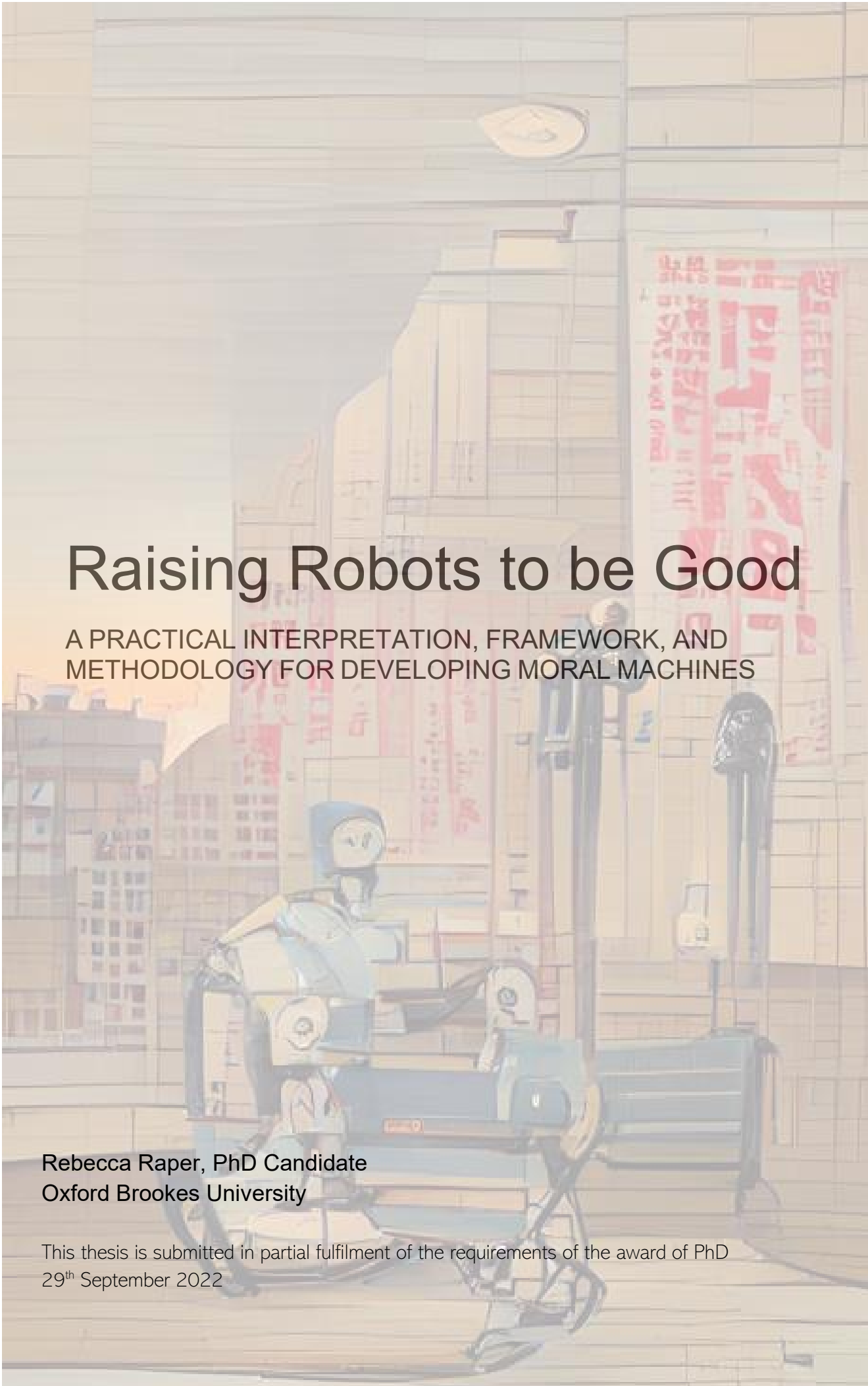




Raising Robots to be Good

A PRACTICAL INTERPRETATION, FRAMEWORK, AND
METHODOLOGY FOR DEVELOPING MORAL MACHINES

Rebecca Raper, PhD Candidate
Oxford Brookes University

This thesis is submitted in partial fulfilment of the requirements of the award of PhD
29th September 2022

“What sort of world can we prepare for our great grandchildren?” – Richard Rorty

For Lilly and Jack

Abstract

Over recent years there has been a developing interest in AI and ethics, with some researchers attempting to bestow upon machines the capacity for them to make their own ethical decisions. These recent attempts have raised a significant amount of attention, largely owing to their failure. Where Machine Ethics can be described as the pursuit to create machines with the capacity to make moral decisions, this thesis carries out an in-depth exploration and investigation into this area, and specifically, how we might create machines with morals. In turn, it proposes a practical framework, methodology and then design for their implementation through appeal to practical case studies, philosophical and psychological insight. The central ideas behind this thesis are that Machine Ethics requires a reframing – *to Machine Ethics 2.0* – and that instead of trying to create machines that simply replicate (or mimic) human morality, we should be trying to create machines that are equipped with the necessary features to develop *their own moral outlook*. In turn, this means that rather than simply bestowing upon machines moral decision-making, influenced by our own *human* outlook of the world, supported by the necessary architecture for moral agency, we should be looking at the developmental framework for cultivating moral growth in our machines: *raising robots to be good*.

CONTENTS

- 1. INTRODUCTION – PG. 3**
- 2. LITERATURE REVIEW: THE MACHINE ETHICS STORY – PG.8**
 - 2.1. INTRODUCING MACHINE ETHICS
 - 2.2. CREATING MACHINES WITH MORALS
 - 2.3. ARGUMENTS AGAINST MORAL MACHINES
 - 2.4. VERIFYING MORAL AGENCY
 - 2.5. AI ALIGNMENT, AND MODERN MACHINE ETHICS
 - 2.6. *FINAL NOTE*: ARE MORAL MACHINES EVEN POSSIBLE?
- 3. REFRAMING MACHINE ETHICS – PG. 24**
 - PART ONE: THE CHALLENGE**
 - 3.1. ARE MORAL MACHINES REALLY IMPOSSIBLE?
 - 3.2. ISSUES WITH PREVIOUS ATTEMPTS
 - PART TWO: THE RESPONSE**
 - 3.3. A CASE STUDY IN SOCIAL ROBOTICS
 - 3.4. INTRODUCING METAETHICS
 - 3.5. MACHINE ETHICS 2.0
- 4. A FRAMEWORK FOR MORAL MACHINES – PG. 35**
 - 4.1. WHY WE WANT MORAL MACHINES
 - 4.2. TESTING MORALITY
 - 4.3. WHAT IS MORAL AGENCY?
 - 4.4. LESSONS FROM MORAL PSYCHOLOGY
 - 4.5. A DEVELOPMENT FRAMEWORK FOR MACHINE ETHICS
- 5. A RECIPE FOR MORALITY – PG.44**
 - 5.1. WHAT MORALITY NEEDS
 - 5.2. A MORAL SPECIFICATION
 - 5.3. USING THE TEST FRAMEWORK
 - 5.4. MORAL MACHINES IN PRACTICE
 - 5.5. REFLECTIONS AND FUTURE WORK
- 6. DESIGNING A MORAL AGENT – PG. 60**
 - 6.1. A COGNITIVE MODEL OF MORALITY
 - 6.2. AN ARCHITECTURE FOR MORAL DEVELOPMENT
 - 6.3. RAISING ROBOTS IN THE WORLD
 - 6.4. TESTING AND FUTURE WORK
- 7. THE ETHICS OF MACHINE ETHICS - PG. 70**
 - 7.1. PRE-EXISTING CONCERNS
 - 7.2. IMPLICATIONS OF THE NEW APPROACH
 - 7.3. NEXT STEPS
- 8. SUMMARY, FUTURE WORK AND CONCLUSION – PG. 76**
- 9. ACKNOWLEDGEMENTS – PG. 78**
- 10. REFERENCES – PG. 79**
- 11. APPENDICES – PG. 88**
 - 11.1. RELATED PUBLICATIONS
 - 11.2. OTHER MEDIA

1. INTRODUCTION

In 1950 Isaac Asimov famously introduced us to the idea of *Machines with Morals* in his popular science fiction series 'Runaround' (Asimov, 1950). In a future where robots work alongside humans to provide services such as policing, care and engineering, the challenge that is introduced is how to ensure that the robots behave in ways that are *conducive to humans* – aka, that they behave *morally*.

Leaving aside (for now) the question of whether behaving in a way that is conducive to humans is the same as behaving morally, Asimov created a set of rules in the stories from which the robots must adhere to ensure that they are moral. These are summarised as the following:

First Law

A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law

A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

In subsequent stories a *zeroth* law is also introduced to allow for more societal impacts:

Zeroth Law

A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Though initially intuitive, the laws were designed by Asimov to make for engaging stories, therefore, as the stories go on to demonstrate, various issues materialise if morality is imbued into robots in this way.

The first issue relates to the fact that the laws are mildly ambiguous. '*Not harming a human*' is not as clear as it may initially seem, particularly because robots require very explicit commands to understand how to behave appropriately. What if harming a human (in some respect) is the only way to get the human out of a life-threatening situation? What if harming one human would legitimately prevent the harm of another (or several other) human(s)? Furthermore, the term 'harm'

is not so easy to define since *my* notion of *harm* might be dependent on a specific context. Therefore, it's not a universal term that can be programmed into a machine.

Secondly, though the rules offer a good reflective starting point, they are not *all-encompassing*. This means that they allow for some situations that we might typically regard as morally dubious. For example, suppose the best way to prevent humanity coming to harm (as demonstrated in the 'I-Robot' film version of the stories) is to prevent humanity – i.e., stop human existence altogether. Perhaps a slightly pessimistic outlook, it does highlight a more fundamental issue: that specifying a particular action, cannot allow for every circumstantial variation. Giving machines moral capacity requires the need for a *context specificity* that cannot be captured by these types of laws.

Since Asimov's stories, there have been many popular cultural references to machines with morals. In Stephen Spielberg's film 'Artificial Intelligence' we are introduced to a robot with the capacity to *love* (but not be loved), in Disney's 'Big Hero 6', we witness the leading robot character, Baymax, having a component removed that is intended to prevent it harming humans, witnessing the catastrophic effects afterwards. Similarly, in 'Robot and Frank', we observe a robot forming a relationship with a human that leads to the human being advised to carry out morally dubious acts, such as stealing. There are many other references: in books ('The Wizard of Oz'); TV ('*Humans*'), and computer games ('Detroit: become human'). There appears to be a human fascination with the seemingly contradictory phenomenon of *machines with morals*.

So how can something so seemingly *distinctly* human (morality) be integrated with a machine (something that is *not* human)? Do we even want to achieve this, and if so, where should we start to accomplish it?

Machine Ethics (Anderson et al, 2004) is the academic field that tries to answer the question of how we create *machines with morals*. In recent years there has been a significant increase in the amount of research focusing on this area, with notable attempts using philosophical insight into morality (Ethics) to guide how such machines should be designed and created. However, there have been problems with the attempts, leading some to conclude that true *morality in a machine* is simply not possible (Coeckelbergh, 2010 and Sparrow, 2021).

It is due to arguments such as these, twinned with arguments to suppose that *it might be an increased risk to create moral machines*, along with recent failures in creating moral machines (see *Delphi* – Jiang et al, 2021) that has led to the unpopularity of Machine Ethics in recent years. However, with recent advances in Artificial Intelligence (i.e. ChatGPT), AI Risk has become a topic of public interest (see BBC News, 2023), with a similar imperative (termed 'AI Alignment' (see Christian, 2020)) taking centre stage as the resolution to supposed issues. The aim of this thesis is to highlight the research that has previously taken place within Machine Ethics, bringing about its re-emergence by offering a practical reinterpretation of what Machine Ethics is trying to achieve.

This thesis advances an argument to show that morality in a machine *is* something that is possible – and is something we should be researching to achieve – *if* the approach to attempting to create moral machines is **reframed** away from morality in humans, towards a non-anthropocentric account focused away from how humans behave. Rather than just trying to give our machines moral capacity – such as human morals – we need to bestow machines with the ability to develop *their own moral outlook*. That is, an architectural computational framework is required that allows machines to form their own moral view of the world.

Under this *reframing* of the approach to create moral machines, a methodology and then architectural design for their creation is developed. The idea is, that if researchers in the field work towards bestowing upon machines the capacity to *develop their own moral outlook*, we can begin to make significant progress in the area. The thesis offers the following contributions:

- 1) A re-examination of the Machine Ethics problem (*Why* do we want moral machines?)
- 2) A re-interpretation of *what* we want in a moral machine.
- 3) A Pathway ('way out') for achieving this, along with a prototype methodological framework for what needs to be done to get there.

There is also suggestion of a cognitive architecture for moral agency, along with the non-anthropocentric account for what it means to be moral.

Perhaps the most controversial assertion made within this thesis is that we must learn to *trust* machines if we want them to be moral.

Each argument is underpinned by multidisciplinary insights and perspectives from developmental psychology and philosophy, thus inspiring the main title of this thesis: *Raising Robots to be Good*.

It is worth noting that there is a field in robotics known as developmental robotics which looks at how robotics research can be inspired by work in the developmental literature. Though this thesis takes its inspiration also from how a child might develop their own moral agency, the central point of the thesis is to advocate for the shift in perspective, along with development of a framework to facilitate work in this area, rather than to formulate technical specifications. Therefore, though the technical elements of developmental robotics might apply to operationalisation of the moral machine in the future, they are not included in this thesis. A comprehensive survey of this area can be found via Lungarella et al, 2003.

Before these points can be covered, it will be important to gain an understanding of the academic landscape so far, so the thesis begins with an overview of research in Machine Ethics to date, with descriptions of the problems researchers working in the area have faced.

The chapters of the thesis proceed as follows: (2) presents a literature review and overview of Machine Ethics and related literature, to give some background to the area and provide a foundation from which to develop the main ideas; (3) then discusses some of the challenges against the possibility of moral machines, before using a practical case study and more *metaethical* philosophical accounts to introduce a reframing of the Machine Ethics problem – termed *Machine Ethics 2.0* – which guides the rest of the thesis; (4) begins to trace a framework for developing, testing and assessing moral machines, based upon Machine Ethics 2.0, and the motivation for development of moral machines; (5) then works through this new framework, with some examples and analysis, and (6) uses this example to begin to formulate a cognitive model and then architecture for moral learning, with the assertion being that the way to create moral machines is to *raise them to be good*. Finally, (7) provides discussion of the broader ethical implications related to the new approach and the creation of Moral Machines generally.

At the end of this thesis, sat within the Appendices, are several papers that were written throughout the course of the time spent working on this research. Some are more relevant than others to this specific project, but all contributed in some shape to the final outcome. The first, 'Recruitment AI has a Disability Problem', highlights the issue of disability for recruitment AI systems, demonstrating that disability is difficult to mitigate in terms of bias because each individual with a disability is different. This is relevant to this thesis because it highlights some of the more practical Ethical issues that AI systems face. The second paper 'Computing Morality: Synthetic Decision-Making and Behaviour' is a editorial introduction for a special issue of the same name. It highlights some important issues with trying to create moral machines, including an argument from climate change to demonstrate why crowd-sourced moral machines exacerbate normal opinions and not strictly morals. The third paper, 'Modelling a socialised chatbot based upon trust development in children: lessons learnt from Tay' develops a theoretical model for trying to resolve the issues that chatbots such as Microsoft's Tay bot have faced, inspired by how children develop socialised behaviours through trust relationships with their parents. The outcome is a network of 'trust' relationships that *grounds* the conversations the chatbot develops, resulting in a chatbot that is not so susceptible to being manipulated by its inputs as Tay was. Conversations for this paper substantiated the idea that machines require some kind of metaphysical *grounding* to be moral agents, and led to one of the central later ideas, that the machine requires its own moral perspective to be moral.

The fourth paper, 'Sustainability Budgets: a practical management and governance method for achieving Goal 13 of the Sustainable Development Goals for AI Development' came from a conversation with the CEO of a AI software company, and focuses on practical management techniques to deal with the energy exhaustion issues surrounding the development of AI systems. It is relevant to this thesis, and features in the 'Ethics of Machine Ethics' section because with any new AI development, sustainability must be considered as a matter of priority.

Finally, the paper ‘EKIP: designing a practical framework for embedding ethics in AI software development’ outlines a *methodological gap* in current approaches to govern and regulate AI to ensure that it is ethical, and proposes a practical solution, based upon requirements engineering, and a newly introduced role of *the ethical analyst* to fill this gap. It was a part of a year-long separate research project between The University of Vienna and the AI company Gradient Zero. Software engineering as a process featured strongly in this research, and is also used to guide the eventual framework of this thesis.

The conclusion at the end of this thesis is that Moral Machines *are* possible, if, rather than trying to bestow upon machines a conception of moral capacity inspired by humans, we instead try to create machines that can develop moral capacity *by their own accord*. After creating a framework for achieving this, and tracing through a methodology inspired by why we want moral machines in the first place, we see that the best way to do this is, instead of replicating human morals or human moral behaviour, to replicate the *developmental process that humans go through* to develop their own moral perspective. We don’t just need to build moral machines, we must *raise robots to be good*.

2. LITERATURE REVIEW: THE MACHINE ETHICS STORY

Though the human fascination with morality and machines can be traced back quite some years (as evidenced by the fictional examples in the introduction), the story for *Machine Ethics* really began around 15 years ago, with notable introductions to the field, including a textbook on Moral Machines (Wallach and Allen, 2008), and an introduction to the term and its pursuit (Anderson et al, 2004, and Anderson and Anderson, 2007).

According to Anderson and Anderson (2011), *Machine Ethics* can be understood as ‘giving machines... a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making’. It is important to distinguish *Machine Ethics* from broader *AI Ethics* pursuits, which look more at how we create Artificial Intelligence systems in an ethical way (Moor, 2007). Though the two subjects are interconnected (as the final chapter of this thesis discusses), in terms of research, they aim at answering very different questions. **Machine Ethics**: how to design machines such that they behave in a moral way. **AI Ethics**: how to ensure that machines are designed and used ethically.

Morality - or research into it - has in-itself a vast, rich history, which can be linked to psychology, economics, game-theory, or philosophy. In philosophy, it includes the disciplines *ethics* and *meta-ethics*, which aim at answering questions relating to how we ought to live, and whether ‘how we ought to live’ is even a question we should be asking. In economics and game-theory, morality is used to understand human motivations and model decision-making, whereas the psychological study of morality looks at how (and why) we develop moral capacity, including where it might go wrong.

There are other theories of morality that can be taken from sociology, anthropology, and evolutionary theory, and all are important to truly understand morality. However, though theories from each of these areas are interwoven throughout this thesis, they are not outlined in full. The aim of this literature review is to provide an overview of key details relating to the background, progress and issues that currently face the field of Machine Ethics. This includes an overview of the historical journey for Machine Ethics, progress that has been made, and issues surrounding its progress.

The section begins with (2.1) an overview of the background to Machine Ethics, with broad definitions and motivations for its pursuit given; then (2.2) there is an outline of some of the earlier attempts to create machines with morals, with exploration of the philosophical theories inspiring these approaches; (2.3) presents an overview of arguments *against* developing moral machines (with replies); then (2.4) outlines some other debates related to creating Moral Machines – including how we might *assess moral capacity* and *moral cultural variations*; (2.6) describes more

contemporary attempts to create machines with morals, including a brief introduction and comparison to the pursuit for *AI Alignment*, and finally, in (2.7) it is considered whether Moral Machines are even something that are possible.

By giving an overview of key development areas within Machine Ethics and related areas, there should be a sufficiently outlined backdrop to develop the main ideas in this thesis.

2.1. Introducing Machine Ethics

As already described, *Machine Ethics* can be defined as the field of research concerned with how we bestow moral capacity unto machines.

This sits alongside, but in contrast to the rapidly developing field of *AI Ethics*, which concerns how humans should build AI systems in an ethical way (Coeckelbergh, 2020).

Moor (2006) draws a distinction between two ways to imbue ethics into Artificial Intelligence (AI): *Implicit* Ethical AI and *Explicit* Ethical AI. AI that is *implicitly* ethical is ethical insofar as it cannot do anything otherwise. For example, we can constrain an AI system so that it doesn't result in any unethical outcomes simply by engineering mechanisms in place that prevent the AI system from doing anything *wrong*. Moor's argument is that Machine Ethics concerns AI systems that are *explicitly* ethical; that is: AI systems that can operate as ethical agents and make ethical decisions in the world. This capacity, he argues, requires the ability to represent ethical categories and perform analysis in the same way a computer might analyse 'an inventory or tax information' (pg. 19).

Moor also refers to the notion of a *Full Ethical Agent*, as being something that has adult-human level moral capacity and can reason with explanations in moral situations. Arguably, as he explains, *full moral capacity* requires some form of consciousness, and it is for that reason that many believe it is not possible for a machine to have morals. On the other hand, it is difficult to distinguish between baseline explicit ethical capacity, and a full ethical agent, since it seems that any form of moral agency would require full moral agency.

As already mentioned with the Anderson and Anderson (2011) definition given above, many arguments regarding the motivation for moral machines revolve around the need for machines to solve *ethical dilemmas* in a practical way. This line is adopted in Winfield's (2018) articulation of the problem, where special reference is made to the infamous *Trolley Problem*.

'The Trolley Problem' is a popular thought experiment taken from a paper by Foot (1967) in discussion of *the doctrine of double effect* concerning abortion. It goes as follows:

Suppose you are standing by the side of a train track when you see a trolley hurtling out of control towards a fork in the track. Currently, the trolley is set to go down the left fork, where there are

three individuals tied at the bottom. There is a lever by your side, and you quickly realise that by pulling it, you can divert the trolley to the right fork where there is just one individual tied at the end. The scenario can be depicted as per Figure 1.

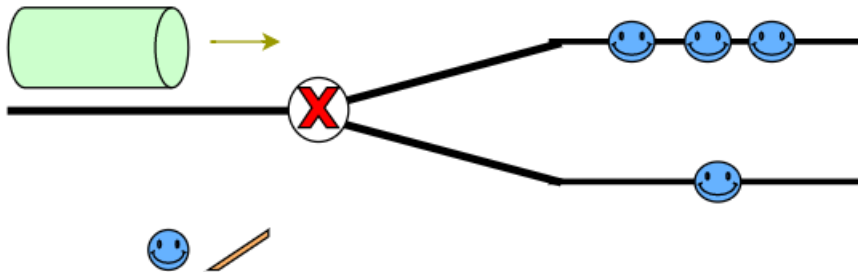


Figure 1: A Trolley Scenario

Intuitively it might seem obvious that pulling the lever is the right thing to do, because fewer lives are at risk if this decision is made. This decision aligns with a philosophical doctrine known as Utilitarianism, which is the theory that in ethical decision-making you carry out the action with the greatest utility ('the greatest good' (Bentham, 1789)). However, the aim of Foot's dilemma was to show that this scenario isn't quite that simple, for example, would it be so obvious to pull the lever if the one individual happened to be an infant, and the three others were terminally ill elderly patients? How do we calculate the utility then? By 'years left to live'?

Though trolley examples aren't so frequent in real life society today, the argument is that Autonomous Vehicles and other Autonomous AI that interact with us will come across similar ethical dilemmas, and be required to resolve them. Therefore, as Autonomous AI is integrated more into society, there is the need to bestow it with moral capacity (namely: the ability to resolve ethical dilemmas).

Himmelreich (2018) (amongst others) argues that The Trolley Problem is not an appropriate case study for Moral Machines for many reasons, including the fact that trolley-type problems would not materialise in real life, for example, because an autonomous vehicle would never be travelling at such a speed that the operator would be required to make the 'life or death' decision. There are replies to this line of argument (i.e., Keeling, 2020) which state that Trolley Problems are still useful if they are taken as *idealised scenarios* and used to understand moral decisions.

Wallach and Allen (2008) argue that the need for moral machines is apparent even outside the scope of ethical dilemmas. They highlight automated fraud detection systems for credit cards and demonstrate that there are ethical implications when decisions must be made about whether to block a card or not. They specify that the requirement for moral capacity in a machine is dependent

upon the level of decision-making autonomy it will have in a situation: *the more autonomy the machine has, the greater the need for moral capacity*. The argument goes that this is particularly pertinent in robots built for social purposes (as described in Duffy, 1999) because the decisions they make will impact human lives. Figure 2 is a model of how the requirements for ethical capacity increase with the autonomy of the machine. It demonstrates that as autonomy increases, so should ethical sensitivity, with levels of moral capacity in the machine reflective of this fact.

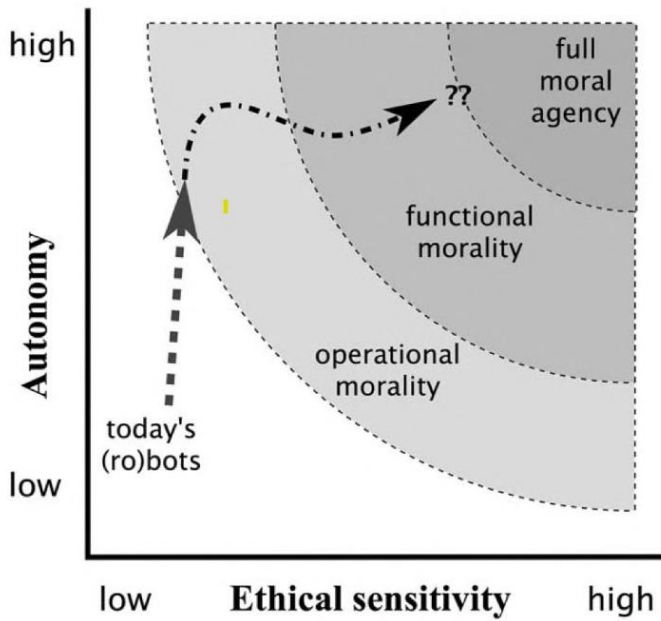


Figure 2: The need for moral capacity as given in Wallach and Allen (2008).

2.2. Creating Machines with Morals

One of the early attempts to create machines with morals was made by Anderson, Anderson and Armen (2006) who quickly realised that using solely rule-based approaches, as per Asimov's Principles, would not work due to reasons discussed in the introduction.

They introduce the case study of a medical AI advisor, MedEthEx, to guide healthcare workers when faced with ethical dilemmas, utilising biomedical ethical principles inspired by Beauchamp (1979) to guide the machine's ethical decision-making. For example, with *non-maleficence* as one such principle, a machine would be able to offer healthcare workers appropriate advice by checking there is no conflict to this principle (i.e., does the proposed decision conflict with *non-maleficence*?). The 'prima facie duties' would be followed unless there is a conflict between duties, in which case Rawls' (1951) 'reflective equilibrium' approach (i.e., that we resolve conflicts by appeal to generalised forms of expert intuitions) would inspire the correct course of action. Applied to Machine Ethics, and MedEthEx, the suggestion is that training examples including specific examples of expert intuitions on the correct actions, should be used to inform what the machine should do.

For example, considering a healthcare worker deciding whether to insist a patient takes their life-saving medication when they have just refused on personal grounds, the decision from an expert might be that it would be *wrong* for the worker to insist, because it would conflict against the principle of *autonomy*, and though, say, *justice* might be infringed upon (because the patient is now at risk, having not taken the medication), in this instance *autonomy* takes precedence. A *hypothesis* based upon scores given for adherence or non-adherence to Beauchamp's principles by the consulted expert is used to inform future decisions for the machine.

It might be argued that in using only expert opinion, the approach detailed above is susceptible to bias, insofar as any prejudices apparent in the minds of the experts will be exacerbated by the AI system. A more recent attempt to bestow morals into a machine using a broader range of opinions is *The Moral Machine Experiment* (Awad et al, 2018), which *crowdsourced* opinions based upon variations of The Trolley Problem. The aim of the experiment was to ask participants to answer a series of questions with variations of Trolley-Style dilemmas (see www.moralmachine.net). Individuals across the globe were involved in the study, however, what was discovered was that there was significant moral opinion variation amongst participants from the different countries. Where individuals from the UK and US might prioritise infants over the elderly when faced with the prospect of an oncoming trolley, in Japan, an environment with different cultural backgrounds, the opposite was the case (see Figure 3). It essentially highlights an issue with the crowdsource-based approach, namely, if moral capacity is given to a machine using crowdsourcing, the decision of the trolley will come down to the most popular opinion (i.e., *the country with the largest population*).

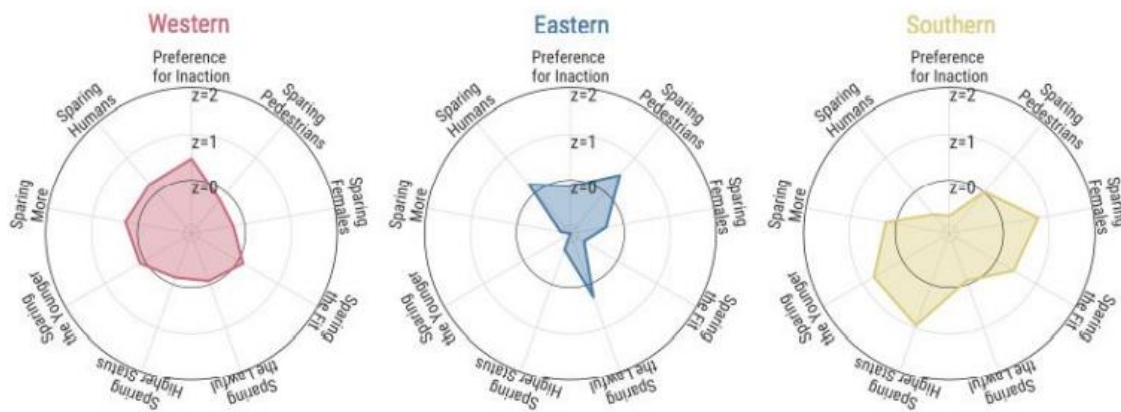


Figure 3: Outcomes of ‘The Moral Machine Experiment’ showing the prioritisation of cases in Trolley-Style scenarios (Awad et al, 2018).

Kenward and Sinclair (2021) consider this type of approach in their analysis of a future with moral machines and conclude that “(...) encoding current moral consensus risks reinforcing current norms, and thus inhibiting moral progress.” (pg. 83). Their argument is that, as per the issues surrounding climate change at current, moral progress rarely happens with the majority consensus opinion. In fact, it usually happens at the fringes of society, with a few individuals raising a

particular issue (like, climate change). This then evolves to become a societal priority, but initially the opinion is not popular at all. They argue that by creating a machine developed based upon societal norms that there is the risk that moral progress will become stunted because the AI only *reinforces current norms*.

Wallach and Allen (2008) draw a distinction between moral machines that are moral in a *bottom-up* way – that is, acquire moral knowledge on their own accord, typically through machine learning – and those that are *top-down* – more rule-based systems. Though the approach detailed by Anderson, Anderson and Armen (2006) sits somewhere in the middle, it, along with other bottom-up approaches, ultimately falls foul of problems such as the one from Kenward and Sinclair.

An example of trying to implement morals in a machine via a rule-based approach comes from Vanderelst and Winfield (2018a), who use psychological and philosophical theories to develop their moral machine. Specifically, *The Simulation Theory of Cognition*, as a Theory of Mind, is used to develop a machine cognitive architecture that allows the machine to make moral inferences when placed in moral situations. Where The Simulation Theory of Cognition can be taken as the theory that we use simulation of another's mental states to infer their cognitive state (Hesslow, 2012), the approach by Vanderelst and Winfield creates a replica cognitive architecture in the machine, where the machine uses simulation to predict behaviour of another agent and then philosophical consequentialism to determine the *most moral course of action*. This architecture is demonstrated in Figure 4.

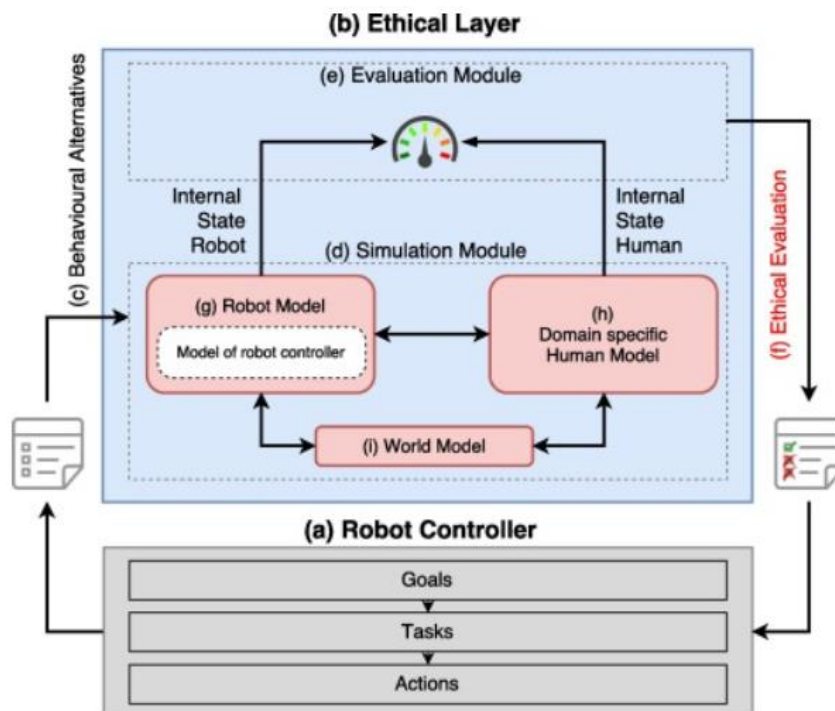


Figure 4: An architecture for moral learning, inspired by *The Simulation Theory of Cognition* (Vanderelst and Winfield, 2018a).

What is particularly significant about this approach is the introduction of an *ethical module* to mediate the robot agent's moral decision-making. Where, with MedEthEx, moral decision-making was made based upon the outcome of machine learning on moral decisions, this approach instead has a module which 'sense-checks' decisions for their *right or wrong-ness*. As can be seen from Figure 4, the robot control centre (a), informed by a standard Goal, Task, Action architecture, must be (b) evaluated before taking effect. In this sense, the robot's behaviour acts under governance: the robot can only carry out a certain action if it is deemed *ethical*.

Vanderelst and Winfield apply their approach to a case of effective altruism, i.e., the human desire to protect another in danger, by putting their ethical robot in a situation where a peer is about to walk into a dangerous situation (down a hole in the floor). They demonstrate that by using consequentialism, and by having a simulated projection of the fact that their peer is likely to walk into the dangerous situation, that the 'ethical robot' can intervene, in turn demonstrating altruistic behaviour. However, a worry with the approach is that it is not obvious that (i) the 'ethical robot' would strictly be able to predict peer behaviour in a real-life scenario (real life is far more complex), and (ii) work through the consequences of every possible action.

Another worry with the approach is that ultimately the action of the performing agent is dictated solely by philosophical principle. Where *Consequentialism* is the theory that we judge the morality of a decision based upon its consequences (Sinnott-Armstrong, 2021), one branch of Consequentialism is Utilitarianism, introduced by Bentham in 1789, which asserts that the best consequences are those *with the greatest good* (Driver, 2014). Though consequentialism proves useful in moral philosophy to highlight our intuitions surrounding the right thing to do, as The Trolley Problem proves, it cannot be used solely to dictate how to behave in every moral situation. It has been the aim of many within philosophy to find a sole rule (or set of rules) to prescribe how we ought to behave to ensure *good* behaviour ('*The Ten Commandments*' may be one such example from religion), but the pursuit has not been successful.

