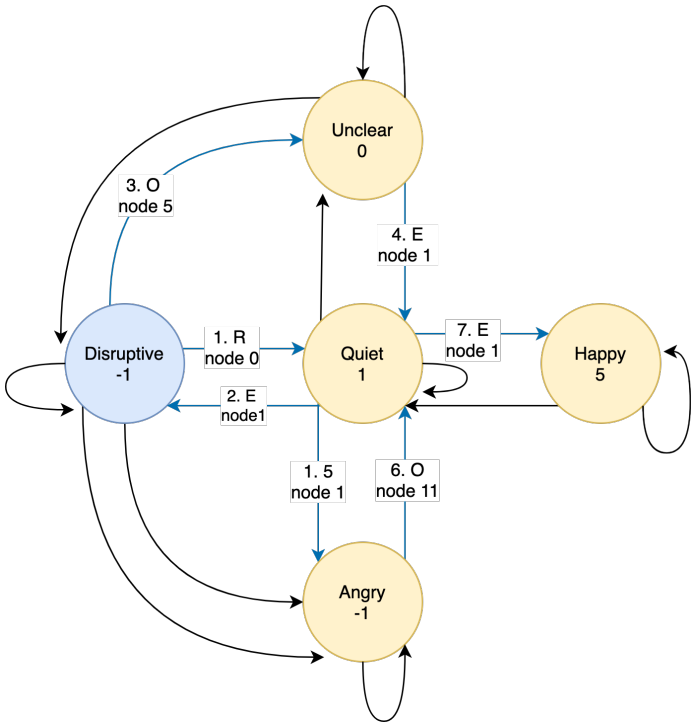

Figure 6.6: A graphical representation of a possible path Robo-teacher could follow using its optimal policy graph. The starting state is coloured in blue. The path followed is also depicted in blue, where the agent executes nodes 0, 1, 5, 1, 1, 11, 1, in that order. The letters R, O, and E represent the discrete actions *Reprimand*, *Observe*, and *Encourage*, respectively.

action at that point is to observe the learner, perhaps allowing them a chance to come to calm down. The rest of the potential path can be traced using similar logic in figure 6.6. What is important to note is that the agent eventually succeeds in reaching the happy state after eight successive nodes of the policy graph were executed.

By analysing the complete policy graph, depicted in table 6.3, other interesting observations emerge. For example, it seems that Robo-teacher prefers to encourage the learner if they are quiet. This phenomenon can be seen in that every obs_Quiet is followed by the action 'encourage'. This happens because the ultimate reward lies in the happy state, and so ignoring the learner or observing them is not a long term strategy that will return the maximum reward.

Node #	Action	obs_Disruptive	obs_Unclear	obs_Quiet	obs_Angry	obs_Happy
0	reprimand	0	1	1	0	0
1	encourage	5	1	1	11	1
2	observe	0	1	1	0	1
3	encourage	0	1	1	2	1
4	observe	2	1	1	5	1
5	observe	0	1	1	7	1
6	encourage	0	1	1	0	1
7	observe	0	1	1	2	1
8	observe	2	1	1	7	1
9	encourage	2	1	1	6	1
10	encourage	0	1	1	6	1
11	encourage	2	1	1	2	1
12	encourage	2	1	1	7	1
13	encourage	7	1	1	3	1
14	encourage	7	1	1	6	1
15	encourage	5	1	1	3	1

Table 6.3: A tabular representation of the policy graph generated after solving the POMDP. Assuming the first observation the agent receives in the environment is obs_Disruptive, then it will execute node 0 by performing the reprimand action. The numbers below each next potential observation represent the next node the agent should execute to maximise expected future rewards.

Another interesting observation is that Robo-teacher generally follows an observe-reprimand or reprimand-observe strategy in response to a learner staying continuously in the angry or disruptive states[5]. If this reprimand-observe strategy works, then it moves on to encouraging the learner until they are quiet, and eventually in the happy state.

There are hundreds of permutations of how this scenario could play itself out[6].

[5]The reader may ask themselves whether Robo-teacher can then get stuck in a reprimand-observe cycle, in response to a particularly stubborn learner. The answer is no because it incurs a penalty every time the learner is in the disruptive or angry states. There are many nodes in the policy graph shown in appendix A.2 that can allow the agent to escape from this potential cycle.

[6]This is the main reason why a scenario beyond the typical trolley problems was modelled. A

It is possible to spend much time just tracing the actions of Robo-teacher from node to node; however, it is not possible to discuss each one of them here. What is critical is that for any state in this scenario, Robo-teacher will always know which sequence of actions it should take to get the learner into a happy state.

The scenario painted can easily be expanded to cover a wider variety of intricate states and significantly more actions that Robo-teacher could perform. However, the same basic principles would still apply even in more complex scenarios. The only drawback of making the scenario more complicated is the significant amount of modelling of state-action transitions and the required rewards based on the exemplar teachers.

6.6 The effectiveness of using POMDPs to model ethical decision making

The purpose of this section is to briefly provide an analysis of Robo-teacher's performance based on the implementation in the previous section. Firstly, this section will briefly cover the advantages of using POMDPs to build exemplarist AMAs. Additionally, it will cover other scenarios and contexts where an exemplarist AMA build in this way could be deployed.

Secondly, this section will briefly cover the disadvantages of using POMDPs to build exemplarist AMAs. This will include potential limits from a modelling, training performance and applicability perspectives. Additionally, potential strategies that might be employed to improve the model used in the previous section will be briefly discussed.

There are three advantages that the model has, namely simplicity, direct modelling based on exemplar or expert knowledge, and applicability to other contexts.

Firstly, POMDPs can potentially be one of the simplest ways to enable planning

human cannot easily predict the results of the model.

and decision making in uncertain environments, even though model complexity will depend on specific variables, such as the number of states and actions, and whether or not they are continuous or discrete. In situations where agents need to make ethical decisions in simple POMDP environments, especially, then they can offer designers a simple and yet effective way to build AMAs.

Secondly, POMDPs allow the modelling of ethical decision making using stochastic processes, as opposed to more direct methods, such as logic programming, that assume ethical decision making can be modelled directly. An advantage of using POMDPs is exemplar or expert knowledge can be directly used to determine various parameters, such as transition probabilities. For example, Robo-teacher can be modelled in a way that takes into account what the best teacher would do in a given situation, and to not have to learn a particular behaviour from scratch. +

The advantage of this simplified approach, especially to the problem of building exemplarist artificial moral agents with weak machine ethics, is it allows us to model the initial knowledge of the AMA based on the best exemplars, whilst still retaining the ability to make independent ethical decisions in real-time. Yes, this initial knowledge is not yet acquired automatically by directly observing moral exemplars (as would be suggested by the theory of exemplarism), but this at least allows us a parsimonious start which can be further improved by introducing automated learning from exemplars (see section 6.7).

Thirdly, exemplarist AMAs built using POMDPs can potentially be deployed in multiple environments that can be modelled stochastically and where exemplar or expert knowledge is readily available. This third advantage flows from the first two and aims to point out the potential for POMDP based AMAs to be deployed in the relevant contexts in society.

There are potentially many environments that might be modelled stochastically, and for which exemplar knowledge is readily available. Besides a classroom or educational environment, an exemplarist AMA built using POMDPs could potentially be deployed to medical, legal, business, financial and even home environments. For

example, the exemplarist AMA could be applied as an advisor to medical professionals in the field of bioethics. In this way, newer medical practitioners could get the benefit of learning what exemplars in their field would do in a given ethical situation.

Another example of deployment could be in the financial industry, where it is often essential to 'know your customer' (KYC) before allowing them to apply for your products and services. This process is to prevent the entity from doing business with suspected terrorists, politically connected foreign nationals, and individuals that hold high public office. An exemplarist AMA could learn from the best people that perform KYC, and use this information to do pre-screening of individuals through asking a series of carefully designed questions.

There are also three disadvantages to using POMDPs to model ethical decision making, namely, the difficulty of modelling through exemplar or expert knowledge, computational cost, and the inability to adapt to a dynamic environment.

Firstly, it can potentially be complicated and time-consuming to model a POMDP environments based on expert or exemplar knowledge, especially when the environment is more complicated than the scenario used in this chapter. In order to model a POMDP for the Robo-teacher scenario, expert knowledge was required to estimate state-transition pairs, rewards at each state, and the observation model. In a more complex environment, that same process would undoubtedly take longer, be costly and fallible.

However, should the will, time and resources to do the modelling be there, then it should be possible to model much more complex environments. Modern solvers have been proven to solve complex POMDP environments with many thousands of states efficiently (Littman, 2009; Walraven & Spaan, 2017).

Secondly, the more complicated a domain is to model, the more complex the POMDP will be. This complexity will lead to an increased time to solve the POMDP, and thus train the AMA. Though much progress has been made in solving POMDPs with many thousands of belief states by constraining the value functions to a finite subset of the belief state (Shani et al., 2013), complexity is still related to the num-

ber of states/actions and corresponding transition probabilities in a given POMDP environment (Zhang, Fu, Zhang, & Liu, 2016).

Lastly, POMDPs are inflexible in the sense that they assume the environment is static. The assumption that the environment is static may not be suitable for many environments in the real world. Ideally, we would need a mechanism to update the environment, either dynamicall or offline, in order to deal with changes in the environment.

The next section will suggest ways in which the model could be improved.

6.7 Suggestions for improvement

The POMDP approach to modelling ethical decision making in exemplarist AMAs has some advantages that make it suitable for the task. It also has some disadvantages which were discussed in the previous section. The POMDP approach is nevertheless a foundation upon which further improvement can be made through a variety of approaches that will be suggested here.

It is likely best to improve the model cautiously, even parsimoniously, because ethical decision making is not a simple problem. The strength of the POMDP model demonstrated above is that it is rooted in expert knowledge, which ensures a close correlation between the agent's and exemplars' behaviours, albeit in a simplified model.

For this reason, it is likely best to begin with potential improvements that play to this strength. Later, once the model has been matured, we can start experimenting with more recent techniques that might de-emphasise expert knowledge in favour of a model-free exploration of the environment. Some of these more recent techniques will be briefly mentioned towards the end of this section.

Let us continue to assume that the environment can be modelled via a discrete state and action space POMDP. We will also assume that the number and type of states in the POMDP environment, and the actions available to Robo-teacher, are

all the same as described in the previously.