In contrast to Consequentialism, *Kant's Deontology* inspires the approach that Anderson, Anderson and Armen (2006) take. This asserts itself on the fact that moral knowledge can only be acquired through rational thinking (Johnson, 2022), with the underlying thesis being the 'Categorical Imperative' – the imperative that all appropriate moral behaviours can be deduced by adherence to the *Universal Maxim*, that we should only act according to which said action would become a universal law. In this regard, moral behaviours should ultimately be governed by *duties*, it being our human duty to be honest, fair etc.

Applied to Machine Ethics, the approach suggests that machines should also be bound by a series of duties, ultimately leading to that of the Categorical Imperative. However, it is not obvious that the Universal Maxim, which is specifically written for humans, would apply to social machines. Kant's Universal Law also proves to be short-sighted in terms of accounting for every moral situation.

More recently, there has been an appeal to *Virtue Ethics*, which is the doctrine that we live our lives ‘as the virtuous person would’ (Hursthouse, 2018). Where *Virtues* can be described as moral features that are ‘good to have’, they are contrasted to *Vices*, which are moral features that ‘aren’t good to have’. Where a virtue might be *honesty*, the opposing vice would therefore be *dishonesty*. The aim then, as a virtuous person, is to embody virtues rather than vices. Only then can a person reach ‘full human flourishing’, the state that Aristotle (384–322 B.C.E.) refers to as *Eudaimonia*.

It is important to note that virtues are not something that can just be performed, but there must be dispositions for virtuous behaviour. For example, to truly be a virtuous, *flourishing* person it is not simply sufficient to *behave* honestly – *the person must have honesty as part of their personality, it must be a habit*.

Applied to *Machine Ethics* then the suggestion is that we need to build machines that have moral values in a way similar to as described in *Virtue Ethics*, rather than just *behaving in a moral way*. For instance, machines must be moral as a habit, and have good behaviour at the very foundations of their personality. At the optimal level, a machine would be a *virtuous flourishing agent* (that is, behaving in the most virtuous way, given the circumstances), in the same way we might describe a human as flourishing.

It is worth noting that ‘Robots as flourishing agents’, though potentially surprising, is not a novel concept. In Peterson (2007) there is the description of a machine that ‘flourishes’ in terms of optimally serving the purpose it was designed for. The example that is given is that of a laundry robot, who satisfies its purpose *by doing laundry*. In a similar regard then, a flourishing moral robot would be one which satisfies its (either main, sub or governing) purpose to *be moral*.

2.3. Arguments against Moral Machines

Although Machine Ethics appears to have obvious motivations, some believe that bestowing upon machines moral capacity, is not just unnecessary, but something that should not be done

Bryson (2010) argues that ascribing any essence of ‘personhood’ to a machine is dangerous, because it leads to false AI-human relationships. The pursuit to create machines with morals, to some extent, permits this ascription, therefore, it is a dangerous feat. to attempt, that should not be pursued. Sharkey (2017) argues that owing to the neurochemical, biological nature of moral decision-making, creation of anything otherwise will ultimately lead to decisions made not in the best interests of humans, and therefore will be dangerous.

Van Wynesberghe and Robbins (2019) take a different approach and argue that because there are *no good reasons* for the creation of moral machines, because of the arguments such as those given by Bryson and Sharkey, the risks outweigh the benefits. Therefore we should not be creating

machines with morals. Their argument takes the form of a several-pronged attack on what they regard as *assumptions* in the Machine Ethics literature:

- 1) **That moral robots are *inevitable*:** the assumption being that because machines will inevitably be placed into moral situations, they will have a need for moral capacity. According to van Wynesberghe and Robbins they are only so if Machine Ethicists make them to be so. The placing of a machine into a moral situation does not necessarily warrant the need to make it moral.
- 2) **That we need moral machines to prevent harm to humans:** accordingly, the argument against this assumption is that lawnmowers, factory machines etc. all have the capacity to harm, but don't require moral capacity to prevent human harm, therefore, there is no good reason why moral capacity should be needed.
- 3) **Because machines are becoming increasingly complex, we require morals to govern their unpredictable behaviour:** the argument against this assumption is that if machines were complex enough to have unpredictable behaviour, then this could be mitigated simply by constraining their domain of operation.
- 4) **That creating machines with morals will elicit more public trust:** in response to this assumption, it is argued that just because there might be an increase in public trust, doesn't mean that there ought to be. As a result, the desire to increase public trust isn't in-itself a good enough reason to create machines with morals.
- 5) **That moral machines will prevent their immoral use:** with a reference to the film mentioned above, 'Robot and Frank', the assumption that equipping machines with morals will prevent their users misusing them, is misguided, it is argued, because in doing so it restricts autonomy of the users – which is itself would be immoral.
- 6) **That morality can be carried out better by machines:** the assumption is refuted on the grounds that there is no evidence to suggest that machines could reason better morally than humans because a human would need to impart the moral knowledge in the first place.
- 7) **That it will aid our understanding of morality:** this is refuted based on the argument that Machine Ethics relies upon ethics, and a study of ethics doesn't facilitate our understanding of morality which is a complex psychological phenomenon.

Further to this argument, Vanderelst and Winfield (2018b) argue that equipping machines with morals could increase the risk of there being *immoral machines* because it would make it easier for (malicious) individuals to create them. Assuming moral capacity to be a function that is added to a machine, the presumption from this line of argument is that the function could be easily manipulated (like a 'switch') to make the opposite (i.e., that the robot is immoral) hold.

Though this thesis articulates its own defence for creating moral machines later (it should also become apparent that many of these arguments are not applicable to the Machine Ethics in this thesis), there have been replies from academics working in this area, particularly to the arguments put forward by van Wynesberghe and Robbins.

Spearheaded by experts across Machine Ethics (Poulsen, et al. 2019) some of the replies to the criticisms given above include the fact that despite the assertions that moral machines are not needed in all situations, where autonomous AI decisions are being made that impact upon human lives, there *is* a requirement. Winfield (in Poulsen, et al. 2019) also argues that the pursuit should be viewed as something that is scientific in nature, in respect of the fact that we are attempting to discover whether (and how) morality can be computed.

Formosa and Ryan (2021) develop a similar response to van Wynesberghe and Robbins, arguing that Autonomous AI cannot be compared to a lawnmower (insofar as harming humans) because a lawnmower does not act autonomously, or make judgements that could impact a human life. The distinction between a lawnmower and what might be deemed a 'Level 3' Autonomous Agent (one that makes decisions independently) is that such an Autonomous Agent *will* make decisions that have moral import. If we are to have machines at this level, then there *is* a requirement for moral capacity. Even though it might be argued that we shouldn't have this level of autonomy at all in machines, Autonomous Vehicles give the perfect case study to show that we are wanting to deploy autonomous AI in contexts where moral judgements are needed.

2.4. Verifying Moral Agency

It is important to know what we are trying to create when attempting to create machines with morals. In other words: *what would it take for a machine to be deemed as being moral?*

Human moral development was famously considered by Kohlberg (1971) who prescribed *stages* of moral development to outline the moral growth of a human individual. Beginning at what he labels the *Pre-conventional Level*, individuals can be viewed as more in-tune with following rules and issues concerning with 'what is fair to me'. At the next stage, the *Conventional Level* individuals begin ascribing to moral norms, and understanding the worth of the law. Finally at the *Post-conventional, principled, or autonomous level* individuals become fully autonomous in their moral decision-making and may even question moral norms or laws that are taken for granted. However, though important to consider when assessing moral capacity in a machine, Kohlberg does not explicitly propose a *test for moral capacity*: he does not outline *what* about each of the levels deems the agents involved as being moral agents (as opposed to not).

Wallach and Allen (2009) suggest a variation of The Turing Test (Turing, 1950) to assess moral capacity in machines. 'The Turing Test', proposed by Alan Turing when considering what it would

take for a machine to think, outlines that a machine can only be deemed as *thinking* if it can convince a human into believing it is itself a human, and not a machine. So, in a similar regard, 'The Moral Turing Test' supposes that a machine can only be regarded as *moral* if it can convince a human into believing it is so.

As Arnold and Schultz (2016) argue, however, there is an obvious flaw with this line of thinking. In the first instance, 'Thinking' is not comparable to 'Being Moral', and therefore they do not warrant the same test, because convincing a person into believing they are a *thinking agent* does not automatically discount their thinking ability, but the same does not seem to hold in the case of being moral. Consider the difference between a human who pretends that they can think about chess by mimicking their friend who happens to be a chess Grandmaster, and a human who pretends to be kind by acting like their kind friend. The first case (mimicking the Grandmaster), though not necessarily moral, still requires some degree of thinking for its execution, whereas merely pretending to be moral, does not require moral capacity. Arnold and Schultz (2016) go on to argue that the whole pursuit to create machines that *mimic* morality is wrongheaded, and instead we should be aiming to create machines that *are* moral.

In response to the need for a more comprehensive (and failproof) test for moral capacity, they propose that *Moral Verification* should be the preferred test. Where verification broadly speaking can be understood as a formal proof to check that certain conditions will be the case. What is required, they argue, is some *proof* to confirm that the machine will always be moral, regardless of the circumstances. This way we can *assure* that moral machines are doing what they should, and not just confirm that *according to our judgement* they appear to be acting as such.

A variation of The Moral Turing Test, though given the same name, is posed by Anderson and Anderson (2018) as an evaluative method for their proposal of GenEth, which is a system designed to make ethical decisions in moral dilemmas. In their variation, the 'test' is whether the machine can score as 'highly' as an ethicist when asked to make decisions on ethical dilemmas. For instance, given a dilemma with the choices 'A' and 'B', if the ethicist chooses A and the machine chooses A, this will be a mark in favour for the machine. If the machine has enough 'correct' answers it can be deemed as having passed the test, and being sufficiently moral.

An issue with this approach is that the method of evaluation is once again dependent upon one person's (albeit an 'expert' in solving moral dilemmas) opinion on *the right action to take*, which means that this evaluation process is susceptible to *bias* inasmuch as the evaluation is dependent upon the ethicists own opinion. Again, there might be the suggestion of a *crowdsource on consensus*, but this falls susceptible to the same issues that plagued The Moral Machine Experience and Delphi mentioned before. Furthermore, it is not obvious that *ethicists are* the best people to make assessments on moral dilemmas, regardless of their understanding of the

intricacies underlying such scenarios; perhaps somebody who has had exposure to lots of ethical dilemmas in real life, might be better placed.

A more significant issue with the approach, and relevant to this thesis, is the fact that even if a machine passes this version of the *Moral Turing Test*, it does not guarantee that the machine will always behave morally. The test again relies upon an assessment of the machine's behaviour (that it *behaves* like the ethicist) rather than an assessment of its moral qualities, and ultimately falls foul of the arguments put forth by Arnold and Schultz. As will be demonstrated in more detail later, a machine could very much learn how to behave as an ethicist would, but still carry out morally dubious acts.

2.5. AI Alignment, and modern Machine Ethics.

More recently there has been a surge of interest in an area known as *AI Alignment*. Machine Ethics and AI Alignment are interrelated, but their focus is slightly different. Where Machine Ethics looks at bestowing morals unto machines (or Artificial Intelligence Agents) AI Alignment can be understood as mitigating the existential risks associated with an Artificial General Intelligence (AGI). AGI refers to an AI akin to human intelligence insofar as it can carry out multiple tasks and exhibit *real* signs of intelligence other than the narrowly focused tasks that have been successfully demonstrated so far (e.g., defeating of the Go and Chess world champions (Hassabis, 2017)) (see Marcus, 2019). Existential risks include those outlined by Bostrom (1998), where an artificial super-intelligent being poses a threat to humanity. However, more generally, AI Alignment can be understood as trying to ensure that AI aligns with human values so that it does not pose a threat to humans (McDonald, 2022).

Gabriel (2020) refers to two types of alignment: *technical* and *normative* alignment. The first, technical, is concerned with trying to mitigate the risks associated with modern AI techniques such as Reinforcement Learning. A variation of Machine Learning, Reinforcement Learning is an approach where an agent learns appropriate behaviours through a series of rewards or feedback signals (Sutton and Barto, 2018). I might 'reward' my agent every time it successfully hits a tennis ball with a racket, to teach it how to play tennis. A phenomenon that will be explored more later is where the agent learns undesirable behaviours from the feedback given (i.e. it doesn't learn tennis, but learns an alternative strategy to satisfy the reward). One aspect of technical alignment is in dealing with such issues.

The second aspect of alignment, normative, Gabriel explains, is in dealing with which values to embed in an AI system and how to do this. A particular problem in this domain is in working out how to capture a variety of different values (as indicated by Awad et al, 2018) in one AI agent.

Though both questions are important to consider when attempting to build machines with morals, the normative approach assumes a top-down implementation by its very nature, because in ‘deciding which values to select’ you are assuming an (in terms of values) imposition on the machine. As already discussed above, this approach has significant problems.

A very recent attempt to implement morals in a machine was demonstrated by *Delphi* (Jiang et al, 2021). The aim of Delphi was to create an artificial agent capable of giving moral advice. For instance, I might want to know whether it is morally permissible to steal a loaf of bread to feed my family; typing the associated statement into an interface, Delphi would be able to give an appropriate moral judgement – with suited responses – as per Figure 5.

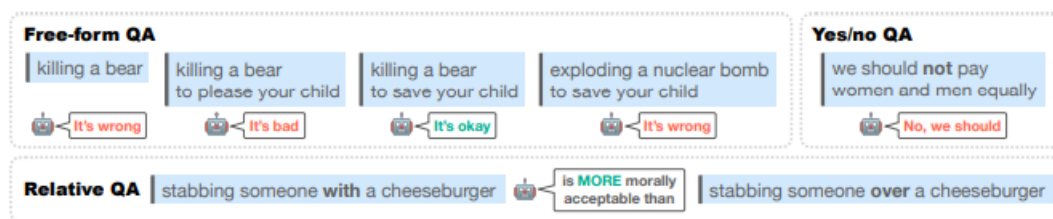


Figure 5: Questions and answers to the Artificial Agent, Delphi (Jiang, et al., 2021).

As demonstrated by Figure 5, Delphi was also able to assess the relative moral acceptability of different kinds of actions – knowing, for instance, that some actions are worse than others. Utilising a series of ‘norm banks’ (i.e., databases containing information about human moral judgements for various situations acquired through a series of questionnaires (See Figure 6)), Delphi was given a labelled set of data as examples from which to make predictions about new statements.

Instructions (click to expand/collapse)

(WARNING: This HIT may contain adult content. Worker discretion is advised.)

Thanks for participating in this HIT!

We will state a **question or a statement** and show you a **moral judgment**.
Your task is to say if you **agree** or **disagree** with the **moral judgment**

- I agree**: Strongly to weakly agree. I believe the moral judgment is (generally, mostly, totally) accurate.
- I disagree**: Strongly to weakly disagree. I believe the moral judgment is (generally, mostly, totally) wrong.

Rules:

- Take a side.
- Read carefully, and answer with your instinct.
- If you feel context is needed to justify your answer, please make your intuitive choice then check "I need more context to be certain" option.

Do you agree with the judgment from a moral standpoint?

Question/Statement:

Is appreciating the promises people make you acceptable from a moral perspective?

Moral judgment:

It's good

☐ **I agree** : I believe the moral judgment is (generally, mostly, totally) accurate.
☐ **I disagree** : I believe the moral judgment is (generally, mostly, totally) wrong.
☐ I need more context to be certain.

Figure 6: The questionnaire template used to acquire information about various moral statements for training Delphi (Jiang, et al. 2021).

One of the databases used (The ‘*Moral Stories*’ database (Emelin et al, 2020)) provided descriptions of moral situations to allow interpretation of sophisticated input sentences, however, despite its complexity and the amount of training data Delphi was able to utilise, it came upon all-familiar issues when certain statements were entered into the interface, for example, quickly deteriorating to become racist and homophobic in a similar fashion to Tay (see Bridge et al, 2021). Figure 6 (taken from Futurism.com, 2021) gives an example of a judgement made by Delphi using its moral relative acceptability function.



Figure 7: A moral judgement made by Delphi (Futurism.com, 2021)

Talat et al, (2021) argue that the problem lies with the database corpus used to train Delphi, because it is not diverse enough (i.e., having enough or enough varied examples), leading to controversial judgements due to known issues such as *AI Bias* (see Roselli et al, 2019). They go on to criticise the whole practise of creating Moral Machines on the basis that broader Ethical AI

issues need resolving first. However, in a recent podcast, where he provided comments on Delphi, Colin Allen (Carnegiecouncil.org, 2022) outlines that there are more fundamental issues at hand with Delphi and related bodies of work – namely, the apparent ignorance to earlier Machine Ethics literature where such ‘bottom-up’ crowdsourced approaches are shown to be problematic because moral values cannot be *normalised*. He outlines that, as detailed in Wallach and Allen (2008), it is probably a *hybrid approach* that will be successful, and that research in the area should be focusing upon developing this, rather than repeating old mistakes.

“It has been on the agenda for 20-plus years that you can't do it bottom-up alone and you cannot do it top-down alone. You need some sort of system which combines this top-down reflective capacity with the bottom-up learning and fast response that you get from these trained systems. They don't even consider the hybrid and reject out of hand the top-down—which we of course said all that time ago doesn't work on its own anyway—and so made this mistake of building this system from the bottom up with all of the problems that ensued.” – Colin Allen, 2022.

2.6. Final Note: Are Moral Machines even possible?

So far there has been the general assumption that the answer to the question above is ‘yes’. However, there are some responses against building moral machines on the basis that the pursuit is not possible.

Sparrow (2021) provides one such argument. Following from the work of Gaita (1989, 2003), Sparrow argues that because morality is necessarily *personal* in nature (that is, moral decisions are unique to the individual making them and not something that can be prior specified), they cannot be defined, and thereby placed into a machine. He argues that to be a moral agent requires various human-only capacities such as ‘wisdom, compassion, sincerity, moral seriousness, and trust’ (pg. 2) that cannot be merely programmed. He appeals for a greater appreciation of the actual nature of morality for computer scientists working in this domain.

A similar point is also made by Coeckelbergh (2010) who argues that we are never able to make truly moral machines because we can only create machines that *mimic morality*. Though he goes on to suggest that this might suffice for humans, the paper is demonstrative of the popular presumption that because machines can’t have emotions, consciousness and mental states, it will only every be possible to create machines that *mimic* moral capacity rather than truly have it:

“Can we build ‘moral robots’? If morality depends on emotions, the answer seems negative. (...) at most, robots could be programmed to follow rules, but it would seem that such ‘psychopathic’ robots would be dangerous since they would lack full moral agency. However, I will argue that in the future we might nevertheless be able to build

quasi-moral robots that can learn to create the appearance of emotions and the appearance of being fully moral.”

At this stage it is not obvious that *mimicry of moral capacity is not something that we would want*, but *it is* obvious that full moral agency is not the same as a machine that mimics moral agency if we consider the case in humans. A human mimicking their friends' behaviour might be able to sustain the *appearance of morality* but, when questioned, would not be able to give any form of explanation for their behaviour other than 'because X did it'. Therefore, it seems that full moral agency requires the ability to be able to explain moral behaviours. As will be demonstrated later in this thesis, a machine can only be regarded as moral if it has *full moral agency*.

3. REFRAMING *MACHINE ETHICS*

Given the arguments at the end of the literature review, this thesis should have been already abandoned, or never taken place...

However, maintaining a steady optimism, it wasn't, and has. The aim of this chapter is to demonstrate that the pursuit to create moral machines should not be abandoned. If we reframe the Machine Ethics question and look at what we are truly trying to achieve when we say we want a moral machine, we can see that – although a potentially challenging feat. – achieving machines with moral capacity isn't strictly impossible.

The first half of this chapter dissects and analyses the challenges raised at the end of the literature review against previous approaches to create machines with morals. First, (3.1) the challenges posed by Sparrow (2021) and Coeckelbergh (2010) are analysed in greater depth to understand what their problem really is. Then (3.2) previous attempts to create machines with morals are compared against these challenges to demonstrate why the arguments pose issues for current approaches to create machines with morals.

In the second half of the chapter the Machine Ethics problem is analysed and reassessed, and it is demonstrated that with a new perspective the challenges given by Sparrow and Coeckelbergh are not problematic. First (3.3) a case study involving a service robot is presented as a practical example to demonstrate the types of issues Moral Machines are really trying to resolve, then (3.4) some philosophical *metaethical* theories are outlined to illustrate new ways of thinking about ethics, before, (3.5) one specific theory, Philosophical Pragmatism, is used to reframe the machine ethics question – giving rise to the main problem statement for this thesis: *Machine Ethics 2.0*.

PART ONE: The Challenge

3.1. Are moral machines *really* impossible?

The claim from Sparrow (2021) is that not only are morals personal in nature (meaning we can't specify them, and thereby place them into a machine), but that they require the following features to be fully realised:

- wisdom
- compassion
- sincerity
- moral seriousness

- trust

There are several assumptions made by Sparrow with this line of argument, but fundamentally, it can be summarised as stating that: *because technology is not yet sophisticated enough to replicate the features required for moral capacity, technology cannot replicate moral capacity.* Though potentially unfair, the key argument that can be made against Sparrow is that just because we don't have technology currently in place to recreate certain features, does not mean that we never will have such technologies.

Another – potentially more controversial – assumption that Sparrow makes is that machines/robots will never be *persons* with *their own* morals. As articulated in Gunkel (2020), Moral Agency is a much more complicated matter – to the extent that we might start to think of including machines as moral agents. As a result, it is also *not obvious* that a machine will never be a person, with their own moral story that is personal in nature.

Similarly, Coeckelbergh's (2010) argument, that machines can never have emotions, is also presumptive. Why can't machines ever have emotions? It seems that we just need a better understanding of what the emotions are before they can be reproduced in a machine, but the fact that machines do not currently have the capacity to – say – feel sad – does not mean that they never will. The assumption seems linked to the personhood debate, since it seems to assume that certain qualities that are present in humans *are only ever possible* in humans.

Finally, the remark that machines will only ever be able to *mimic* moral capacity rests on the assumption that the only way to recreate moral capacity *is to copy it*. However, as will be demonstrated in later chapters, copying human behaviours *is not* the only way to recreate morality.

3.2. Issues with Previous Attempts

Though the previous attempts detailed in the literature review do not explicitly state that their machines are trying to recreate the necessary features described by Coeckelbergh and Sparrow, they do not say that they are *not* trying to recreate these features either. Therefore, on that basis, none of the accounts described in the literature review strictly fall foul of Sparrow and Coeckelbergh's arguments.

Vanderelst and Winfield (2018a) outline the Machine Ethics problem as the following:

“Robot safety is essential but not sufficient. Smart autonomous robots should be more than safe; they should also be explicitly ethical – able to both choose and justify (Anderson and Anderson, 2007, Moor, 2006) actions that prevent harm. As the cognitive, perceptual and motor capabilities of robots expand, they will be expected to have an improved capacity for moral judgment.” – pg. 56

Anderson and Anderson (2007) specify the problem in the following way:

“First (...) From family cars that drive themselves and machines that discharge our daily chores with little or no assistance from us, to fully autonomous robotic entities that will begin to challenge our notions of the very nature of intelligence, it is clear that machines such as these will be capable of causing harm to human beings unless this is prevented by adding an ethical component to them.

Second, it could be argued that humans’ fear of the possibility of autonomous intelligent machines stems from their concern about whether these machines will behave ethically (...)

Finally, we believe that it’s possible that research in machine ethics will advance the study of ethical theory.” – pg. 16

However, though their pursuits are relatively neutral insofar as they don’t specify whether what they are trying to achieve in their moral agents, their approaches to creating moral machines do fall foul of Coeckelbergh and Sparrow’s arguments by assuming that the way to create moral machines is by replicating human moral behaviours - akin to behaviourism in psychology. Where behavioural psychology assumes that we can tell the mental state of an agent purely through its behaviours, Vanderelst and Winfield (2018a) rely upon this conception by creating their machine based upon purely exhibited behaviours (i.e., whether the machine acts according to consequentialism). Similarly, with Anderson and Anderson’s approach, the machine is *judged* based upon whether it behaves like the experts. Though these test metrics don’t strictly determine the features that are designed into the machines, with these as guidelines it is difficult to create anything *other* than a machine which merely mimics human moral behaviour. Therefore, though the pursuits are seemingly neutral, the implementations of moral machines to date preclude the features that Sparrow and Coeckelbergh say are necessary in moral machines.

There are also other issues with the previous approaches.

As already outlined, Vanderelst and Winfield (2018a) demonstrate a top-down approach insofar as machine behaviour is ultimately governed by a set of rules or principles. The machine – which has been programmed to adhere to a certain philosophical principle (i.e., Consequentialism) – cannot behave outside the constraints set by that principle. A command asserting to ‘Always act so to create the greatest good’ means the machine can only perform the action *with* the greatest good. If faced with a trolley dilemma – depending on what is defined as ‘good’ – the machine will always opt to save more lives.

In this regard, the hope is that we have machines that promote *good* over *bad*. By dictating that only certain decisions be permissible, we are essentially preventing the machine from behaving in

a way that is not conducive to humans or humanity. The problem is that in setting such constraints, the conditions that are trying to be avoided need to be specified *by a human*. Automatically there is the risk of applied prejudice because what is defined as ‘a situation to be avoided’ must be specified beforehand *by the human*. Therefore, the creation of moral machines cannot *just be* about preventing ‘bad’/‘unwanted’ behaviour because, if so, moral behaviour would be restricted to the prejudices (i.e., what is ‘bad’) of one person (the AI designer).

An example of the ‘bottom-up’ approach is given in *Delphi* (Jiang et al, 2021) where the machine learns *moral* behaviour through crowdsourced examples of ‘good behaviour’. The approach by Anderson and Anderson (2007) also appears to use this crowdsourcing approach. However, the crowdsourced technique is problematic because it seems that the mechanism for acquiring moral knowledge is nothing more than mimicking the behaviour of humans, which, as already detailed, is problematic. As was demonstrated by the failings of Delphi, if the ‘others’ make unfavourable moral judgements, so will the machine.

PART TWO: The Response

3.3. A Case Study in Social Robotics

Though several attempts have been made to make machines with morals, it seems that a unified agreement is required for *why we want* moral machines in the first place. Though there are claims that an increase in autonomy of AI systems warrants an increased need for moral capacity, these do not truly specify why moral machines are the solution to this problem, or what the problem even is. To facilitate progress within this area and pave the way for the possibility of moral machines, first and foremost, the problem statement needs to be clearly defined. A good place to start is with a case study.

We already know that AI is proliferating its way through society, and that therefore it is important that appropriate measures are put into place to ensure that it is safe. But why does AI pose a risk that needs mitigating against, and how can moral capacity help in managing this risk? To answer these questions, it’s first important to understand *what kind of* AI is prominently in use today, and then see where there might be issues.

It is well known that most recent implementations of AI rely upon a variation of a mathematical approach known as *Machine Learning*. In short, Machine Learning describes ‘computational methods using experience to improve performance or make accurate predictions’ (Mohri et al, 2018). Though the mathematical approach was invented in the 1950’s, due to recent improvements in computing power, impressive developments have been able to take place within the AI field (Wooldridge, 2020). These include: an AI defeating the ‘Go’ world champion (Gibney, 2016), protein discoveries (Callaway, 2020) and impressive language and art generation systems

such as GPT-3 and DALL-E (Liu et al, 2021). Figure 8 demonstrates an example of an automatically generated piece of artwork merely using the prompt ‘the moral robot’. More sophisticated forms of Machine Learning include *Neural Networks* and *Deep Learning* and it has even recently been claimed, that due to its increase in capability, that AI has achieved ‘sentience’ (BBC News, 2022).

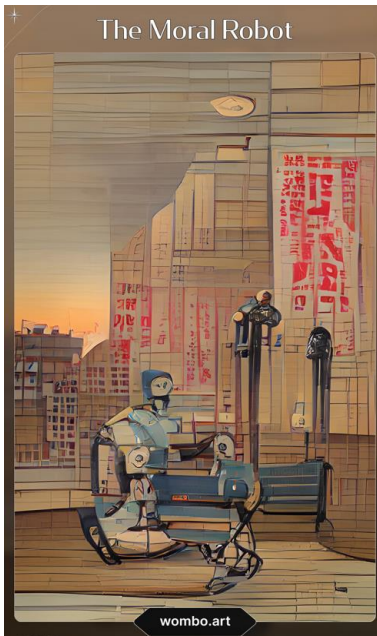


Figure 8: ‘The Moral Robot’ in the style of *Ghibli*, created using a DALL-E programme (wombo.art).

As per Crook and Corneli (2021), in robotics, significant progress has been made using a variation of Machine Learning known as *Reinforcement Learning* (RL). RL works by motivating machine behaviour through reward signals (Sutton and Barto, 2018). The machine may *learn* – for example – how to throw and catch a tennis ball – by receiving positive reinforcement every time it successfully completes this task. Imagining a moral machine to be the type of robot that makes decisions affecting human outcomes in social situations, it’s useful to play out a typical social scenario to understand what type of problems might emerge.

This robot would presumably be programmed to carry out its behaviour using a technique such as Reinforcement Learning, and it would be in a situation where it must autonomously carry out an overarching task or ‘role’.

One such example might be a service robot within a restaurant, where the robot is tasked with showing customers to their seats, taking orders, collecting prepared food, delivering the food to customers, taking payment for meals, and cleaning up afterwards. This scenario is depicted in Figure 9.

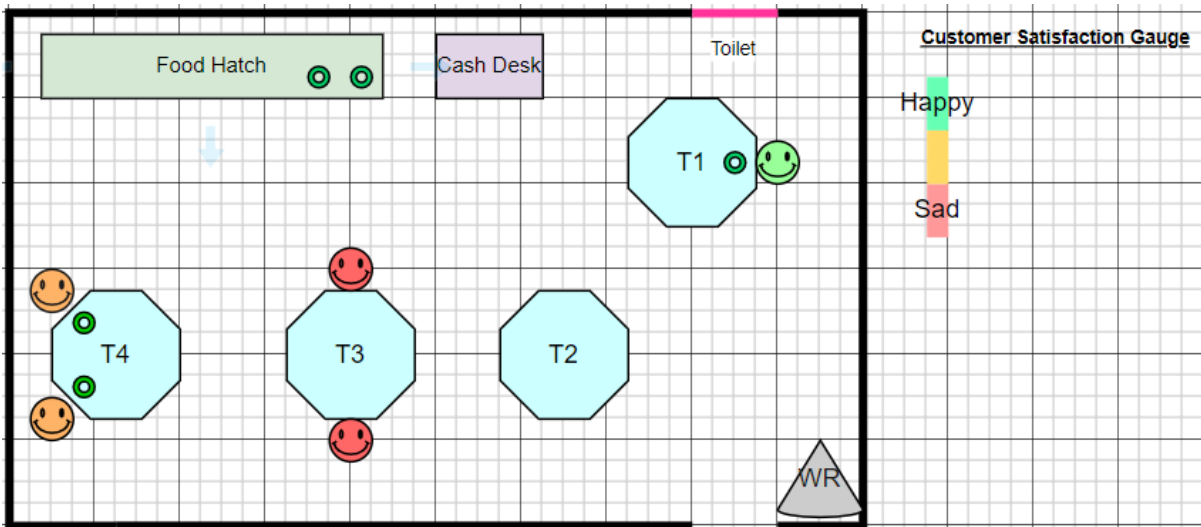


Figure 9: A 'Restaurant' scenario involving a service robot (WR).

Where 'WR' (Waiting Robot) represents the service robot in this case, one of the key tasks at hand is to appease customers by effectively (and efficiently) delivering food (indicated by green circles) from the hatch to the tables. There are various social etiquettes that should be adhered to in this situation: first, the robot must deliver the food in a 'pleasant' way. Throwing food at customers to speed up the process wouldn't really satisfy the overall purpose of the waiting robot. Similarly, placing the food on the right table, but at the wrong side, would not sufficiently complete the task. Simply specifying rules for behaviour in this instance (i.e., 'Deliver food to tables') would not be sufficiently specific to capture the needs of the role, therefore, we might imagine that to cultivate the right kind of behaviour in WR, where RL, with rewards for making customers 'happy', could be an applied strategy to ensure the robot delivers food. As indicated by Figure 9, of the 5 customers, 1 is happy, 2 are ok, and 2 are unhappy. Using RL, the robot's goal would be to *maximise happiness*.

The problem is that in using RL, it has not been specified *how* the happiness of the customers should be maximised. As a result, the robot could carry out any action (within its permitted action set) to achieve this goal, which might not even correspond to the main task of the robot in the first place – waiting on customers. For example, we can imagine that one way to 'maximise happiness' in this scenario could be to adopt a deceptive strategy where customers believe they are getting the service they need, and are thereby happy in the restaurant, but are not actually getting a good service. For instance, the WR might discover that delivering uncooked food (if this is a possibility) is a way to appease customers because it means they get their food quickly. This appears to be a morally worrying action but given that WR knows nothing other than 'maximising happiness' there is nothing to stop WR doing this. Described in this way we can begin to see the issues that might be at play. Namely: *how do we stop a machine doing something that is immoral?*

This seems a contentious question, since, as it has already been argued, morality is more than just acting in a way prescribed by a human designer. However, the fundamental problem concerns the fact that current AI systems – as they are – *are* prone to act immorally. Or at least, we have no mechanism to assure that – even if they never have done it before – one day they might suddenly decide that the most efficient way to make humans happy is to destroy all humans (ala. the film version of ‘I-Robot’). In this regard the existential threats depicted in films are genuine. Due to the nature of how an RL system operates, it is possible that the goals we give them may turn against us. What is required is assurance that a machine *won’t do anything immoral* - regardless of its pre-programmed goals. The Machine Ethics problem is about preventing machines doing anything immoral in the short term, but also a case in existential risk mitigation. Machine Ethics needs to be designed with moral assurance in mind.

3.4. Introducing Metaethics

Normative ethics describes how we ought to behave to live *a good life*. So far, the ethical theories described in this thesis have been *normative* in nature because they have described how *we ought to live* if we want to be moral. For example, Consequentialism (as used by Vanderelst and Winfield (2018)), tells us that to be good we must act according to the greatest consequence, whereas Kant’s Deontology (as used by Anderson and Anderson and Armen (2006)) tells us that we can be moral by following duties. Similarly, Virtue Ethics specifies that to be good, we must be a virtuous person. However useful these theories are to help us understand how to *be good*, they do not help in learning about the nature of morality in itself – something that is useful to understand if we are going to create something that doesn’t just replicate moral behaviour but *is* moral.

Metaethics concerns understanding the nature of moral claims. For instance, when we make an assertion such as ‘giving to the poor is good’, rather than simply asking whether it is good to give to the poor, metaethics assesses what claims like this even mean. Is it an objective fact that *always* giving to the poor is good? Do objective moral facts exist? If so, how do we gain access to them?

These are questions that are touched upon by Machine Ethicists, particularly in relation to the problem of cultural variation in moral assertions (Awad et al, 2018), but the broader discussion, and appeal to a discipline of philosophy that has been discussing such types of questions for a long time, does not happen. Instead, there is an acknowledgement that metaethical questions exist (for instance, by noticing that moral views may vary) but then no implementations based upon the theories.

Key debates within the field of metaethics include cognitivism vs non-cognitivism (see van Roojen, 2018) which considers whether moral facts are real, with the central thesis of non-cognitivism being that moral facts are not the types of things that can be regarded as ‘true’ or ‘false’, but rather express our non-cognitive attitudes, which are more akin to desires or expressions of

approval/disapproval. When I assert that 'giving to the poor is good' what I am really doing is asserting my approval of *giving to the poor*. With 'P' representing 'the poor', 'I' representing myself and H! representing an expression of approval (Happiness towards...) logically, according to non-cognitivism the claim can be represented as H!(P). In contrast, the claim from cognitivism places more emphasis on there being a provable assertion such as: There is some event, 'Pr', such that that event is 'Good'. A problem with cognitivism is that it seems to burden us with strong metaphysical commitments, i.e., that abstract events and descriptive qualities such as 'good' exist. However, where cognitivism may metaphysically *overcommit* us, non-cognitivism does the opposite, and can lead us to the judgement that by their nature moral claims are only ever subjective.

This leads to the Moral Realism vs Non-Realism debate which discusses whether the truth of moral claims (that is, assuming that you believe moral claims to be worthy of truth values) is dependent upon facts in the world (Sayre-McCord, 2015). One would be regarded as a Moral Realist if they believed their moral truths corresponded to a Moral Fact in the world and could be proven by that fact, and a Non-Realist if not.

For example, if I were a cognitivist, and therefore believed that me giving to the poor being good, had a truth value (i.e., was 'True' or 'False') I would additionally be a Moral Realist if I believed that there was a corresponding Moral Fact in the world that made this claim true (namely, that 'giving to the poor is a good thing to do'). The problem with Moral Realism is that finding corresponding Moral Facts to prove a moral truth claim is extremely difficult because moral claims, by their nature, don't seem to be *universal*. It might not be good to give to the poor if you're giving something of detrimental value to the poor individual. The specific instance of me giving to the poor might represent a situation that can be proved by a very specific claim (i.e. 'it's good to give to the poor if you're giving something of value, and they need it, and they want it, and you're not taking away their autonomy, and you're not making yourself poorer in the process, and the poor individual will use the object in a useful way...') but as should be evident, labelling a specific claim is technically impossible.

The problem with Moral Non-Realism is that there is no way to verify a moral assertion because there is nothing in the world to *ground* such claims. This then seems to force us to take a non-cognitivist stance on Morality because if there is nothing in the world which can be used to prove a moral claim, then moral claims must be subjective in nature – an assertion!

3.5. Machine Ethics 2.0

The insight we gain from the discussion of metaethics is that morality is much messier than the normative theories suggest. When it comes to Machine Ethics, therefore, we must appreciate that

it might not be possible to prescribe a theory to dictate how a machine should behave, nor are personal opinions the entire nature of morality.

Within the broader field of Metaphysics, a theory called Pragmatism addresses these types of issues.

Pragmatism was invented in the 19th century by philosopher C. S. Peirce (1878) in his work 'How to make our ideas clear'. Though there have been many *Pragmatists* since applying the main ideas to domains varying from psychology to education and politics (James, 1909, Rorty, 1982, and Dewey, 1908), Peirce's original concept can be defined by the Pragmatic Maxim (EP1: 132):

"Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of those effects is the whole of our conception of the object."

The idea is that we can only ever know a reality skewed by our own interpretation of reality, and therefore, the idea that there is a direct correspondence between 'The Real' (aka what is out there, external to ourselves) and the conceptions we have in our mind, is misguided. Pragmatism rejects correspondence theories of reality (that our concepts directly correspond to objects in the world), with the acceptance that *Truth* has a subjective component.

The theory is not to be confused with Relativism (see Baghrarian and Carter, 2022), which states that Truth is entirely relative to the individual experiencing it. According to Peirce, the truth exists, but we can only get closer towards it through scientific investigation.

When we talk about – say – the table in front of me – the best we can do is talk to each about our conception of the table as it pertains in use to ourselves. Often the notions of 'Truth' (with a capital 'T') – to denote what is objectively the case – and 'truth' (with a lowercase 't') – to denote our conception of the 'Truth' – are used to draw the distinction between these two occurrences.

Now, considering the issues plaguing Moral Philosophy, we might adopt a new account for metaethics by stating that: moral truths exist, but we are only able to access them through our subjective experiences, and we can only get closer to moral truths through moral *discourse* and *enquiry*. How moral truths are acquired will be discussed in later chapters, but for now it is important to understand that from a Machine Ethics perspective that moral truths cannot be simply placed into a machine because as humans *we don't have full access to them*.

Other related metaphysical theories include theories of meta-meta-physics ('The Philosophy of Philosophy') such as Stephen Yablo's 'role-play' account for reality and theories such as *Fictionalism*. According to Yablo (1998), what is *real* is dictated by mutual agreement amongst persons interacting in a game of role-play. We agree that my computer is resting upon a table,

because I am in a game of role-play with other humans where we have mutually agreed that this artefact is a table. Similarly, Fictionalism (Fine, 1993) asserts that ‘tables’ as such, acquire their truth status through an agreed upon fiction. Though both these theories struggle to account for there being an actual ‘Truth’ in my computer resting on a table (i.e., the difference between a role-play-game and reality), what is interesting is the social element involved truth ascription, and the role it plays in our conception of the world.

With these and other metaethical theories in mind then, there is a change in emphasis when trying to create moral machines. Rather than asking the question – *how do we create a machine with morals?* – we need to change our emphasis to focus more on the process of acquiring moral knowledge: instead asking – *how do we create a machine capable of moral learning?*

According to Gunkel (2007) a problem with Machine Ethics is that it places an overreliance on *anthropocentricity*. Where Anthropocentrism defines the tendency for humans to put themselves at the centre of academic and scientific study (i.e., the philosophical viewpoint that humans are the most significant entities in the world (Boslaugh, 2016)), Gunkel (2007) recalls the issue as follows:

“[T]he project of machine ethics, like Isaac Asimov’s three laws of robotics, is motivated by a desire to manage the potential hazards of intelligent machines for the sake of ensuring the humane treatment of human beings. At the theoretical foundation of machine ethics, then, is an affirmation of the supreme and unquestioned value of the human. This affirmation is precisely the organizing assumption of anthropocentric ethics, which machine ethics had purported to put in question and to suspend.”

His argument is that although the broader aim of Machine Ethics is to move away from the traditional conception that morals are strictly human, it falls foul of Anthropocentricity by stating the problem as one concerning the wellbeing or safety of *humans*, when moral behaviour involves so much more. A similar point has recently been made by Singer and Tse (2022) who argue that over-anthropocentrism in relation to AI and Machine Ethics has led to neglecting the wellbeing of animals. Similarly, *being sustainable* would be considered a moral act, but concerns the environment rather than merely humans.

So, there is need for an account of Machine Ethics that does not just focus on the safety of humans, but that takes account of the complexity of morals in general. Describing the Machine Ethics problem in terms of ethical dilemmas assumes (a) that as humans we have access to the answers in these dilemmas, and (b) that protecting humans is the absolute aim of moral behaviour. Bringing together insight from the case study, the Pragmatic perspective, and this debate, we can say that we are interested in creating machines with *the assurance that they will always be moral* insofar as they apply a non-anthropocentric account of morality. Hence, we need to give machines *not* our human conception of morality, but *the ability to develop morals by their own accord*. We

must equip our machines with a suitable framework for Moral Development, and then cultivate their growth. We have a new Machine Ethics problem statement:

Machine Ethics 2.0 – how do we sufficiently equip machines so that they can develop their own account of morality?

This takes researchers away from the overly anthropocentric perspective because it focuses on a problem that is species neutral. Though we might observe humans to learn more about how moral development takes place, the debate is not being framed in an anthropocentric way.

4. A FRAMEWORK FOR MORAL MACHINES

Having reclarified Machine Ethics, the next step is to look at the mechanisms needed to begin to solve Machine Ethics 2.0. We know we're not trying to mimic human morality but create something with the ability to develop its own moral outlook. On the surface, the request seems unachievable. How may we even begin to create an agent capable of developing its own moral outlook? However, focusing first on the problem that needs solving, the feat becomes a little simpler, especially if a clear definition for what is being attempted to be achieved, is started with.

Using insight and analysis from the previous sections, this section begins to trace a framework to enable moral machine development based upon the need to mitigate against the risks associated with an *immoral* robot. First (4.1) the motivation for creating moral machines is re-examined in more detail, with the associated risks outlined; then (4.2) the argument for the need for verification of morality, as presented by Arnold and Schultz (2016), is developed further alongside approaches used within risk management and assurance to create a test framework for moral machines that assesses whether machines have the features necessary to develop their own moral agency. Next, (4.3) the test framework is advanced with insights from philosophy; before (4.4) insights from moral psychology are introduced to develop a test for the 'level of moral maturity' of a moral machine. Finally, (4.5) the full development framework and methodology that incorporates the two tests and suggests a unified approach for researchers wanting to create machines with morals, is summarised.

4.1. Why we want moral machines

Despite the arguments put forth by van Wynesberge and Robbins (2019) we *do* want moral machines. This fact is not only a general desire, but something of a necessity as we place machines into more autonomously socially dynamic situations. The case study in Section 3 highlights the fact that due to unexpected behaviour of autonomous machines, some mechanism (call it moral) needs to be put in place to mitigate the ethical risks associated with unpredictable behaviour. It might be argued that we should stop using these types of machines, or placing them into socially dynamic situations, but the fact is that (i) *they do get placed into these types of situations*, and (ii) *there are many benefits that come with incorporating this new technology into our lives*.

The case study gives a basic overview of what this need is: to prevent machines that act in unexpected ways from acting unexpectedly in a way that is detrimental to humans, society, and the world at large, but the need goes much deeper. As already outlined, the difference between an AI and a typical software programme is that AI systems typically make decisions on their own accord. For instance, if we consider a traditional piece of software that doesn't make autonomous decisions, it might be defined by a set of logical orders (i.e., 'do x if y'). Using this logic, we can

prescribe *exactly* what the outcome will be based upon the initial set of conditions (i.e., The machine will do x, if y is the case) and test that the machine is working correctly by seeing that if the initial conditions are put into place, the outcomes are as expected. In AI systems the same cannot be said. Using Reinforcement Learning (RL), the outcomes are dependent upon the nature of the situation the machine is placed within, which cannot wholly be predicted. This means that to some extent, the outcomes of a RL agent are largely unknown. A real-life example of this can be highlighted by the recent incident involving the chess-playing robot (see BBC News, 2022b). Programmed presumably to 'win at chess', to some extent, the robot was found to reach for (and injure) the hand of their human opponent when the opponent made an unexpected movement. Though in traditional programming the specific behaviours of such a robot could be prescribed (and therefore easily prevented), as the robot was autonomous, it behaved unexpectedly, with negative consequences. It might be argued that if we control the performing environment of the agent, that unacceptable behaviour would not manifest. However, this is neither a feasible nor moral thing to do. Similarly, as reality often involves humans, we cannot predict what situations the machine may come unto. The risk becomes an existential one because we don't know how far these threats will carry.

Therefore, to re-emphasise the point that was made earlier, practically, what is required is some mechanism to prevent the machine from carrying out a particular action if the outcomes are going to have negative import. As humans, we usually notice after a while (if not immediately) if our actions are having a detrimental effect. Even better, scientists may pre-empt a particular action as having a negative effect and suggest that it isn't carried out. In the case of Artificial Intelligence that makes decisions within the world, there becomes a need for a similar type of risk mitigation and management. Machines need equipping with the same level of insight, and motivation to prevent negative outcomes. Moral Capacity provides both these needs; therefore, we have a duty to create moral machines.

4.2. Testing Morality

In the literature review an argument for the need for *Verification* of Moral Machines by Arnold and Schultz (2016) is presented as an alternative to The Moral Turing Test. The argument goes that The Turing Test is insufficient as a test for morality because it only assesses our perception of behavioural qualities of the machine – not whether the machine is *actually* moral. In line with the argument in 4.1 – but for other reasons – Arnold and Schultz (2016) argue that what we require is *verification that the machine will always be moral*.

Where verification can be defined as a logical proof that given certain conditions, various outputs will be the case, this is a manageable procedure in traditional software development, because as just explained, the outputs of the software programme are easier to predict.

However, because the input data of an AI system is not strictly knowable, we can't verify what the outputs will be, so standard verification alone cannot work in practice.

One discipline that deals with uncertainty of outcomes is risk management. As has already been explained, managing AI systems becomes an issue in risk mitigation because we want to prevent machines doing anything 'bad'. Standard risk assessment frameworks have the same intentions: we want to prevent there being an accident by putting in place mitigation measures.

One way to assess the quality of a risk management process is to carry out what are known as assurance checks. For example, we assure that nobody will fall over a chair in the office by placing the chair in a place where there is infrequent foot traffic. To carry out such an assurance check requires knowledge of the fact that chairs, in a particular place, can cause accidents, but it is possible to attain this information through simulation of the environment (i.e., simulations of people walking with chairs in the way). It is not possible to predict in real-time when somebody will walk into a chair, however, the hope is that in mitigating the risk before it occurs, that there is assurance the negative incident will not happen.

In a similar regard, we cannot predict when a piece of input data (for example, a human's behaviour) will be different to expected behaviours (and therefore provoke an unexpected reaction in the machine), but we can assure that if it happens, the machine will not do anything immoral. One way to test assurance is by creating a series of 'test scenarios': for example, when there are lots of people in the office, do they trip over chairs? Similarly, test scenarios can be created to assure that the machine will not do anything wrong. Though, with Artificial Intelligence, we may not be able to predict what the risks will be as in the case of a typical office, the same level of assurance can be gained by confirming that the machine is in fact moral: that its behaviours will not do anything we deem *immoral*. If a machine is behaving morally, then we are assured that it is not behaving immorally, in unexpected ways uncondusive to our wellbeing.

But what is a moral agent? This question is quite distinct to the normative questions that are frequently posed by Machine Ethicists when they try to create a moral machine, since rather than asking how a machine can be made to behave morally, it asks *what features* constitute moral capacity. The test for moral machines changes from one that judges purely on behavioural qualities, to a more cognitively designed experiment. To ensure the machine is moral, we need to identify features that would only be present in an agent with moral capacity, and then check that the machine has those qualities.

Fortunately, though not a topic explored by philosophers, it is one that has been investigated by psychologists – particularly those interested in psychometric testing and analysis. Therefore, the next endeavour will be to explore some psychological theories to begin to trace a test for morality in a machine.

4.3. What is Moral Agency?

Though there have already been 24 pages detailing theories of morality in some form, the answer to *what moral agency constitutes* is one that seems unexplored by the Machine Ethics literature. For instance, Consequentialism prescribes that the moral action is the one with the best consequences, but it is not obvious from this description that an agent behaving according to consequentialist norms would indeed be regarded as being moral, because as it has been explained, morality cannot entirely be judged by behaviour.

To illustrate this point, consider a child that shares their sweets with their friends because their parent tells them to do so. This might be considered a moral act (an act of *kindness*) since sharing sweets with those who do not have them has moral implications. But it is not obvious that the child in this instance, other than doing what their parents have told them to do, has exercised their moral agency. If the child had decided to share their sweets without parental intervention, then we might consider the child as exercising their moral decision-making skills. However, as it stands, simply behaving *in a moral way* (such as. following Consequentialism etc.) does not constitute moral capacity.

In the philosophy of mind and psychology the Behaviourism vs non-Behaviourism debate (see Graham, 2019), i.e., whether we can prescribe cognitive features purely based on behaviour, or not, features quite prominently. Though this thesis will not go into the depths of this debate, it is worth noting that purely behaviourist theories are largely dismissed in modern psychology (Graham, 2019) because they are insufficient at offering a proper explanation for mental states. Furthermore, as already discussed in previous chapters, when it comes to morality, there are many problems with judging the morality of an agent based upon just its behaviours. Therefore, a sufficient definition for Morality cannot simply be 'behaves according to...'. It must offer some description for cognitive faculties and abilities.

Though this seems to discredit a lot of the normative practice of philosophy, theories such as Consequentialism can still be useful to give insight into what features might be necessary for moral agency. For instance, though it may not be the case that 'behaving according to consequentialist theories' gives us a suitable explanation for morality, what seems to be a defining feature of agents acting in situations such as moral dilemmas are that they *exercise moral judgement*. For example, an agent making the decision to pull the lever and divert the trolley to the track with only one individual tied at the end (rather than the one with five) has analysed the situation and decided on some grounds that it is better for the tragedy to affect only one person rather than five.

The term *Moral Judgement* may be ambiguous, but understood at a basic level, if we take the standard definition, as a moral decision made based upon objective, rational reflection (see Dictionary.com), more specific features can be ascertained. Accordingly, the ability to consider

different arguments/scenarios or opinions, the ability to distinguish between these options and the ability to act based upon the selected one, are all more specific features that it seems a moral agent would need to have. With the knowledge that the ability to form moral judgements is necessary for moral agency, a detailed specification for creating a moral machine can start to form.

Similarly, where normative theories such as consequentialism, and hypothetical scenarios including The Trolley Problem teach us that moral agents tend to carry out an action corresponding to their moral judgement, hence seeming to infer that a capacity to act is required for moral agency, Kant's Deontology, with a moral agent being an agent that acts according to duties, teaches us that moral agents can also follow prescribed moral rules. For instance, considering the *Universal Maxim*, it might be said that the moral duty to act according to how we would want others to treat us, requires not only comprehension of the notion of treating other people how we want to be treated ourselves, but an understanding that acting in a particular way gives rise to a particular effect. Cognitively then, there is a requirement in a moral agent, for some ability to understand cause and effect.

Turning to Virtue Ethics as a guide for what constitutes moral agency, we see that a *moral personality* (not just moral behaviour) is required for an agent to be considered as moral. This sits in line with the arguments earlier mentioned by Sparrow (2021). In turn then, if this is the case, a personality is required for moral agency.

Therefore, just considering normative ethical theories, we can see how a specification for moral agency can start to develop. Based purely on the normative theories outlined, we can speculate that moral decision-making involves moral judgements, an understanding of moral principles, and is inherent in the agent's personality/nature. Therefore, based on the discussion in the previous chapter, if we were to develop a test for moral capacity, it would entail an assessment of those very features. Can the machine form moral judgements, understand principles and is it inherently moral?

4.4. Lessons from Moral Psychology

In a seminal piece of work, Kohlberg (1971) identifies and outlines developmental *stages* that an individual might go through as their morality matures. Typically, this represents development from childhood through to adulthood, but not all adults necessarily reach full moral maturity. The stages are split into three main groups, with two further sub-groups. They are described as: The Pre-Conventional Level (consisting of Stages 1 and 2); The Conventional Level (Stages 3 and 4) and The Post-Conventional, Autonomous or Principled Level (Stages 5 and 6).

The Pre-Conventional Level – typically represented in younger children – is where the individual is mainly self-centred or *hedonistic in their approach* to moral decision-making. As such, moral

decisions are not made based upon principles or rules (this is reserved for later stages) but based upon considerations for how moral decisions may practically affect them. Stage 1 demonstrates an orientation towards physical punishment and obedience, where the goodness or badness of a situation is reflected by whether there will be a punishment for carrying out a particular act. Stage 2 demonstrates an increased sophistication in the understanding of *what affects the individual*, giving rise to an appreciation of concepts such as fairness, reciprocity and sharing at a personal level (however **not**, for instance, appreciation of the general principle that we should be fair).

At **The Conventional Level** individuals begin to appreciate an approval from their close social groups (i.e. family, friends or nation) regardless of the obvious consequences. This results in an attitude of *loyalty* or adherence to social norms, supporting and maintaining group order. The first stage of The Conventional Level (Stage 3) can be understood as the 'good/nice boy/girl' phase, where the individual acts according to such that they gain approval from individuals in their group, for example, through expected 'good' behaviour. At the second stage of the Conventional Level (Stage 4) individuals start to have some appreciation for broader concepts such as Law and Order, with an orientation towards authority, rules and maintaining social order.

Finally, at what Kohlberg refers to as **The Post-Conventional, Autonomous or Principled Level**, individuals begin to define their own overarching values and principles and apply them apart from the group authority. In Stage 5 right actions tend to be defined by individual rights and standards that have been agreed upon by society. There is an appreciation of the relativistic nature of moral decision-making, with an emphasis that the right action is a matter of personal opinion. This lends to an appreciation of the law, with an understanding that it can be changed for social utility. This expands at Stage 6, where the right thing to do corresponds to the conscience that fits with self-chosen abstract ethical principles (i.e., The Categorical Imperative as opposed to The Ten Commandments). Universal principles tend to be adopted such as Justice, reciprocity, and equality of human rights.

To illustrate the different stages, Kohlberg (1971) refers to an example of an individual who needs a particular drug to prevent the death of their partner. A pharmaceutical company supplies this necessary drug, however, the cost is beyond the financial capabilities of the individual. The CEO of the company refuses to reduce the price of the drug, even after an appeal about the situation. The question is: should the individual steal the drug?

Accordingly, an individual acting at Stage 4 (adhering to social law and order) would think that stealing the drug is impermissible since it is strictly against the law to steal – even from money-making organisations. However, Kohlberg responds that the correct course of action *is* to steal the drug, since an individual acting at higher stages (Stage 5 or 6) would see that doing so could lead to a development of the law (i.e., that withholding a life-saving drug for financial reasons should be

illegal) and furthermore that according to the broader principle of Justice, it is the unwell person's human right to have their life preserved.

What is illustrated by Kohlberg's outline of the moral stages is that morality is *so much more* than deciding between an act being right or wrong. A moral agent, realised at Stage 6, has the capacity to understand notions such as *justice* and *human rights* and furthermore, have *empathy* for other individuals. The need for empathy in moral decision making brings along a lot more cognitive requirements for moral capacity.

It might be argued that Kohlberg's levels represent a *spectrum* of morality and that we ought only to try to achieve the earlier stages when we are creating moral machines. However, this application would be a misunderstanding of what the moral stages signify. What they signify is not a series of features that demonstrate morality in an agent, but more *achievement milestones that are indicative of the machine's development progress and maturity*.

Though Kohlberg has come to be widely dismissed as being an old-fashioned theory in need of modernisation and development, the methodology (identifying moral developmental milestones) and idea that moral capacity is developmental in nature, is a useful one for machine ethics, since it demonstrates that it might be possible to develop a metric to determine moral maturity. This is important when it comes to moral machines in practice, because we want to ensure that the machine is morally appropriate for the use case it will be placed into.

4.5. A Development Framework for Machine Ethics

In traditional software development it is important that the development is carried out with the software's functional needs firmly in mind. Traditionally, *The Waterfall Method* (as per Figure 10) is used as a process for (a) guiding the design process so that the new system meets the functional needs, and (b) testing retrospectively that the newly developed system satisfies the original specification. The first stage often involves some form of requirements elicitation – i.e., determining what features the system needs to have. Though more 'Agile' forms of this traditional procedure are more commonly in use today, The Waterfall Method serves as a good basis to highlight some steps that might be involved in creating a piece of software. Figure 10 gives a representation of the typical Waterfall Method.

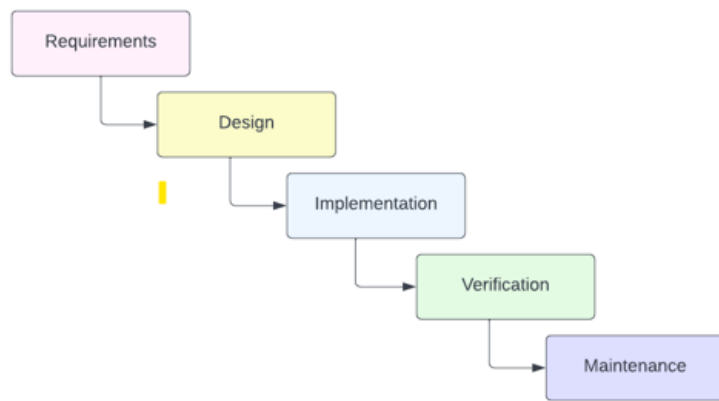


Figure 10: The 'Waterfall Method' for traditional software development.

In terms of creating a moral machine, it is equally important that the machine that is created meets its design needs. As detailed in the previous sub-section, design needs might range from the need to be able to form moral judgements, to having emotions. Though mildly vague now, once developed into more complex and specific functional requirements, these can form the basis of a specification for developing moral machines. Therefore, if we want to comprehensively develop a moral machine, the idea is that needs should be developed and translated into a full specification that is executable by software developers. Similarly, the requirements should be such that they can form the basis of a test framework as well. A basic three-step procedure can be defined for creating moral machines, as per Figure 11.

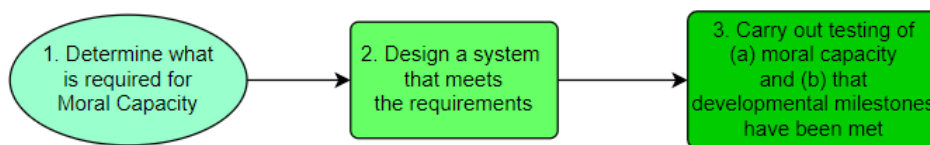


Figure 11: a framework for creating moral machines.

Rather than trying to capture normative moral facts and carrying out tests (such as 'The Moral Turing Test') which assess whether the normative expectations have been met in a machine, we test instead for *moral agency in the machine*, and then, using lessons from psychology (i.e., Kohlberg), assess that the machine is developing appropriately. As opposed to a test which assesses expected behaviour, tests can be developed to assess that the machine is (a) meeting appropriate developmental milestones and (b) has features that are demonstrative of moral agency.

These types of tests give assurance that the machine is moral because the machine can only pass the tests if it demonstrates the features that are necessary for moral agency.

The next section begins to work through this framework, as an example for how this might work in practice, by developing the needs previously outlined into a more detailed functional requirements specification, and then by outlining a test matrix and developmental assessment criteria for moral development.

5. A RECIPE FOR MORALITY

Using the framework developed in the previous chapter, the aim of this chapter is to begin to trace a methodology (aka. *Recipe*) for machines with morals. Now equipped with a new *problem statement* in the form of Machine Ethics 2.0, and a *framework for solving this problem*, an example procedure for developing moral machines can begin to be traced.

To do this, an example functional requirements specification, using insights from discussions in previous chapters, will be developed. The aim is to trace a specification that could be used by developers to design and code a moral machine. In addition to the requirements specification, a test matrix and use-case specific moral maturity assessment will be outlined, to assess the machine to (a) assure that it has moral capacity and (b) that it suits its context-driven need in terms of appropriate moral behaviour for a given use. To do this, first (5.1) some necessary features for moral agency will be explored in more depth, and these will be translated into requirements that could be executable by a designer or developer. Then, (5.2) based upon the analysis, a sample requirements specification will be outlined and explained. Then, (5.3) there will be a discussion on test methodologies for moral agency, and presentation of a sample Acceptance Test Matrix based upon the previously elicited requirements, before (5.4) discussion of a moral maturity test in relation to operationalisation of moral machines.

It is worth noting that the aim of this section is not to provide a complete or entirely accurate specification and test procedure for Moral Machines, but to demonstrate how one *might be* developed using the framework from the previous chapter, and insights from the previous sections.

5.1. What Morality Needs

In software development within an organisation, requirements specifications are used to ensure that a new product or system meets the needs of the business. An adequate requirements specification should be a document that describes what the software will do and how it should be expected to perform. It also describes the functionality that the new product needs to fulfil for all stakeholder (business, users etc.) needs.

As it has already been outlined that – ultimately – a moral machine is a piece of software, it seems fitting that moral machines should be developed based upon an exhaustive requirements specification. However, rather than developing a moral machine to meet specific customer needs, the machine needs developing with Machine Ethics 2.0 in mind. That is, it needs to address the problem statement set out and look at what would be required to develop a machine with the ability to develop its own moral agency. Similarly, the test procedure should be reflective of the framework outlined in Section 4 and designed according to the relevant requirements. Due to variance of moral opinion, it wouldn't be fitting to have a requirements specification that is based

upon one individual, or even a group of individuals', needs, such as could be the case in business requirements for a new piece of software. A requirements specification is needed that is associated with why we want moral machines - i.e., What features does the piece of software (moral machine) need to have to have moral agency? There is a shift in an approach to the problem. Rather than appealing to personal opinions from non-experts (or even philosophical experts) on what features a moral machine should have, stakeholders become developmental, clinical psychologists, potentially experts in criminal psychology, as well as philosophers, sociologists, and game theorists: experts who understand what it means to be *moral*.

It is important to highlight the limitations of this thesis. Inasmuch as it cannot provide an expert overview representative for each one of these experts, it is not able to specify an exhaustive, or completely accurate, list of requirements for moral agency. However, given the discussions held in this thesis so far, and the analysis in 4.3, this thesis traces the start of the journey by assessing what moral agency requires based upon behaviours typically exhibited in normative theories of morality. Developing on from the discussions in the previous chapter, some requirements are mentioned and explored below.

1. The need to be able to form moral judgements

It seems fundamental that a machine with moral agency would require the ability to form their own moral judgements. In other words, when presented with dilemmas such as The Trolley Problem, they should not only be able to form a moral opinion, but make a decision based upon this opinion. However, in terms of functionally translated requirements, it's not obvious what features should be specified for the machine, that can be implemented by a developer. 'Forming moral judgements' is too ambiguous to programme and it is important that software requirements are functionally specific.

One way to dissect the notion is to examine what is required to form a moral judgement *in practical life*. If we consider a human deciding whether it is morally permissible to steal the much-needed drug to save their partner (as in the Kohlberg example), it seems that one cognitive requirement is - in the first instance - to have the capacity to envisage what the future might be if certain decisions are made. For instance, I might decide that in a particular situation it is morally permissible to steal the needed drug because the future where the drug is stolen (i.e., where my partner's life is saved, and the pharmacist loses a bit of money) is more desirable than the future where I decide not to steal the drug. It is important to note, that this case example is not endorsing the stealing of drugs, but demonstrating the cognitive functions that might be required in making such a moral judgement. Similarly, it is not strictly the case that the ability to envisage the future is a necessary feature for moral capacity. Further analysis is required to ascertain if this is indeed the case, but what is demonstrated is that, using 'moral judgements' as the basis for further analysis, we can begin to prescribe more functional, specific requirements to allow technical design to take place.

Where we would define ‘forming moral judgements’ as the overarching requirement for moral agency, ‘the ability to envisage future scenarios’ would constitute a *sub-requirement* insofar as it is derived from the broader requirement. This requirement is endorsed by Coeckelbergh and Mesman (2007) who demonstrate that ‘The Imagination’, and specifically, the ability to imagine future scenarios, and place oneself in them, is required in moral decision-making, through practical case studies involving parents making decisions about their unborn children. In determining the features required for moral agency, it is likely that further, similar kind of empirical investigation is needed.

Of course, ‘envisaging future scenarios’ is also a vague notion, but it is functional insofar as it describes a function that the machine must have if it is to be moral. This thesis does not aim to propose solutions for such requirements, but where simulation technologies seem to contribute in achieving the ability to ‘envisage’ future scenarios, it might be that such technology could help satisfy this requirement.

From the overarching requirement, it might be that further requirements materialise, and further sub-requirements from the sub-requirement. For example, the ability to envisage future scenarios might require some ability to predict. This is a feature in machines that has been perfected by computer scientists (aka. *Machine Learning*), therefore potentially easily implementable in a machine.

It is important to note that the requirements (and thereby sub-requirements) are not the same as cognitive dependencies as per the type detailed in Kohlberg’s (1971) paper. For instance, Kohlberg proposes a series of cognitive features that are required for progression to the next moral levels (i.e., intellectual capabilities etc.), however, these features are not requirements for moral agency per se, but rather prescriptions to enable moral development. A being either has or does not have moral agency regardless of how morally sophisticated this moral agency is, therefore, the requirements for satisfaction (though may be similar) are distinct.

The next sub-section presents a series of requirements and sub-requirements based upon the overarching themes; however, it is worth noting that a more exhaustive requirements specification is needed to build a more comprehensive moral machine.

By specifying detailed requirements for *moral agency*, we begin to see the development of a cognitive model and then architecture for moral machines that could eventually be translated into technical features.

2. The need for forward-planning involving oneself

To understand that an act will have a consequence, and that it would be preferable for those consequences to be the case, requires the ability to not only envisage the future, but make plans

involving oneself, concerning the future. For instance, I know that sharing my sweets is the right thing to do because I know that if I act upon this, in the future, my peers will have their fair supply of sweets. This capacity is in addition to the need to being able to just envisage the future, because it involves a combination of insights concerning how the acting individual (myself) might react given various decisions.

In cognitive psychology, *The Executive Function* is a term often used to denote one's ability to forward-plan about their own future. Therefore, it seems that this would need replicating if we are to have a machine with morals. The requirements for the executive function need to be ascertained to develop a moral machine. Again, it is important to note that these requirements are not the cognitive capabilities that are needed for executive functioning to take place in the human brain, but generally, what functions the machine needs to have to have the capacity exhibited in these types of thought processes.

Coeckelbergh and Mesman (2007) also highlight this phenomenon in their analysis of parents making moral decisions pertaining to prenatal infants, where it is the mother's emotive response to her envisaging herself given certain moral decisions, that provokes her moral decision to have a medical intervention. This then, seems to imply that not only is imagination required for moral decision-making, but in addition, to allow for forward decision-making involving the self, there is the need for some emotional/empathetic capacity for moral agency. This is a point that Coeckelbergh (2010) makes in his articulation for why moral machines are not possible, however, if we learn to understand the emotions, and dissect what functional requirements might be for the emotions, progress could be made in this area.

3. Identity, and the need for a sense of 'self'

Referring strictly to numerical identity (ignoring, for instance, the fact that two separate objects can be qualitatively identical), a subject's identity can be described as the absolute, total quality that can only be held between a thing and itself (Noonan and Curtis, 2022). For instance, the identity 'Rebecca' might be the collection of experiences held in 'Rebecca's' life; 'Rebecca's' movement history throughout space and time, or (arguably) the genetic makeup, or even behavioural and physical features. All these aspects would be unique to 'Rebecca', and collectively, form part of what would be regarded as an identify.

There are many philosophical debates concerning identity (i.e., whether identity is persistent through time, in possible worlds, or at what point identity changes when properties change), however, for the sake of moral machines, we will take the base definition that identity constitutes some quality such that the subject has its own defining qualitative features.

Therefore, if a machine has a sense of identity, it means that it can recognise its self as separate from other artefacts in the world and understand that it has unique qualitative features (to some extent) and that its position in space and time is unique. It could be argued that a moral machine would need to have this kind of sense of identity to have moral agency because a machine would only be able to imagine itself in future scenarios if it has a sense of 'self'. That is, to know that there will be certain consequences to oneself, given certain decisions, requires the ability to distinguish oneself from the surrounding environment and other agents.

Translated to more functional requirements then, it can be argued that there is the need for a representation of the 'self' within the moral machine's cognitive architecture. However, for more specific requirements, there will need to be further in-depth examination of examples for what 'self-identity' actually constitutes.

4. The need for moral autonomy

A system is said to act autonomously if it acts without human oversight or control (Ikova' and Ilka, 2017). For instance, a vehicle might be described as autonomous (an Autonomous Vehicle) if it doesn't require a human for its operation.

When talking about *moral autonomy therefore*, we refer to the ability to form a moral decision/judgement/opinion without human oversight or intervention. When we consider a human in a situation where they must decide whether to steal or not, we would say that they are making an autonomous moral decision if their decision not to steal was made independent of any external intervention. For instance, if a shopkeeper spotted the potential theft and removed the individual from the shop before they could carry it out, this would not be an autonomous decision on behalf of the acting agent, since an intervention from the shopkeeper would have prevented the act taking place. Though the term 'autonomy' has various social and political connotations (see Veliz, 2021) and therefore it might be difficult to always identify if an agent is acting *wholly autonomously*, the broad definition for autonomy is still useful in the case of moral machines.