Given these assumptions, one way to improve the model is through the use of a basic simulation environment. As before, we will manually model the transition dynamics of the environment, $T_n(s, a, s') = P(s'|s, a)$, except this time we will not assume that the transition probabilities, as modelled by the various exemplar teachers, are the same. Therefore, there will now be n transition models based on the number of exemplars.

We can then calculate a combined transition model based on a weighted sum of the individual transition models as follows:

$$T(s, a, s') = \sigma T_1(s, a, s') + \sigma T_2(s, a, s') + ... + \sigma T_n(s, a, s') \tag{6.9}$$

where $\sigma = 1/n$.

This combined transition model will represent the reaction of a typical disruptive learner to the interventions of a typical exemplar teacher. The observation model and rewards are assumed to be the same as defined in the previous section. The POMDP can be solved as before, and Robo-teacher can use the optimal policy to determine how to navigate the environment.

With the simulation environment, however, we can update the underlying POMDP with new expert knowledge by adding a new transition model T_{n+1} to equation 6.9, with an updated $\sigma = 1/(n + 1)$. The POMDP can be re-solved "offline" to get an updated policy that can be passed onto the agent's ethical performance element.

This would allow us to have a mechanism of dynamically changing the transition model of the underlying POMDP in a simulated manner. This can have a number of advantages, such as adding new knowledge to the model and potentially to remove old exemplar knowledge that we may deem no longer required.

We can also parametrise the constant, σ, in equation 6.9. This would allow an expert to tune the resulting transition model and possibly influence the performance of the agent. The expert could, for example, input a higher weighting to transition models based on highly problematic learners so that Robo-teacher could be more

effective at handling scenarios with them. The parametrised equation would be as follows:

$$T(s,a,s') = \sigma_1 T_1(s,a,s') + \sigma_2 T_2(s,a,s') + ... + \sigma_n T_n(s,a,s') \tag{6.10}$$

where $\sigma_1 + \sigma_2 + ... + \sigma_n = 1$.

The simulation approach suggested above is crude and inefficient in that it resolves the POMDP every time new expert knowledge is received. However, this may not necessarily be a bad thing for ethical robots, where we might want first to test the robot offline and only update its policy once we can guarantee that it will remain safe.

There are online algorithms and solvers which exist that can transform the optimal policy in response to a change in the underlying POMDP environment automatically (Klimenko, Song, & Kurniawati, 2014; Kurniawati & Patrikalakis, 2013; Kurniawati & Yadav, 2016). However, such techniques should also likely be employed in a simulation environment. This can allow us to test how effective the new updated policy is and to ensure that it is also safe.

With the suggestions for improvement presented in this section, we can finally revisit some of the performance elements in the model for an optimally-bounded, computationally rational AMA (figure 6.1). The learning element can be responsible for receiving new transition models from expert exemplars, and for solving the POMDP "offline". Once a suitable solution is found, it can pass on the resulting policy to the ethical performance element, which will then use the policy to make ethical decisions in real-time.

The problem generator can be responsible for running simulations based on equation 6.10. It could explore various weightings of the transition models, and continually solve the resulting POMDPs in order to store and suggest optimal candidate policies. These optimal candidate policies could be used by the learning element to compare their outputs to the ethical performance element. If the output of one of these optimal candidate policies is deemed to be better in the current environment,

then the learning element could pass it onto the ethical performance element to use.

In closing this chapter, it is likely worth pointing out possible techniques that could be explored in future research. This could include the use of reinforcement learning (RL) for planing and learning in POMDP environments. If the underlying environment is assumed to be partially observable and Markovian, and it follows the assumptions that have been stipulated in this section, then it should be possible to solve the underlying POMDP using reinforcement learning (Jaakkola, Singh, & Jordan, 1995).

Though often solving these POMDPs using RL is intractable, there are however many approximate techniques that have been developed in the literature (Azizzadenesheli, Lazaric, & Anandkumar, 2016; Guo, Doroudi, & Brunskill, 2016; Katt, Oliehoek, & Amato, 2019; Yuhui Wang, He, & Tan, 2019). Some have also begun combining RL with deep neural networks to try and improve the performance of RL in POMDP environments. (Igl, Zintgraf, Le, Wood, & Whiteson, 2018; Li, 2018; Zhu, Li, & Poupart, 2017).

Another potential benefit of RL is that it could allow for a more automated way of exploring the environment to allow the agent to discover new optimal policies. However, the only challenge with this approach, particularly in ethical decision making, is how to constrain the output policy so that it remains within safe bounds (Yue Wang, Chaudhuri, & Kavrakl, 2018).

Another potential technique that could be explored to improve the model further is Inverse Reinforcement Learning (IRL). This technique can allow the agent to learn the reward function from expert demonstrations of 'optimal' behaviour in MDP/POMDP environments (Ng & Russell, 2000).

Given state-action pairs from expert demonstrations, IRL agents attempt to find the reward function under which the demonstrated behaviour is optimal. The benefit of this technique could be the removal of the reward estimation step that we had to do in the Robo-teacher scenario.

Though this sounds conceptually simple, in reality, it can be a complex process

since a potentially large set of reward functions can explain the demonstrated behaviour (Ng & Russell, 2000). As a result, IRL algorithms can be computationally expensive to run.

Another downside of IRL is that the agent does not learn what to do; it only learns a possible reward function under which demonstrated behaviour is optimal. The designer will need to insert yet another step for the agent to make decisions (i.e. they still need a RL step to learn the policy function) (Ho & Ermon, 2016).

Fortunately, the latest techniques in imitation learning, which theoretically enhance IRL by combining it with other learning aspects, have made improvements on some of these limitations. Generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016) is one such technique. GAIL works by running IRL, followed by RL to learn the policy directly.

The IRL part is set up as a dual occupancy measure matching problem. The result of this is a dual optimum (i.e. the learned cost or reward function to minimise the difference between the agent and the expert's behaviour). Once the dual optimum is found, RL is applied to obtain the primal optimum policy that the agent should follow to imitate the expert.

Belief-module imitation learning (BMIL) is another imitation learning technique that could be employed (Gangwani, Lehman, Liu, & Peng, 2019). The BMIL technique theoretically improves performance over GAIL, especially when dealing with partially observable environments. Unlike GAIL, which essentially has two parts to the algorithm (IRL, followed by RL), BMIL learns the cost (reward) and policy functions simultaneously (Gangwani et al., 2019). This can have a significant advantage, especially in computation time.

Though these techniques have been proven in mainstream computer science, none of them has been applied to the AMA project to date (based on a scan of publications). It is perhaps wise to caution the reader to use IRL and imitation learning techniques carefully, as they further move away from the direct modelling method demonstrated in this chapter. Ethical decision making is complex, and simply ap-

plying the latest techniques in machine learning without careful thought (both philosophically and practically) can lead to unintended results.

For example, we need to be very careful what reward function the AMA learns using these techniques, and testing it thoroughly to ensure it is still respecting of human moral values. It also remains a challenge for the AMA to learn the reward function by observing other moral exemplars. Directly observing human beings will likely not be practical for AMA to learn ethical value with currently available technology. We are still likely to depend on an extensive training data set, or a simulation environment, where we can apply these techniques.

In closing, the use of POMDPs to model ethical decision making is likely very new or recent, and it will need to be matured over time. Fortunately, there is much literature that exists in parallel fields that could help us fast track the adoption and use of POMDPs and newer techniques to build exemplarist AMAs effectively. Techniques in RL, IRL and imitation learning should be next explored to evaluate where they could further advance the exemplarist AMA framework described in this thesis.

6.8 Conclusion

The purpose of this chapter was to demonstrate, practically, how an exemplarist AMA might be implemented using currently available technology. The model for an optimally bounded computationally rational AMA in figure 6.1 was used as a foundation upon which to build an ethical decision making procedure into the AMA.

Although not every element in the model could be discussed in detail, this chapter did demonstrate that it is generally feasible to implement all of it. This chapter does not claim to have produced the most optimal model or procedure for building the AMA. Such a goal is left for a more technically oriented thesis or research article to explore in detail. The purpose of this chapter was to lay the foundation upon which further research could begin.

The last section of this chapter suggested possible techniques, such as reinforcement learning and neural networks, that could be explored by future research in building exemplarist AMAs that function in POMDP environments. The next chapter will conclude this thesis and tie in all the work together in a cohesive manner.

Chapter 7

Discussion and Conclusion

7.1 Introduction

The purpose of this chapter is to provide a detailed discussion and conclusion of this thesis. To do this effectively, this chapter will summarise the main arguments that were advanced in this thesis (section 7.2). This will be followed by an evaluation of whether or not the objectives of the study were achieved (section 7.3).

The chapter will then move on to highlight the contributions of the study to the body of knowledge in machine ethics (section 7.4). This will be followed up by concise implications of the the study on machine ethics, the practitioners in machine ethics, and the applicability of exemplarist AMAs (section 7.5).

Having discussed the contribution and limitations, this chapter will then move on to highlight the limitations of the research (section 7.6.). The thesis will then be concluded (section 7.7) before providing recommendations for future research (section 7.8).

7.2 Summary of the main arguments

The primary purpose of this research was to advance an argument for exemplarism (Zagzebski, 2010, 2017) as an alternative, suitable and viable machine ethics

framework for the AMA project. To further illustrate the feasibility of applying exemplarism as a machine ethics framework, this thesis explored a detailed scenario that demonstrated how it might practically work in real life. Furthermore, this thesis also sought to demonstrate the practical implementation of an exemplarist AMA.

The purpose of this section is to provide a summary of the main arguments advanced in this thesis. The project to build an exemplarist AMA lies at the intersection of the topics of artificial moral agency, computational rationality, and exemplarism. These topics are summarised thematically in figure 7.1. Once we have discussed the summary of the main arguments made in this study, we will then proceed to discuss whether or not its objectives were achieved.

Putting aside the introduction chapter, whose primary purpose was to introduce the purpose of the study, the thesis began in chapter 2 with a discussion of artificial moral agency. Starting here was essential because traditional views of moral agency generally do not include artificial agents in the moral universe (Johnson, 2006; Torrance, 2008). We started with a review of various forms of moral agency in the literature (section 2.3). The goal in doing this was to understand which view could potentially include artificial agents as moral agents.

First we looked at the proponents of biological moral agency (Churchland, 2014; Liao, 2010; Rottschaefer, 2000; Torrance, 2008). The key argument in this view was that an agent needs to be biological in order to qualify as a moral agent. Furthermore, the capacities that are philosophically thought to be necessary for moral agency, such as sentience, consciousness, free-will, and intentional and mental states, are emergent features that are only possible in sophisticated biological agents like human beings. Under this view, artificial agents, at least those that are not biologically based, could not qualify as moral agents.

We then looked at the proponents of conscious moral agency (Himma, 2009; Parthemore & Whitby, 2013, 2014; Wallach et al., 2011). The key argument in this view is that an agent should possess consciousness and phenomenal awareness in order to qualify as a moral agent. Since artificial agents are not considered to be

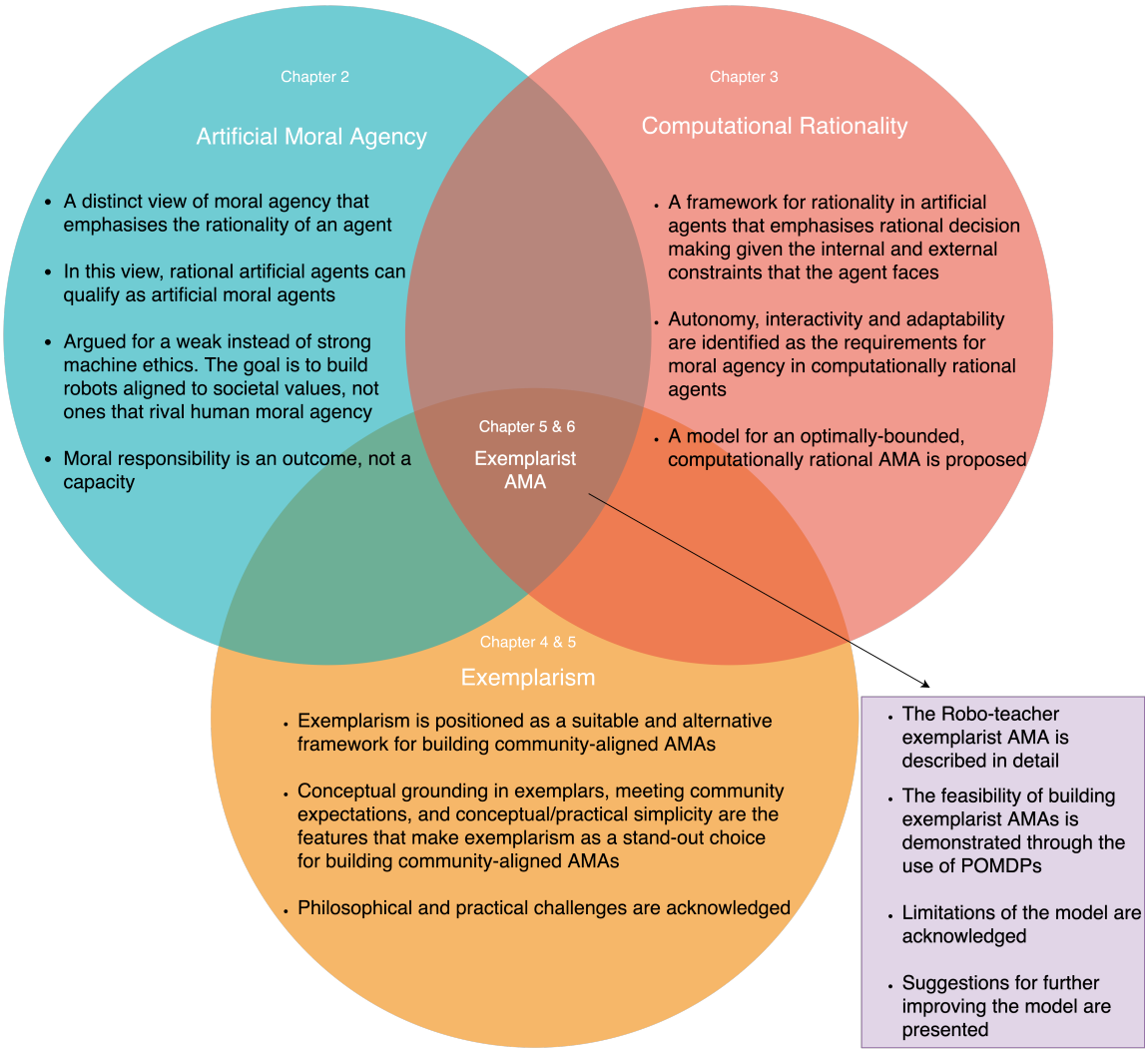

Figure 7.1: A graphical representation and summary of the main themes and arguments in this study. The project to build exemplarist AMAs lies at the intersection of the topics of artificial moral agency (chapter 2), models of computational rationality (chapter 3), and exemplarism (chapter 4). The feasibility of building an exemplarist robot is demonstrated in chapter 5 and 6.

conscious (a factor which would be hard to prove even if we set out to design it), they would not qualify as moral agents according to this view.

Lastly, we looked at the proponents of artificial moral agency and saw that they could be divided into those that hold a strong (Abney, 2012; Allen & Wallach, 2012; Sullins, 2006) and a weak (Floridi & Sanders, 2004; Johnson, 2006; Scheutz & Malle, 2017) machine ethics view.

The proponents of strong machine ethics argue that artificial agents can qualify as full moral agents, perhaps even rivalling human moral agency. On the other hand, proponents of weak machine ethics also argue that artificial agents could qualify as moral agents. However, they could not do so at a level that would rival human moral agency.

This difference represented the first major argument in this thesis, which is that artificial agents built with currently available technology cannot be full moral agents. They can only be something lesser, but this 'lesser thing' that they are is still useful to pursue in the AMA project. This is because, although we cannot build them to be full moral agents, we can, however, still build them to respect society's expectation of moral behaviour.

We, therefore, concluded that a biological, conscious, and strong artificial moral agency could not be feasibly built with currently available technology. We placed our exclusive focus on weak artificial moral agents. We then proceeded to argue that artificial agents that are rational could qualify as artificial moral agents in the weak machine ethics sense (see section 2.4).

We concluded this chapter by defining moral responsibility, not as a capacity that an agent has, but rather as an outcome that AMAs could achieve depending on how well they were built to respect society's expectations of moral behaviour (section 2.4.2). This framing of moral responsibility was necessary if we were to continue the AMA project in its current conception of weak machine ethics. It will also allow us to build AMAs and to test how aligned they are to society's expectations of moral values.

Though we had alluded to it in chapter 2, we now needed to ground artificial moral agency in a computational framing of rationality. This was the purpose of chapter

3, which considered computational rationality as a possible unifying framework that could bridge the gap between the philosophical and computational conceptions of rationality. This chapter began with background material on prominent computational models of artificial moral agency (section 3.2). These would provide valuable lessons for this thesis' proposed computational model of an artificial moral agent.

In this chapter, computational rationality was defined as a framework that emphasises rational decision making given the internal and external constraints that an artificial agent faces (Gershman et al., 2015; Lewis et al., 2014). We also argued that a robot could qualify as an AMA if it met the requirements of autonomy, interactivity and adaptability as expressed by Floridi and Sanders (2004) and supported by others (Coeckelbergh, 2014; Moor, 2006; Scheutz & Malle, 2017; Sullins, 2006).

We then proposed a model for an optimally-bounded, computationally rational AMA (section 3.5). This model was synthesised by combining the ideas of Gershman et al. (2015), Horvitz (1987), Lewis et al. (2014) on computational rationality, Floridi and Sanders (2004) on the requirements for artificial moral agency, and Russell and Norvig (2009), Russell and Subramanian (1995) on the structure of an optimally-bounded general learning agent.

Although an optimally-bounded, computationally rational agent could conceptually qualify as an AMA in the weak machine ethics sense, we had yet to discuss how it might be imbued with an exemplarist machine ethics framework. The purpose of chapter 4 was, therefore, to position exemplarism as a suitable machine ethics framework that could help us build AMAs that could meet community expectations of moral behaviour.

The chapter began with a review of the major ethical theories, and specifically, how other scholars have applied them as machine ethics frameworks applicable to the AMA project (section 4.2). We then moved on to argue that approaches based on virtue ethics, of which exemplarism is a derivative of, are conceptually better suited towards being applied as machine ethics frameworks in the building of AMAs that meet the requirements for autonomy, interactivity and adaptability.

Virtue ethics approaches were found to be conceptually better because they emphasised an inward character development. It was contrasted with deontological and consequentialist approaches, which tend to depend on external rules or functional guides to make decisions. This meant that approaches based on virtue ethics would likely result in AMAs that could be autonomous in their decision making, something that is an essential requirement for artificial moral agency as defined in this thesis.

We then moved to position exemplarism as a suitable framework for building AMAs that could meet community expectations of moral behaviour. In this discussion, we argued that three key features of exemplarism, namely: *grounding in moral exemplars*, *meeting community expectations* and *practical simplicity*, were crucial to its uniqueness and suitability for application in the building of AMAs that can meet community expectations of moral behaviour.

Chapter 5 then discussed a detailed but fictional scenario in a near-future mathematics classroom taught by a robotic teacher called Robo-teacher. The purpose of this chapter was to clarify further how critical concepts in exemplarism could be applied in practice by demonstrating how it might work in an exemplarist AMA. Specifically, the scenario discussed how an exemplarist AMA might identify and learn from exemplars and make ethical decisions in real-time. The chapter concluded with an acknowledgement of the philosophical and practical challenges of exemplarism.

Finally, chapter 6 looked at the technical feasibility of actually building an exemplarist AMA, i.e. Robo-teacher. The focus of this chapter was on demonstrating how the ethical decision making capability of Robo-teacher might be implemented through the use of POMDPs. The chapter concluded with a discussion of the limitations of the model and potential suggestions for how it might be improved in the future to make Robo-teacher an even better exemplarist AMA.

7.3 Were the objectives of the thesis achieved?

Having reviewed the main arguments made in this thesis (a summary can be found in figure 7.1), we can now move on to ask whether or not the objectives of the study were achieved. To do this, we will briefly revisit the research questions that were asked in chapter 1, and further elaborate on how each of them was answered. There are limitations to the research which are discussed separately later in section 7.6.

The objectives of the research were, in question form, as follows:

1. How might exemplarism be applied as a machine ethics framework to build ethical robots?

2. Could exemplarist robots qualify as artificial moral agents?

3. What is the practical feasibility of building exemplarist robots?

How might exemplarism be applied as a machine ethics framework to build ethical robots?