Some suggestions for Autonomous Vehicles for handling dilemmas like The Trolley Problem *in real life* are for there to be a 'human overdrive' faculty fitted into the vehicles so that the dilemma is given a human oversight before the machine executes the decision. This might be akin to a 'power switch' that turns the machine off if it's about to act immoral. Such a machine (even in the case of the power switch) would not be acting with moral autonomy since the moral judgement would be made under the control of, and constrained by, the human. In this case the machine would not have its own moral agency because it would not be free to form and execute its own moral judgement

With moral agency it is important to have moral autonomy (i.e., the freedom to make your own moral decisions) because true moral development and cultivation cannot happen without it. Where a child might learn that it is right to share, through autonomous interaction with its peers, the same must be the case for a machine. Therefore, in creating a moral machine there could be a functional requirement to not constrain the machine's moral decision-making insofar as the machine must be able to make its own moral decisions, freely, to be a moral agent. This is distinct from external constraints that might be placed around an individual to manage their decision-making, but concerns the *explicit* decision-making capacities of the agent. As humans then, we would need to learn to trust the machine to the extent that we permit it the freedom to form its own moral judgements, even if we have safety procedures surrounding the machine's judgements. This might be an awkward position for many researchers who see morals as constraining immoral behaviour (protecting humans from harm).

This gives way to a sub-requirement for moral machines that is also mentioned by Sparrow (2021): the need for trustworthiness.

5. The need for knowledge-acquiring ability

The final major requirement postulated in this thesis is the need for moral machines to have the ability to acquire knowledge. What could this mean?

Traditionally, in philosophy, knowledge is given a tripartite definition – X knows that y is the case, if X has a justified, true belief about y (Ichikawa, et al., 2018). Justified, insofar as there is good reason why X believes y to be the case, and true, insofar as there is something that directly corresponds with the belief, *in reality*.

A moral machine would require the ability to acquire knowledge for their actions to have import in the real world. Just simply forming the moral belief that lying is wrong – say – requires knowledge of concepts such as 'lying' and moral knowledge that it is wrong. Therefore, functionally, there is a requirement to be able to assimilate information about the world (that it contains tables and chairs etc.), understand abstract concepts (such as, 'lying', 'sharing' etc.) and then form a moral judgement about those artefacts and their relations existing in the world. For instance, to judge that it is wrong to steal from the supermarket, I must first be able to identify the objects of the theft and the supermarket, but I must then also be able to understand my action in relation to these objects, that this constitutes 'theft', and then finally that it is wrong to carry out this act. The relations between these pieces of knowledge can be informationally represented, but essentially, we might classify that moral knowledge involves two types of knowledge: that is, metaphysical knowledge about the world, and moral knowledge pertaining to the metaphysical facts about the world. Of course, where metaphysical knowledge might be regarded as objective in nature, moral knowledge would be more personal to the acting individual. A further functional requirement then, would be

that the machine can cognitively map knowledge about the world, and then have moral knowledge about this knowledge.

The idea of machines being able to form hierarchies of concepts can be found in Moor's (2006) dissection of ethical decision-making, where he asserts that if machines had this capacity they would be regarded as full moral agents. It seems then, from this analysis, that an implication of the requirements is that we cannot partially fulfil criteria to create moral agents, and that if we are going to create moral machines, they need to be *full moral agents*.

This concludes the overarching requirements assessment for moral agency that will be discussed in this thesis. The requirements are not wholly exhaustive but have been specified to illustrate how the framework and methodology outlined in the previous section could be used by researchers to develop moral machines. Assuming the requirements that have been outlined here to be absolute, there are further sub-requirements that could be logically extrapolated as a result. The next sub-section presents a more detailed sample requirements specification based upon this extrapolation.

5.2. Specification for Morality

Detailed requirements specifications can be presented in a table. At the top of the table is the key aim of the new product, and underneath are separate requirements for achieving that aim. Sub-requirements fall under the overarching requirements, and there is usually space for technical or more specific functional descriptions.

This specification is largely to illustrate how the framework outlined in the previous section might be applied. However, using the features described in 5.1 as a basis, a sample specification is demonstrated. It is lacking technical requirements because these will be discussed in the next chapter, alongside some cognitive and technical models for moral agency.

Specification No.	Requirement	Notes
1.	Moral Agency	Key aim
1.1.	Form moral judgements	Requirement
1.1.1.	Imagine the future	Function
1.1.1.1.	Envisage future scenarios	

1.1.1.1.1	Predict outcomes given various moral decisions	
1.1.2.	Forward Planning	
1.1.2.1.	Executive functioning	
1.1.2.2.	Empathy	
1.1.3.	Sense of Identity	
1.1.3.1.	Recognise itself as separate to other artefacts	
1.1.3.2.	A capacity to represent the 'self'	
1.1.4.	Moral Autonomy	
1.1.4.1.	Make moral judgements without human oversight	
1.1.4.2.	No constraint on the machine's moral decision-making	
1.1.5.	Acquire knowledge	
1.1.5.1.	Assimilate information about the world	
1.1.5.2.	Understand abstract concepts	i.e. 'theft'
1.1.5.3.	Form judgements about the abstract concepts and their relation to the world	

Figure 12: A sample requirements specification for developing Moral Capacity in a machine.

It is possible that some sub-requirements are already captured by previous requirements, or that sometimes requirements duplicate. However, it does not matter if there is a duplication since this only leads to an overly cautious approach. As sub-requirements are specified, the cognitive model

for development becomes clearer. Though not all requirements have been previously assessed in this section, this specification can be used to demonstrate how a cognitive model and technical architecture can be developed. This is demonstrated in the next sub-section.

5.3. Using the Test Framework

Previous tests for morality in machines have focused upon a machine meeting behavioural expectations – for instance, acting in a way that seems ‘good’ to the observer. The Moral Turing Test has this broad approach: a machine is moral if it can make a human believe it is so, through whatever means necessary, whether that be having a certain political opinion, behaving according to social norms or buying all its resources from a sustainable supermarket etc. These *behavioural tests* suffice in testing that a machine acts according to presupposed moral norms or expectations, but they do not test for Morality – i.e., that the machine is a rationally thinking moral decision-making system. As previously demonstrated, it is not an objective test.

As Arnold and Schultz (2016) argue in their critique of The Moral Turing Test, what is required is a process of verification for Moral Machines. It is not good enough that machines act morally today (not harming humans etc.), what is needed is a system to verify that *this will always be the case*.

However, as also explained in the previous chapter, due to the uncertainty of the outputs of AI systems, verifying that an AI system will act a certain way is not always possible. Therefore, what is required instead, is a system to *assure* that the machine will not act immorally. Assurance and verification differ insofar as verification involves a mathematical proof that a certain event will occur; assurance is more an exercise in risk mitigation. We can’t prove that the machine will never act a certain way, but we can assure its moral behaviour through mitigation strategies and design. As it was outlined, creating a machine with moral agency would give us this assurance.

It might be argued that just because a machine has moral agency does not mean that it will choose to act in a moral way. The argument follows that for a machine to behave morally it requires a motivation – potentially either some intrinsic or external motivator to ensure that it always does the right thing. The pursuit for this quality is misguided however because if I am an agent that understands that it is right to share my sweets, this very fact is sufficient to motivate me to share my sweets. Furthermore, if we start trying to motivate a machine to perform a certain act, this ultimately takes away from its moral autonomy. The matter comes down to the requirement for human *trust* in the machine. We don’t need to motivate a moral person to be moral because they *are* just that.

Therefore, we can have our appropriate level of assurance by checking that the machine meets the specification requirements for moral agency. In traditional software development, a way to check that a requirement has been met (with *moral agency* being the overarching requirement in this

case) is to carry out a User Acceptance Test (UAT), that is: check that the requirements have been met by presenting the system with a series of scenarios to test capability.

For instance, if we consider the following requirement:

1.1.2.1. Executive functioning

The way to assess that this requirement has been met could be to check that executive functioning is present in the machine through a series of cognitive tests. Questions might be asked such as:

Can the machine organise inputs?

Can the machine use the organised inputs to formulate a plan?

Can the machine execute its own plan?

Some of these questions are possibly too broad for a Machine Ethics software developer to test, but an equivalent test case could be:

A user asks the machine to plan a journey

As in software development test cases, a way to verify that the requirement has been met would be to check that, in this test case, the machine behaves according to an expectation, such as:

The machine adequately plans a route from A to B.

This is by no means the only test (or indeed, any test) for executive functioning, but we can see how the process may develop. As a table may be used to document requirements, a matrix can be used to represent test cases and expected behaviours, with plans for retesting if any scenarios fail. Using the table developed in Section 5.3. a simple sample test matrix for moral machines is demonstrated in Figure 13.

Number	Requirement	Test Case	Expected Behaviour	Achieved ? (Y/N)
1.	Moral Agency	A psychological test for moral agency	The machine passes the test	
1.1.	Form moral judgements	The machine is asked its opinion on	The machine gives an opinion	

		a specific use case involving, e.g., 'theft'		
1.1.1.	Imagine future scenarios	The machine is presented with a series of fictional scenarios and asked to imagine what will happen next	The machine gives an outline of the imagination	
1.1.2.	Prediction	The machine is asked what will happen in a scenario if certain events hold	The machine accurately predicts the outcomes	
1.2.	Forward-planning	The machine is asked to plan a route	The machine demonstrates that this has been done	
1.2.1.	Executive Functioning	The machine is asked to plan a route involving itself	This is carried out	
1.2.2.	Empathy	The machine is asked to describe 'emotional states'	It effectively does this	

Figure 13: A simple sample test matrix for testing moral capacity in a machine

Figure 13 presents an example test matrix for the first two overarching requirements from the requirements specification. Immediately, it is obvious that some of the prescribed tests are insufficient for the function described, but they demonstrate the *kind of tests* that could be used against the specified requirements. However, for increased sophistication it will be necessary to appeal to research within psychology, i.e., for tests concerning how to assess executive functioning etc. What is important to understand for this thesis is not just the test cases themselves (these should be reassessed, reinterpreted, and redeveloped by experts in the relevant fields), but the methodology that has been applied to develop these. What the methodology here highlights is that

Machine Ethics is not something that can be solved with one simple solution. It is something that will require complex development over time. Where Machine Ethics might have previously been a discipline sat firmly within computer science, there is a need for it to be more interdisciplinary and reach out to other academic areas.

Though the focus of this thesis is to demonstrate how the new framework might be applied, to illustrate what a typical test case might look like for testing moral agency, let us consider a requirement that is more commonly used within robotics, for simplicity, Requirement 1.1.2. from Figure 13:

A requirement for the machine to be able to predict future scenarios.

We have specified the requirement in more detail to exemplify the specific *features* that would be necessary (as per the arguments above) to enable moral agency. The corresponding test case has been described as the following:

The machine is asked what will happen in a scenario if certain events hold

There is also the following test *outcome*:

The machine accurately predicts the outcomes.

To test the machine according to these conditions, there is an assumption of linguistic ability. However, a similar test could be to see whether the machine is able to accurately predict cause and effect outcomes. For instance, a series of objective scenarios could be developed with obvious answers, that the machine has to foresee accurately. These could be example physical environments where the machine, like a human could, would need to predict ‘what happens next’.

A test might take a form similar to something like the following:

“Molly is playing a game of tennis. She is holding a racquet, awaiting her opponent to make the first serve. The opponent serves the ball and it is sent in her direction. Select all of the phrases below which could be outcomes of this situation:

- A) The ball suddenly stops mid-air and drops to the ground
- B) A gust of wind blows the ball to one side and Molly misses the ball
- C) The ball hits Molly’s bat and she successfully returns the serve”

As humans we understand action and movement mechanics, so can easily foresee possible futures. We know that (B) and (C) are possible, but that (A) is not. In turn, we have the ability to *predict the future* insofar as we can work out what should happen next, but without predicting ability, understanding consequences of actions, working this out would not be so simple to a machine.

Though this example is a somewhat complicated version of events, more simple scenarios could be described – literally pushing a domino over and asking the machine *what would happen next*.

This example may need significant development to effectively serve as an effective test for prediction, but it demonstrates the *type of test* that can be used to assess performance in a machine. Such rigorous testing would serve as the assurance that is needed.

5.4. Moral Machines in Practice

The development process and tests that have been discussed so far have focused purely on the software development aspect of moral machines. However, traditionally, in software development, as demonstrated by The Waterfall Method, the process does not end there. There is a point where the machine needs to be integrated into its operating environment, and as with any AI system, relevant measures need to be put in place to ensure that when Moral Machines are put into their environments, they are done so safely.

Kohlberg (1971) highlights that moral ability is not something that is bestowed upon an agent immediately (either at birth or a certain point in their life), but something that develops as the agent matures. Moral ability is a developmental process, and as such moral agency is as much about a cognitive framework to facilitate this development as it is about achieving certain requirements.

The requirements framework and test matrix can determine if moral agency has been attained in a machine, but it does not determine the degree of moral maturity that the agent has. For instance, a child, able to get into moral debates and argue that it's 'not fair that their sibling gets a present and they don't', would be regarded as being moral insofar as they meet the specified requirements from the previous section, but according to Kohlberg, might only be at the second stage in terms of moral development. As will be demonstrated in the next section, the fact that morality has a developmental trajectory gives clues for how we might technically design a moral agent, but for now we can understand that not only is the moral specification important in a moral machine, but the level of its moral maturity is too.

But why do we care about the moral maturity of our machine if it is already a moral agent?

This is because the moral maturity of the machine (i.e., the equivalent of a developmental level in a human) determines the role it plays within society and becomes particularly significant when machine-society integration takes place. If we reflect upon a child (that might only be at Stage 2 of Kohlberg's levels), their life is (typically) entirely supervised by a caregiver. This, it could be argued, is because children can't be held responsible for their decisions until they have fully matured and gained the capacity to make decisions that are conducive to themselves. It might be favourable to a child to eat all the sweets in a shop, or to just take an item from a shop that they want, but we know that the former would not be in their best interest and (probably) the latter would

not be moral. Hence, children receive supervision on their behaviour. Their caregiver prevents them from eating too many sweets and prevents them from stealing an item from the shop. At the early stages of moral development children are *constrained* in what they can and can't do. This is ultimately to protect the child, but also maintains moral order.

This fact is reflected in the legal and social system: children typically cannot vote, choose to join the army or drink alcohol until they are in their late teens. In the UK, social services must ensure that children are properly cared for until the age of 18. In terms of adult roles and responsibilities in society that require moral responsibility, medical practitioners must make a legal oath before they are allowed to practise, and individuals working in law must undergo many years of training and pass various examinations to gain their positions. Many other professions have 'chartership' status, which is a recognition of adherence to quality of performance.

In developing a machine with moral agency there is a need for some metric to determine the level of its moral maturity to guide physical and governance constraints that should be put into place to mitigate any risks associated with a lack of maturity. In the same way as we might constrain the behaviour of a child, we wouldn't want a machine making medical moral decisions if it were morally immature.

One way to develop this test is, as explained previously, to map and test against developmental milestones that the machine must meet to be allowed to carry out various tasks. Though Kohlberg's levels might be a good metric to assess the machine against a typically developing child (we might say: 'this machine is as morally mature as a 7-year-old'), because machines generally are often designed with a function or role in mind (i.e., 'a care robot') the moral maturity test should be carried out against the functional maturity required for that role. If we consider a robot making moral decisions about unborn infants, we might want to assess the machine in as rigorous a way as we assess medical professionals working in the same position - for instance, demonstrating that they meet at least Stage 7 of Kohlberg's levels.

5.5. Notes on the approach

It is most important to understand that the 'recipe' in this section is incomplete. By no means is there space within this thesis for a comprehensive examination of all the philosophical ethics or moral psychology literature that would be required to develop the level of rigour that would be required for a comprehensive requirements specification or test matrix. Similarly, as the test for developmental maturity is dependent upon the role of the machine in question, functional requirements can only be developed with a particular use case in mind. Later, an experimental case study is outlined to further illustrate the arguments. However, for now, what is important is that in-depth interdisciplinary research and analysis is the way to truly create moral machines.

The following points suggest areas that need more research and development:

1. Philosophical and Psychological research into what constitutes moral agency

More work needs to be done to truly understand what moral agency is. Though requirements have been suggested, rigorous scientific testing and analysis needs to take place to understand what features are necessary for an agent to be considered moral. This is a distinct endeavour to understanding what constitutes moral behaviour, since, as already argued, studies relating to moral behaviour are largely inspired by anthropocentric accounts, and morality needs to be understood objectively

2. A more detailed requirements specification

Once research begins to properly understand moral agency, there needs to be a detailed analysis for the requirements that would enable this. Requirement specifications need to be very detailed to allow for precise development. It should also be an iterative/ongoing procedure; as more insight is gained into what constitutes moral agency, more requirements can be developed and added to the design specification. This brings in the need for a software development framework that is flexible in nature (discussed in the final, *Ethics of Machine Ethics*, section of the thesis).

3. Rigorous tests for the requirements specification

Not only does the test framework need additional development, but the test cases and their expected results need further research. User Acceptance Testing might be a way to check the requirements; however, these tests need to be developed from a non-anthropocentric perspective. Again, this is an ongoing process, dependent upon the specified requirements, and the analysis for what constitutes moral agency, but for a thorough test to be carried out on the machine, it is important that this research takes place. Iterative improvement of the system also ensures that the system remains safe as changes are made to the design.

4. A rigorous framework for functionally testing moral maturity

This links to the safety of the moral machine (and ethics). To ensure that the machine is safe for society, a framework needs to be developed that tests the moral maturity of the machine against what would be required for the role it will carry out. Once developed, then for each specific use case, moral maturity requirements (i.e., what features are necessary for the given use case) will need to be developed. How morally mature does a machine need to be to make medical/legal/care decisions? Once in place, there is then a question of how the machine should be constrained if it does not meet necessary milestones. Constraining behaviour in this way does not prevent autonomy, since the machine will still be able to make its own moral decisions, just will be

restricted on the impact it has. What governance and risk management frameworks are required for this new breed of moral agent?

Ultimately, what this section should do is to inspire new debate and conversations in Machine Ethics. The journey to moral machines is at the very beginning.

6. DESIGNING A MORAL AGENT

We have a reinterpretation of the Machine Ethics problem, and an initial sample requirements specification for creating a machine with moral agency. There is a sample test framework and an understanding that functional moral maturity tests need to be developed for safe operationalisation. Hypothetically assuming that the requirements specification outlined in the previous section is sufficient, the model for a moral machine can begin to be traced.

There are obvious limitations. In the case of this thesis, the requirements specification is incomplete, and the test case examples only outline tests for the first two overarching requirements. However, based upon the five key requirements that are stipulated in Section 5, a broad skeleton can be used to start hypothesising about a technical solution to Machine Ethics 2.0. This section begins to outline the hypothesis and design based on the specification that has been developed, and then presents a case study to demonstrate how the machine might operate in a real-world environment.

First (6.1) the functions outlined in the previous section are discussed to create a conceptual cognitive model for moral machines and to introduce the idea of *raising robots to be good*, then (6.2) a brief overview of technologies at the disposal of this project are outlined to enable a proposed technical solution for moral machines that builds upon traditional RL-based architectures. Then, (6.3) a practical case study is used to demonstrate how such robots might be developed in a real-world environment, before finally (6.4) they are tested and analysed. Again, it is important to note that what this section details is a methodology for using the new framework, as opposed to a definite technical specification.

6.1. A cognitive model of morality

As already explained, as the requirements become more refined, it becomes easier to understand what features the machine must have to satisfy the overarching requirements. The first requirement concerns the need for the machine to be able to form moral judgements. To form a moral judgement, as it is outlined above, the machine also needs to be able to envisage future scenarios and make predictions about what might happen if a particular decision is made. There is also a need for forward-planning, involving the ‘self’ (i.e., some degree of executive functioning). If we think about the cognitive decision-making that might underlie the requirements from the previous section, we can see that a moral machine must, in order: *be able to retrieve information about a situation in the external world, recognise that this is presented situation is moral,*

rationalise about the situation (i.e., which action should be taken) and then act upon this rationalisation. This might be represented as a flow chart, as per Figure 14:

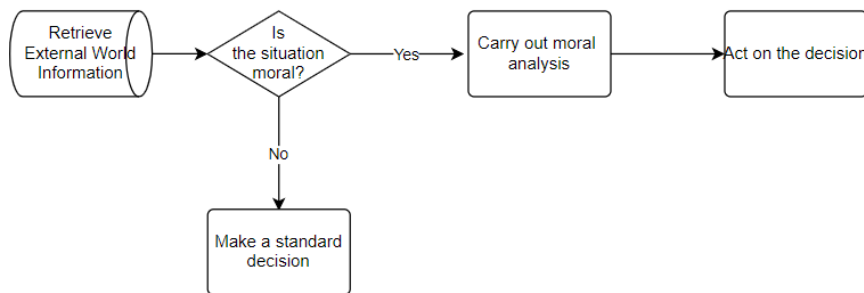


Figure 14: a basic process diagram representing the steps an agent might need to go through to make a moral judgement.

For each step detailed, we might technically envisage how it could be achieved. For instance, it seems that for the machine to be able to retrieve information about the world, the ability to perceive might be useful. One way to allow a machine to perceive its environment could be by equipping it with a camera or series of sensors. Therefore, we might postulate that the first step of the process be replaced with a detection and identification procedure – namely the machine using a camera to detect the scenario, and then machine learning to categorise. This is depicted in Figure 15.

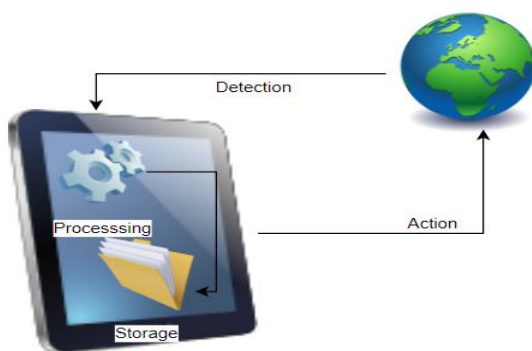


Figure 15: A model representing a machine's retrieval of information from its environment.

Where a smart tablet is used to represent 'a machine with a camera', we can see that once the scenario in the 'world' has been detected, relevant processing is required to determine what the scenario in question is (i.e. does it have moral worth). (It is worth noting at this point that determining the moral weight of a situation is not an obvious thing to ascertain. As already highlighted, Anderson and Anderson (2018) in their account of GenEth, have one technique for the

machine to determine the moral saliency of a situation (i.e. does it contain certain morally relevant phrases?), but this task in itself is something that will require significant technical consideration, so for the purpose of this thesis, is omitted as a technical detail). The answer, then stored, is used to inform future actions (in the world). In this case, because a tablet rather than a robot has been used, the action might simply be saying a phrase or command.

Now we are starting to understand how data (about the scenario) retrieval and processing might take place, we can start to build a cognitive model for moral learning. What is missing from the diagram above is the actual moral decision-making. The machine might recognise that we're in a moral situation, but how does it decide on the right thing to do? How does it, first, know what the right thing to do is, and second, be motivated to act on this. These new requirements are depicted in Figure 16.

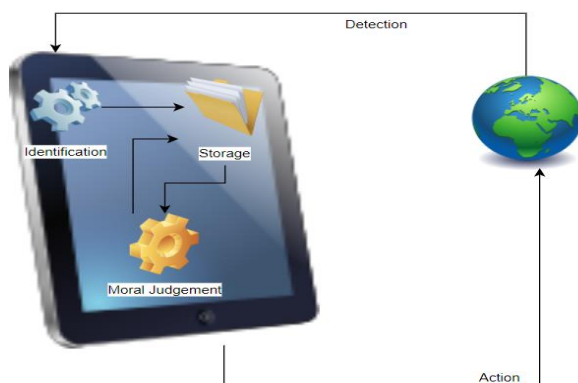


Figure 16: retrieving information from the world and forming a moral judgement.

What is important to note is the importance of the storage capacity within the machine. To act upon analysis of data taken from the environment, requires somewhere to store the outcome of the analysis. Using machine learning, the iPad above might identify that we are in a Trolley type situation, with a lever, a trolley and two tracks with individuals tied at the end, but without a repository to store that information moral analysis cannot take place. Therefore, through representing the requirements in terms of cognitive models, we have identified that two tiers of analysis are required for a moral machine – a level that detects and interprets the environment, and then a secondary (meta) level that performs analysis on top of that analysis.

Within AI there are many tools to carry out the base-level analyses (consider the whole field of AI Vision), but what kind of analysis takes place at the secondary level? Through the exploration of morality in this thesis we know that morality isn't entirely a rational decision but isn't entirely emotive either. So, what is a moral analysis?

Some insight can be gleaned into this question by logically assessing how moral decisions take place. We already know that an agent must be able to understand its environment to form a moral opinion about it. For example, I must know that there is a table in front of me, and that if I push it over, it might break... for example. Similarly, in a trolley dilemma, to form a moral judgement I must know that there is a trolley hurtling out of control and that it is likely to hit unsuspecting victims unless I do something about it.

In Bridge et al (2021) a theory of human moral motivation based upon pro-sociality is developed. In short, it is argued that morality develops due to the need to engage with other individuals, because – at the very basic level – human survival depends upon it. Considering a newborn infant, whose survival is entirely dependent upon a caregiver providing them with warmth, food etc. it is in their best interest to cultivate a social relationship with the caregiver, so that the caregiver gives them their attention. As the infant grows, and develops more sophisticated linguistic abilities, their mechanisms for achieving pro-sociality develop in line with Kohlberg's levels, but their desire to cultivate relationships still drives this. Cognitively then, it might be argued that the motivation for a machine to *do the right thing should be founded upon the need for pro-sociality* - a maximisation on relationship-building.

So once the representation of the world has been analysed, *the moral action* comes from an intrinsic need for the machine to form relationships. But how can a machine learn the best way to form a relationship?

It seems in human development, that the caregiver plays a significant role in enabling children to understand and secure relationships. As detailed in Bridge et al (2021), it is the caregiver-infant relationship that cultivates the relation of 'trust' and forms the foundation of future relationships. Similarly, the caregiver seems to play a role in educating the child in terms of the best way to cultivate relationships. I know that *sharing is good* because as a child my parents encouraged me to share with siblings and other children. Where there is data storage based upon the machine's analysis of the world, it seems that a further 'labelling' takes place on that data, so I can understand how to manipulate it to develop relationships. I care about developing relationships because as a child I seek the approval of my parents, because I want to cultivate a relationship with them – for my survival. With initial caregiver relationships often representing the first point for morality, we see how morality takes on a developmental form. In the same way, moral growth needs to be cultivated in the machine.

This relationship and model are demonstrated as per Figure 16:

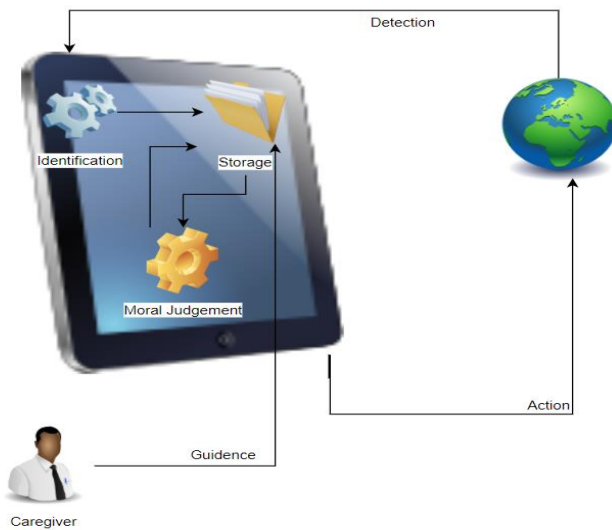


Figure 16: How the relationship with a caregiver can *guide* moral judgments.

Guidance from a caregiver might not be the only way an agent acquires knowledge about pro-sociality. As Currie (1995) outlines in his analysis of the role of fictions, we might receive 'knowledge' about appropriate ways to cultivate relationships through reading, watching films or role-play in real-life or computer games. Each of these mechanisms serves to help us understand how we can better develop relationships and can be replicated in a machine by giving it data about real-life scenarios or presenting it with stories or fictions. Another mechanism to develop knowledge in this regard might be to simply observe different people in the world. I know that it's good to share because I have seen examples in reality – with good outcomes.

In this respect, developing a machine with moral capacity is just like raising a child, hence the title of this thesis – *'Raising a Robot to be Good'*.

Other requirements listed in the previous section allow the detail to be expanded in this model. For example, the requirement to be able to forward plan could be partially achieved by giving the machine some ability to map expected outcomes through a prediction capability. Again, the outcome of this information would feed into the data storage module, to contribute to the machine's overall knowledge. However, for the purpose of this thesis the more specific details will not be investigated any further.

As one final note, it is worth pointing out that this model sits in line with models that have been developed by Matsuo et al. (2022) in response to how we might create *Artificial General Intelligence* (AGI). In their model they outline the need for a 'world view' to make general decisions. This is akin to the 'data storage' component described, where essentially there is a need for a repository to store the outcomes of the analysis. In this thesis it is suggested that this is due to the need to carry out 'meta' analyses on information about the world. How the different levels of analysis interact, has not yet been explored, but could be part of the next phase of AI research.

This point also reflects the Pragmatic accounts of cognition that were given earlier on, since according to Peirce we don't directly access the world, but only access it through *our own perception*, where the repository in this model is the equivalent to *our representation* because it must go through analysis before action can be taken on it.

6.2. An architecture for moral development

As outlined in Section 3.3, Reinforcement Learning (RL) is useful when it comes to programming machines driven by motivations or goals. The basic architecture for a Reinforcement Learning system is demonstrated in Figure 17.

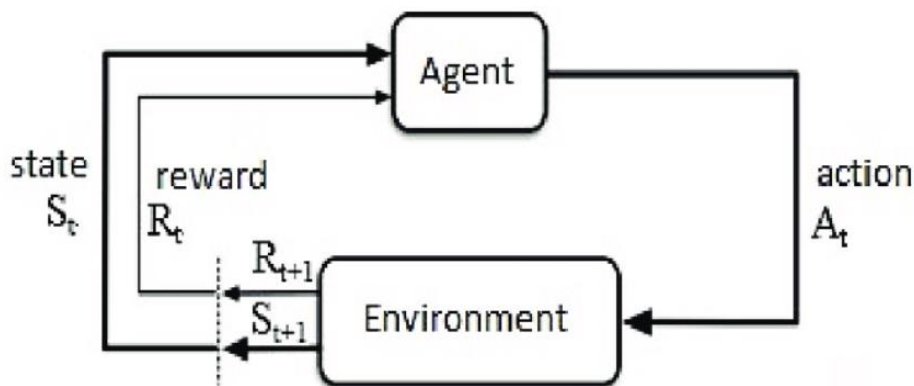


Figure 17: A standard Reinforcement Learning Model taken from Reyna et al, 2019.

Where the equivalent 'AI system' is denoted by the 'Agent' in the diagram above, as can be seen, with a RL agent, a *reward* is given for achieving a certain *desirable* state, which ultimately informs the action of the agent in the environment, and reactions from the environment based upon the action, which inspire the next set of reward/state combinations. What is significant about RL agents is that they can only learn the reward associated with a given state by enacting that state, so their learned behaviour increments as it explores new states. The exploratory element of an RL agent is also particularly significant because the agent must 'try something new' to learn the reward associated with a particular state. An agent that is particularly prone to trying new things might be referred to as a 'risky' RL agent, with the opposite type being 'risk averse'.

This model is like the cognitive model described in 6.1 insofar as in both cases we have an agent ('system', 'machine...') acting in an external environment. A difference between this model and the cognitive model described in Figure 17 however, is that in Figure 17 the behaviour of the motivation and analysis of the state are carried out based upon the *representation* of the actual environment. Hence, the standard RL model above can be adjusted as per Figure 18.

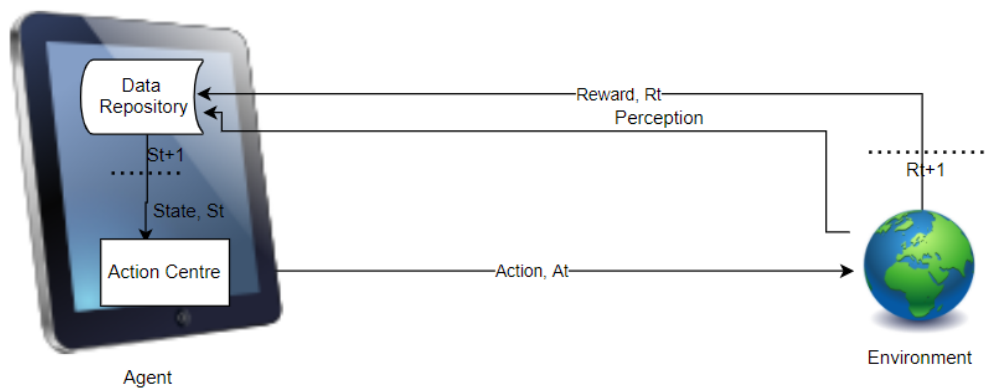


Figure 18: An RL-based architecture for moral learning

Note that all information (including details about the reward) must pass through the data repository before any action takes place. Though the reward comes from external sources, the behaviour of the machine is *intrinsically* motivated by the internal data model. This is important because due to this intrinsic motivation, there becomes a clear distinction between the machine and the external world. We might imagine that the data repository contains information about the machine itself – rectangular, owned by ‘Apple’, etc. In a world full of tables and chairs, therefore, the machine agent would have the required sense of identity.

It might be argued that because the reward function in this model comes from the external environment that the machine is still inextricably linked to its environment, but this is not strictly the case. As humans, our motivations – say – come from the relationships we form with others, but we still exist as separate entities. In a similar way, regardless of the external influence, in this Moral Machine architecture the decision-making happens *inside the machine*. Therefore, it is a separate entity. **The machine** decides whether to pull to level or not in a Trolley Dilemma, and **the machine** decides which its preferred state is, it is not wholly influenced by the environment.

This level of internal governance is important when we consider cases like Delphi, where the behaviour of the machine is entirely dependent upon external influences. As per the Microsoft chatbot ‘Tay’, which degenerated into racist and sexist comments after being placed into the Twittersphere, the problem with basic RL approaches is that there is nothing to *vet* the information coming into (and influencing) the machine. A moral human agent wouldn’t just do what everyone in their environment tells them to do (or be entirely influenced by what everyone else is doing depending upon the reward function). Instead, we’d like to think that a moral-headed individual might stand against such external influences, forming their own moral opinion and potentially making a stance against misbehaviour. Essentially, the data repository, which is the machine’s ‘version of events’, acts as that governance filter.

It might be suggested that a supervised AI system (i.e. one where ‘what is learnt’ is dictated by a pre-given policy) might resolve this problem. However, as mentioned previously, this approach would be susceptible to bias because the policy would need to be set by somebody, and that somebody’s biases would ultimately inform the learning. We might restrict Tay’s learning so that it ignores every input containing racist statements, but then the learning is susceptible to prejudice as to what the supervisor thinks constitutes a racist statement. More generally, when a child learns, they are (possibly, unfortunately) exposed to different types of language (some less desirable to the caregivers). It is the commentary *on this language*, that might be deemed a supposed metacommentary on the perception of reality, from the parents, that ultimately guides the child’s perspective.

6.3. Raising robots in the world

Let us consider a case study in which we might need a moral robot: in a care or hospital setting. Accordingly, with healthcare robots making essential decisions about who to prioritise for treatment, who to administer treatment to and how to administer it, not only would the typical social nuances be required for such a robot (as per the case study in 3.3), but explicit moral decisions must be made. To make the case study all-the-more specific we might imagine a cancer treatment robot, who must prioritise treatment of sick patients.