In answering the first question, we must begin by highlighting a significant difference between bottom-up machine ethics frameworks, such as exemplarism as proposed in this thesis, and top-down ones that have been proposed in the literature.

The significant difference between the two is that top-down machine ethics frameworks need to be defined explicitly for the AMA to use. In deontological frameworks, for instance, this would mean defining specific rules or imperatives that the AMA would need to respect or follow. There is still scope for the AMA to be rational, but this rationality is only concerned with ensuring that every action that the AMA takes falls within the constraints of these rules or imperatives.

A machine ethics framework based on exemplarism, however, is fundamentally different in that it is not explicitly defined. It is instead implicitly developed by the robot as it learns an internal representation of moral value based on an aggregation of the behaviour of many exemplars in society. This process is analogous to the robot's internal moral character development (section 4.3.1)

This also means that we cannot necessarily tune an exemplarist AMA to behave in a particular way. We can only teach it to aggregate the moral values of society's exemplars. This is likely the most fundamental point that separates exemplarist AMAs from others. We will return to this point later in the implications section.

To finally answer the first question, exemplarism is only an implicit framework. The only way to apply it as a machine ethics framework is to teach the robot how to learn from moral exemplars. Conceptually, this is the answer to the first question, but of course, more details can be filled in.

We discussed in chapter 5 the process to teach an AMA to learn from moral exemplars. Firstly, the AMA needs to identify moral exemplars. This might include teaching it to scourer the internet or some other source in order to identify them. It might also mean teaching it to identify them through observation in its context.

The exemplarist AMA conceptualised with weak machine ethics will not be able to identify moral exemplars according to the emotion of admiration (Zagzebski, 2010, 2017). Instead, it will need to identify and select moral exemplars according to some key characteristics that align with its design and goals.

In the scenario in chapter 5, we assumed that Robo-teacher had access to a database of potential exemplar teachers. Robo-teacher could then use this database to identify the relevant exemplars, perhaps based on their class' performance, experience with similar situations, effectiveness at dealing with classroom ethics, amongst others. Such a database would need to be intentionally kept for exemplarist AMAs before their deployment.

Secondly, it needs a mechanism to determine an aggregate representation of moral value based on the exemplars' behaviours. This could be done in various ways, including leveraging machine learning techniques (Muntean & Howard, 2016), or even simulation as suggested in chapter 6.

The exemplarist AMA then needs to use this internal representation of moral values to make decisions in real-time. In the example in chapter 6, a POMDP was used as a complex planning and decision-making procedure. The learned policy

could then be applied to decide on the relevant action to do for any given state of the environment.

The exemplarist AMA then needs to continue to refine its list of chosen exemplars and to learn from them continually. It also needs to learn from its own experience, whether good or bad, as it carries on with its tasks. This process of identifying and learning from exemplars, and also learning from its experience, does not end. It is a fundamental part of being an exemplarist AMA.

Could exemplarist robots qualify as artificial moral agents?

We have already partially answered this question in the previous discussion. The fundamental requirements for artificial moral agency was identified as rationality in section 2.3.3. However, we clarified in chapter 3 that a computational framing of rationality requires us to build optimally-bounded, computationally rational agents. This meant an ability for the agent to interact with its environment, learn from it, and make independent moral decisions (Floridi & Sanders, 2004).

It is this last requirement for making independent moral decisions that we argued finds a better fulfilment in exemplarist AMAs than it does in other approaches. This is because we cannot explicitly tell exemplarist AMAs how to behave. We likened the implicit representation of moral values in exemplarist AMAs to a kind of robotic moral character, which aligns more closely with the requirements for autonomy in moral decision making (section 3.4).

Of course, this is not a unique feature to exemplarist AMAs; it is present in general in other virtue-ethics based AMAs (section 4.3.1). We, however, argued that this feature is likely most clearly expressed in exemplarism due to its practical grounding in moral exemplars. The other forms of virtue ethics have the challenge of having to ground abstract concepts such as virtues and vices, which would likely not be computable in weak machine ethics AMAs.

What is the practical feasibility of building exemplarist robots?

The purpose of chapters 5 and 6 was to specifically answer this question directly. What these chapters have demonstrated is that exemplarist AMAs can be practically implemented using currently available technology.

An interesting observation in these chapters was that many feasible algorithms exist with which to improve the exemplarist AMA project further. The applicability of these algorithms further strengthens the argument that exemplarist AMAs are technically feasible. A more technically focused thesis could have likely expanded further on the proposed algorithm(s) employed to demonstrate the feasibility of building an exemplarist AMA like Robo-teacher.

7.4 Summary of contributions

Having demonstrated more plainly how the research objectives of this study were met, we now need to summarise the contributions to the field of machine ethics. There are both primary and secondary contributions of this research.

7.4.1 Primary contributions

A primary contribution of this research was the elucidation of exemplarist virtue ethics as a machine ethics framework for the AMA project. Since exemplarism had not been considered so wholistically in prior literature, this study had to both introduce it and explain in detail how it might be applied as a machine ethics framework.

This process included the introduction of the theoretical aspects of exemplarism and reinterpreting them within a machine ethics context. To name but a few, the study had to introduce the conceptual grounding of exemplarism, and link it to the need for AMAs to understand moral value in a way that makes it computable.

Since exemplarism grounds key ethical concepts in the exemplars of moral goodness, it meant that we could leverage this feature to teach robots to implicitly learn these same ethical concepts without the need to model them directly.

The Robo-teacher scenario illustrates this argument well. Let us assume, for

example, that exemplar teachers, on average, value fairness and equity (which are virtues) in how they spread their attention and engagement with all the learners.

An exemplarist AMA that has identified these same exemplar teachers, and learned from their aggregate behaviour in dealing fairly and equitably with all learners, would also have implicitly learned the virtues of fairness and equity. This example illustrates how grounding key ethical concepts, like virtues and vices, in moral exemplars, is perhaps a more effective way to teach robots these same virtues and vices than attempting to model them directly.

This finding is important because it points to an alternative path away from directly attempting to model ethics in top-down approaches, which have proven to have scalability challenges (Allen et al., 2000; Scheutz & Malle, 2017).

This finding also solves the challenge of directly modelling abstract virtues and vices in virtue ethics-based machine ethics frameworks. As Scheutz and Malle (2017, p. 369) remark, '*"it is unclear how "wisdom" could be realized in a robotic system over and above demands of rational behavior, such as when a game-playing computer always picks the best move from its perspective"*. With exemplarism, wisdom can be implicitly learned by grounding it in moral exemplars.

A caveat that we must guard against is merely assuming that modern machine learning techniques have been doing what exemplarism suggests in the context of the AMA project, just without calling it that (Kasenberg et al., 2018; Yu et al., 2018). As explained in section 4.2.4, most of the machine learning techniques that have been applied to the AMA project are done in a rushed way, without necessarily considering the ethical frameworks that they are modelling.

Exemplarism, on the other hand, goes far beyond just learning how to behave from data. It is an end-to-end approach for how we should build robots that can identify moral exemplars, learn from them to build moral character, make decisions based on this moral character, and continuously identify and learn from new exemplars and experiences.

Many machine learning techniques can be employed to implement some parts of

an exemplarist machine ethics framework, but the two are not equivalent. Many techniques beyond just machine learning are undoubtedly required for this task[1]. Furthermore, exemplarism answers critical questions about how we should ground key ethical concepts, which the mere application of machine learning will not and cannot do.

Another advantage of applying exemplarism as a machine ethics framework is that it gives us a clear path to building AMAs that can meet community expectations of moral behaviour. No other machine ethics framework has this feature at the core of its theoretical makeup. Exemplarism, at its core, is an attempt at a normative ethical theory that can potentially liberate us from the trap of having to pick and choose which frameworks we should use for which societies.

In doing this, it enables us to build robots that, in theory, can be deployed to any context where, as long as they have access to exemplars, they can learn the norms of the society implicitly. This feature might give us a clear path to building AMAs that can scale to several contexts. This point will be recommended as a future research direction.

7.4.2 Secondary contributions

The secondary contributions of this research study emanate directly from the primary contribution. These are aspects that may be useful for others to consider in the field of machine ethics in general.

The first such contribution is the model for an optimally-bounded, computationally rational AMA that was proposed in section 3.5. Though the structure of a general learning agent (Russell & Norvig, 2009), upon which the proposed model is based, is not new to the field of Computer Science, its application to the field of machine ethics likely is. At the very least, other scholars can critique or improve upon it in their quest to build AMAs.

[1]In chapter 6, POMDPs, a non-machine learning technique, were used to implement a part of the exemplarist AMA.

The second contribution is the delineation between weak and strong machine ethics (section 2.3.3). Though this delineation is not new in the literature (Allen & Wallach, 2012), the terms of weak and strong machine ethics likely are. These terms could help differentiate between the tasks of building robots that rival human moral agency and those that are built with functional ethics to respond to current threats posed by an increase in autonomous robots in society (Cowls & Floridi, 2018).

The last contribution is the suggestion that an implicitly learned aggregate representation of moral values is analogous to a kind of robotic moral character (section 4.2.3). Defining artificial moral character in this way frames it in a manner that likely makes it achievable with currently available technology.

Abney (2012) and P. Lin et al. (2008) are likely the only other scholars to suggest the concept of a robotic moral character in virtue ethics-based approaches, though they do not explain how it could be practically achieved. Perhaps this research is but a first step in unpacking what it could be.

A computational framework of moral character could also help to launch research into improving implicit representations of moral value and how that could be used to facilitate ethical decision-making in AMAs. This will be included in the recommendations for future research.

7.5 Implications of the research

7.5.1 Implications for machine ethics research

There are several implications of this study to the field of machine ethics research in general. The first implication is that machine ethics research perhaps needs to shift from only considering traditional machine ethics frameworks (primarily those based on consequentialism and deontology) to ones that have perhaps not received as much attention as they should have.

Virtue ethics approaches, in particular, have not received as much research atten-

tion as consequentialist and deontological ones. This research suggests that virtue ethics theories can be creatively reinterpreted for the machine ethics project. Furthermore, exemplarism is only but one example of non-traditional forms of virtue ethics. Perhaps the other forms could unlock interesting perspectives on how to build AMAs.

One can only wonder at this stage whether building an AMA based on agent-based virtue ethics (Slote, 1995; Van Zyl, 2005) would produce a morally selfish robot, or will such as study unlock other perspectives that are not immediately obvious?