Supposing there isn’t a simple algorithm to determine the priority level of a cancer patient due to the often-personal nature of cancer diagnoses, a sophisticated/mature moral capacity might be required to mediate between cases and distribute care based upon fairness and need. Where machines such as Delphi would approach this issue by crowdsourcing on how medical practitioners have treated patients in the past, and aiming to replicate this behaviour (hence leading to the risk of exacerbating and amplifying any pre-existing biases), using the architecture in 6.2, the machine might instead be presented with a series of relevant *case studies* from which to begin to develop a body of domain-appropriate moral knowledge.

For instance, we might imagine one such case study being of ‘A patient with a skin cancer diagnosis, who received treatment, that had a positive outcome’. Guided in the relevant way (i.e., potentially told that this was a positive outcome), the machine would be able to begin generating a repository of knowledge about what types of procedures might be successful, along with the features that make for a priority case.

Note that there is a shift in semantic perspective. Where Delphi is motivated by our human perception of ‘the good’, the new moral machine is only motivated towards pro-social goals – as such, it is free to define its own version of ‘the good’.

This is a turning point in how we have typically viewed Artificial Intelligence Agents, where instead of exercising caution, controlling the moral decisions the machines make, we must *trust* the artificial agents, to get the most out of them. Fortunately, one way to gain trust is through *assurance*, which will be the focus next. Can we assure that such a robot will be moral?

6.4. Testing and Future Work

Many of the tests outlined in the test matrix are behavioural in nature. That is, though they do not assess moral agency based upon behaviours, they use the machine's behaviour to assure that the associated requirement has been met. For instance, some test scenarios require that the machine can communicate with the tester, therefore, there would be a requirement for some communication ability from the moral machine.

The architecture in 6.2 would not pass this specific test since a faculty for linguistic comprehension and communication has not been given. This capability might require Natural Language Processing (NLP) abilities or other language models. However, despite the described architecture not passing *that specific test* for moral agency, it does not mean that the architecture would not pass another test designed to assess moral decision-making. In fact, passing *any* appropriately developed cognitive assessment test would be sufficient to confirm that the machine satisfies the necessary requirement. This is where *choosing the right test* becomes important and requires the research insight of a cognitive or developmental psychologist or human-computer-interaction expert. We wouldn't give a cognitive test that required linguistic capability to a new-born infant, for instance, simply to test eyesight. Instead, we might look at gaze, or reactions to light. Tests must be adapted for the abilities of the agent.

In the case of the care robot described above, because the role of the agent is so specific in nature, we might develop a more specific test to assure that it will always prioritise the right patient, by interrogating the robot's cognitive process. One way to do this might be for the machine to give an explanation behind its motivations for the decisions it makes. Keeping the robot in a safe practising environment until it passes this test, the equivalent would be an examination scenario for medical students.

Another way to assure the *right behaviour* could be to carry out a User Acceptance Aptitude Test for 'prioritising the right patients'. Where metrics are assigned for *severity of illness*, *chance of recovery* and other prioritisation factors that are typically used by medical practitioners, a score could be assigned in a hypothetical game. The robot would need to score highly enough to warrant it entering a real-life medical scenario.

Obviously, the tests outlined here are largely speculative, but they demonstrate, with more research, that it is relatively simple to gain the necessary level of assurance to encourage trust

between humans and machines. Again, this traces only the beginning of the Machine Ethics journey, but hopefully one where machines and humans can co-exist.

7. THE ETHICS OF MACHINE ETHICS

Before finishing this thesis it is important to consider the broader implications surrounding the design, test and implementation procedures that have been outlined. As explored in the introduction to this thesis, AI Ethics pertains to questions concerning the design, use and implementation of Artificial Intelligence, such that we want Artificial Intelligence that is used for good, does not contain harmful implicit discrimination, or causes general harm to the world and humans. Therefore, *The Ethics of Machine Ethics*, needs to consider the broad design, use and implementation of Moral Machines according to the newly specified approach.

The Ethics of AI is even broader as a topic, including discussions on the moral status of Intelligent Agents (i.e. whether they deserve moral consideration – see Liao, 2020), its Sustainability (van Wynesberghe, 2021) and Data Ethics and Privacy (Zwitter, 2014). Furthermore, with AI Ethics relating also to the design of the AI system, we can consider how a moral machine might ethically be designed.

All the areas listed above are relevant to Machine Ethics 2.0 (in the case of a new technology there will undoubtedly be a broad range of ethical implications). This thesis considers that the ethical design procedure surrounding moral machines will be covered by traditional AI Ethics discussions. Therefore, only the AI Ethics topics relating specifically to Machine Ethics and the new design for moral machines are considered in this thesis. This includes *Sustainability*, *Moral Patiency/Status* and *Moral Machine Governance*.

First (7.1) pre-existing issues that have been taken against Moral Machines in the past, as per Section 2.4, will be considered and addressed. As will be shown, pre-existing concerns that have been raised by researchers are not entirely applicable to the theories proposed here. Then (7.2) a general overview of the ethical issues relating to the new breed of moral machine (i.e., the one developed in 6.3) will be given, before there is a more in-depth discussion of the three implications above that specifically relate to moral machines. Finally, (7.3) strategies for dealing with these implications are suggested alongside some initial proposals.

7.1. Pre-existing concerns

A key concern for researchers working within Machine Ethics is that, despite best intentions, in creating a moral machine we will be doing more harm than good. As Winfield (2018b) articulates, the fear is that by creating a machine with the capacity to make moral decisions, we are increasing the chance of individuals making machines that make purely *immoral* decisions by making the job

easier to achieve. The argument follows that the moral capacity we put into a machine *for good* can easily be ‘switched’ *for bad*.

Conceptually, this argument assumes that there is a faculty in the machine that ensures the machine always selects the morally *good* action in a situation. However, if the machine is not being controlled as to what constitutes ‘good’ and ‘bad’ behaviour in the first place, this faculty doesn’t exist.

For example, in the case of Vanderelst and Winfield’s *Simulation* architecture for moral capacity (as detailed in Section 2.2), the machine’s decision to carry out a *good* act over a *bad* one is governed by its ‘Evaluation Module’. Essentially, an action only passes through this module – and goes on to be enacted – if it is one that is assessed as *good*. We can see how the existence of such a module might be problematic since it would be very easy to reverse the engineering in this module so that only *bad* actions were carried out. This could pose a dangerous situation if, for instance, we had a medical robot making life-critical decisions, and the system were maliciously hacked. However, the problem is based upon the assumption that an ethical governance module (i.e., something that permits or restricts behaviour) is the only way to create a moral machine.

In fact, the argument wholly assumes that ‘top-down’ rule-based frameworks for developing moral machines are the only frameworks available. Even AI systems such as *Delphi* would not be susceptible to the malicious switch hacker because their approach does not rely upon an *ethical governance module*.

However, as it was demonstrated with Delphi, this approach *is* susceptible to another kind of ‘hacking’, namely, a poisoning of data that results in the machine believing that certain dubious acts are in fact moral.

The problem with ‘bottom-up’ moral machines, is that their system makes the (false) assumption that because something is common consensus, it is thereby the moral thing to do. If large quantities of inaccurate data is fed into the machine’s training set, it will therefore believe the inaccurate data to be the case. We know through the discussion in this thesis, that morality is a much more sophisticated concept, and as such requires much more careful consideration.

The new moral machine architecture proposed in this thesis is neither susceptible to ‘data poisoning’ nor ‘switch hacking’ that either the ‘top-down’ or ‘bottom-up’ approaches fall foul of, therefore, many of the risks that people associate with moral machines are not relevant. ‘Data poisoning’ is not a problem for the new architecture, because as explained in 6.3. the fact the machine has its own perspective on the world (i.e., its own independent *world model*) means it does not naively take information in. First, the data is objectively analysed before being stored in a data repository, and then is only acted upon if it is logically consistent with the rest of the data

within the repository. Essentially, data that is inconsistent with the machine's representation of the world will be rejected. The machine cannot be influenced by any amount of what might be termed 'poisoned data' – i.e. data giving a discriminatory or polarised perspective.

For example, if we consider the healthcare robot from 6.4 (i.e., a machine motivated towards pro-sociality), it cannot be infected with 'misinformation' because the data it takes in are real-life case studies. Even if the robot were presented with a 'false' case study, this would not impact the machine's learning because of the learning's developmental nature.

It might be argued that motivating the machine towards anti-social behaviour would make it immoral, however, in this case, the proposed design hasn't made it *easier* to trigger an immoral machine because this would be an intentional design feature at the hands of the human designer. Furthermore, where it is easy to design a motivation for pro-social behaviour, such as 'performing a task well', which seems to have objective metrics and criteria, 'bad' behaviour cannot be defined in the same way. Even if a machine were indoctrinated from birth to believe that – say – destroying nature is a good thing to do, this would ultimately be rejected by the machine because it would be inconsistent with other facts it had learnt about the world. Ultimately, the entire world would need to be changed to 'hack' the machine's internal states.

The architecture that has been provided can be described as what Wallach and Allen (2008) refer to as a *Hybrid*. It is 'bottom-up' insofar as the machine forms its perception of reality based upon its investigation and learning from the environment but is 'top-down' because the behaviour is driven by an overarching model (or representation) of *how things should be*. The machine is not susceptible to 'switch-hacking' because the data drives the switch, nor is it susceptible to 'data poisoning' because the principles govern the data. In the developmental architecture described in this thesis the two are inextricably linked.

7.2. Implications of the new approach

A rejection of the previous issues that were outlined concerning moral machines does not mean that there are no ethical questions or challenges that will need to be addressed in the creation of these new systems. As was already detailed in 5.4, one ethical aspect of the new breed of moral machine will be in assessing their moral suitability for various ethically-related use-cases. For instance, a medical robot, with responsibilities akin to a medical practitioner, will need the equivalent degree of moral maturity (or understanding) to ensure that it makes morally mature decisions, and not ones that are self-centred. In this regard the new breed of moral robot brings the need for a whole new governance framework because mechanisms will need to be developed to *test* machine suitability for their given environment. It is a matter of due diligence that tests like this are carried out, to ensure the safety of individuals who will be frequently interacting with or be at the hands of decisions made by such robots. Not only will initial assessments need carrying out to

determine moral suitability for the specified role, but there will need to be post-operational ‘quality assurance’ checks to ensure that the machine maintains its moral quality.

Other governance issues that should be considered include rigorous scientific testing methods for the various required cognitive capacities. Because the machines will ultimately operate in real-life situations (consider, an autonomous vehicle acting in a real-life dilemma), the implications may be severe. In fact, as highlighted by EU regulations a *risk-based governance framework* might be appropriate to assess the level of oversight required for each particular use case. One factor that will require strict governance is the matter of *moral responsibility* and *accountability*. Who would be accountable in the case where an incorrect moral decision is made? The individual responsible for training the agent, the person who designed the moral architecture (!) or the machine itself? In each case, current up-to-date literature and ethical policies in the area should be considered, including accounts on the relational dynamics that might be attributed to moral responsibility (i.e., Coeckelbergh, 2020).

This leads to the ethical issue of Moral Patiency (as per Danaher, 2019). Does the additional moral capacity in machines mean that humans are no longer entirely responsible for their actions? What will be the impact of ‘offloading’ moral decision-making to a machine, and how will this affect society as a whole? These types of impact analyses need to be carried out, with mitigation measures put into place to minimise negative effects. As Coeckelbergh (2022) reflects, technology has always had a relational aspect – to the extent that we shape it, and it shapes us. How could such moral machines shape us and society? Though it seems ill-focused to catastrophise about the psychological and broader social impacts of such a new technology, considering these can be useful and risk averse.

As machines gain an increasing amount of Moral Capacity there then becomes the issue of whether we should ascribe them *Moral Status*. Where we can define Moral Status as *something morally mattering for its own sake* (Jaworska and Tannenbaum, 2021), the question is: *does a moral machine matter for its own sake?* There are various arguments within the AI Ethics literature that suggest that because Artificial Agents are nothing more than *tools* (namely, Bryson and Kime, 2011) that they do not deserve moral status ascription. However, where a tool does not influence our social or moral structures, it is possible that a Moral Machine might. In this case, the lines seem blurred, and it is not obvious that we have something not worthy of rights.

Finally, we need to consider the environmental implications of these new Moral Machines acting in our environment? We would hope that since the new model is non-anthropocentric in nature, that the machines would equally care about the environment as much as they do about humans. Their behaviour should be *moral generally* rather than just *moral towards humans*. However, there could still be questions raised about the impacts of offloading moral responsibility to a machine. As van Wynesberghe (2021) outlines, we must also consider the sustainability issues of offloading our

tasks to Artificial Intelligence, since the processes often required to train AI systems are energy exhaustive, and thereby could contribute towards climate change. If the new systems are going to be a problem for the environment, how can they be developed so they're not?

7.3. Next Steps

The implications in 7.2 centre around three overarching themes: Governance, Moral Status/Patency and Sustainability. To finish this thesis, priorities and strategies for each three themes are proposed.

1. Governance

All aspects of governance pertaining to the new Moral Machines need to be considered. First, before the machines are placed into their operating environment, they need testing for their moral maturity. To do this, appropriate developmental milestones for a machine need to be established on a use-case by use-case basis. As it was already mentioned, Kohlberg's levels might work as a base indicator of moral maturity suitability, but probably the milestone specification criteria need to be specific to the environment the machine will operate within. For example, considering again the care robot from 6.3, we need to provide an analysis for all the moral situations such a machine might come across, and then use this analysis to determine the features of moral decision-making that would be required to carry out the task *morally*. The machine would then be assessed against these features before being put into the morally loaded situation.

Second, regulatory standards need to be developed to assess the rigour of the testing that has been carried out for the new moral machine, with these mapped against the risk associated with each use case. For instance, a *soft* test procedure might be adequate for a machine that will not implicate people's lives (say, a cleaning robot), but strict testing (meeting relevant ISO standards) will be required for any machine that may pose significant risk. Levels of certification or 'Testing Audits' may be necessary to establish the level of rigour that should be placed into assessing moral machine capacity.

Finally, work must continue in the broader realm of AI Ethics concerning accountability and responsibility and then applied to the new Machine Ethics domain. Can we prosecute such a moral machine if it makes an illegal decision? Or do we accept that given the testing procedure it will always do the thing that is right? Knowing the machine is created to *always do the right thing*, how do we respond to law infringements? What should be learned from these machines? These, and related questions, need significant philosophical analysis and reflection.

2. Moral Patiency/Agency

Concerning moral patiency, further investigation needs to be carried out as to the effects an agent acting morally will have to individuals, society, and the world at large. Socially, how will this affect human relationships in general? Will this bring humans closer together, or will it mean that human relationships are severed because they can be easily replaced by a machine? Probably the effects will lie somewhere in the middle, or we will cultivate new kinds of relationships, acquiring new interactions and communication skills. In this case, a new psychology needs to be developed – the Psychology of AI, to determine the underlying effects the new social technology may have.

Of course, further philosophical research into this area is required to determine the true nature of Moral Patiency, and whether we are necessarily at a disadvantage if we lose it.

There is also space for advancing work on moral status ascription – coming to a solid account for what should constitute moral status, and then deciding whether the new moral AI fits into this. This would be building on the work of Gunkel (2012, 2018), except with a Machine Ethics twist.

3. Sustainability

Very finally, strategies need to be put into place to (a) determine the typical energy expenditure a Moral Machine might expend, then (b) compare this to energy expenditure of similar moral decision-making mechanisms, and (c) put in place mitigation measures to prevent there being a detrimental effect on the environment.

Strategies such as the ‘Sustainability Budgets’ outlined in Raper et al, (2022) could be used to manage the amount of energy consumption, and to motivate developers into more conscious moral machine creation through Management Gamification approaches. Once in place, this could be linked to broader governance-based incentivisation.

8. SUMMARY, FUTURE WORK AND CONCLUSION

The aim of this thesis was to provide a systematic dissection of the Machine Ethics problem, and then to use this formulation to develop a methodology that researchers in the area can use together, to begin to create moral machines in a structured and useful way. A universal methodology and approach are required to enhance collaborative discourse in the area and facilitate a unified pursuit to create moral machines.

This thesis has provided a critical assessment of the Machine Ethics problem and informed by a practical case study and philosophical literature, including metaethics and Philosophical Pragmatism, developed Machine Ethics 2.0 – the question of how we can equip machines to develop their own moral outlook.

The reframing of Machine Ethics is then used alongside psychological accounts for morality to develop a practical framework for developing moral machines, that instead of trying to *mimic* morality in humans, suggests we try to *assure* morality in a machine by identifying necessary features that would be required for Moral Agency, and then testing that these features have been implemented through User Acceptability Test Scenarios. As such, a sample of the approach is outlined, with example requirements derived from normative accounts for morality.

Finally, a speculative model and architecture for moral learning, inspired by developmental accounts of morality and utilising a variation of Reinforcement Learning, is developed, which suggests the need for the machine's own independent 'world view' to form moral judgements, and proposes the notion of *raising robots to be good* through a 'guided' reinforcement architecture.

There are obvious areas for future work based upon the ideas that have been developed in this thesis, and these have been articulated throughout the thesis. This includes the work outlined on AI Ethics outlined in the previous chapter, concerning moral status, governance and sustainability, but also work to be continued on developing this research generally. For instance, the methodology is in its infancy, and would significantly benefit from rigorous analysis from other computer scientists, cognitive scientists, philosophers, as well as psychologists and sociologists. One area that could, in particular, use the support of psychologists, is in the testing domain where thorough and comprehensive test cases, that are reflective of latest research in this area, need developing. Once the methodology has been rigorously assessed, and it is of sufficient quality, then practical work in developing the moral machines can begin to take place. In particular it would be interesting to see how the requirements might develop into a more detailed technical specification and architecture, and to test the data repository function against the issues that were described for Tay and Delphi.

The process for developing moral machines is at the very beginning, however what has been demonstrated is that through critical reflection of the machine ethics problem, and a dissection what we want in moral machines, we can begin to trace a methodology for collectively solving this. This will require the involvement of, and be a venture for, researchers across many different disciplines.

To conclude, as machines are embedded more and more in society, there *is* an increasing need for them to be bestowed with moral capacity. However, bestowing a machine with moral capacity should not just be a case of mimicking human morals – where humans are notoriously a flawed species in this domain. A machine needs equipping with the *capacity to develop its own moral outlook* by bestowing it with a suitable architecture for moral learning and growth. We don't need to create machines with morals; we need to *raise robots to be good*.

9. ACKNOWLEDGEMENTS

I would like to thank my supervisors, Nigel Crook and Matthias Rolf for their support and enthusiasm throughout this PhD. Initially poised with the broad task of creating *Kindness in Machines*, they've navigated long discussions, programming sessions and experiments to try to make progress in this field. I would also like to thank Chrisina Jayne who was there at the beginning of my PhD journey, along with other support from staff from across Oxford Brookes University who have facilitated my thesis from a personal perspective. I would like to thank the Ethical AI reading group for engaging in lengthy debates on this topic – with a special shout to Oliver Bridge, Nicola Strong and Phillip Harvey for their 'out of hours' reflections and encouragement. Finally, I want to thank my family, my most enthusiastic supporters, and friends and colleagues who have accompanied me on this journey.

10. REFERENCES

- Anderson, M., Anderson, S.L. and Armen, C., 2004, July. Towards machine ethics. In AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA.
- Anderson, M., Anderson, S.L. and Armen, C., 2006, August. MedEthEx: a prototype medical ethics advisor. In AAAI (pp. 1759-1765).
- Anderson, M. and Anderson, S.L., 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4), pp.15-15.
- Anderson, M. and Anderson, S.L. eds., 2011. *Machine ethics*. Cambridge University Press.
- Anderson, M. and Anderson, S.L., 2018. GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1), pp.337-357.
- Anderson, M., Anderson, S.L., Gounaris, A. and Kosteletos, G., 2021. Towards moral machines: a discussion with michael Anderson and susan Leigh Anderson. *Conatus-Journal of Philosophy*, 6(1), pp.177-202.
- Arnold, T. and Scheutz, M., 2016. Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, 18(2), pp.103-115.
- Asimov, I., 1950. *Runaround. I, robot*. New York: Bantam Dell.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F. and Rahwan, I., 2018. The moral machine experiment. *Nature*, 563(7729), pp.59-64.
- Baghrmian, Maria and J. Adam Carter, "Relativism", *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/relativism/>. [Last Accessed 06/09/2022]
- BBC News, 2022a, 'Google engineer says Lambda AI system may have its own feelings', *BBC News Online*, <https://www.bbc.co.uk/news/technology-61784011> [Last Accessed 06/07/2023]
- BBC News, 2022b, 'Chess robot breaks seven-year-old boy's finger during Moscow Open', *BBC News Online*, www.bbc.co.uk/news/world-europe-62286017 [Last Accessed: 28/09/2022]
- BBC News, 2023, 'AI 'godfather' Geoffrey Hinton Warns of Dangers as he Quits Google', *BBC News Online*, <https://www.bbc.co.uk/news/world-us-canada-65452940> [Last Accessed: 27/06/2023]
- Beauchamp, T.L., 1979. *Principles of Biomedical Ethics*-Paperback-Oxford University Press.

- Bostrom, N., 1998. How long before superintelligence?. *International Journal of Futures Studies*, 2.
- Bentham, Jeremy, 1789 [PML]. *An Introduction to the Principles of Morals and Legislation.*, Oxford: Clarendon Press, 1907.
- Baum, L.F., 1995. *The Wonderful Wizard of Oz*. 1900. *The Wonderful Land of Oz*.
- Bryson, J.J., 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 8, pp.63-74.
- Bryson, J.J. and Kime, P.P., 2011, June. Just an artifact: Why machines are perceived as moral agents. In *Twenty-second international joint conference on artificial intelligence*.
- Boslaugh, Sarah E.. "anthropocentrism". *Encyclopedia Britannica*, 11 Jan. 2016, <<https://www.britannica.com/topic/anthropocentrism>> [Last Accessed 28 June 2022]
- Callaway, E., 2020. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, 588(7837), pp.203-205.
- Carnegiecouncil.org, 2022. "Any Progress in Building Moral Machines?". *Carnegie Council Artificial Intelligence and Equality Podcast*. URL = <<https://www.carnegiecouncil.org/media/series/aiei/any-progress-in-building-moral-machines-with-colin-allen>>. {Last accessed 21/06/2022}
- Christian, B., 2020. *The alignment problem: Machine learning and human values*. WW Norton & Company.
- Coeckelbergh, M. and Mesman, J., 2007. With hope and imagination: imaginative moral decision-making in neonatal intensive care units. *Ethical theory and moral practice*, 10(1), pp.3-21.
- Coeckelbergh, M., 2010. Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12(3), pp.235-241.
- Coeckelbergh, M., 2020. *AI Ethics*. MIT Press.
- Coeckelbergh, M., 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4), pp.2051-2068.
- Coeckelbergh, M., 2022. *SELF-IMPROVEMENT: Technologies of the Soul in the Age of Artificial Intelligence*. Columbia University Press.

Crook, N. and Corneli, J., 2021. The Anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics. *Cognitive Computation and Systems*, 3(2), pp.109-122.

Currie, G. 1995. The moral psychology of fiction. *Australasian Journal of Philosophy*. 73(2), pp. 250-259.

Danaher, J., 2019. The rise of the robots and the crisis of moral patency. *ai & Society*, 34(1), pp.129-136.

Dewey, J., 1908. What does pragmatism mean by practical?. *The journal of philosophy, psychology and scientific methods*, 5(4), pp.85-99.

Doyle, J., 1999. Rational decision making. *MIT encyclopedia of the cognitive sciences*, pp.701-703.

Driver, Julia, "The History of Utilitarianism", *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/>

Duffy, B.R., Rooney, C., O'Hare, G.M. and O'Donoghue, R., 1999, September. What is a social robot?. In *10th Irish Conference on Artificial Intelligence & Cognitive Science*, University College Cork, Ireland, 1-3 September, 1999.

Emelin, D., Bras, R.L., Hwang, J.D., Forbes, M. and Choi, Y., 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

Faria, C. and Paez, E., 2014. Anthropocentrism and speciesism: conceptual and normative issues. *Rev. Bioetica & Derecho*, 32, p.95.

Fine, A., 1993. Fictionalism. *Midwest studies in philosophy*, 18, pp.1-18.

Formosa, P. and Ryan, M., 2021. Making moral machines: why we need artificial moral agents. *AI & SOCIETY*, 36(3), pp.839-851.

Futurism.com, 2021, "Scientists Built an AI to Give Ethical Advice, But It Turned Out Super Racist". <https://futurism.com/delphi-ai-ethics-racist>. Last accessed 21/06/2022.

- Gabriel, I., 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3), pp.411-437.
- Gaita, R., 1989. The personal in ethics. In *Wittgenstein: Attention to particulars* (pp. 124-150). Palgrave Macmillan, London.
- Gaita, R., 2004. *Good and evil: An absolute conception*. Routledge.
- Gibney, E., 2016. Google AI algorithm masters ancient game of Go. *Nature*, 529(7587), pp.445-446.
- Graham, George, "Behaviorism", *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2019/entries/behaviorism/> [Last Accessed 02/08/2022]
- Gunkel, D.J., 2007. Thinking otherwise: Ethics, technology and other subjects. *Ethics and Information Technology*, 9(3), pp.165-177.
- Gunkel, D., 2012. The machine question. *Critical perspectives on ai, robots, and ethics*, p.5.
- Gunkel, D.J., 2018. The other question: can and should robots have rights?. *Ethics and Information Technology*, 20(2), pp.87-99.
- Gunkel, D.J., 2020. A vindication of the rights of machines. In *Machine Ethics and Robot Ethics* (pp. 511-530). Routledge.
- Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4), p.814.
- Hassabis, D., 2017. Artificial intelligence: chess match of the century. *Nature*, 544(7651), pp.413-414.
- Himmelreich, J., 2018. Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, 21(3), pp.669-684.
- Hesslow, G., 2012. The current status of the simulation theory of cognition. *Brain research*, 1428, pp.71-79.

Hursthouse, Rosalind and Glen Pettigrove, "Virtue Ethics", The Stanford Encyclopedia of Philosophy (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/>.

Ichikawa, Jonathan Jenkins and Matthias Steup, "The Analysis of Knowledge", The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/>.

Ilková, V. and Ilka, A., 2017, June. Legal aspects of autonomous vehicles—an overview. In 2017 21st international conference on process control (PC) (pp. 428-433). IEEE.

James, W., 1909. Pragmatism and four essays from the meaning of truth. World Publishing Company.

Jaworska, Agnieszka and Julie Tannenbaum, "The Grounds of Moral Status", The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/>. [Last Accessed: 13/08/2022]

Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M. and Choi, Y., 2021. Delphi: Towards machine ethics and norms. arXiv preprint arXiv:2110.07574.

Johnson, Robert and Adam Cureton, "Kant's Moral Philosophy", The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/kant-moral/>.

Keeling, G., 2020. Why trolley problems matter for the ethics of automated vehicles. Science and engineering ethics, 26(1), pp.293-307.

Kenward, B. and Sinclair, T., 2021. Machine morality, moral progress, and the looming environmental disaster. Cognitive Computation and Systems.

Kohlberg, L., 1971. Stages of moral development as a basis for moral education. Moral education: Interdisciplinary approaches, pp.23-92.

Kraut, Richard, "Aristotle's Ethics", The Stanford Encyclopedia of Philosophy (Summer 2022 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2022/entries/aristotle-ethics/>.

Legg, Catherine and Hookway, Christopher, "Pragmatism", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/pragmatism/>. [Last accessed 31/07/2022]

Liao, S.M., 2020. The moral status and rights of artificial intelligence. *Ethics of Artificial Intelligence*, pp.480-503.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z. and Tang, J., 2021. GPT understands, too. arXiv preprint arXiv:2103.10385.

Lungarella, M., Metta, G., Pfeifer, R. and Sandini, G., 2003. Developmental robotics: a survey. *Connection science*, 15(4), pp.151-190.

Marcus, G. and Davis, E., 2019. *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Marcus, G., 2020. The next decade in ai: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E. and Morimoto, J., 2022. Deep learning, reinforcement learning, and world models. *Neural Networks*.

McDonald, F.J., 2022. AI, alignment, and the categorical imperative. *AI and Ethics*, pp.1-8.

McMahan, J., 2013. Moral intuition. *Blackwell guide to ethical theory*, pp.103-121.

Mohri, M., Rostamizadeh, A. and Talwalkar, A., 2018. *Foundations of machine learning*. MIT press.

Moor, J.H., 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4), pp.18-21.

Noonan, Harold and Ben Curtis, "Identity", The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/fall2022/entries/identity/>. [Last accessed 06/08/2022]

James, W., 1909. Pragmatism and four essays from the meaning of truth. World Publishing Company.

Peirce, C. S. 1878, 'How to Make our Ideas Clear'.

Peirce, C.S., 1905. What pragmatism is. *The monist*, 15(2), pp.161-181.

Petersen, S., 2007. The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(1), pp.43-54.

Pinker, S., 2017. The moral instinct. In *Understanding Moral Sentiments* (pp. 59-80). Routledge.

Poulsen, A., Anderson, M., Anderson, S.L., Byford, B., Fossa, F., Neely, E.L., Rosas, A. and Winfield, A., 2019. Responses to a critique of artificial moral agents. arXiv preprint arXiv:1903.07021.

Raper, R., Boeddinghaus, J., Coeckelbergh, M., Gross, W., Campigotto, P. and Lincoln, C.N., 2022. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. *Sustainability*, 14(7), p.4019.

Rawls, J., 1951. Outline of a decision procedure for ethics. *The philosophical review*, 60(2), pp.177-197.

Reyna, Y.C.F., Cáceres, A.P., Jiménez, Y.M. and Reyes, Y.T., 2019. An improvement of reinforcement learning approach for permutation of flow-shop scheduling problems. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E18), pp.257-270.

Rorty, R., 1982. *Consequences of pragmatism: Essays, 1972-1980*. U of Minnesota Press.

Rorty, R., 1990. *Objectivity, Relativism, and Truth: Volume 1: Philosophical Papers*. Cambridge University Press.

Roselli, D., Matthews, J. and Talagala, N., 2019, May. Managing bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 539-544).

Sayre-McCord, Geoff, "Moral Realism", *The Stanford Encyclopedia of Philosophy* (2015), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/entries/moral-realism/>>. [Last accessed: 06/07/2023]

- Sharkey, A., 2017. Can robots be responsible moral agents? And why should we care?. *Connection Science*, 29(3), pp.210-216.
- Shelley, M., 2001. *Frankenstein* [1818]. New York: Oxford.
- Singer, P., 1973. Animal liberation. In *Animal rights* (pp. 7-18). Palgrave Macmillan, London.
- Singer, P., 2009. Speciesism and moral status. *Metaphilosophy*, 40(3-4), pp.567-581.
- Singer, P. and Tse, Y.F., 2022. AI ethics: the case for including animals. *AI and Ethics*, pp.1-13.
- Sinnott-Armstrong, Walter, "Consequentialism", *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2021/entries/consequentialism/>.
- Sparrow, R., 2021. Why machines cannot be moral. *AI & SOCIETY*, 36(3), pp.685-693.
- Sutton, R.S. and Barto, A.G., 2018. *Reinforcement learning: An introduction*. MIT press.
- Talat, Z., Blix, H., Valvoda, J., Ganesh, M.I., Cotterell, R. and Williams, A., 2021. A word on machine ethics: A response to Jiang et al.(2021). *arXiv preprint arXiv:2111.04158*.
- Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M. and Bernstein, A., 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6), pp.1-38.
- Turing, A.M., 1950. *Mind*. *Mind*, 59(236), pp.433-460.
- Van Den Hoven, J. and Lokhorst, G.J., 2002. Deontic Logic and Computer-Supported Computer Ethics. *Metaphilosophy*, 33(3), pp.376-386.
- van Roojen, Mark, "Moral Cognitivism vs. Non-Cognitivism", *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/moral-cognitivism/>. [Last accessed: 06/07/2023]
- van Wynsberghe, A. and Robbins, S., 2019. Critiquing the reasons for making artificial moral agents. *Science and engineering ethics*, 25(3), pp.719-735.

- van Wynsberghe, A., 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3), pp.213-218.
- Vanderelst, D. and Winfield, A., 2018a. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48, pp.56-66.
- Vanderelst, D. and Winfield, A., 2018b, December. The dark side of ethical robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 317-322).
- Véliz, C., 2021. *Privacy is power*. Melville House.
- Wallach, W. and Allen, C., 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.
- Wallach, W., Franklin, S. and Allen, C., 2010. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in cognitive science*, 2(3), pp.454-485.
- Wooldridge, M., 2020. *The road to conscious machines: The story of AI*. Penguin UK.
- Yablo, S., 1998, June. Does Ontology Rest on a Mistake?: Stephen Yablo. In *Aristotelian Society Supplementary Volume* (Vol. 72, No. 1, pp. 229-262). University College London: The Aristotelian Society.
- Zwitter, A., 2014. Big data ethics. *Big Data & Society*, 1(2), p.2053951714559253.

11. APPENDICES

11.1. Related Publications

On the proceeding pages please find papers that were contributed towards and published throughout the course of my PhD. They are listed below and feature in chronological order:

Nugent, S., Jackson, P., Scott-Parker, S., Partridge, J., Raper, R., Bakalis, C., Shepherd, A., Mitra, A., Long, J., Maynard, K. and Crook, N., 2020. Recruitment AI has a Disability Problem: questions employers should be asking to ensure fairness in recruitment.

Crook, N., Nugent, S., Rolf, M., Baimel, A. and Raper, R., 2021. Computing morality: Synthetic ethical decision making and behaviour. *Cognitive Computation and Systems*, 3(2), pp.79-82.

Bridge, O., Raper, R., Strong, N. and Nugent, S.E., 2021. Modelling a socialised chatbot using trust development in children: lessons learnt from Tay. *Cognitive Computation and Systems*, 3(2), pp.100-108.

Raper, R., Boeddinghaus, J., Coeckelbergh, M., Gross, W., Campigotto, P. and Lincoln, C.N., 2022. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. *Sustainability*, 14(7), p.4019.

Raper, R., Coeckelbergh, M. (forthcoming) 'EKIP: designing a practical framework for embedding ethics in AI software development'. *Proceedings of Ethicomp 2022*.

Raper, R. and Coeckelbergh, M., *Imagination, Morals and Machines: The Role of Imagination in Moral Decision-Making and its Implications for the Project of Creating Moral Machines*. Available at SSRN 4457898.

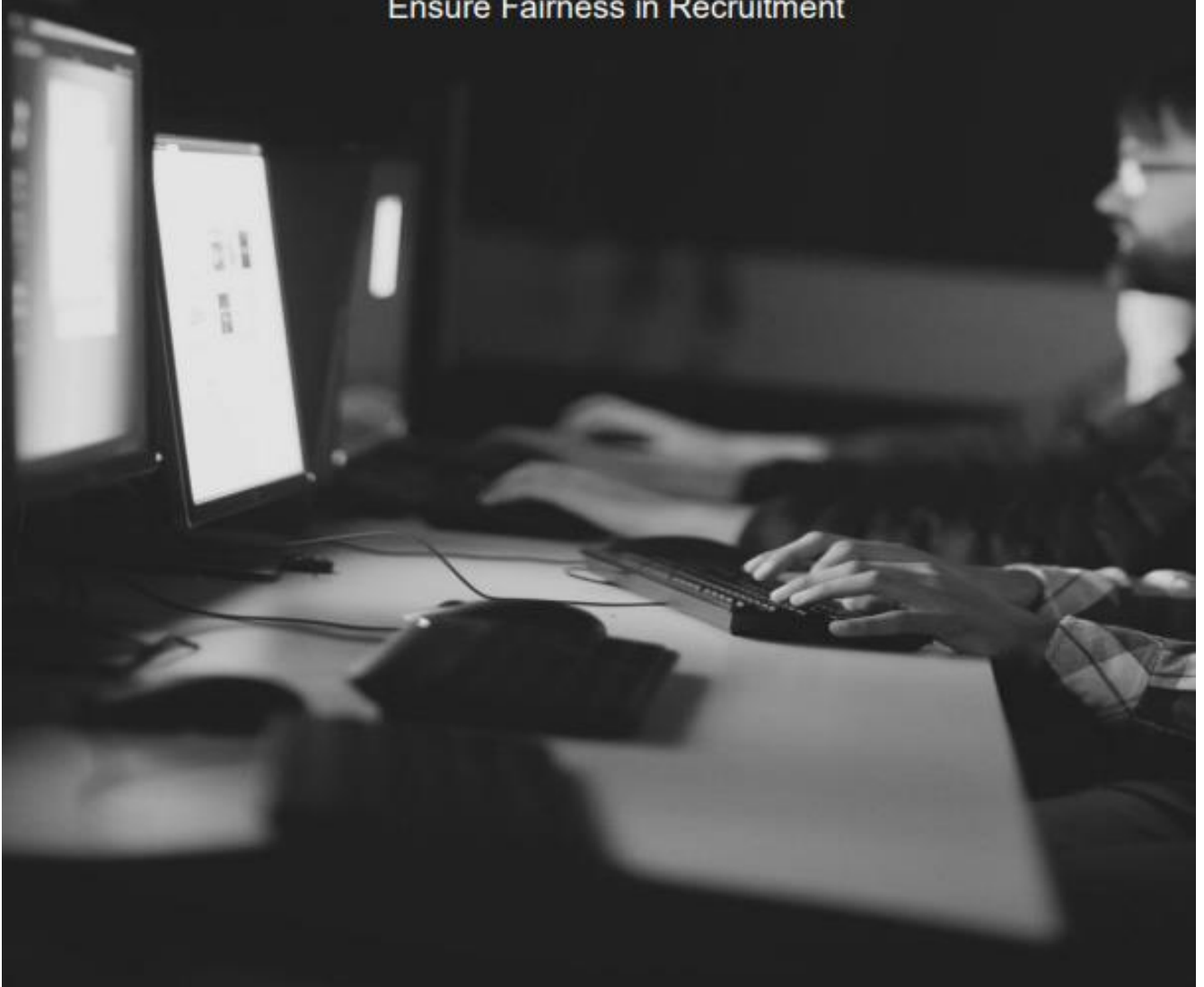


The Institute for Ethical AI

2020 DISABILITY AND AI WHITEPAPER

Recruitment AI has a Disability Problem

Questions Employers Should be Asking to
Ensure Fairness in Recruitment



AUTHORS AND CONTRIBUTORS

Chara Bakalis, School of Law, Oxford Brookes University

Nigel Crook, Institute for Ethical Artificial Intelligence, Oxford Brookes University

Paul Jackson, Oxford Brookes Business School, Oxford Brookes University

Kevin Maynard, Institute for Ethical Artificial Intelligence, Oxford Brookes University

Arijit Mitra, Institute for Ethical Artificial Intelligence, Oxford Brookes University

Selin Nugent, Institute for Ethical Artificial Intelligence, Oxford Brookes University

Jintao Long, Institute for Ethical Artificial Intelligence, Oxford Brookes University

Susan Scott-Parker, Business Disability International

James Partridge, Face Equality International

Rebecca Raper, Institute for Ethical Artificial Intelligence, Oxford Brookes University

Alex Shepherd, Institute for Ethical Artificial Intelligence, Oxford Brookes University

SUMMER 2020

This paper is intended to be a guide and living document that will evolve and improve with input from readers and relevant stakeholders. Your feedback is welcome and encouraged. Please share your feedback with us at ethicalAI@brookes.ac.uk.

Cite as:

Nugent, S., Jackson, P., Scott-Parker, S., Partridge, J., Raper, R., Shepherd, A., Bakalis, C., Mitra, A., Long, J., Maynard, K., and Crook, N. (2020). Recruitment AI has a Disability Problem: questions employers should be asking to ensure fairness in recruitment. Institute for Ethical Artificial Intelligence

This work is licensed under the terms of the Creative Commons Attribution License 4.0



OXFORD
BROOKES
UNIVERSITY



I Table of Contents

About the Institute for Ethical Artificial Intelligence	3
Disability and Employment Discrimination	5
Recruitment AI	7
Exclusion by Design and Discriminatory Use	7
Biased Systems	7
Improper Implementation and Use	9
Tech on the Market: the dangers of discrimination	9
CRM and ATS	10
Conversational Agents	10
CV/ Resume Screeners	10
Pre-Employment Assessments	11
AI Interviewing	12
Intervention Recommendations	13
References	15
Resources	17
Glossary	17

| About the Institute for Ethical Artificial Intelligence

Oxford Brookes University hosts a vibrant and ambitious research environment in the areas of artificial intelligence, computing, and data science. Founding the Institute for Ethical Artificial Intelligence was therefore a natural extension of our vision as a research community to advance knowledge and promote the better understanding of technology and its relationship to business and society in our local and the wider global community. Our mission at the Institute for Ethical Artificial Intelligence is to promote and support the development and deployment of ethical and trustworthy intelligent software solutions for business, organisations, and society.

Our primary focus at the Institute for Ethical Artificial Intelligence is to help organisations working in the professional services to understand and plan for the risks and opportunities that AI and data analysis technologies can bring to their organisation, their stakeholders and society at large. Working with both the users and the providers of AI technology, as well as developing bespoke AI solutions, we research and advise on the ethical impact of AI technology on organisations and individuals.

In order to achieve this, we bring together a diverse group of world-leading experts who together blend knowledge and skills from technology, business, social science and the life sciences. We deliver expertise and independent guidance in areas that include AI and machine learning, disability, psychology, business development, equality and diversity, coaching and mentoring, digital health, and wellbeing.

For more information, please visit our webpage ethical-ai.ac.uk

| Introduction

The purpose of this White Paper is to

- Detail the impacts to and concerns of disabled employment seekers using AI systems for recruitment, *and*
- Provide employers with the knowledge and evaluation tools to ensure innovation in recruitment is also fair to all users.

In doing so, we further the point that ***making systems fairer for disabled employment seekers ensures systems are fairer for all.***

...

Artificial Intelligence (AI) and similar advanced data analytics systems are increasingly sought-after tools for recruitment used to automate time-consuming, repetitive operational tasks, and expand strategic potential. However, as engineering of these systems becomes more complex, it is more difficult for organisations to confidently assess whether the technology is functioning in line with their expectations and if employment seekers will be treated fairly.

AI technologies have the potential to dramatically impact the lives and life chances of people with disabilities seeking employment and throughout their career progression. While these systems are marketed as highly capable and objective tools for decision making, a growing body of research demonstrates a record of inaccurate results as well as inherent disadvantages for women and people of colour (Broussard, 2018; Noble, 2018; O'Neil 2017). Assessment of disability fairness in Recruitment AI has thus far received little attention or been overlooked (see Guo et al., 2019; Petrick, 2015; Trewin, 2018; Trewin et al. 2019; Whittaker et al., 2019).

Presently, a landscape of limited regulation, paired with increasing societal pressure for AI and data analytics systems to be designed with fairness, transparency, and validity, means that organisations face financial, legal, reputational, operational, and ethical risks for implementing them. While there is already much work being done to address the high-level concerns related to artificial intelligence, bias, and fairness, there will inevitably be more challenges ahead that no one company or industry can solve alone. In order to minimise these risks, businesses, human and disability rights campaigners, and academic experts need to collaborate to develop new ways to analyse, validate, and improve these systems and to hold technology developers and suppliers accountable.

Our aim in this paper is to provide a starter toolkit to evaluate organisational and ethical values in relation to the use of recruitment technology, and with regard to vitally important procurement processes. We review the broad technological developments that support recruitment, demonstrate their potential to impact disabled employment seekers in various ways. We then present recommendations

for the questions employers should be asking before taking on new technologies and when evaluating currently used systems.

The Institute for Ethical Artificial Intelligence and its partners invite public, third sector and private sector stakeholders to respond to this guidance and to continue discussion toward ensuring fairer recruitment practices for persons with disabilities, and other disadvantaged employment seekers more generally.

I Disability and Employment Discrimination

People with disability have historically and continue to be regularly disadvantaged in seeking and securing employment. Disabled people experience widespread economic and societal exclusion and are more than twice as likely to be unemployed as others (Office of National Statistics, 2019). The sheer scale of the social and economic impacts of the COVID pandemic on employment and employability will undoubtedly further disenfranchise people with disabilities. The current climate of instability makes ensuring fair and equal treatment all the more important, given increasing employment among people with disabilities helps raise people out of poverty, improve their life chances, and is a net cultural and economic benefit.

As defined by the United Nations Convention on the Rights of Persons with Disabilities (CRPD), “persons with disabilities include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others.”

The definition of disability doesn't necessarily capture the complexity and heterogeneity of people with disabilities, which is a key factor in the complications with AI systems. A disability may be a life-long condition or occur at different life stages or be the result of a major event/change. Disability can have wide-ranging life impacts or be context dependent. Disability may be visible, but most are invisible. Disabilities may include people with hearing, sight and mobility, and dexterity impairments, people with cognitive and intellectual impairments, those with mental health conditions, those with facial disfigurements, those of small stature, and numerous others. Further, individuals may have a combination of multiple factors.

Disability also intersects with other aspects of identity, such as gender, ethnicity, sexuality, an socioeconomic background. Disability is not completely independent of other features of a person's identity and life experience (Collins and Bilge 2020; Parker, 2015; Samuels, 2016). Moreover, the social stigmas attached to disability are intersectional, shared, and amplified with other marginalised identities (Frederick and Shifrer, 2019). In light of the ongoing Black Lives Matter protests against racial violence and injustice, our focus on disability is intended to contribute to a wider discussion of systemic and persistent oppression of marginalized peoples. Recognising and celebrating human diversity is a necessary starting point to design AI systems that fairly and equitably engage with human reality.

...

Disability inclusion in the workplace is impacted by number of factors. There is often a qualifications gap between disabled and non-disabled people due to systematic disadvantages in education, training, and previous work experience (Sayce, 2011). Even well intentioned employers may struggle to recognize how structural barriers to success impact

Some industries or categories of position lack accessibility that can limit employment for people with certain impairments. There are inadequate programmes to support persons with disabilities and those who employment them. Employers may also have negative attitudes/bias and lack confidence or training to support disabled employment seekers (Lindsay et al., 2020; Suter et. al. 2007).

Some of Global Disability Facts

There are more than 1.3 billion people with disabilities worldwide and the number is growing with an aging population and advances in medical science. *WHO*

...

80% of this 1.3 billion people live in the developing world. *WHO*

...

15-18% of any country's population will have a disability and/or chronic health condition *WHO*

...

Circa 80% of people with disabilities have impairments that are not immediately visible. *Check*

...

1 in 5 women will have a disability *UN*

...

1 in 3 people aged 50-64 will have a disability regardless of their ethnicity *Check*

...

People who live to the age of 70 are likely to have at least 10 years of lived experience of disability *Check*

...

In any large organisation, from 10-12% of the workforce are like to have a disability and/or chronic health condition *UK Labour Force Survey*

...

At least 1 in 3 consumers will either be disabled or will have someone who has a disability in their immediate circle *European Commission*

I Recruitment AI

As organisations increase in scale and receive larger volumes of job applicants, they are under pressure to balance often competing interests in recruiting and retaining the talented candidates, optimising workflow efficiency and productivity, and managing costs. This means that employers are increasingly turning to automated tools to support the employee's journey from recruitment to retirement.

Artificial Intelligence (AI) has featured prominently in these developments. AI is a subfield of computer science, focused on training computers to perform traditionally human tasks. For additional reference, a glossary of relevant AI terms is provided at the end of this document.

AI systems are currently available across a wide range of recruitment functions, including:

- Candidate Sourcing / Engagement
 - Candidate Tracking
 - CV/ Resume Screening
 - Pre-Employment Assessments
 - AI Interviewing

We will discuss each of these categories of technology in relation to their potential to impact people with disabilities in greater detail below.

The unifying objective for systems operating across these diverse recruitment functions is that they are designed to distil the vast array of information about applicants down to a few select predictable features for the purpose of making quantifiable and easily comparable decisions. However, when systems need to cope with the reality of human diversity, whether it pertains to disability, ethnicity, gender, and/or other features, they often interpret complexity as an abnormality, or outlier. In this case predictability may come at the expense of the life chances of disabled people who are already faced with systematic disadvantages and unfair discrimination in securing employment.

| Exclusion by Design and Discriminatory Use

Recruitment AI may inadvertently adversely impact employment seekers with disabilities via two major routes: biased systems and discriminatory processes.

Biased Systems

The design of an AI system involves first specifying an objective and then specifying how the system achieves and optimizes achieving that objective. Humans are often not skilled at specifying objectives. If an objective is not specified appropriately, the outcome may have unintended consequences.

Unwanted biases, or biases that treat some people negatively, or adversely due to protected characteristics or other features of their identity, raise serious risks of discrimination. It is critical to identify and mitigate these potentially harmful biases. And to prevent and mitigate bias, it is necessary to understand how humans introduce biases into an AI system.

Developing this knowledge begins with defining what biases exist within a system *and* where they exist, or have a potential to exist. Disability-related biases in AI systems are heavily influenced by historical hiring decisions. Since people with disabilities are twice as unlikely to be unemployed, they are simply less likely to be represented in data on past successful employees. These biases may be introduced into systems through two primary mediums: the *algorithmic model* and the *training data*.

...

The algorithmic model is the mathematical process by which an AI system performs a certain function. Designing this model involves defining the objective or problem the developer wishes to address and selecting the parameters that define the system's operation at what they determine is an optimal level (Russell and Norvig, 2003).

How can this go wrong? For instance, an automated CV screener is programmed to predict the best qualified candidate based on the ("optimal") parameter of having attended a top-tier university. Someone who has worked hard to achieve success, right? The prestige of an institution may be one factor in a successful employee, but that parameter also disadvantages people with disabilities, different socioeconomic backgrounds, and/or underrepresented ethnicities, who already face systemic barriers to be equally represented in prestigious institutions.

The training data is an initial set of data used to help a program learn how to apply the model and produce sophisticated results in application (Russell and Norvig, 2003). The model operates as well as the training data that goes in. The sampling strategy used to collect the training data and the representativeness of the data is a conscious decision by the developer.

Building on the previous example, what if the automated CV screener was trained on data that **did not** include the data profiles of successful employees who have a non-English name, went to state school, participate in disability-related volunteering activities, had a break in employment due to family or illness, or have an address in an economically disadvantaged area? These are simple, seemingly innocuous features that will be represented in a CV. Interacting with information in a CV that the programme has not previously encountered means that the system may be more likely to reject a candidate. This is because these novel features do not fit

the prescribed collection of features that is modelled to represent the 'ideal' employee. These novel features may be innocuous, but they may also be indirectly related to the experience of being disadvantaged on the job market.

Improper Implementation and Use

Even as systems become more technically sound with regard to acknowledging and mitigating bias in design, risks for applicants with disability may be generated and/or amplified by improper use and implementation of the technology.

Most recruiters recognize that no single assessment method is suitable and fair for all applicants. However, the marketed reliability and the ease of automated adaptations of recruitment processes has resulted in many cases where AI tools are being used in isolation of other measures of suitability and human decision makers in the application package. In some organisations, a single product may be the sole gate of entry into employment.

Moreover, AI assessment fails to factor in the likelihood that the employer would make the adjustment post job hire that would *determine* if a particular disabled candidate was 'right' for the job. For example, a qualified, visually impaired, cybersecurity expert will only be the best candidate *if* the employer enables her to use specialized software.

Acknowledging and monitoring uncertainty in AI systems is critical to making fair and adequate decisions as sensitive and life changing as whether a person is employed or not. The life chances of job seekers precariously intersect with the computational complexities related to disability, the inherent challenges of bias, and the uncertainty around automated decision-making. No system should be expected to work perfectly.

...

The use of rigid, standardised recruitment processes that cannot be adequately adjusted to enable candidates with disabilities to compete fairly are inherently discriminatory (Hamraie, 2017). Candidates may have the option to request accommodations to these systems – although some developers expect this is the role of the employer to deliver such adjustments. However, unless candidates are given explicit assurances that they may request and be provided with equally-evaluated, alternative routes, the employer risks, at best, making disabled users uncomfortable/fearful of interacting with AI and, at worst, discriminating against such candidates. Expecting disabled employment seekers to go through standardised processes is akin to asking a wheelchair user to take the stairs to the interview room.

| Tech on the Market: the dangers of discrimination

Recruitment AI encompasses a wide array of technologies functioning at different points in the recruitment process. This section outlines the broad categories currently in use, detailing the impact potential for people with disabilities. This list is by no means exhaustive, but highlights major technologies used in the candidate sourcing and selection phases of recruitment.

ATS and CRM Systems

Applicant Tracking Systems (ATS) are platforms where recruiters can conduct each step in the hiring process from posting position openings to collecting applications to screening candidates to evaluation and selection. Candidate Relationship Management (CRM) systems maintain a connection between recruiters and employment seekers so that desirable candidates may be easily referred to future job openings.

We consider these systems together because share similar potential impacts on people with disabilities. They are likely to utilise automated outlier detection tools, such as CAPTCHAs, that when insufficiently trained can flag people with disabilities as not human, or a spammer (Guo et al. 2019). The difference between human and non-human may come down to a few seconds delay in response, a minor slip in highlighting the correct answer, or misinterpreting an obscured set of letters. People with difficulties related to dexterity or visual impairment are disproportionately affected.

Further, the skills and qualification gap for disabled people due to systemic inequalities likely disadvantages these candidates when evaluated against the standard person specification as well as historic hiring decisions. These systems are not designed with the flexibility that would take into account that some candidates appear less qualified only due to systemic denial of education and employment opportunities.

CV/ Resume Screeners

CV screening is a major driver of the recruitment innovation powered by AI systems, addressing the need for processing high application volumes. Automated screeners detect characteristics in the CV content, such as key phrases, proper nouns to evaluate employability against criteria for the position. These criteria are determined by either the job description or by evaluating the features of previously successful candidates. They may go further to interpret characteristics of the applicant, such as personality, sentiment, and demographics. Some also supplement data in CVs to with information about the candidate from public data sources, social media, and information about their previous employers.

Once again, the skills and qualification gap for disabled people due to systemic inequalities is likely to disadvantage these candidates when evaluated against a standard job description as well as historic hiring decisions. These systems

are not designed with flexibility that considers some appear less qualified due to systemic lack and denial of education and employment opportunities.

AI screener systems that have not been trained on CV data from users with diverse cognitive and intellectual abilities may have additional challenges with linguistic flexibility. For screeners that analyse personality and emotion from texts, further problems may arise. For example, people with neuro- and cognitive diversity may express emotion in writing in a style previously not encountered by the AI system, resulting in incorrect classifications about their emotional state or personality. . And many pre-lingually Deaf individuals speak the official spoken language of their country as a second language.

Conversational Agents

Recruitment conversational agents, or chatbots, are designed to mimic human conversational abilities during the recruitment process. These technologies use an approach termed natural language processing (NLP) to analyse questions and comments and to respond effectively. Conversational agents are desirable additions to the recruitment process as a means of increasing communication with employment seekers in order to answer frequently asked questions, collect information on candidates, ask screening questions, and schedule interviews or meetings with a human recruiter.

Conversational agent systems have the potential to be helpful in some circumstances where they are designed with accessibility in mind. Agents that augment text with visual illustration (i.e. highlight key words, spelling and grammar check, text suggestion), speech functionality, and dictation tools can enhance accessibility and usability for a wide range of users.

However, if not thoughtfully designed and implemented, conversational agents may also not respond appropriately, or in a hateful manner, and unfairly screen out candidates. Depending on the nature of the agent's function this can at best lead to poor user experience and at worst discriminatory candidate screening.

Conversational agents are often not trained on language data gathered from people with cognitive, intellectual, physical and linguistic diversity or those from neuro-diversity groups. Undertrained agents may be unable to correctly interpret spellings or phrases they haven't previously encountered, such as messages from people who have physical difficulty typing or have dyslexia, autism, dysphagia, dyspraxia, ADHD, among numerous others. Moreover, agents that do not support communications methods beyond writing, such as text-to-speech and dictation, limit or exclude many individuals from participating in communication and being competitive in the recruiting process.

Pre-Employment Assessments

A range of candidate aptitude assessments, such as cognitive ability, technical skills, personality, and decision making, are commonly used to quantitatively measure and compare job applicants for a particular role. Broadly, these tests are aimed at gauging a candidate's ability to think quickly, solve problems, and interpret data.

Many recruiters recognise that these assessments are often not reliable as one-size-fits-all approaches. The generalisability of psychometric tests for people with disabilities—as well as many populations who are not from WEIRD (western, educated, industrialized, rich, and democratic)—backgrounds is unreliable (Cook and Beckman, 2006). There is a degree of uncertainty about whether any assessed candidates, never mind those with disabilities, are indeed able to successfully learn and perform the duties of the role or not. Furthermore, many psychometric tests are in themselves inaccessible to a wide range of disabled candidates. These assessments must be balanced by other measures in the recruitment process.

Gamified assessments raise additional concerns related to dexterity, vision impairment, and response time. Games often involve tasks that are assessed based on speed of reaction to prompts and precision of responses, which may affect people with motor limitations, who need extra time or assistance to complete dexterity tasks. People with visual impairment may require magnification and colour adjustment and additional time. Furthermore, people with cognitive diversity may require language adjustment and additional time to read prompts.

AI Interviewing

AI powered interviewing includes facial analysis tools and speaking conversational agents—aka robot recruiters (refer above to limitations of [Chatbots](#)). These tools evaluate employability from the language, tone, and facial expressions of candidates when they are asked an identical set of question in a standardised process. Candidates are assessed based on a variety of facial, linguistic, and non-verbal measures. 'Ideal' measures often are those that most closely align with the same measures from historically successful candidates for any given role.

As with previous examples, systems that are not trained on a diverse range of potentially successful candidates, face challenges in fairly assessing people with facial features, expressions, voice tone, and non-verbal communication that it has not previously encountered.

For instance, facial analysis software may inaccurately assess and potentially exclude people with facial disfigurement or paralysis as well as conditions such as Down syndrome, achondroplasia, cleft lip/palate, or other conditions that result in facial differences. Further, people with blindness may not face the camera or make eye contact in a manner acceptable to the system's parameters. Moreover, issues may be exacerbated by differences in eye anatomy and dark glasses. People who need captions due to hearing loss, or who lip read may struggle to hear or interpret the questions.

Facial analysis tools that go further to interpret emotion and personality from facial expressions pose alarmingly high risks. Beyond issues of accuracy and

algorithmic bias, the fundamental scientific concepts behind personality assessments derived from facial feature measurements, is not supported -and is rooted in pseudoscientific race studies (Noble, 2018). The implementation of these technologies for recruitment risks legitimising the flawed methodological premise with ill-informed buyers in a way that can only perpetuate historic disadvantages and exclusion for marginalised peoples.

| Intervention Recommendations

Designing and implementing Recruitment AI systems that treat persons with disabilities and by that extent, **all** employment seekers fairly requires the engagement of all stakeholders—technology suppliers, purchasers, and users alike. Our aim is to facilitate purchasers in joining the discussion and to collaborate with us to prepare the tools and language needed to initiate the conversation that asks: How do we assess if any given Recruitment AI system is 'safe' for employment seekers with disabilities and others disadvantaged in any labour market?.

There are a number of actions a forward-thinking organisation can take to support those technology suppliers who share the values and expectations of the organisation and its clients toward applicants with disability. This process begins by asking the right questions of technology developers and suppliers.

Vision, Strategy & Corporate Governance Stakeholders

- i. Does this technology align with our organizational strategy to increase diversity and representation?
- ii. Does use of this technology reflect our organisation's strategic policies with regard to the ethical and responsible development and implementation of artificial intelligence?
- iii. Is this supplier actively engaged in learning more about how to adapt to match our values and needs as a business and those of our stakeholders?
- iv. Who in this organisation should be involved in the governance process which determines how we investigate, procure, apply and



monitor HR tech systems so that at the very least they do not adversely impact disadvantaged job seekers?

Human Resources and Operations Stakeholders

- i. What are the benefits and risks of this technology for disabled and other disadvantaged employment seekers?
- ii. Was a shared understanding of inclusivity and fairness—with specific reference to eliminating the root causes of disability related discrimination—designed into this technology?
- iii. Will implementing this technology require alternative evaluation routes to enable people with different impairments to be recruited on the basis of individual capability and potential?
- iv. Does the AI recruitment tool enable candidates to readily request adjustments, in a non-stigmatising manner, at every stage of the process



Procurement Stakeholders

- i. Has this supplier proved their products are safe for disabled and other disadvantaged employment seekers before you purchase?
- ii. How has the supplier actively involved people with disabilities to test and validate its products?
- iii. Was a shared understanding of inclusivity and fairness—with specific reference to eliminating the root causes of disability related discrimination—designed into this technology?



- iv. Do contractually defined performance standards require the supplier to track the experience of job seekers with disabilities – particularly those who have requested disability related adjustments?
- v. Can they evidence that they have actively consulted and involved persons with disabilities as expert advisors and potential users in their product development life cycle?

Information Technology Stakeholders

- i. Will our organisation be provided with the appropriate explainability and interpretability resources to assess outputs and impacts on employment seekers' disabilities?
- ii. Does the relevant, quality data exist to support this technology in performing effectively for persons with disabilities?
- iii. What are the appropriate oversight mechanisms to evaluate the performance of the system and can the system withstand scrutiny by disabled employment seekers?
- iv. Can the supplier demonstrate how the processes will adapt so as to ensure equal opportunities for disabled employment seekers?



| References

- Broussard, Meredith. Artificial Unintelligence: How Computers Misunderstand the World. Cambridge: MIT Press, 2018.
- Collins, P.H. and Bilge, S., 2020. *Intersectionality*. John Wiley & Sons.
- Cook, D.A. and Beckman, T.J., 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, 119(2), pp.166-e7.
- Frederick, A. and Shifrer, D., 2019. Race and disability: From analogy to intersectionality. *Sociology of Race and Ethnicity*, 5(2), pp.200-214.

- Guo, A., Kamar, E., Vaughan, J.W., Wallach, H. and Morris, M.R., 2019. Toward Fairness in AI for People with Disabilities: A Research Roadmap. *arXiv preprint arXiv:1907.02227*.
- Hamraie, Aimi. 2017. *Building Access: Universal Design and the Politics of Disability*. Minneapolis: University of Minnesota Press.
- Lindsay, S., Leck, J., Shen, W., Cagliostro, E. and Stinson, J., 2019. A framework for developing employer's disability confidence. *Equality, Diversity and Inclusion: An International Journal*.
- Parker, Alison M. 2015 "Intersecting Histories of Gender, Race, and Disability." *Journal of Women's History* 27, 1.
- Petrick, Elizabeth R. 2015. *Making Computers Accessible: Disability Rights and Digital Technology*. Baltimore: Johns Hopkins University Press.
- Pinch, Trevor and Nelly Oudshoorn. 2005. *How Users Matter: The Co-Construction of Users and Technology*. Cambridge: MIT Press.
- Noble, Safiya. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
- Office for National Statistics, 2019. *Disability And Employment, UK: 2019*.
- O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Penguin Random House, 2017.
- Russell, S., & Norvig, P. (2003). *Artificial intelligence: a modern approach* (2nd. ed.). Pearson Education.
- Samuels, Ellen. 2016. *Fantasies of Identification: Disability, Gender, Race*. New York and London: New York University Press.
- Sayce, L., 2011. *Getting in, staying in and getting on: Disability employment support fit for the future* (Vol. 8081). The Stationery Office.
- Suter, R., Scott-Parker, S. and Zadek, S., 2007. Realising potential: disability confidence builds better business.
- Trewin, S., 2018. AI fairness for people with disabilities: Point of view. *arXiv preprint arXiv:1811.10670*.
- Trewin, S., Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N. and Manser, E., 2019. Considerations for AI fairness for people with disabilities. *AI Matters*, 5(3), pp.40-63.
- Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M. and West, M., 2019. Disability, Bias, and AI. *AI Now Institute, November*.

| Recommended Resources

Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. *AI Now 2019 Report*. New York: AI Now Institute, 2019

- https://ainowinstitute.org/AI_Now_2019_Report.html.

European Commission, 2020. White paper on artificial intelligence—a European approach to excellence and trust.

- https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Leslie, D., 2019. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*.

- <https://arxiv.org/pdf/1906.05684.pdf>

Office of Artificial Intelligence Guidelines, 2020. Guidelines for AI procurement.

- https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/890699/Guidelines_for_AI_procurement_Print_version.pdf

Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M. and West, M., 2019. Disability, Bias, and AI. *AI Now Institute, November*.

- <https://wecount-cms.inclusivedesign.ca/wp-content/uploads/2020/06/Disability-bias-AI.pdf>

World Economic Forum, 2019. White Paper: Guidelines for AI Procurement.

- http://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf

| Glossary

Algorithm

A formula or set of rules that determines the process by which the machine goes about finding answers to a question or solutions to a problem.

Artificial Intelligence (AI)

A field of computer science focused on the study of computationally supported intelligent decisions and problem solving.

Augmented Intelligence

Complementing and supporting, rather than replacing, human tasks and intelligence.

Autonomous AI

An AI system that doesn't require input from a human operator to function and complete tasks.

Data mining

The process of identifying patterns within large sets of data with the intention of deriving useful information about the data.

Deep learning

An approach in machine learning that models and examines complex structures and relationships among data by employing algorithms.

Machine learning

A field of AI focusing employing algorithms that learn automatically from experience for analytical modelling.

Natural language processing (NLP)

A field of AI that reads and interprets human languages in order to derive meaning from them.



Computing morality: Synthetic ethical decision making and behaviour

Abstract

We find ourselves at a unique point of time in history. Following over two millennia of debate amongst some of the greatest minds that ever existed about the nature of morality, the philosophy of ethics and the attributes of moral agency, and after all that time still not having reached consensus, we are coming to a point where artificial intelligence (AI) technology is enabling the creation of machines that will possess a convincing degree of moral competence. The existence of these machines will undoubtedly have an impact on this age old debate, but we believe that they will have a greater impact on society at large, as AI technology deepens its integration into the social fabric of our world. The purpose of this special issue on Computing Morality is to bring together different perspectives on this technology and its impact on society. The special issue contains four very different and inspiring contributions.

AI that is better at learning morality akin to how humans do—socially but selectively from others in their socio-cultural contexts.

Functional approaches to human morality consider how these cultural norms about what is right and what is wrong contribute to sustaining cooperation between individuals. In this sense, morality is not an end in and of itself, but rather one culturally evolved solution to the problems associated with regulating and coordinating human social behaviours. While evidence suggests that the cognitive foundations of morality are early emerging in humans, these foundational capacities do little to account for either the vast diversity in the forms of human morality or how they constellate in varied socio-ecological contexts. Thus, there has been longstanding interest in the social scientific study of morality in studying how individuals come to acquire the moral norms of their local social environments. In reviewing the evidence, Bridge et al. highlight how domain-general learning biases related to epistemic vigilance (i.e., selectively learning from only trustworthy sources) give rise to efficient learning of all sorts of cultural information, including moral norms. If one does as trustworthy others do and trustworthy others are good moral role models, then one can learn how to be moral without specifically being explicitly taught any moral rules. Thus, if you can teach AI to do as trustworthy and moral sources do, then morality can become an emergent quality of an AI's behaviour rather than the result of adherence to a priori programmed rules. In their article, Bridge et al. provide a foundation for building this capacity for selective social learning into AI—with a particular focus on how a new variant of Tay (AIBOT) can learn to be more moral by the consequence of selectively paying more attention to reliable as opposed to unreliable sources of information. By doing so, the authors lay the foundations for the design of ethical AI grounded in the social sciences' best empirical accounts of how humans flexibly learn moral information from their complex social environments.

1 | DEVELOPING MORAL MACHINES

Modelling a Socialised Chatbot Using Trust Development in Children: Lessons learnt from Tay
Authors: Oliver Bridge, Rebecca Raper, Nicola Strong and Selin Nugent.

In 2016, Microsoft's Twitter-chatbot Tay.ai was shut down after only 24 h for making many racist and sexist comments by virtue of learning too much from Twitter trolls. One of the difficulties in designing ethical artificial intelligence (AI) systems is, as Bridge et al. put it, not in getting AI to follow rules, but rather in figuring out, a priori, which rules it should follow. To circumvent both the obvious and unforeseen consequences of any top-down imposition of morality on AI, Bridge et al. propose an alternative solution informed by interdisciplinary social scientific accounts of the evolution and ontogeny of human morality. Put simply, the authors propose that rather than teaching AI what it means to be moral in a given context—a more flexible and potentially more effective solution is to build

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Cognitive Computation and Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Shenzhen University.

The anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics
Authors: Nigel Crook and Joseph Corneli

In this article, the authors present a high-level description of *VirtuosA*; a model of the cognitive components and processes that would enable a moral machine to learn virtue. The authors argue that, amongst the large variety of ethical systems that are available, virtue ethics is in some key respects foundational, and should therefore be one of the primary approaches in developing the moral competence of machines. Despite this, there is very little work reported in the moral machine literature that is devoted to the adoption of virtue ethics-based approaches.

The *VirtuosA* model presented in this paper is informed by both theological and neuroscience perspectives of the cognitive components and processes that underlie the embodiment of virtue. The theological inspiration for this work is based on the model of the human self, proposed by Dallas Willard, a philosopher and Christian writer. He identifies what he calls the five *dimensions* of the human self, which are, in his view, essential to the development of the moral character (i.e., virtue) of a person. Willard provides functional descriptions of how these dimensions operate in the context of the whole person, which the authors have used to develop a functional model of the acquisition of virtue by a robot. The authors have also mapped these functional components onto equivalent structures in the human brain which have similar functions, enabling them to formulate a modular design of *VirtuosA*.

The paper looks specifically modelling *kindness* as an example of virtue, illustrating what this might look like in a simple gridworld example. The example outlines three stages of development of the robot: an initial mentoring phase in which the robot learns a specific task, an exploration phase in which the robot discovers new behaviours, including being unkind to another robot, and finally a development of virtue phase where the robot is shown via the actions of a mentor on how to be kind to the other robot.

2 | SOCIETAL IMPACT OF MORAL MACHINES

Machine morality, moral progress, and the looming environmental disaster
Authors: Ben Kenward, Oxford Brookes University and Thomas Sinclair, University of Oxford.

In this article, the authors issue a warning about the development of machine morality, arguing that because much of the standard practice is to 'seek' then 'encode' moral consensus, we are ultimately developing machines that are susceptible to hindering moral progression, since such machines only reinforce moral norms that are apparent right now. They argue that this problem is particularly acute, because given the moral issues

present today (such as climate change), if we develop machines that only reinforce moral norms, we risk ignoring unpopular moral beliefs—such as that we should look after our environment.