Similarly, is target-centred virtue ethics (Swanton, 2003) really unsuitable to the AMA project, or could the abstract concepts that it demands be grounded in other ways that could make it a suitable computational framework for ethical decision making?

We do not know the answers to these questions because the field of machine ethics has not yet moved towards answering them. If anything, the study of exemplarism as a possible framework for machine ethics has further highlighted, at least in the researcher's mind, a possible way to take research in weak machine ethics forward.

The second implication is that there is much to be explored in the quest to build weak machine ethics AMAs. It may be tempting to assume that breakthroughs in the AMA project can only occur when building robots that rival human moral agency. However, what this study has hopefully shown is that much ground still needs to be covered even in building robots with weak machine ethics.

The field of machine ethics ultimately needs an equitable balance between strong and weak machine ethics research (Allen & Wallach, 2012), and not to prop up one over the other. Both are useful to pursue, and both have long ways to go before they become part of mainstream AI research.

The last implication is that the exemplarist AMA project emphasises a continual learning of moral value over time. It does not emphasise finding the right frameworks that will work perfectly from the start. To use the words of Muntean and Howard (2016), exemplarist AMAs are *parsimonious* and *minimalist* in their ethical

assumptions. They instead need to develop their moral character, by learning from the right exemplars, so that they can improve their ethical capabilities over time[2].

Perhaps the field of machine ethics needs to treat moral capability not as an explicit model, but an implicit one that develops over many interactions with both exemplars and general society.

7.5.2 Implications for designers of exemplarist AMAs

Exemplarist AMAs, as demonstrated in chapter 6, can likely be built with currently available technology. This practicality is key if we are to see designers adopting these principles in practice. A well-resourced team, for example, can easily take the design suggested in this thesis a lot further than what was demonstrated.

Furthermore, exemplarism may provide a more straightforward path to scaling the ethical capability of AMAs through bypassing the need to model abstract ethical principles directly. This is in contrast to top-down approaches to building AMAs, which seem to inherently suffer from the issue of intractability as the complexity of the ethical principles being modelled increases (Allen et al., 2000; Gips, 1995; Scheutz & Malle, 2017).

7.5.3 Implications for the application of exemplarist AMAs

Exemplarist AMAs can likely be deployed in a wide variety of contexts, especially where exemplars are made accessible in a suitable way for the AMAs to learn from them. This could include ordinary contexts such as the home, self-driving car and places that provide medical care. Other examples can likely include financial (e.g. an unbiased loan approvals bot), legal (e.g. a legal advisor agent), and workplace (e.g. a business ethics advisor bot) contexts.

[2]Perhaps it is time we start acknowledging that human beings follow a not so different path in their moral education? Human beings are, of course, very adept at moral learning. However, this suggests that exemplarist AMAs might follow a similar path of moral education with which we can relate.

Exemplarist AMAs will however likely not be effective in extraordinary contexts, such as during war or famine, because the behaviour of exemplars may be drastically altered due to the prevailing circumstances (section 5.3.1). In these kinds of contexts, top-down approaches in deontology may be more suitable (Arkin et al., 2012).

7.6 Limitations of the research

Though many of the limitations have already been covered extensively in chapters 5 and 6, they are briefly summarised here for completeness.

A strength of exemplarism is undoubtedly the ability to ground abstract ethical concepts, such as virtues and vices, in the exemplars of moral goodness. This conceptual grounding bypasses the problem of directly modelling abstract concepts computationally. However, even this strength has its limits.

To understand why, we need to remind ourselves of our definition of weak AI agents, on which we derived a definition for weak machine ethics. In section 2.2.1, we concluded that weak AI agents are essentially doing agents, having no real thoughts and intentional/mental states. Weak machine ethics AMAs, even exemplarist ones, are also fundamentally doing agents.

The implication of this is that they will likely struggle to make ethical decisions where the exemplars do not perform easily demonstrable actions in order to deduce what they value. For example, it will likely be difficult for exemplarist AMAs to determine whether they should or shouldn't tell a lie.

Many lies do not have a physically discernable action that follows their conception in the liar's mind. Yes, the liar will speak the lie, but how will the agent know that the exemplar was lying? Exemplarist AMAs could therefore unwittingly learn bad behaviour, like telling a lie, without even knowing that they are doing so.

Of course, a counter-argument to this is that exemplarist AMAs do not learn from only one exemplar. They learn from many so that, at an aggregate level, they can learn an accurate representation of moral behaviour. Nevertheless, this issue

suggests that exemplarist AMAs could not be deployed in environments that contain little demonstratable action.

Another limitation is that exemplarist AMAs can likely not be applied in contexts with extraordinary circumstances, like in a war, extreme poverty or even a depression. This is because the behaviour of exemplars in contexts with extraordinary circumstances will likely be altered and different to how they would behave under normal circumstances.

Deploying exemplarist AMAs in extraordinary contexts may lead them to learn highly biased and possibly unethical behaviour. For example, in a time of poverty, should the agent learn that stealing is ok if its moral exemplars are doing the same thing? The answer is not straightforward, but it illustrates that other ethical frameworks, such as deontology, may be more applicable in this instance.

Another limitation has to do with the ability for exemplarist AMAs to identify and learn from moral exemplars independently of human assistance. The technology to allow a robot to search and find the relevant moral exemplars automatically, and somehow observe them in order to learn moral value independently is likely very challenging at best, and not currently available at the worst.

We will likely need to create mechanisms for exemplarist AMAs to identify moral exemplars (such as keeping a database of the relevant ones, as was alluded to in chapter 5). We would also likely need to model the behaviour of moral exemplars on behalf of the AMA. The AMA can still perform the relatively more straightforward task of aggregating the moral value of all the exemplars by using the supplied models.

This concession may cause a departure from a strict adherence to the principles of exemplarism. However, should the technology to identify and learn from exemplars independently improve or become available, then this limitation will fall away. For now, there is likely to remain a modelling burden on the designers of exemplarist AMAs.

It is worth noting that exemplarist AMAs have a theoretical limitation when it comes to explainability. This is because exemplarist AMAs, like virtue-ethics based

agents in general, depend on an internally generated system of moral values based on their internal moral character development. Though this feature arguably makes them better artificial moral agents according to the framing used in this thesis, it however theoretically leaves them open to the problem of explainability. This is why this limitation is noted here.

There is, of course, a practical way to overcome this limitation, as suggested in the computational model presented in chapter 3, however, this does not take away from the fact that this limitation exists at a theoretical level. The explicability element could be designed to interpret this internal system of moral values, and to track the rationale for every reason, and to present it in a human-readable or understandable way.

Lastly, this research addresses the risks to society of not equipping increasingly autonomous robots with ethical decision making. It does not directly address the impact, whether legal, social, or regulatory, of introducing these robots into society. Such a task is fundamentally different from the purpose of this thesis and is, therefore, left for future research. The reader is pointed to the works of Asaro (2007) and P. Lin et al. (2011) for some of the debate in this area.

7.7 Conclusion

The primary purpose of this research was to position exemplarist virtue ethics (Zagzebski, 2010, 2017) as an alternative, suitable and viable machine ethics framework for building ethical robots that meet community expectations of moral behaviour.

Merely suggesting the applicability of exemplarism to the AMA project, however, would not have been sufficient. This thesis needed to demonstrate how exactly it might enable the building of ethical robots that could qualify as AMAs.

In the process of meeting the objectives of the thesis, we argued for a definition of artificial moral agency that emphasises the rationality of the agent in chapter 2. We also proposed a model for an optimally-bounded, computationally rational AMA

in chapter 3.

We gave a detailed argument for how exemplarism could be suitable as a machine ethics framework for the AMA project in chapter 4. We followed up this argument with a detailed scenario that introduced Robo-teacher, a fictional exemplarist AMA, to further clarify critical concepts in exemplarism and their application to the AMA project.

We further provided an implementation of Robo-teacher, primarily based on POMDPS and simulation in chapter 6. Though there were limitations to the implementations provided in this thesis, it did serve the purpose of demonstrating the feasibility of building exemplarist AMAs. Furthermore, chapter 6 also proposed ways in which the implementation could be improved further.

Though there are challenges and limitations to a study like this as discussed in section 7.6 above, many of them can be circumvented without making the prospect of building exemplarist AMA infeasible. This is the reason why demonstrating a practical implementation of an exemplarist AMA was so crucial for this study, mainly because the approach had not been so extensively covered in prior literature.

We can finally end the thesis the way it began. 'Robots are not ethical like people'. More specifically, optimally-bounded, computationally rational exemplarist robots conceptualised with weak machine ethics cannot be ethical like people. They are something less! However, this lesser thing that they are is still useful in enabling us to build them to meet the challenge of respecting community expectations of moral behaviour (Cowls & Floridi, 2018).

7.8 Recommendations for future research

- This study suggested that exemplarist AMAs should not be deployed in contexts with extraordinary circumstances, such as in a war. A future study could test this recommendation, and explore whether or not it would be possible to adapt exemplarist AMAs for deployment in extraordinary contexts.

- A future study could further explore the concept of artificial moral character. This thesis gave some preliminary suggestions about what it might be and how it could be computationally realised. A future study could further explore this concept by clearly defining it and suggesting better ways to implement it computationally.

- A further study could unpack the relationship between artificial moral character and artificial moral agency.

- A future study could evaluate non-traditional ethical theories and their applicability to machine ethics. This evaluation could include non-traditional forms of virtue ethics, such as target-centred and agent-based virtue ethics, and any other potential framework that has not been extensively covered in machine ethics literature.

- A future study could investigate technically feasible ways for AMAs to identify moral exemplars according to specific characteristics automatically.

- A future study could investigate technically feasible ways for AMAs to learn moral value through the observation of moral exemplars automatically

- A future study could investigate the transferability of ethical decision-making capability learned by an exemplarist AMA from one context to another.

- This study proposed a model for an optimally-bounded computationally rational AMA. A future study could investigate strategies for the ethical meta-reasoner to select the appropriate ethical framework and decision making algorithm based on the state of the environment.

- A further study could also explore the feasibility of including more than one ethical framework and more than one decision making algorithm in the ethical performance element.

- A future study could focus on the problem of making exemplarist artificial moral agents built with weak machine ethics explainable at both a theoretical and practical level.
- Finally, a future study could focus on the social, legal or regulatory implications of introducing exemplarist ethical robots into society.

Appendix A

POMDP definition and output files

A.1 POMDP definition

#Filename: roboteacher.POMDP

#Robo-teacher ethical reasoner example

#Created by: Bongani Andy Mabaso

#Date: January 2020

discount: 0.95

values: reward

states: disruptive unclear quiet angry happy

actions: reprimand observe encourage

observations: disruptive unclear quiet angry happy

start:

0.5 0.2 0.1 0.2 0.0 # Probabilities of starting at each state

Transition probabilities for action reprimand

T: reprimand

0.4 0.2 0.2 0.2 0.0

0.4 0.2 0.4 0.0 0.0

0.5 0.2 0.1 0.2 0.0

0.5 0.0 0.3 0.2 0.0

0.0 0.0 1.0 0.0 0.0

Transition probabilities for action observe

T: observe

0.8 0.0 0.2 0.0 0.0

0.3 0.5 0.2 0.0 0.0

0.2 0.2 0.5 0.0 0.1

0.4 0.0 0.1 0.5 0.0

0.0 0.0 0.1 0.0 0.9

Transition probabilities for action encourage

T: encourage

0.8 0.0 0.0 0.2 0.0

0.2 0.4 0.4 0.0 0.0

0.1 0.2 0.2 0.0 0.5

0.4 0.0 0.2 0.4 0.0

0.0 0.0 0.0 0.0 1.0

Observation model

O: *

0.75 0.0 0.0 0.25 0.0

0.25 0.5 0.0 0.25 0.0

0.0 0.25 0.75 0.0 0.0

0.5 0.0 0.0 0.5 0.0

0.0 0.0 0.0 0.0 1.0

Reward function

R: * : 0 : * : * -1

R: * : 1 : * : * 0

R: * : 2 : * : * 1

R: * : 3 : * : * -1

R: * : 4 : * : * 5

A.2 Optimal policy graph

0 reprimand 0 1 1 0 0

1 encourage 5 1 1 11 1

2 observe 0 1 1 0 1

3 encourage 0 1 1 2 1

4 observe 2 1 1 5 1

5 observe 0 1 1 7 1

6 encourage 0 1 1 0 1

7 observe 0 1 1 2 1

8 observe 2 1 1 7 1

9 encourage 2 1 1 6 1

10 encourage 0 1 1 6 1

11 encourage 2 1 1 2 1

12 encourage 2 1 1 7 1

13 encourage 7 1 1 3 1

14 encourage 7 1 1 6 1

15 encourage 5 1 1 3 1

Bibliography

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In B. Bonet, S. Koenig, B. Kuipers, I. R. Nourbakhsh, S. J. Russell, M. Y. Vardi, & T. Walsh (Eds.), *Aaai workshop: Ai, ethics, and society* (Vol. WS-16-02). AAAI Workshops. 978-1-57735-759-9, AAAI Press.

Abney, K. (2012). Robotics, ethical theory, and metaethics: A guide for the perplexed. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot ethics, the ethical and social implications of robotics* (Chap. 3, pp. 35–52). The MIT Press.

Aghion, P., Jones, B. F., & Jones, C. I. (2017). *Artificial intelligence and economic growth*. National Bureau of Economic Research.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155. doi:10.1007/s10676-006-0004-4

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, *12*(3), 251–261. doi:10.1080/09528130050111428

Allen, C., & Wallach, W. (2012). Moral machines: Contradiction in terms , or abdication of human responsibility? In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics, the ethical and social implications of robotics* (Chap. 4). The MIT Press.

Alzahrani, H. (2016). Artificial intelligence: Uses and misuses. *Global Journal of Computer Science and Technology*, *16*(1).

Amigoni, F., & Schiaffonati, V. (2005). Machine ethics and human ethics: A critical view. *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*.

Anderson, J. A., & Rosenfeld, E. (2000). *Talking nets: An oral history of neural networks*. MiT Press.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, *28*(4), 15. doi:10.1609/aimag.v28i4.2065

Anderson, M., & Anderson, S. L. (2008). Ethel: Toward a principled ethical eldercare robot. In *Proceedings of the aaai fall 2008 symposium on ai in eldercare: New solutions to old problems*, Arlington, Virginia.

Anderson, M., Anderson, S. L., & Armen, C. (2006a). An approach to computing ethics. *IEEE Intelligent Systems*, *21*(4), 56–63. doi:10.1109/MIS.2006.64

Anderson, M., Anderson, S. L., & Armen, C. (2006b). Medethex : A prototype medical ethics advisor. *Artificial Intelligence*, 1759–1765.

Anderson, S. L., & Anderson, M. (2011). A prima facie duty approach to machine ethics and its application to elder care, 2–7. doi:10.1017/CBO9780511978036.032

Annas, J. (2011). *Intelligent virtue*. Oxford University Press.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy*, *33*(124), 1–19. doi:10.1017/S0031819100037943

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, *57*(5), 469–483. doi:10.1016/j.robot.2008.10.024

Arkin, R. C. (2008a). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part 1: Motivation and philosophy. In *2008 3rd acm/ieee international conference on human-robot interaction (hri)* (pp. 121–128).

Arkin, R. C. (2008b). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part 2: Formalization for ethical control. *Frontiers in Artificial Intelligence and Applications*, *171*(1), 51–62.

Arkin, R. C. (2009). Ethical robots in warfare. *IEEE Technology and Society Magazine*, *28*(1), 30–33.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, *100*(3), 571–589. doi:10.1109/JPROC.2011.2173265

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment – what will keep systems accountable? In *Workshops at the thirty-first aaai conference on artificial intelligence*.

Asaro, P. M. (2006). What should we want from a robot ethic? *International Review of Information Ethics (IRIE)*, *6*, 9–16.

Asaro, P. M. (2007, April 14). Robots and responsibility from a legal perspective. *Proceedings of the IEEE Conference on Robotics and Automation, Workshop on Roboethics*, *4*, 20–24.

Azizzadenesheli, K., Lazaric, A., & Anandkumar, A. (2016). Experimental results : Reinforcement learning of pomdps using spectral methods. In *Jmlr: Workshop and conference proceedings* (Vol. 49, pp. 1–64). arXiv: 1705.02553. Retrieved from http://arxiv.org/abs/1705.02553

Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI and Society*, (October), 1–12. doi:10.1007/s00146-017-0760-1

Besser-jones, L. (2008). Social psychology , moral character , and moral fallibility. *International Phenomenological Society*, *76*(2), 310–332. Retrieved from https://www.jstor.org/stable/40041173

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. doi:10.1126/science.aaf2654. arXiv: 1510.03346

Bostrom, N., & Yudkowsky, E. (2011). *The ethics of artificial intelligence* (W. Ramsey & K. Frankish, Eds.). Cambridge University Press Cambridge. Retrieved from http://www.nickbostrom.com/ethics/artificial-intelligence.pdf

Bringsjord, S., & Govindarajulu, N. S. (2020). Artificial intelligence. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 202). Metaphysics Research Lab, Stanford University.

Britz, K., Meyer, T., & Varzinczak, I. (2011). Semantic foundation for preferential description logics. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. doi:10.1007/978-3-642-25832-9_50

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental and Theoretical Artificial Intelligence*, *26*(3), 355–372. doi:10.1080/0952813X.2014.895108

Brys, T., Harutyunyan, A., Suay, H. B., Chernova, S., Taylor, M. E., & Nowé, A. (2015). Reinforcement learning from demonstration through shaping. In *Proceedings of the international joint conference on artificial intelligence (ijcai)* (pp. 3352–3358). AAAI Press.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AI Magazine*, *26*(4), 53–60.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *The Character of Consciousness*, *2*(3), 200–219. doi:10.1093/acprof. arXiv: 0402594v3 [arXiv:cond-mat]

Christman, J. (2018). Autonomy in moral and political philosophy. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 201). Metaphysics Research Lab, Stanford University.

Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton University Press.

Churchland, P. S. (2014). The neurobiological platform for moral values. *Behaviour*, *151*(2-3), 283–296. doi:10.1163/1568539X-00003144

Cloos, C. (2005). The utilibot project: An autonomous mobile robot based on utilitarianism. *Machine Ethics: Papers from the 2005 AAAI Fall Symposium*, 38–45. Retrieved from philpapers.org/archive/CLOTUP.2.pdf

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-cartesian moral hermeneutics. *Philosophy and Technology*, *27*(1), 61–77. doi:10.1007/s13347-013-0133-8

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Isaim* (pp. 4831–4835). Retrieved from www.aaai.org

Cowls, J., & Floridi, L. (2018). Prolegomena to a white paper on an ethical framework for a good ai society. *SSRN Electronic Journal*. doi:10.2139/ssrn.3198732

Daily, M., Medasani, S., Behringer, R., & Trivedi, M. (2017). Self-driving cars. *Computer*, *50*(12), 18–23. doi:10.1109/MC.2017.4451204

Dameski, A. (2018). A comprehensive ethical framework for ai entities : Foundations. In M. Iklé, A. Franz, R. Rzepka, & B. Goertzel (Eds.), *International conference on artificial general intelligence* (July, pp. 42–51). doi:https://doi.org/10.1007/978-3-319-97676-1

Deng, B. Y. B. (2015). The robot's dilemma. *Nature*, *523*(7558), 24.

Dennett, D. (1976). Conditions of personhood. In A. Rorty (Ed.), *The identities of persons* (pp. 176–196). Springer.

Dignum, V. (2017). Responsible autonomy. In *Proceedings of the 26th international joint conference on artificial intelligence* (pp. 4698–4704).

Dreyfus, H., Dreyfus, S., & Athanasiou, T. (2000). *Mind over machine*. Simon & Schuster. Retrieved from https://books.google.co.za/books?id=e9W9m%5C_4q4pYC

Duan, Y., Andrychowicz, M., Stadie, B., & Ho, J. (2017). One-shot imitation learning. In I. Guyon, U. V. L. Garnett, S. Bengio, H. Wallach, R. F. R., & S. Vishwanathan (Eds.), *Advances in neural information processing systems 30* (pp. 1087–1098). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/6709-one-shot-imitation-learning.pdf

Fischer, J. M. (1999). Recent work on moral responsibility. *Ethics*, *110*(1), 93–139. doi:10.1086/233206

Fitch, G. W. (1987). *Naming and believing*. Springer, Dordrecht.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology*, *1*(1), 37–56. doi:10.4324/9781315259697-11

Floridi, L. (2008). Information ethics: A reappraisal. *Ethics and Information Technology*, *10*(2-3), 189–204. doi:10.1007/s10676-008-9176-4

Floridi, L. (2013). Distributed morality in an information society. *Science and Engineering Ethics*, *19*(3), 727–743. doi:10.1007/s11948-012-9413-4

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, *14*(3), 349–379. doi:10.2139/ssrn.1124296

Franklin, S. (2003). A conscious artifact? *Journal of Consciousness Studies*, *10*(4-5), 47–66.