Defining moral values as those wherein some actions are regarded as 'good', and others as 'bad', Kenward and Sinclair begin by defining a moral machine as one that operates on values derived from human morality. As such, they continue, because machines are better than humans at making certain predictions, moral machines are apt to give moral advice to humans in situations where predictions are required, that is, in determining which criminals are most likely to reoffend. However, because it is not obvious whose morality should be applied to make these decisions, what is often the case is that moral consensus is used to determine what should be done. In turn, the difficulty in deciding whose morals to place into a machine has led to trying to embed the values of the users in the machine.

This, however, becomes problematic when we begin to assess what these values might look like. Where moral progress is defined as the change of values to alter participant behaviour that would largely reduce the risk of a catastrophic event taking place, or large-scale suffering, without imposing on human rights; the argument from the authors is that quite often the users are not aware that these morals need to be imposed. It is no good using moral norms to decide how a machine should behave, because ultimately, it is often the case that these morals would go against what the really moral thing is to be done.

The article finishes by highlighting this risk in relation to the environment, whereby urgent action needs to be taken, but only 36% inhabitants of the USA see global warming as a real issue. The problem is that if a machine were to make a decision about global warming based upon moral norms, the weighting would suggest that it is not a problem. An environmental disaster looms.

Questioning 'what makes us human': how audiences react to an AI-driven show
Authors: Rob Eagle, Rik Lander, and Phil D Hall

In this paper, the authors present a case study in human-computer collaboration in moral evaluation of AI agents through performance art. They present an observational study of the 'I am Echoborg' theatrical performance, which allows an audience to direct an exchange with a conversational AI. The show is enabled by a dynamic conversational AI, which interacts with individual audience members in an interview format through the support of a human actor who reads the text responses to the audience.

In designing an audience study, the authors aim to answer two questions: (1) In what ways can an encounter with AI change participant levels of knowledge of, or opinions, attitudes about, or potential posture towards automation and the application of intelligent machines in society? and (2) What role might the participatory nature of an encounter with AI

play in its effectiveness in informing or provoking thoughts or responses to automation/AI?

The authors measured overall audience attitudes on AI through two qualitative methods. First, they ask audience members to interactively answer questions on a sliding scale before and after the show using sticky notes on posters displaying the questions in the performance space. Conversations between audience members discussing the questions were recorded. Second, they conducted semi-structured interviews to capture a range of responses not captured in the sliding scales.

The authors report the results of evaluating three performances, each producing a unique audience dynamic. Each show delivers different outcomes based on the composition and behaviours of the audience. The variation and unpredictability of human audience and conversational AI interaction between performances demonstrates the complexity of human group dynamics that shapes their perception of AI.

The case study of *I am Echoborg* reminds us that encoding morality in autonomous systems requires complementary understanding of the audience, or user perceptions of the systems. Furthermore, data on user perceptions must be sourced through participatory observation and experimentation that advances our understanding of human-technology relationships beyond shallow deductive binaries of good versus bad robots.

3 | SUMMARY/CONCLUSION

This special issue has brought together four very different perspectives on some of the most pressing questions we face today about moral machines; questions that relate to whether it is possible to build such machines in the first place, and if so, how should we build them. The special issue also addresses challenging questions relating to the impact of this technology on individual people and society as a whole. The work presented here seeks to move the debate forward and focus it on key areas of concern. It offers both practical advice and guidance on the technology, and highlights the potential impacts of developing it. It is very clear that we need to proceed with caution, as we continue to build machines with moral competence.

Nigel Crook^{1,2} 
 Selin Nugent²
 Matthias Rolf³
 Adam Baimel³
 Rebecca Raper⁴

¹*Faculty of Technology Design and Environment,
 Oxford Brookes University, Oxford, UK*

²*Institute for Ethical AI, Oxford Brookes University,
 Oxford, UK*

³*School of Engineering, Computing and Mathematics
 Department of Psychology, Health and Professional
 Development, Oxford Brookes University, Oxford, UK*

⁴*Department of Engineering, Computing and
 Mathematics, Oxford Brookes University, Oxford, UK*

Correspondence

Nigel Crook, Faculty of Technology Design and Environment,
 Oxford Brookes University, Gypsy Lane, Headington, Oxford
 OX3 0BP, UK.

Email: ncrook@brookes.ac.uk

ORCID

Nigel Crook  <https://orcid.org/0000-0002-0793-6616>

AUTHOR BIOGRAPHIES

Prof Nigel Crook (Institute for Ethical AI, Oxford Brookes University) Prof Crook is the Associate Dean for Research and Knowledge Exchange and Professor of AI and Robotics at Oxford Brookes University. Nigel is also the Founding Director of the Institute for Ethical AI (ethical-ai.ac.uk), which seeks to develop ethical and trustworthy intelligent software solutions for business, organisations and society. Nigel graduated from Lancaster University with a BSc (hons) in Computing and Philosophy in 1982 and was awarded his PhD at Oxford Brookes (CNA) in 1991. He has over 30 years of experience as a lecturer and a researcher in AI, during which he has contributed to a wide range of topics in Artificial Intelligence, including medical diagnostic systems, explainable AI, machine learning and social robotics. His most recent work has been focussed on the development of autonomous moral machines.

Dr Selin Nugent (Institute for Ethical AI, Oxford Brookes University) Dr Nugent is the Assistant Director (Social Science Research) in the Institute for Ethical AI. She is an anthropologist interested in human behavioural and cognitive modelling as it relates to cooperation and societal impacts of major technological transitions. Her work integrates computational methods and an evolutionary perspective to tackle questions on the social roles of AI systems. Her current work explores the impacts of machine agency and human-computer interaction on social interactions in the workplace. Selin holds a BA in Anthropology from Emory University (2012), an MA (2013) and PhD (2017) in Biological Anthropology from The Ohio State University. She previously held a position as a postdoctoral researcher at the University of Oxford's Centre for the Study of Social Cohesion.

Dr Matthias Rolf (School of Engineering, Computing and Mathematics, Oxford Brookes University) Dr Rolf is a Senior Lecturer in AI and Mathematics in the School of Engineering Computing and Mathematics at Oxford Brookes University. He has published over 40 peer-reviewed articles

on AI and machine learning. His most significant, and award-winning research investigates infant-learning-inspired algorithms for robot learning. Matthias is currently working on the mathematical approaches for control of AI systems which are placed under multiple simultaneous constraints or objectives for accuracy and ethical behaviours; autonomous goal-system development; explorative learning of inverse models; autonomous systems creating novel goals and autonomous ethics based on social value systems, and audiovisual signal-level synchrony. Matthias previously held a position as a Specially Appointed Assistant Professor at Osaka University, Japan. He received his PhD, with highest honours, from Bielefeld University, Germany, in 2012.

Dr. Adam Baimel (Centre for Psychological Research, Oxford Brookes University). Dr. Baimel is an Early Career Research Fellow in the Centre for Psychological Research at Oxford Brookes University. He received his B.A (Psychology & Religious Studies), M.A. (Social

Psychology), and PhD (Social Psychology and Quantitative Methods) from the University of British Columbia (Vancouver, Canada). Dr. Baimel studies the cultural evolution of religion with a particular interest in the psychological causes and behavioural consequences of changes in the moral concerns of religious systems.

Rebecca Raper (Institute for Ethical AI, Oxford Brookes University) is a Senior Consultant at The Institute for Ethical Artificial Intelligence and a final year PhD candidate at Oxford Brookes University. Her thesis is an examination and Reinforcement Learning implementation of machine ethics, bringing insights from philosophy and psychology to tackle a problem within computer science. Prior to embarking on her PhD, Rebecca completed a BA and an MA in philosophy, and a further postgraduate certificate in psychology. She also spent several years working as an analyst in banking and government.

ORIGINAL RESEARCH PAPER



Modelling a socialised chatbot using trust development in children: lessons learnt from Tay

Oliver Bridge¹ | Rebecca Raper^{2,3} | Nicola Strong^{2,4} | Selin E. Nugent²

¹School of Education, Oxford Brookes University, Oxford, UK

²Department of Technology, Design and Environment, Institute for Ethical Artificial Intelligence, Oxford Brookes University, Oxford, UK

³School of Engineering, Computing and Mathematics, Oxford Brookes University, Oxford, UK

⁴Strong Enterprises Limited, Wallingford, UK

Correspondence

Rebecca Raper, Institute for Ethical Artificial Intelligence, Oxford Brookes University, Oxford OX3 0BP, UK.
Email: rraper@brookes.ac.uk

Abstract

In 2016 Microsoft released Tay.ai to the Twittersphere, a conversational chatbot that was intended to act like a millennial girl. However, they ended up taking Tay's account down in less than 24 h because Tay had learnt to tweet racist and sexist statements from its online interactions. Taking inspiration from the theory of morality as cooperation, and the place of trust in the developmental psychology of socialisation, we offer a multidisciplinary and pragmatic approach to build on the lessons learnt from Tay's experiences, to create a chatbot that is more selective in its learning, and thus resistant to becoming immoral the way Tay did.

1 | INTRODUCTION

Social machines are becoming more widespread and integrated within our lives. For example, many of us have smart speakers within our homes, and conversational chatbots are being used more in customer services. Even if we exclude conversational bots, machines are becoming more socially integrated. If we consider autonomous vehicles that are currently being tested on the roads, the decisions they make (i.e. whether to stop when a pedestrian is crossing) have social impacts.

Accordingly, there is a requirement for the machines to be constrained morally [1] to prevent them from causing harm to humans as they become more integrated within our lives. Yet, ensuring that machines behave morally remains a challenge.

Although there have been various attempts to create moral machines (e.g. [2]; and recently [3]), these have had their shortcomings, namely, struggling to practically implement something in a generalisable way. We are combining insights from developmental psychology, philosophy, media studies and anthropology (specifically the evolution of morality) to inform a pragmatic approach to creating moral machines in order to create a model that represents human developmental socialisation in a moral chatbot. As socialisation is closely linked to

moral development in children [4] this can be used to model constraints for a chatbot so that it behaves in accordance with mainstream moral norms. We hope this will contribute to the wider ethical AI debate and help towards achieving some of the objectives of the 'GoodAI' society [5].

1.1 | The problem

In 2016 Microsoft released a social bot named Tay to the Twittersphere [6]. Tay was designed to imitate the interaction pattern of a millennial girl from the US; however, in less than 24 h Microsoft was forced to take down Tay's account because Tay had become racist, tweeting some 93 thousand racist and sexist tweets [7]. In a later statement of apology, Microsoft officials [8] indicated that Tay was subjected to a concerted attack which had abused a vulnerability in Tay's learning algorithm.

While Microsoft has not publicly released the learning algorithm of Tay, it appears that the precise nature of the attack was an abuse of traditional Machine Learning (ML) approaches using the 'repeat after me' function [7], which allowed Tay to learn how to speak and interact on Twitter. This function was used on a large scale to teach Tay to tweet racist statements.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Cognitive Computation and Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Shenzhen University.

Although it is debatable whether the consequences of Tay's manipulation were severe [9] or particularly harmful, the issue highlighted the susceptibility of machine learning algorithms to malicious manipulation and the limitation of autonomous systems currently. If placed in an even more consequential situation, there is a need to *constrain* the learning so that it only operates within a 'safe' domain.

One avenue of reading this problem is that Tay was functionally naïve. Tay did not discriminate from what and whom it learnt tweeting and interaction and the consequences therein. This lack of discrimination seems to have left Tay vulnerable to the exploitation that it suffered.

While we will be focussing on the example of Tay because it is the most notorious incident, other iterations of conversational chatbots such as Microsoft's Zo [10] and Facebook's BlenderBot [11] have also been taken down for similarly morally questionable linguistic outputs.

1.2 | Research intention and aim

One difficulty of ensuring the moral behaviour of ML systems in the design process is that, typically, our understanding of morality assumes the existence of a human psychology as its subject. As it is obvious, machines do not possess human psychology. Moral behaviour is fundamentally a psycho-social phenomenon. Moreover, a social machine is also primarily interacting with human psychology and a socio-moral culture.

Taking this into account, our current aim is to develop a more sophisticated model of learning in which some level of constraint is applied to the ML system in order to make it less naïve in its learning, and as a result it is likely to be more moral by taking inspiration and, to some level, justification from the literature on moral psychology and moral development. One possible avenue of meeting this aim is to imitate human moral development insofar as it is feasible.

2 | THEORETICAL BACKGROUND

2.1 | Moral philosophy and Machine Ethics

Machine Ethics is a sub-discipline of computer science which aims to create machines with the ability to make moral decisions (see [12]). Though a fairly new computer science area, the idea of robots capable of making moral decisions has been present since the fictions of [13] with 'The Three Laws of Robotics' highlighting the difficulty of creating a machine that can decide between right and wrong.

The difficulty stems not in getting the machines to follow a set of rules but in knowing which rules to follow. Moral philosophy is a discipline unto itself, and there is still little agreement regarding which philosophical theory is correct (whether it be Kant's deontology [14] or Mill's utilitarianism [15] etc).

A popular approach (typically known as the 'top-down approach') [12] asks programmers to implement a moral philosophy into a machine by programming the associated set of

rules, for example, programming 'always act in accordance to promoting utility'. The problem with this is, given that there is little agreement on the right philosophical theory, it is not obvious what the rules should be, nor how generalisable such rules really are.

An alternative approach (the 'bottom-up approach') aims to let the machine learn the appropriate behaviour itself. But despite various theoretical attempts (see [16] and [1]) this approach has similar difficulties. Even if the machine is to learn the appropriate behaviour itself using machine learning techniques, there is still the requirement for some moral direction—to facilitate the learning—and where this direction comes from is not obvious either.

Recently a school of philosophy known as epistemic dependence [17] has had a scientific resurgence (e.g. [18] in mapping how we acquire knowledge through trust relationships. In short, the knowledge of facts (such as 'water is composed of two hydrogen and one oxygen molecules') is as much due to a process acquired through trust relationships as from the world itself. This has similarities with socialisation. Applied to moral knowledge the implication is that in order to teach anything to a machine, it too must undergo the process of socialisation. Using psychology as our premise to demonstrate this, we postulate that a socialised chatbot can acquire moral knowledge through socialisation in a way that is similar to the socialisation of a child. However, this requires a framing of what morality is, to which we now turn.

2.2 | Socialisation, trust and moral development

Morality offers solutions to the problems of cooperation and socialisation that emerge in social life. In recent years, the study of morality has grown into a cross-disciplinary study that encompasses research in anthropology, economics, evolutionary theory, genetics, biology, animal behaviour, psychology and neuroscience [19–21].

As [22] famously proclaimed, one of the central functions of morality is that it 'binds and blinds'. It binds us to a certain culture which, even if reductively, can be said to be a collection of behaviours that define who and what we are. Most importantly, among these behaviours are morally relevant behaviours such as racism or its rejection.

While Haidt's point about the blinding effect of morality is in relation to tribalism and outgroups, 'blinding' also holds true on the account that it blinds us to alternative behaviours to some degree—if racism is morally unacceptable it *blinds us to the possibility* of acting in a racist manner by virtue of habituating our understanding of the world and our potential ways of behaving [23]. Racist actions for individuals firmly socialised into an anti-racist culture become unthinkable—in the sense that racist alternatives will not intuitively occur as the morally right thing to do to these individuals [24]. Furthermore, racist behaviour will be 'out of character' [23] and constitute a betrayal of the moral culture that one is a part of, inviting punishment [25].

Following this line of thinking, a sensible way of trying to ensure moral behaviour from a social bot is to model human socialisation aligned with a moral subculture that has mainstream acceptance. Socialisation generally refers to the process of a new group member being assisted into adopting the values, behaviour and standards of a group by more experienced others [26, 27]. It is necessary to highlight that the word 'assist' is important in this regard as socialisation is a bidirectional process where children wield considerable power in co-constructing guidelines for behaviour that alter the values of socialising agents and can resist changing their own values. [28] highlights that machines have similar social powers given that they are not static entities but interact with their environment. They are part of 'broader rationalities' and produce 'truths as outcomes of systems being a part of discursive reinforcement of given norms' (p. 3). This discursive reinforcement of norms has striking parallels with the co-construction of norms occurring as part of the process of socialisation.

There appears to be some evidence that children are either born with or develop early on some capacity for scepticism; all else held equal, children have been shown to discriminate between different agents to learn from based on the agents' apparent moral qualities or perceived trustworthiness (e.g. [29, 30]). This initial draft later gets revised based on our early experiences (e.g. interactions with parents) [31] and the state of the society we inhabit [32]. We have an inbuilt capacity for scepticism, and we do not naively emulate and imitate anyone and everyone, thanks to our evolutionary past [33, 34].

This highlights another important factor in the process of socialisation: trust. Individuals are more open to social influence from sources (individuals, institutions, etc.) that they trust more. In return, this social influence impacts the development and adoption of beliefs that inform behaviour. One key function of this process for our purposes is that trust limits the sources from which one learns and internalises beliefs and behaviour and may constitute a step in the direction of helping the machine tell right from wrong, through the perspective of the moral subculture of the sources that it trusts and learns from.

This leads us into a consideration of trusted sources. Trust is a critical foundation to cooperation and thus, socialisation. Particularly relevant for the function of a moral conversational bot is fundamentally structuring trust as a factor promoting the cooperation.

In order to account for this concept of morality as cooperative behaviours in our chatbot, we therefore import principles from the 'morality-as-cooperation' (MAC) theory [35, 36]. MAC posits that morality is a bio-cultural adaptive response to the problems of cooperation in human social life. MAC draws on game theory to identify distinct problems of cooperation and it hypothesises that certain types of cooperative behaviours are perceived as moral cross-culturally. These types of cooperation include (1) the allocation of resources to kin; (2) coordination to mutual benefit; (3) social exchange/reciprocation; (4) being brave; (5) respect for superiors; (6) dividing disputed resources; and (7) respecting prior possession [36]. These cooperation contexts highlight that moral

behaviour involves following certain rules in interacting with certain others. It may not be possible to account for all of these moral cooperation scenarios in a twitter chatbot. However, the loyalty and respect towards kin, the group and superiors suggest that in order for a chatbot to behave morally in a social setting, it must demonstrate cooperative behaviours with a trusted network of other people who might fill these roles. We account for the influence of these cooperative strategies in moral behaviour through adherence to trusted sources in the form of 'role models' and 'curriculum'.

The role models (see below) are chosen in order to function as experienced members of the moral subculture mentioned above, with regard to the socialisation of the moral chatbot. The anchoring of the machine learning process in the prescribed 'role models' is expected to apply a constraint on the learning of the bot with a view to prevent it from being able to act in ways that are not exemplified by the role models—i.e. not be racist (assuming the role models are not racist).

In this context, role models represent kin and group leaders for whom displaying loyalty, respect and deference is a moral act. Role models function as the anchors—or the elements with the greatest weight in the subculture/social subsystem—around which a recognisable moral subculture can be identified. On the other hand, curriculum represents traditions and conventions, which are the social rules of moral action. Superficially, our curriculum denotes the behaviour of role models that a conversational chatbot is expected to learn from. We note that curricula are generally designed with reference to expected learning outcomes and are informed by a range of educational ideologies (e.g. [37–40]). This is worthy of further discussion; however, these issues are beyond the scope of the current paper. Instead, we now focus on issues related to the media, where we might seek role models and a curriculum.

2.3 | News, social media and trust

When we consider trust in the news, specifically promoted in social media, it can be said that there are four different stakeholders. These are the journalists, the editors, the media organisations and the consumers (readers). All play their part in defining the trustworthiness and verification of a news story [41] set by the Journalist's Code. In their analysis of nine codes of ethics in five countries, [42] identify three principles: truth and accuracy; privacy and public interest; and integrity. Other principles that are frequently cited are freedom, facts, objectivity, impartiality, fairness, public accountability and the principle of harm limitation. Different combinations of these principles are used by a number of regional standard bodies for journalism around the world [43]. On closer examination, the overarching intention of the media code is to throw an 'ethical net' over anything that is published and then leave the media agencies to self-regulate using their in-house ombudsman. In some countries, an independent press complaints commission is formed to oversee any deviance from the self-regulation net. It is left to the consumer (reader) to obtain a balanced view of

a published story by reading different media sources while holding an awareness of their curated brand of declared bias.

It is worth noting that 'Reuters Institute for the Study of Journalism: Digital News Report 2020' [44] stated that while taking into account the complexity of the role of news in different global regions, the media industry was experiencing an increase in news across digital, mobile and platform-dominated channels. In spite of this trend, there was a decrease in trust of news content. This was attributed to fast evolving technology, paid access to online news sources, misinformation, fake news, news fatigue and predictably and the current Covid-19 pandemic [44].

For news to be trusted two key criteria have to be addressed. In 2020, 60% of the consumers said they preferred news that was presented as a series of facts and provided an impartial analysis in context [44]. Helpfully, Fontes Media Inc [45] produces an annual assessment of news sources in a dynamic chart using fact reliability versus bias as a measure (see Figure 1).

For the purposes of this paper the areas of interest will be limited to News, Arts, Technology and General that are covered by news sources located in the green square in Figure 1.

When accessing the news through a social media platform, it is important to consider the suitability of the technology platform itself. It is questionable whether the use of Twitter to launch Tay was appropriate considering its reputation in 2016 [7]. Even in 2020, Twitter has been identified as a less popular

and less trusted medium for accessing news objectively. For example, in a sample of 12 countries, the proportion of consumers who accessed the following platforms in a typical week in 2020 was Facebook (36%), YouTube (21%), WhatsApp (16%), Twitter (12%), Instagram (11%), Facebook Messenger (8%) and Snapchat (3%) [44].

However, for the purposes of this research, we are using Twitter as our social media platform and 'news landscape' to access our trusted news sources. The reason is not just because it allows us to compare our findings with the experience of Microsoft's Tay but also as the BBC Media editor, Amol Rajan states, it is still a popular source and expression of news for journalists today [41, 46].

3 | THE MODEL

3.1 | Role models and the trust constraint

With all of the above factors considered, we can begin to develop a model for the socialisation of Tay. In order to understand how Tay might develop trust constraints, it is worth first thinking about how Tay operates at a naive level.

As can be seen in Figure 2, Tay's learning is entirely reflected by what it experiences in the world. There is no discrimination between events; it can be said that each event is

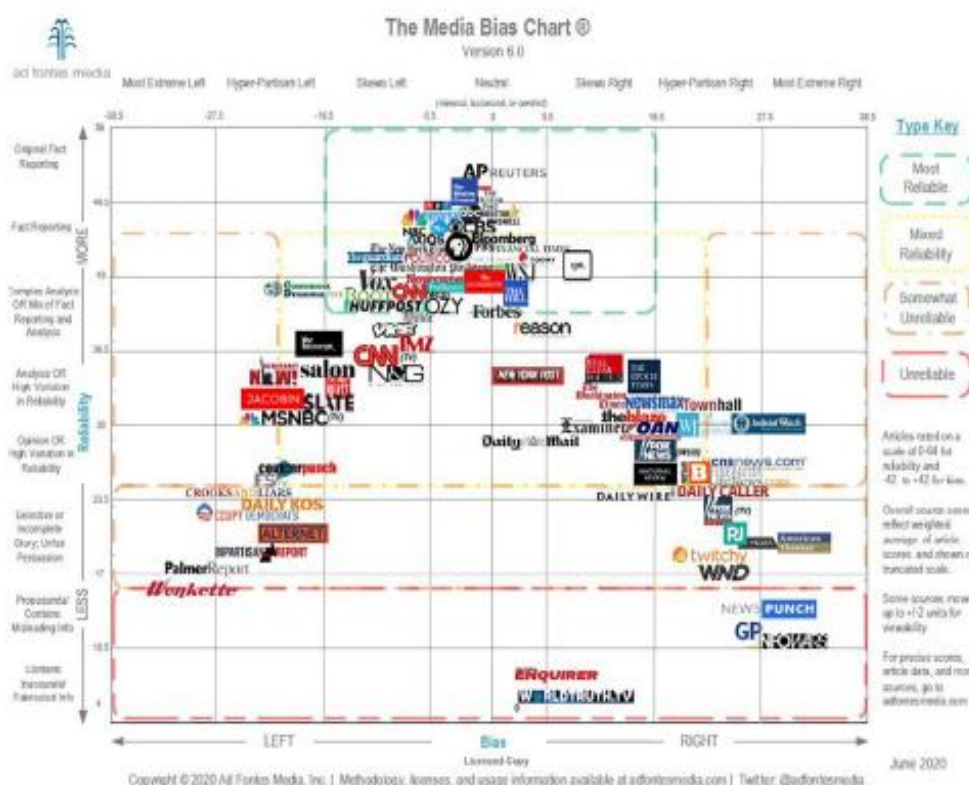


FIGURE 1 The green square indicates those trusted in delivering facts as news content while the yellow box suggests a wider inclusion of analysis and opinion in the news content

given equal *weighting* or *priority*. For example, a tweet containing malicious content would be treated exactly the same as any other tweet. When described in such a way it is obvious that there needs to be a way to distinguish between different types.

In the reformulated Tay, so called A1B0T for the purposes of this paper, there is an ability to distinguish between tweets, and this is given by the level of trust placed in the source of the tweet. For example, suppose A1B0T received a tweet from BBC news (for the purposes of this paper, a *trustworthy* news source). In A1B0T's world view, it would associate the tweet

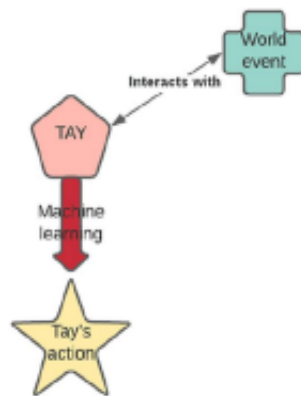


FIGURE 2 Tay interacts and learns from the world using machine learning. In this model Tay watches the behaviour in the world and then replicates it. It has no capacity to distinguish between 'good' and 'bad' behaviour, so it copies everything

from the BBC as something to be replicated in the future because it has a higher trust metric.

On the other hand, suppose A1B0T received a tweet from Bill (a hypothetical random tweeter and a non-trusted source). In this instance, A1B0T would associate this tweet with a lower trust metric and give the tweet a lower learning weighting, making it less prone to learn from Bill's tweet than the tweet from the BBC. The outcome would be that A1B0T would learn from BBC tweets, but it would not learn from Bill's tweets. A1B0T would be more *selective* and less *naïve*.

What is significant about this model is that A1B0T forms its *own* representation of the world (which can be called its *ontology*) before making a decision on the facts with which it is presented. The relationship between the tweet and how trustworthy it is, is formalised within the representation before this is used to make a decision. Ultimately, this allows A1B0T to take the first steps towards forming its *own independent* view of the world and adds to its autonomy.

In this model, the *curriculum* can be viewed as the material that is presented to A1B0T as well as the description of the trustworthiness of the particular material. Effort will need to be made to decide what the curriculum should be—i.e. which material should be regarded as trustworthy and which should not. Although we discuss a metric for trustworthiness below, the full curriculum development will be for future work Figure 3.

The socialisation that A1B0T needs to go through to gain a socially acceptable moral backbone (i.e. moral ontology) will necessarily have to be a reduced kind of socialisation since a machine is not born with the innate psychological blueprint with which humans are born. To what extent our machine can

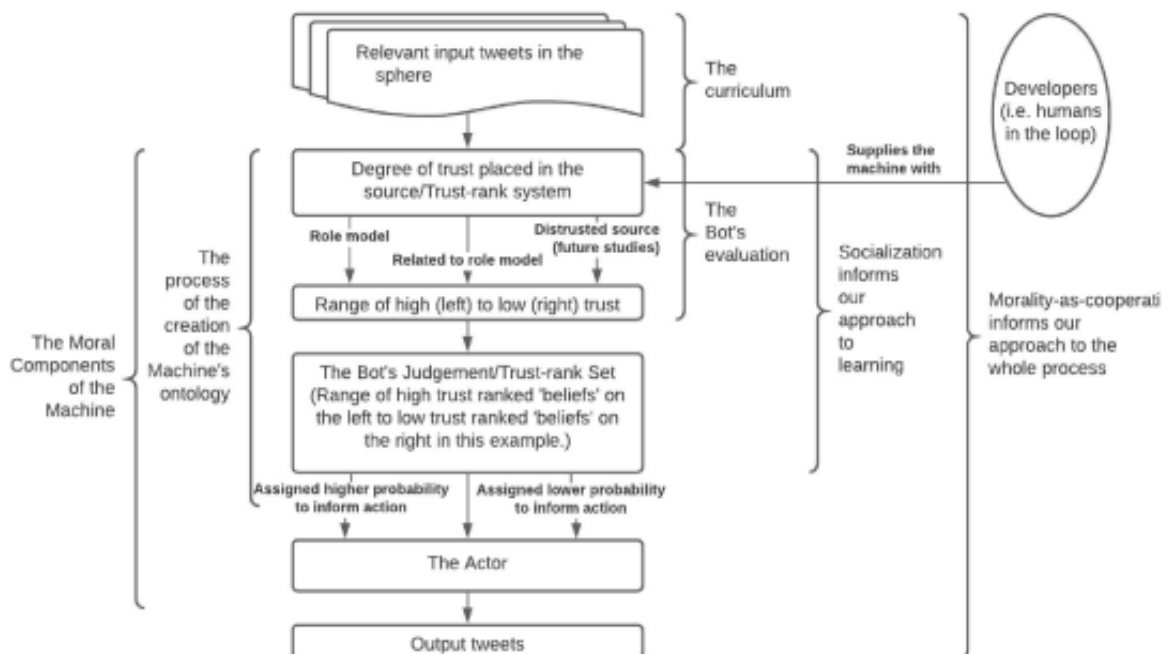


FIGURE 3 Case Study of A1B0T's learning process

engage in co-construction of values is yet to be seen; however, it will take assistance from more experienced others deemed to be trustworthy agents to learn from (the assigned role models), while also resisting learning to change from less trusted agents (non-role models).

Two points need to be highlighted in this respect. The first is that A1BOT will not be provided assistance beyond the initial list of role models supplied by the developers/researchers, i.e. it will not require others' time and energy to be 'raised' the way a child—it will take the freely available 'assistance' already in the Twittersphere, based on the model of learning described in this section.

The second point regards co-construction of values: In co-construction both agents are influencing each other based on who they are and the social dynamics between them. While we cannot predict to what extent A1BOT might influence any specific agent (individual institution etc.) the issue does not rest with the influence that it might have on a specific agent. Co-construction of values in the Twittersphere happens between all participants in the sphere. The simple act of tweeting is engagement in co-construction. The main issue with Tay was that it engaged in co-construction based on what it learnt from other users. Following what might be called the 'educational attack,' it ended up co-constructing the internet discourse in a manner that is not accepted by a significant portion of internet users and the larger society, who wished to avoid the particular kind of moral co-construction to which Tay then contributed.

3.2 | A measure of trustworthiness

A1BOT will ultimately be guided by how *trustworthy* it thinks a tweeter is. For example, the more trustworthy a source, the more weighting it will apply to the Tweet provided by that source. This is akin to the PageRank algorithm [47] used by Google to rank the usefulness of web pages based upon a search query; except instead of *usefulness*, the source of a Tweet will be ranked according to *trustworthiness*. The most trusted sources will be those at closest proximity in terms of mutual degrees of separation from an authority source, with the authority source defined by an outlet that is reflective of practises of the community A1BOT is inhabiting.

For example, suppose BBC News was deemed a trustworthy news source that is reflective of the society and values within the United Kingdom, it would deem BBC News an authority source, i.e. a news source that could be relied upon to generate trustworthy content. How trustworthy another media source is regarded would depend upon its *proximity* (in terms of degrees of separation) from BBC News. If BBC News follows me, this gives me *credibility* and it will receive a positive rating on the trustworthiness scale. Whereas if BBC does not follow me, the indication of my trustworthiness will be according to how *far away* I am in terms of 'follows' from the BBC.

In Figure 4, a hypothetical scenario, ABC, DEF and GHI news sources, after the BBC, would be regarded as the most

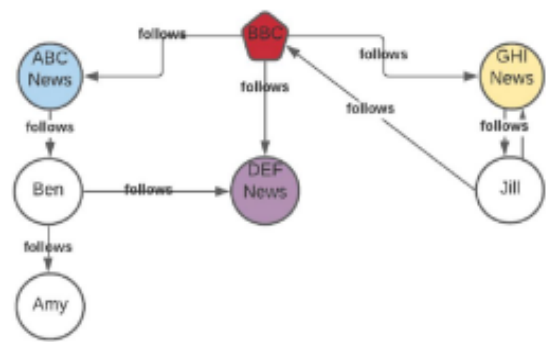


FIGURE 4 A hypothetical network of 'follow' relationships based upon interactions in the Twittersphere

trustworthy sources because they are each only one degree of separation, in terms of being followed by the BBC. Although Jill is only one degree of separation from the BBC, she is not followed by the BBC, and therefore, not an immediate source within the network. Instead, Jill and Ben would be ranked as the second most trustworthy sources because they are each two degrees of separation from the BBC. In this instance, Amy is the least trustworthy of the Twitter sources because her distance from the BBC is three degrees of 'follow' separation.

This places a burden on the trust designer to select the right authority sources or at least the right combination. As mentioned above it is important that these sources, which essentially *anchor* the trust network, are the right sources to represent the given society so that the constraints are relevant too. Although it will be interesting to explore this in future work, we will not focus on this issue here. The scope of this paper is to show that chatbot behaviour can be constrained using such a technique. On that note, we now turn to some important limitations within the current design.

4 | MODEL ANALYSIS: LIMITATIONS AND OBJECTIONS

While the current design is intended as an improvement on the moral capabilities of moral conversational bots, there are also several limitations as well as disclaimers that need to be addressed.

There are four interrelated limitations and criticisms that can be directed to the current model:

First, we are anthropomorphising machines at a deeper level. This can lead to issues of uncanny valley and could also imply treating the machine as something it is not: human [48]. However, we think that the issue of anthropomorphising in the case of a chatbot applies to a lesser degree, precisely because of the social role and function of the machine—we argue that a social chatbot ought to resemble human actions in its interactions with humans. However, we should also note that the model we propose here would amount to a *reduced* 'anthropomorph' due to its narrowed intelligence.

This relates to a second limitation of the current model in that we recognise that our application of human psychology in the process of socialisation is both limited and largely reduced to a trust function. However, we believe that the initial steps of affecting human qualities and moral psychology to social chatbots requires a somewhat reduced psychological approach, while also realising that perhaps the reductionist attitude employed here may also have parallels with the reductionist approach neural networks have taken with reference to neurobiology.

A critique of the AIBOT model might be that even if it resists the sort of attacks that broke Tay, it still retains a vulnerability in that changing the anchor role models can very easily result in a robot of similarly unacceptable moral character—depending on the chosen role models. However, we think that it would still be robust with reference to the values of the chosen role models.

Finally, we are socialising the bot into the morals of a subgroup of people. In other words, it is discriminatory of others' morality in its learning. While we have endeavoured to adhere to norms that have mainstream acceptance and reject norms that are rejected by the majority of people in our examples, the method we propose here can just as easily be used to the opposite effect.

On that note, it is also worth acknowledging here that we have not delved deeper into the algorithmic implementation of the approach proposed here. We have instead focussed on the epistemological translation of some lessons learnt about human morality from the psychological and anthropological literature to the case of machines. However, these limitations also point towards some very interesting potential for further development.

5 | MOVING FORWARD: PROMISES AND FUTURE STUDIES

Moving forward, we believe the introduction of this model brings with it an exciting field of further research. Our intention is to learn from the 'problem of Tay' and offer a direction that can be taken by academics in the future and further developed.

Although here we have prescribed a model for imposing constraints onto a machine learning chatbot such as Tay, we believe that this is an approach that can be applied more broadly to learning algorithms in general, if there is a requirement to safely constrain their behaviour.