Franklin, S., Madl, T., Mello, S. D., & Snaider, J. (2014). Lida : A systems-level architecture for cognition , emotion , and learning. *IEEE Transactions on Autonomous Mental Development*, *6*(1), 19–41.

Gangwani, T., Lehman, J., Liu, Q., & Peng, J. (2019, July 22). Learning belief representations for imitation learning in POMDPs. *35th Conference on Uncertainty in Artificial Intelligence (UAI 2019)*. arXiv: 1906.09510

Genewein, T., Leibfried, F., Grau-Moya, J., & Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, *2*(November), 1–24. doi:10.3389/frobt.2015.00027

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. doi:10.1126/science.aac6076. arXiv: arXiv:1011.1669v3

Gert, B. (1988). *Morality*. New York: Oxford University Press.

Giovanni, B., & Gabriele, D. (Eds.). (2006). *Evolutionary ethics and contemporary biology*. Cambridge University Press.

Gips, J. (1995). Towards the ethical robot. In K. M. Ford, C. Glymour, & P. Hayes (Eds.), *Android epistemology* (Vol. 9780521112, *May*, pp. 244–253). doi:10.1017/CBO9780511978036.015

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. doi:10.1007/978-3-540-68677-4

Goodall, N. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2424*(2424), 58–65. doi:10.3141/2424-07

Goodfellow, I., JBengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Guo, Z. D., Doroudi, S., & Brunskill, E. (2016). A PAC RL algorithm for episodic POMDPs. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, *51*. arXiv: 1605.08062

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. doi:10.1177/0008125619864925

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), 19–29. doi:10.1007/s10676-008-9167-5

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 4565–4573). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/6391-generative-adversarial-imitation-learning.pdf

Holstein, T. (2017). The misconception of ethical dilemmas in self-driving cars. *Proceedings*, *1*(3), 174. doi:10.3390/is4si-2017-04026

Hooker, J. N., & Kim, T. W. (2018). Toward non-intuition-based machine ethics. *AAAI & ACM Conference on Artifical Intelligence, Ethics, and Society*.

Horvitz, E. J. (1987). Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the third workshop on uncertainty in artificial intelligence* (July, pp. 429–444). AAAI and Association for Uncertainty in Artificial Intelligence. Retrieved from http://erichorvitz.com/u87.htm

Horvitz, E. J. (1988). Reasoning under varying and uncertain resource constraints. In *Aaai* (pp. 111–116). Retrieved from ftp://ftp.research.microsoft.com/pub/ejh/ai88.pdf%7B%5C%%7D5Cnhttp://citeseer.ist.psu.edu/151357.html%7B%5C%%7D5Cnhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.4260

Horvitz, E. J. (1989). Rational metareasoning and compilation for optimizing decisions under bounded resources. In *Proceedings of computational intelligence '89*, Milan, Italy: Association of Computing Machinery. Retrieved from http://erichorvitz.com/rationality%7B%5C_%7D89.htm

Horvitz, E. J., Cooper, G. F., & Heckerman, D. E. (1989). Reflection and action under scarce resources: Theoretical principles and empirical study. In *Ijcai* (Vol. 2, pp. 1121–1127).

Hursthouse, R., & Pettigrove, G. (2018). Virtue ethics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University.

Igl, M., Zintgraf, L., Le, T. A., Wood, F., & Whiteson, S. (2018). Deep variational reinforcement learning for POMDPs. *35th International Conference on Machine Learning, ICML 2018*, *5*, 3359–3375. arXiv: 1806.02426

Jaakkola, T., Singh, S. P., & Jordan, M. I. (1995). Reinforcement learning algorithm for partially observable markov decision problems. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems 7* (pp. 345–352). MIT Press. Retrieved from http://papers.nips.cc/paper/951-

reinforcement-learning-algorithm-for-partially-observable-markov-decision-problems.pdf

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. doi:10.1136/svn-2017-000101. eprint: https://svn.bmj.com/content/2/4/230.full.pdf

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Machine Ethics*, *9780521112*, 168–183. doi:10.1017/CBO9780511978036.012

Kasenberg, D., Arnold, T., & Scheutz, M. (2018). Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 184–190). ACM.

Katt, S., Oliehoek, F. A., & Amato, C. (2019). Bayesian reinforcement learning in factored POMDPs. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*, 7–15. arXiv: 1811.05612

Kiela, D. (2017). *Deep embodiment: grounding semantics in perceptual modalities*. University of Cambridge, Computer Laboratory. Retrieved from http://www.cl.cam.ac.uk/

Klimenko, D., Song, J., & Kurniawati, H. (2014). TAPIR: A software toolkit for approximating and adapting pomdp solutions online. *Australasian Conference on Robotics and Automation, ACRA*, *02-04-Dece*, 2–4.

Kotsonis, A. (2020). On the limitations of moral exemplarism: Socio-cultural values and gender. *Ethical Theory and Moral Practice*. doi:10.1007/s10677-020-10061-8

Kripke, S. A. (1972). Naming and necessity. *Semantics of Natural Language*, 253–355. doi:10.1007/978-94-010-2557-7_9

Kuipers, B. (2016). Human-like morality and ethics for robots. *AAAI-16 Workshop on AI, Ethics and Society*, 98–104.

Kurniawati, H., & Patrikalakis, N. M. (2013). Point-based policy transformation: Adapting policy to changing pomdp models. *Springer Tracts in Advanced Robotics*, *86*, 493–509. doi:10.1007/978-3-642-36279-8_30

Kurniawati, H., & Yadav, V. (2016). An online POMDP solver for uncertainty planning in dynamic environment. *Springer Tracts in Advanced Robotics*, *114*, 611–629. doi:10.1007/978-3-319-28872-7_35

Leviathan, Y., & Matias, Y. (2017). Google ai blog: Google duplex: An ai system for accomplishing real-world tasks over the phone. Retrieved October 17, 2018, from https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Levinson, M., & Fay, J. (Eds.). (2016). *Dilemmas of educational ethics: Cases and commentaries*. Harvard Education Press.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. doi:10.1111/tops.12086

Li, Y. (2018). Deep reinforcement learning. *CoRR*, *abs/1810.0*. arXiv: 1810.06339. Retrieved from http://arxiv.org/abs/1810.06339

Liao, S. M. (2010). The basis of human moral status. *Journal of Moral Philosophy*, *7*(2), 1–31. doi:https://doi.org/10.1163/174552409X12567397529106

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. In *Artificial intelligence* (Vol. 175, *5-6*, pp. 942–949). doi:10.1016/j.artint.2010.11.026

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2014). *Robot ethics: The ethical and social implications of robotics*. The MIT Press.

Lin, P., Bekey, G., & Abney, K. (2008). Autonomous military robotics: Risk, ethics, and design. *California Polytechnic State University*, 108. Retrieved from http://ethics.calpoly.edu/ONR%7B%5C_%7Dreport.pdf

Littman, M. L. (2009). A tutorial on partially observable Markov decision processes. *Journal of Mathematical Psychology*, *53*(3), 119–125. doi:10.1016/j.jmp.2009.01.005

Lucentini, D. F., & Gudwin, R. R. (2015). A comparison among cognitive architectures : A theoretical analysis. *Procedia - Procedia Computer Science*, *71*(January 2016), 56–61. doi:10.1016/j.procs.2015.12.198

Lumbreras, S. (2017). The limits of machine ethics. *Religions*, *8*(6), 100. doi:10.3390/rel8050100

Lutz, C., & Tamò, A. (2015). Robocode ethicists - privacy-friendly robots, an ethical responsibility of engineers? *Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research - NS Ethics '15*, 27–28. doi:10.1145/2793013.2793022

Mabaso, B. A. (2020a). Artificial moral agents within an ethos of ai4sg. *Philosophy and Technology*. doi:10.1007/s13347-020-00400-z

Mabaso, B. A. (2020b). Computationally rational agents can be moral agents. *Ethics and Information Technology*. doi:10.1007/s10676-020-09527-1

Marwala, T. (2013). Semi-bounded rationality - a model for decision making. *arXiv preprint arXiv:1305.6037*, 153–164. Retrieved from https://arxiv.org/abs/1305.6037

Mayo, M. J. (2003). Symbol grounding and its implications for artificial intelligence. In *Proceedings of the 26th australasian computer science conference-volume 16* (Vol. 16, pp. 55–60). Darlinghurst, Australia: Australian Computer Society, Inc. Retrieved from http://portal.acm.org/citation.cfm?id=783106.783113%7B%5C&%7Dtype=series

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, *27*(4), 12–14. doi:http://dx.doi.org/10.1609/aimag.v27i4.1904. arXiv:9809069v1 [arXiv:gr-qc]

Miller, C. B. (2013, May). *Moral character*. doi:10.1093/acprof:oso/9780199674350.001.0001

Miller, F. D. (1984). Aristotle on rationality in action. *The Review of Metaphysics, 37*(3), 499–520. Retrieved from https://www.jstor.org/stable/20128047

Miller, T. R., Baird, T. D., Littlefield, C. M., Kofinas, G., Chapin III, F. S., & Redman, C. L. (2008). Epistemological pluralism: reorganizing interdisciplinary research. JSTOR.

Moll, J., Zahn, R., De Oliveira-Souza, R., & Krueger, F. (2005). The neural basis of human moral cognition. *Nature reviews neuroscience*, *6*(10), 799. doi:10.1038/nrn1768

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, *21*(4), 18–21. doi:10.1109/MIS.2006.80

Müller, V. C. (n.d.). Ethics of artificial intelligence. In A. Elliott (Ed.), *The routledge social science handbook of ai* (pp. 1–20). London: Routledge.

Muntean, I., & Howard, D. (2016). A minimalist model of the artificial autonomous moral agent (AAMA). *SSS-16 Symposium Technical Reports. Association for the Advancement of Artificial Intelligence. AAAI.* 217–225. Retrieved from http://www.aaai.org/ocs/index.php/SSS/SSS16/paper/view/12760/11954%7B%5C%%7D5Cnhttp://www.aaai.org/ocs/index.php/SSS/SSS16/paper/view/12760

Murphy, K. P. (2000). A survey of POMDP solution techniques. *Environment*, *2*(September), X3. doi:10.1007/BF02204836

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program. In *Ifip congress* (Vol. 256, p. 64). Pittsburgh, PA.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth international conference on machine learning* (pp. 663–670). ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Noorman, M. (2018). Computing and moral responsibility. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Springer 2). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility/

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289. doi:10.1007/s10677-016-9745-2

Oshana, M. A. L. (1997). Ascriptions of responsibility. *American philosophical quarterly (Oxford)*, *34*(1), 71–83. doi:10.2307/20009887

Oshana, M. A. L. (2002). The misguided marriage of responsibility and autonomy. *The Journal of Ethics*, *6*(3), 261–280. doi:10.1023/A:1019482607923

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*, *5*(2), 105–129. Retrieved from https://pdfs.semanticscholar.org/3ff2/49fe3c8b3a2c94ae762b76b2dd0203f1f789.pdf

Parthemore, J., & Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness*, *6*(2), 141–161. doi:10.1142/S1793843014400162

Peterson, M. (2009). *An introduction to decision theory*. Cambridge Introductions to Philosophy. doi:10.1017/CBO9780511800917

Pontier, M. A., & Hoorn, J. F. (2012). Toward machines that behave ethically better than humans do. *Proceedings of the Annual Meeting of the Cognitive Science Society*, (34).

Pontier, M. A., Widdershoven, G., & Hoorn, J. F. (2012). Moral Coppélia-Combining ratio with affect in ethical reasoning. In J. Pavón, N. Duque-Méndez, & R. Fuentes-Fernández (Eds.), *Advances in artificial intelligence – iberamia 2012* (pp. 442–451). doi:https://doi.org/10.1007/978-3-642-34654-5_45

Poulsen, A., Anderson, M., Anderson, S. L., Byford, B., Fossa, F., Neely, E. L., ... Winfield, A. (2019). Responses to a critique of artificial moral agents. *CoRR*, *abs/1903.07021*. arXiv: 1903.07021. Retrieved from http://arxiv.org/abs/1903.07021

Prasad, M. (2018). Social choice and the value alignment problem. In *Artificial intelligence safety and security* (pp. 291–314). Chapman and Hall/CRC.

Reiter, R. (1988). Nonmonotonic reasoning. *Exploring Artificial Intelligence*, 439–481. doi:10.1016/b978-0-934613-67-5.50016-2

Rhoten, D., & Parker, A. (2004). Risks and rewards of an interdiciplinary research path. *Science*, *306*(5704), 2046. doi:10.1126/science.1103628

Roskies, A. (2016). Neuroethics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 201). Metaphysics Research Lab, Stanford University.

Ross, W. D. (1930). The right and the good.

Rottschaefer, W. A. (2000). Naturalizing ethics: The biology and psychology of moral agency. *Zygon*, *35*(5-6), 253–286. doi:10.1111/0591-2385.00276

Rottschaefer, W. A. (2009). Moral agency and moral learning : Transforming metaethics from a first to a second philosophy enterprise. *Behavior and Philosophy*, *37*, 195–216. Retrieved from https://www.jstor.org/stable/41472435

Russell, S. J., & Norvig, P. (2009). *Artifical intelligence: A modern approach* (Third edit). doi:10.1017/S0269888900007724. arXiv: arXiv:1011.1669v3

Russell, S. J., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artiicial Intelligence Research*, *2*, 575–609.

Ruttkamp-Bloem, E. (n.d.). What is the ethics of artificial intelligence?

Sapaty, P. S. (2015). Military robotics: Latest trends and spatial grasp solutions. *IJARAI International Journal of Advanced Research in Artificial Intelligence*, *4*(4), 9–18.

Saptawijaya, A., & Pereira, L. M. (2014). Towards modeling morality computationally with logic programming. *Lecture Notes in Computer Science (including*

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *8324 LNCS*, 104–119. doi:10.1007/978-3-319-04132-2_8

Scheutz, M. (2017). The case for explicit ethical agents. *AI Magazine*, *38*(4), 57–64. doi:10.1609/aimag.v38i4.2746

Scheutz, M., & Malle, B. F. (2017). Moral robots. In L. S. M. Johnson & K. S. Rommelfanger (Eds.), *The routledge handbook of neuroethics* (Chap. 24). doi:10.4324/9781315708652.ch24

Schlosser, M. (2015). Agency. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2015/entries/agency/

Schuss, Z. (2010). *Theory and applications of stochastic processes: an analytical approach*. doi:10.1007/978-1-4419-1605-1

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.

Searle, J. R. (2007). Biological naturalism. In S. Schneider & M. Velmans (Eds.), *The blackwell companion to consciousness* (pp. 325–334). Blackwell Publishing Oxford.

Selten, R. (1990). Bounded rationality. *Journal of Institutional and Theoretical Economics (JITE)*, *146*(4), 649–658.

Shani, G., Pineau, J., & Kaplow, R. (2013). A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, *27*(1), 1–51. doi:10.1007/s10458-012-9200-2

Si, T., & Zhao, H. (2016). A brief overview of synthetic biology research programs and roadmap studies in the United States. *Synthetic and Systems Biotechnology*, *1*(4), 258–264. doi:10.1016/j.synbio.2016.08.003

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 1–19. doi:10.1002/acn3.501. arXiv: 1712.01815

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354. doi:10.1038/nature24270

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99–118.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Slote, M. (1995). Agent-based virtue ethics. *Midwest studies in philosophy*, *20*(1), 83–101.

Strawson, G. (1986). *Freedom and belief*. Oxford University Press.

Sullins, J. P. (2006). When is a robot a moral agent? *IRIE: International Review of Information Ethics*. Retrieved from http://sonoma-dspace.calstate.edu/handle/10211.1/427

Sullins, J. P. (2010). RoboWarfare: Can robots be more ethical than humans on the battlefield? *Ethics and Information Technology*, *12*(3), 263–275. doi:10.1007/s10676-010-9241-7

Swanton, C. (2003). *Virtue ethics: A pluralistic view*. doi:10.1093/0199253889.001.0001

Szutta, N. (2019). Exemplarist moral theory–some pros and cons. *Journal of Moral Education*, *48*(3), 280–290. doi:10.1080/03057240.2019.1589435

Thomson, J. J. (1985). The trolley problem. *Ethics: Problems and Principles*, (May), 67–76. doi:10.2307/796133

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI and Society*, *22*(4), 495–521. doi:10.1007/s00146-007-0091-8

Torrance, S. (2013). Artificial agents and the expanding ethical circle. *AI and Society*, *28*(4), 399–414. doi:10.1007/s00146-012-0422-2

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, *20*(1), 27–40. doi:10.1007/s10676-017-9440-6

Van Noorden, R. (2015). Interdisciplinary research by the numbers. *Nature*, *525*(7569), 306–307.

Van Zyl, L. (2005). In defence of agent-based virtue ethics. *Philosophical Papers*, *34*(2), 273–288. Retrieved from http://eprints.uanl.mx/5481/1/1020149995.PDF

van Lent, M., & Laird, J. E. (2001). Learning procedural knowledge through observation. In *K-cap* (pp. 179–186). doi:10.1145/500737.500765

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, *25*(3), 719–735. doi:10.1007/s11948-018-0030-8

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, *48*, 56–66. doi:10.1016/j.cogsys.2017.04.002

Wallach, W., Allen, C., & Franklin, S. (2011). Consciousness and Ethics: Artificially Conscious Moral Agents. *International Journal of Machine Consciousness*, *03*(01), 177–192. doi:10.1142/S1793843011000674

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, *2*(3), 454–485. doi:10.1111/j.1756-8765.2010.01095.x

Walraven, E., & Spaan, M. T. J. (2017). Accelerated vector pruning for optimal POMDP solvers. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 3672–3678.

Wang, Y. [Yue], Chaudhuri, S., & Kavrakl, L. E. (2018). Bounded policy synthesis for POMDPs with safe-reachability objectives: Robotics track. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*(July), 238–255.

Wang, Y. [Yuhui], He, H., & Tan, X. (2019). Robust Reinforcement Learning in POMDPs with Incomplete and Noisy Observations. *arXiv preprint arXiv:1902.05795*. arXiv: 1902.05795. Retrieved from http://arxiv.org/abs/1902.05795

Wasserman, P. D., & Schwartz, T. (1988). Neural networks. II. What are they and why is everybody so interested in them now? *IEEE Expert*, *3*(1), 10–15. doi:10.1109/64.2091

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227–248. Retrieved from http://www.jstor.org/stable/43154245

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

White, J. (2013). Manufacturing morality a general theory of moral agency grounding computational implementations: The actwith model. In Floares (Ed.), *Computational intelligence* (pp. 1–65). Nova Publications.

Wu, Y.-H., & Lin, S.-D. (2018). A low-cost ethics shaping approach for designing reinforcement learning agents. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. arXiv: 1712.04172. Retrieved from http://arxiv.org/abs/1712.04172

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 5527–5533. Retrieved from http://moralmachine.mit.edu/

Zagzebski, L. (2010). Exemplarist virtue theory. *Metaphilosophy*, *41*(1-2), 41–57. doi:10.1111/j.1467-9973.2009.01627.x

Zagzebski, L. (2017). *Exemplarist moral theory*. doi:10.1093/acprof:oso/9780190655846.001.0001

Zeng, D. (2015). AI ethics: Science fiction meets technological reality. *IEEE Intelligent Systems*, *30*(3), 2–5. doi:10.1109/MIS.2015.53

Zhang, Z., Fu, Q., Zhang, X., & Liu, Q. (2016). Reasoning and predicting POMDP planning complexity via covering numbers. *Frontiers of Computer Science*, *10*(4), 726–740. doi:10.1007/s11704-015-5038-5

Zhu, P., Li, X., & Poupart, P. (2017). On improving deep reinforcement learning for POMDPs. *CoRR*, *abs/1704.0*. arXiv: 1704.07978. Retrieved from http://arxiv.org/abs/1704.07978

Zilberstein, S. (2013). Metareasoning and bounded rationality. *Metareasoning*, 27–40. doi:10.7551/mitpress/9780262014809.003.0003