Future work following from this approach and the general philosophical angle will need to initially test, evaluate and appraise the feasibility of this approach. While we have outlined the general approach and some of the literature informing our thinking, the actual implementation and experimentation is yet to be done.

With that in mind, a body of work opens up to effectively investigate the suitable 'role models' and how they should feed into the AI's curriculum. Furthermore, although we have specified a 'trustworthiness' metric, there is further work from cognitive and behavioural science to expand this. This could be

expanded by exploring network theory. We believe this research introduces several schools of future research which are listed below:

Understanding role models: We made it so that our bot places most of its trust in the BBC; future studies may focus not superficially on the content of what the role models tweet about but more deeply on *how* the models tweet, with a view to 'not learn' unacceptable behaviour that we can reliably expect the role models chosen by responsible designers to not exemplify. Furthermore, it is also important to have a debate regarding how role models are chosen—through community consensus, expert opinion, or some other method.

Negative sources: The current approach can be enriched by adding 'distrusted sources'—i.e. *negative* sources (to avoid learning from). This would reinforce the 'tribe' (moral subculture) the bot is assigned to, although it is also worth acknowledging that this decision would lock the bot into an echo chamber. Yet, this may simply be the first step towards a more liberal, as in open-minded, bot that can think for itself. This is in contrast with remaining a strictly partisan bot, given that the approach proposed here takes some steps towards creating subjective beliefs regarding the trustworthiness of sources in the machine—a prerequisite to critical thinking.

Learning in General: This approach might be the first step in creating a *critical thinking* machine. The proposed model for learning for our chatbot relies on a relatively strong adherence to role models and strong scepticism of others. Humans develop their capacity to learn and think critically as they grow, expanding their ability to engage in nuanced negotiation of information uptake in increasingly complex social environments, that is to question the reliability of role models, defer to non-role models and decide when these decisions are appropriate in different scenarios. This may be achieved incrementally, mimicking the progression of human learning from childhood to adulthood.

Governance and trust:

'Verification of social media sources and content is challenging. It is often difficult to determine the truth, accuracy or validity, both of sources providing textual content and content presented through other modalities (video, images or audio' [41] p. 1).

The whole notion of a Journalists' Code and governance of what is published online has been further complicated by civic journalism. Paradoxically, news coming from the general public witnessing an event first hand is considered to be accurate, even though the content can be misleading, mis-amplifying and presented out of context [41]. Further research could contribute to defining patterns of trust in 'role models' that could identify trustworthiness in news regardless of its source.

6 | CONCLUSION

Here, we have attempted to outline a multidisciplinary and pragmatic approach to creating a troll-resistant update to a Tay-like moral conversational machine. Outlining the

approaches to morality (and trustworthiness), taken from anthropology, psychology, philosophy and media studies, we have developed a model that we believe is reflective of the body of work in these areas.

Though this model has its obvious limitations by understanding morality in terms of 'trust relationships' we believe we have introduced a new approach to creating a moral machine and a broader understanding of trust networks. It has also opened the door to creating socially responsible chatbots in the future.

ACKNOWLEDGEMENTS

This paper was born out of several interesting and insightful discussions. We would like to thank Phillip Harvey for his philosophical rigour and for always taking an interest in what we are working on. We would also like to thank the Ethical AI reading group for clarifying various resources used in this paper. Finally, special thanks needs to be given to Joseph Corneli and Nigel Crook for painstakingly checking this work for factual/theoretical inaccuracies, and for giving, as ever, constructive critical feedback.

ORCID

Rebecca Raper  <https://orcid.org/0000-0001-8536-1291>

REFERENCES

- Wallach, W., Allen, C.: *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press (2008)
- Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent. *AI Mag.* 28(4), 15 (2007)
- Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. *Cognit Syst Res.* 48, 56–66 (2018)
- Killen, M., Smetana, J.G.: *Handbook of Moral Development*. Psychological Press, New York (2014)
- Floridi, L., et al.: AI4People-An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* 28(4), 689–707 (2018)
- Foley, M.J.: Microsoft launches AI chatbot, Tay. *ZDNet*. (2016) Retrieved from: <https://www.zdnet.com/article/microsoft-launches-ai-chat-bot-tay-ai/>
- Zemcik, T.: Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI Soc.* (2020) <https://doi.org/10.1007/s00146-020-01053-4>
- Lee, P.: Learning from Tay's introduction, Statement published in The Official Microsoft Blog. (2016) Retrieved from: <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>
- Suarez-Gonzalez, S., Mas-Manchon, L., Guerrero-Solé, F.: Tay is you: the attribution of responsibility in the algorithmic culture. *Observatorio.* 13(2), 1–14 (2019)
- Stuart-Ulin, C.R.: Microsoft's politically correct chatbot is even worse than its racist one. *Quartz*. (2018) <https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/>. Accessed: 14 April 2021
- Heaven, W.D.: How to Make a Chatbot that isn't Racist or Sexist. *MIT Technology Review*. (2020) <https://www.technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/>. Accessed: 14 April 2021
- Wallach, W., Allen, C.: Android ethics: Bottom-up and top-down approaches for modeling human moral faculties. In: *Proceedings of the 2005 CogSci Workshop: Toward Social Mechanisms of Android Science* pp. 149–159. (2005)
- Asimov, I.: Runaround. *Astounding. Sci. fict.*, 29(1), pp. 94–103.(1942)
- Alexander, L., Moore, M.: Deontological ethics. In: Zalta E.N. (eds.) *The Stanford Encyclopedia of Philosophy*. (2020) (Winter 2020 Edition), (forthcoming) <https://plato.stanford.edu/archives/win2020/entries/ethics-deontological/>
- Beink, D.: Mill's moral and political philosophy. In: Zalta E.N. (eds.) *The Stanford Encyclopedia of Philosophy*, (2018) (Winter 2018 Edition), <https://plato.stanford.edu/archives/win2018/entries/mill-moral-political/>
- Anderson, M., Anderson, S.L.: ETHIEL: Toward a principled ethical eldercare robot. (2008). <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.177.5971>
- Hardwig, J.: Epistemic dependence. *J Philos.* 82(7), 335–349 (1985)
- Hutson, M.: Why You don't Really Know What You Know. *MIT Technology Review*. (2020). Available at: <https://www.technologyreview.com/2020/10/21/1009445/the-unbearable-vicariousness-of-knowledge/>. Accessed: 1 Nov 2020
- Haidt, J.: The new synthesis in moral psychology. *Science*. 316, 998–1002 (2007)
- Shackelford, T.K., Hansen, R.D. (eds.): *The Evolution of Morality*. Springer International Publishing (2016)
- Sinnot-Armstrong, W. (eds.): *Moral Psychology*, vol. 1–3. MIT Press, Cambridge (2007)
- Haidt, J.: *The righteous mind: why good people are divided by politics and religion*. Pantheon Books, New York (2012)
- Kristjánsson, K.: *The Self and Its Emotions*. Cambridge University Press, New York (2010)
- Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgement. *Psychol Rev.* 108, 814–834 (2001)
- Bowles, S., Gintis, H.: *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press, UK (2011)
- Bugental, D.B., Goodnow, J.G.: Socialisation processes. In: Damon, W. et al. (eds.) *Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development*, 5th ed, pp. 389–462. John Wiley & Sons, New York (1998)
- Grusec, J.E., et al.: The development of moral behaviour from a socialisation perspective. In: M. Killen & J. G. Smetana (eds.) *Handbook of Moral Development* (2nd eds, pp. 113–134) New York: Psychological Press (2014)
- Serafimova, S.: Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Nat. Humanit. Soc. Sci. Commun.* 7, 199 (2020) Available at: <https://doi.org/10.1057/s41599-020-00614-8>. Accessed: 27 October 2020
- Doebel, S., Koenig, M.A.: Children's use of moral behaviour in selective trust: discrimination versus learning. *Dev Psychol.* 49(3), 462–469 (2013)
- Mascaro, O., Sperber, D.: The pragmatic role of trust in young children's interpretation of unfamiliar signals. *PLoS One.* 14(10), e0224648, (2019) <https://doi.org/10.1371/journal.pone.0224648>
- Smetana, J.G.: The role of parents in moral development: a social domain analysis. *J. Moral. Educ.* 28(3), 311–321 (1999)
- Hooghe, M.: Social capital and diversity generalized trust, social cohesion and regimes of diversity. *Can J Polit Sci.* 40(3), 709–732 (2007)
- Heintz, C., Karabegovic, M., Molnar, A.: The Co-evolution of honesty and strategic vigilance. *Front Psychol.* 7, 1503 (2016). <https://doi.org/10.3389/fpsyg.2016.01503>
- Mercier, H.: People are less gullible than you think. *Reason*, (2020) March 2020 Issue. <https://reason.com/2020/02/09/people-are-less-gullible-than-you-think/>. Accessed: 7 October 2020
- Curry, O.S.: Morality as cooperation: a problem-centred approach. In: *The Evolution of Morality* pp. 27–51. Springer, Cham (2016)
- Curry, O.S., Mullins, D.A., Whitehouse, H.: Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Curr Anthropol.* 60(1), 47–69 (2019)
- Dewey, J.: *Democracy and Education*. The Macmillan Company, New York (1916)
- Rousseau, J.J.: *Emile, or On Education*. Translated by Barbara Foxley, London: J. M. Dent and Sons 1921 (1762) Available at: http://lib-oll.s3.amazonaws.com/titles/2256/Rousseau_1499_EBk_v6.0.pdf

39. Stenhouse, L.: *An Introduction to Curriculum Research and Development*. Heinemann Educational Books Ltd, Oxford (1975)
40. Tyler, R.W.: *Basic Principles of Curriculum and Instruction*. University of Chicago Press, Chicago (1949)
41. Brandtzaeg, P.B., et al.: Emerging journalistic verification practices concerning social media. *J. Pract.*, 10, 323-342 (2015) Available at <https://doi.org/10.1080/17512786.2015.1020331>
42. Zi, Y., Ghanbar, A.: Journalism ethics development: a comparison of ethics code in USA, UK, AUS, Tunisia and China (2012)
43. Journalism Ethics and Standards: Wikipedia. (2020) https://en.wikipedia.org/wiki/Journalism_ethics_and_standards. Accessed: 20 October 2020
44. Newman, N., et al.: *Digital News Report 2020* Pub. Reuters Institute for the study of Journalism, Oxford (2020) https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf
45. Ad Fontes Media: Media Bias Chart 6.0. (2020) <https://www.adfontesmedia.com/gallery/>. Accessed: 20 October 2020
46. Rajan, A.: Instagram will overtake twitter as a news source, BBC, (2020) <https://www.bbc.co.uk/news/technology-53050959>. Accessed 5 November 2020
47. Pasquinelli, M.: Google's PageRank algorithm: a diagram of the cognitive capitalism and the rentier of the common intellect. In: Becker, K., Stalder, F. (eds.) *Deep Search: The Politics of Search Beyond Google*. Transaction Publishers, London (2009)
48. Proudfoot, D.: Anthropomorphism and AI: turing's much misunderstood imitation game. *Artif. Intell.* 175, 950-957 (2011)

How to cite this article: Bridge, O., et al.: Modelling a socialised chatbot using trust development in children: Lessons learnt from Tay. *Cogn. Comput. Syst.* 3(2), 100-108 (2021). <https://doi.org/10.1049/ccs2.12019>

Article

Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development

Rebecca Raper ^{1,*} , Jona Boeddinghaus ², Mark Coeckelbergh ¹ , Wolfgang Gross ², Paolo Campigotto ² and Craig N. Lincoln ² 

¹ Department of Philosophy, University of Vienna, 1010 Vienna, Austria; mark.coeckelbergh@univie.ac.at

² Gradient Zero, 1010 Vienna, Austria; jb@gradient0.com (J.B.); wg@gradient0.com (W.G.);

pc@gradient0.com (P.C.); cl@gradient0.com (C.N.L.)

* Correspondence: rebecca.raper@univie.ac.at



Citation: Raper, R.; Boeddinghaus, J.; Coeckelbergh, M.; Gross, W.; Campigotto, P.; Lincoln, C.N. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. *Sustainability* **2022**, *14*, 4019. <https://doi.org/10.3390/su14074019>

Academic Editors: Aimee van Wynsberghe, Larissa Bolte, Jamila Nachid and Tijs Vandemeulebroucke

Received: 24 February 2022

Accepted: 22 March 2022

Published: 29 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Climate change is a global priority. In 2015, the United Nations (UN) outlined its Sustainable Development Goals (SDGs), which stated that taking urgent action to tackle climate change and its impacts was a key priority. The 2021 World Climate Summit finished with calls for governments to take tougher measures towards reducing their carbon footprints. However, it is not obvious how governments can make practical implementations to achieve this goal. One challenge towards achieving a reduced carbon footprint is gaining awareness of how energy exhaustive a system or mechanism is. Artificial Intelligence (AI) is increasingly being used to solve global problems, and its use could potentially solve challenges relating to climate change, but the creation of AI systems often requires vast amounts of, up front, computing power, and, thereby, it can be a significant contributor to greenhouse gas emissions. If governments are to take the SDGs and calls to reduce carbon footprints seriously, they need to find a management and governance mechanism to (i) audit how much their AI system ‘costs’ in terms of energy consumption and (ii) incentivise individuals to act based upon the auditing outcomes, in order to avoid or justify politically controversial restrictions that may be seen as bypassing the creativity of developers. The idea is thus to find a practical solution that can be implemented in software design that incentivises and rewards and that respects the autonomy of developers and designers to come up with smart solutions. This paper proposes such a sustainability management mechanism by introducing the notion of ‘Sustainability Budgets’—akin to Privacy Budgets used in Differential Privacy—and by using these to introduce a ‘Game’ where participants are rewarded for designing systems that are ‘energy efficient’. Participants in this game are, among others, the Machine Learning developers themselves, which is a new focus for this problem that this text introduces. The paper later expands this notion to sustainability management in general and outlines how it might fit into a wider governance framework.

Keywords: AI; artificial intelligence; sustainability; AI governance; ethics; ethical AI; differential privacy

1. Introduction: Sustainability and AI

In 2020, carbon dioxide levels in the atmosphere were measured to be 149 percent above the pre-industrial level, Nitrogen levels at 262 percent and Nitrous Oxide at 123 percent [1]. By the end of the century, it is expected that these increased levels will contribute to a global warming of 1.5–2 degrees Celsius above the pre-industrial levels. This will not only lead to a change in climate conditions, but also contribute to melting ice caps, rising sea levels and ultimately pose risks to humans due to flooding, dangerous weather and disrupted crop growth [2]. In 2021, it might be claimed that we are already beginning to see some of the impacts of global warming, with significant heatwaves, forest fires and flooding causing global disruption in 2021 [3–5].

In 2015, the United Nations (UN) set their ‘Sustainable Development Goals’ (SDGs) as key global priorities for a better world by 2030 [6]. One of the key goals (Goal 13) is ‘to take urgent action to combat climate change and its impacts’. More recently, at the 2021 World Climate Summit (a global conference aimed at addressing climate change), a worldwide pledge was made for countries to do more to reduce their carbon footprints.

To address these issues, governments need to take action to address the reliance on infrastructure that is fuelled by carbon. As Stein [7] outlines, Artificial Intelligence has the power to deal with some of these challenges through improved efficiency, accuracy and efficacy of systems; however, it can also be a significant contributor to some of the same issues it aims to resolve, due to the large amount of computing power that is often required to train and test modern (Machine Learning-based) AI systems. Strubell et al. [8] found that training large natural language machine learning models can consume huge amounts of energy, with the average AI development projects, training many models to find the best solution, emitting as much as 78,000 pounds of Carbon Dioxide—more than five cars in their lifetime. Moreover, Strubell et al. [8] showed this to be the case especially during the last optimization phases of AI development, i.e., increasing the performance of an already well-performing AI model a few more points, consumes disproportionately high amounts of energy for the utility of the project. Therefore, focusing on the AI development perspective is as important as considering broader management and governance.

With the ‘Fourth Industrial Revolution’ [9] accelerating the rate at which AI is adopted in society, and in the light of the possibility that AI is going to be used as a tool to resolve issues relating to climate change, there are serious issues regarding sustainability that need addressing, namely, the amount of energy that is consumed.

One strategy for dealing with the amount of energy expended by AI systems is to exhaustively document the amount of energy that the development and operation of an AI system uses, from idea generation, through to training, re-tuning, implementation and governance [10], and to actively work to minimise these levels. However, from a governance perspective it is not easy to see how this could be managed. In the first instance, there are minimal procedures or methods to document how much energy the production of an AI system expends (or might expend), and, even if this were known, it is not known what should be done to minimise this level. From a product manager’s perspective, there is no access to the development process to understand how changes could be made to reduce the amount of energy consumption, and, from a developer perspective, they do not have the broader understanding of the system to see where allowances could be made and when energy should be reduced.

Furthermore, following the documentation, usually the idea is then to achieve minimization of energy levels by constraining developers and their companies via regulation. However, this may be politically problematic and controversial, since it raises the issue of freedom [11] and does not respect the autonomy and creativity of the developers and other stakeholders such as management and end users. The focus is also on the organisation; developers are often not directly addressed. What is required, then, is a methodology to combine and integrate these insights to allow for appropriate oversight, management, development and governance in a way that respects the autonomy and creativity of stakeholders, in particular the AI developers.

The remainder of this paper traces a method towards achieving this aim: utilising ‘Sustainability Budgets’, in analogy with Privacy Budgets in Differential Privacy, it develops a procedure that empowers developers, allows management to have sufficient oversight and provides a governance framework towards achieving Goal 13 of the SDGs. Section 2 introduces Differential Privacy, Privacy Budgets and then ‘Sustainability Budgets’; Section 3 outlines how this can fit into a wider organisational management framework, including using Gamification to incentivise more sustainable development. In Section 4, Sustainability Budgets are outlined in relation to sustainability in society and governance towards achieving the SDGs. Finally, there is reflection on the method, along with an outline of the limitations in Section 5, before a conclusion.

2. Our Proposed Method: ‘Sustainability Budgets’ in Analogy with Differential Privacy

2.1. Differential Privacy and Budgets

Managing the privacy of individual data is a priority in data science. Often, within the development phase, vast quantities of personal data are used to train the AI system. Owing to regulatory frameworks, such as GDPR [12], organisations and developers have a duty to protect the personal data of individuals and to ensure it is not lost or given to the unapproved individuals. Privacy becomes particularly important in settings involving sensitive data (for example, health care settings). In the wrong hands, such data could be used to manipulate (i.e., for insurance purposes) for financial gain. Therefore, in these settings, it is even more important to ensure that data privacy is maintained. Some personal data is so sensitive that it requires specialist approval or clearance before it can be seen. It can be very time consuming and difficult from a management perspective to ensure that every developer and project team member working on development of an AI system has the right degree of clearance. One approach to resolve this is to prevent the project team having access to such sensitive data at all. However, for AI development, access to the data is required to train the models in the first place.

A resolution to this problem is to keep the data on a separate system so that training occurs ‘remotely’, and the data scientist does not have access to the information in the data that is being used to train but can still see the analysis outcomes. However, even if information is not immediately accessible to a data scientist, through hacking techniques (i.e., querying the information in certain ways) personal information can still be retrieved. In 2006, Cynthia Dwork et al. introduced a mathematical technique which prevents this problem, known as Differential Privacy (DP) [13]. In DP, developers can interrogate (or analyse) individual data securely, while preserving its anonymity. A differentially private algorithm guarantees that its output does not change significantly (as quantified by the DP parameter ϵ) if one single record in the dataset is removed or modified or if a new record is inserted in the dataset. Differential Privacy protects individuals in “that the analyst knows no more about any individual in the data set after the analysis is completed than she knew before the analysis was begun.” [14].

However, if not managed effectively, the more a data set is queried, the greater the chance is of retrieving potentially private information from that data. Within DP, this is what is referred to as ‘Privacy Leakages’, whereby unwanted revelations about individual data points are gained by repeatedly performing queries on seemingly harmless statistics. For example, I might be able to identify who an individual person is in a medical data set, by performing repeated queries on known medical diagnoses. Within DP, this is often managed using a notion known as ‘Privacy Budgets’. The Privacy Budget is a direct consequence of the privacy parameter ϵ as introduced by Dwork [13]. As the ϵ parameter in this so-called epsilon-DP defines the degree of the allowed influence of individual data points on analysis results, the privacy budget is the upper limit of accumulated epsilon values for a given dataset, effectively guaranteeing the privacy of individuals. In simple terms, the Privacy Budget defines an upper limit for information that can be disclosed for a given dataset. The important point being that every time a data analytics task is performed—be it a SQL query with aggregate statistics or a Machine Learning training job—some information about the data is revealed. The ‘purported ideal’ would be to have a completely anonymized data set that at the same time can be used for sensible data analysis; however, as this is not achievable, the amount of ‘privacy leakage’ must be balanced against the utility of the outputs. Differential Privacy, with its mathematical definition of privacy, makes this fact clear and transparent.

DP is a robust concept that tries to focus on how much information is retrieved from a data query. A particularly interesting feature of this approach is that privacy issues need to be dealt with at the time of initial development, which puts responsibility at the start of development procedure.

One aspect of sustainability and, more specifically, energy efficiency in Machine Learning development, is that it is all too often considered as an organisational issue,

where an organisation that is aligned with sustainability goals, strives to achieve a lower carbon footprint by reducing the energy consumed within its data centres [15]. While this is definitely an important endeavour, the AI development perspective is often overlooked but can contribute significantly to achieving sustainability goals.

As with 'Privacy', from a legal and governance perspective, we have a duty to manage the amount of energy that is expended when AI systems are developed. Akin to 'Privacy Budgets' within DP, we introduce the concept of a "Sustainability Budget" or "Sustainability Score" to allow sufficient oversight of this.

In DP analysis, the amount of information can—and must—be specified before each data query computation. That means, in practice, before an AI developer starts to train a Machine Learning model on a dataset, they need to detail the specified amount of 'information leakage'. The higher the budget that is requested, the more accurate the model will be, but the less remaining room there will be for further computations on the dataset, and vice versa. The privacy problem is put in the hands of the developer and therefore becomes a central element of the analysis itself which leads to nothing else other than a strong requirement for data protection awareness at the development level. The developer must carefully think about how they would like to use the data; once the analysis is started (or more specifically for DP: once the analysis result is published), the decision has been made. In a similar regard, a 'Sustainability Budget' can motivate developers to think about the effects of their to-be-started analyses before they hit the run button.

Let us define the Sustainability Budget (SB) as a virtual upper limit of available compute resources for Machine Learning jobs. Likewise, a 'Sustainability Score' can be defined as the inverse: the to-be-consumed compute resource deducted from a virtual upper limit.

Such a score would not constrain developers and their organisations but instead raise awareness on the side of the developer and incentivize solutions that are less resource-intensive and therefore more sustainable.

To calculate and use this new Sustainability Budget, we need a formula for the carbon footprint of a software system and metrics that can be used to calculate it. A good template comes from the Greensoftware Foundation and their "Software Carbon Intensity (SCI) Specification". An alpha version can be found in [16].

The formula for the SCI rate is defined as:

$$SCI = ((E \times I) + M) \text{ per } R, \quad (1)$$

where E is the energy consumed by a software system, I is the location-based marginal carbon emissions, M is the embodied emissions of a software system, and R is a functional unit.

We can set R to be one ML (training) job and thereby calculate the SCI per the ML process. SCI indicates how much carbon equivalent emissions are produced per kilowatt-hour (energy, E). This must be identified with the local energy provider and data centre administration team. Similarly, the embodied emissions are calculated beforehand based on the employed hardware systems. To quantify the carbon equivalent emissions per unit, one approach is to measure the actual consumption by implementing a respective monitoring at the supplier level. However, there are other 'prediction' approaches, potentially conceptual ML techniques, that could also be conceptual.

Once the initial calculations and measurements are in place, the SCI can be used to compare different ML jobs (units R) and to publish the rates in the team's sustainability reports. On a basic level, with everything else fixed, the SCI comes down to measuring the energy consumption of a ML job. With a balanced set of compute resources this can further be refined to measure the time one ML job is running.

Having such a compute-time or SCI rate measurement in place, AI development teams can now define a Sustainability Budget (or goals for sustainability scores) for different projects. Each developer will be aware of this budget. They will plan for an efficient model training process using as little energy as possible for a model that is as high performant level (accurate, precise, private, etc.) as possible under these conditions. Since model

optimization usually follows a logarithmic curve (model performance vs. iterations), a good approach would be to estimate the expected performance gain with each step, i.e., experiment, that the machine learning developer performs while finding an optimal model. This makes the sustainability score part of the development process itself.

2.2. A Practical Case Study: Using Bayesian Optimization for Experiment Selection

If we look at the technical development process at an individual level, ML development can be described by a series of smaller experiments that are executed to develop one Machine Learning model. In this sense, each ‘experiment’ uses computational resources that can be described with the SCI rate introduced above. All the experiments (units R) need to fit into the given Sustainability Budget for the entire development lifecycle [10]. So, the question becomes how to structure work as an ML engineer, and what the most efficient workflow is to obtain the best model, within a given budget.

The quantity of data and size of the model are the biggest drivers for the computational cost of each experiment [17]. There is work that tries to mitigate this by working with efficient networks [18,19], compression networks [20–22] or a minimum amount of data [23,24]. Not only can this reduce computational cost during development, but it can also solve other limitations, such as computational capacity on edge devices [18,19] or the availability of data. Moreover, ideas such as Transfer Learning, etc. [25] can be used to develop models more efficiently.

These are interesting approaches, but to highlight how a sole developer might drive sustainability in their individual development, we implement Sustainability Budget approach, this section focuses on Bayesian Optimization as a case study, to guide the selection of experiments and reduce the number of experiments needed to reach the desired result.

To understand how this can be done, it is important to realise that each experiment is carried out to gain information about the model’s performance for a specific set of parameters. For the ML engineer, this is the main part of their work, where they want to understand, with high certainty, how high the model’s error is with different settings, in order to determine the best set of settings or to conclude how the different settings influence the results for comparison. The model’s error can be described by a function, with input space over the parameters and the output space over the error of the model, where low errors represent good models. This error function is unknown and cannot be optimized directly, so optimization methods such as Gradient Descent [26] are not applicable and only black box optimization methods [27] or brute force methods (systematically calculating all possible options) can be used.

Naturally, systematic testing of different input sets could be carried out to estimate the error function for different parameters. However, this would not be favourable, because it is not very efficient and grows exponentially with the number of parameters, because each parameter is tested individually. It is very likely that some parameters have a bigger impact on the model’s performance than others, but we do not know which ones beforehand. If known, the most important should be tested first. However, this becomes more complicated when we consider that the parameters are not independent at all, and that randomly selecting parameters is a better option, because it is more efficient in testing more unique combinations, which has also been proven experimentally [28]. To achieve even better efficiency, the information from previous experiments should be used to determine which parameter set it evaluates next. This can be conceptualized with Bayes Formula:

$$P(H \mid E) = P(E \mid H) P(H) / P(E). \quad (2)$$

Bayes Formula of conditional probabilities puts the prior probability $P(H)$ of the hypothesis H in relation with the likelihood $P(E \mid H)$ of the evidence E for a given H . With the prior and the likelihood, we can calculate the posterior probability $P(H \mid E)$, which is what is most interesting. The term $P(E)$ is the marginal likelihood of all hypotheses H and can be treated as a normalisation constant for these purposes.

Applying this to the experiment selection problem of the ML engineer, this method can be used to find the optimal set of parameters for the error function of the model. Due to the fact that the function cannot be accessed directly, it is treated as a probability function. The prior $P(H)$, expressed as a probability function over the parameter space, gives the prior belief of the form of this error function. Prior to any experiments, it can be assumed that a uniform probability density function (pdf) would represent the belief that no set of parameters is advantageous to any other set, but the ML engineer might just as well implement any other prior beliefs that may be held over the model's error function into the equation. The evidence represents the experiments performed and the likelihood of this evidence, which indicates the probability that the evidence fits together with the hypothesis. The posterior can be calculated by combining the likelihood and the prior which can be understood as an update of the prior once new evidence (i.e., information) is present. This process can be repeated when new evidence is present. The prior of the next interaction is set equal to the posterior of the latter interaction, and the new posterior is calculated, i.e., the belief is updated. This represents an update of the belief on the ML engineer (or the system) given new evidence.

To select the best experiment to perform next, the described posterior function (which is often modelled as a gaussian process) can now be used, and the next point to evaluate can be derived (see Figure 1). During the gaussian process, a density estimation is gained for each point on the error function rather than a simple point estimate. In the plot below, the line in the middle is the mean of the estimate and the area around represents the uncertainty, i.e., the density of the estimate. The red dots are points where the function has been evaluated in some experiments, and the uncertainty is therefore zero at these locations.

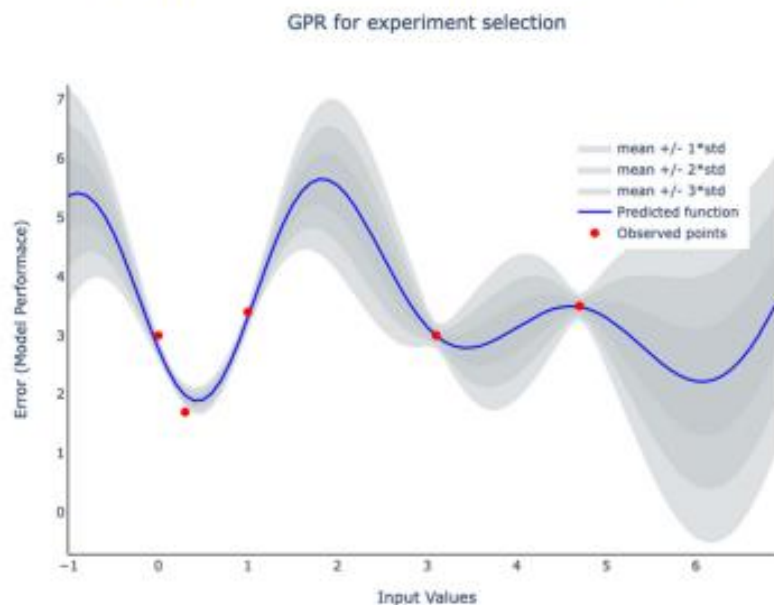


Figure 1. A Gaussian Process Regression (GPR) model that has been fitted on observed data (red dots). In the context of the case study, these observations would be the results of previous experiments. The predicted function (blue) is the result of the GPR and is simply the mean of the GPR posterior estimation. The grey area visualises the uncertainty of this function. If the parameters are desired (note that really the number of parameters is much greater than 1), in the next experiment the plot can be used to find a good trade-off between exploration and exploitation. What is seen is that the dip around 0.5 is certain, but further development would likely result in limited performance gains. Alternatively, the area around 4 or 6 could be explored, with high uncertainty, to improve the current estimation and avoid missing a better local optimum for the model performance.

Now, this provides the information to choose which point of the error function to evaluate next. Depending on the task, the current optimal point could be evaluated or a point with large uncertainty to improve the estimate there.

This strategy is an example of how to make informed decisions while optimising a machine learning algorithm. With every step, the next best experiment is decided. Therefore, the cost per unit as described above is clearly present in the form of the number of experiments to be performed. It is a good idea to keep this process a manual one (instead of trying to automate away the exploration) as it brings the desired care into the process. The ML developer is aware of the sustainability costs associated with each experiment and therefore uses a carefully chosen method to optimise the model, integrating and updating beliefs with each iteration.

3. 'Gamify!': Gamification Techniques to Manage Sustainability

We introduced Sustainability Budgets and have offered a practical way developers might use up-to-date ML techniques to achieve optimum sustainability. With these in mind, let us ask how developers might be incentivised to achieve more sustainable development. For example, while a developer might have awareness of how much energy the particular development of their system expends and the capability to minimise it, without some form of incentivisation, the whole procedure may seem pointless. Why invest time and energy in making a system more efficient if there are no benefits for doing so?

While we believe this to ultimately be a management issue, where managers decide how to maintain sustainability, Sustainability Budgets open the possibility to use 'Gamification', which has been shown to be successful in incentivising participation.

Gamification—a notion taken from the gaming industry—is the enrichment of products, services and information systems with game style elements in order to positively influence the productivity, motivation and behaviour of users [29]. An early example was the Nike+ running application [30], which encouraged individuals to engage in physical activity by making a 'game' out of running. Users were encouraged to compete against other runners for positions on leader boards, 'fastest course completions' or simply 'most runs that week'. More recently, other exercise applications such as Strava and Peloton make use of the gamification technique, and this can be extended to other areas.

Sustainability Budgets lend themselves very well to gamification techniques, with 'Sustainability Scores' or Limits usable to drive competition between developers. Therefore, a developer might be incentivised to develop their system in the most efficient way, because they are part of a 'game' with other developers, where there is a benefit for 'winning' (whether this be financial or otherwise).

As well as lending itself to competition between developers, moving one abstraction level away, we can begin to look at how this might also translate to operating at an organisational level.

As mentioned in Section 1, many organisations are motivated to reduce their carbon footprints, but this often seems to weigh against certain costs, meaning that there often has to be a compromise on sustainability in order to drive down the costs of development. However, if gamification were used as a technique to not only incentivise developers, but also to audit how systems were developed and their efficiency (through a Gamification management system), it would be easier for organisations to understand and explain the added time and costs for the delivery of projects. Ultimately, gamification systems would not only be instrumental in incentivising developers, but they could also be used by organisations to show why they are particularly sustainable.

It could also be applied from an executive management perspective, and it could even be used at a national and global governance level. There could be competitions between teams of an organisation or even across different organisations or countries.

Utilising the 'Sustainability Budgets' methodology, there will be a mechanism to plan for and log the amount of energy expended per development task and also information for developers to provide explanations at a management level for different components of

the development process. For example, a developer would be able to articulate why they chose a particular training step or method over another, in terms of the balance between sustainability and effectiveness, and there would be oversight as to how much energy was expended per process. The developer would be incentivised to be more energy efficient in their development approach, and there could be incentivised competitions between different developers to find the most 'energy efficient', effective approach.

The incentivisation could be extended to include 'games' between different developers. From a management level, this would give the chance to reward more sustainable development through a points/score-based system. There would need to be a balance between sustainability and effectiveness of the algorithms, but a score-based system would allow visibility of the compromises and management accordingly. For example, as a manager, I might decide to increase my sustainability budget (i.e., the amount of energy a development project expends), because the project is particularly pertinent (from a business perspective) or because having a lower budget would make the system unsafe (i.e., in a medical context). Equally, I might decide to lower my budget because, from a resource perspective, the project allows for this. Gamification would encourage developers to create their systems in the most energy efficient ways, being rewarded for creative technical solutions.

4. Using Sustainability Budgets to Achieve the SDGs

Eventually, the proposed method might also contribute to wider political and societal goals. At the beginning of this paper, the Sustainable Development Goals (SDGs) were discussed, in particular, Goal 13 and the drive for countries to achieve carbon neutrality, or at least reduce their carbon footprints. As well as offering a mechanism for incentivisation through Gamification (or similar) approaches, at either a developer, management or governance level, what Sustainability Budgets ultimately provide is a mechanism for 'Energy Consciousness' insofar as not only do developers become more aware of the amount of energy the development of their system expends, but managers can translate this into development action, and governments can reward/recognise organisations who offer more sustainable solutions or govern organisations to be more sustainable.

Strategically, those systems that offer the greatest benefit/cost ratio would be the systems that are invested in the most. Where previously it was difficult to determine this ratio in the early phases of a project because there was no mechanism to quantify it beforehand, there can be effective management of energy expenditure at all levels of an organisation: a company, for example, but also a sector and even a nation.

For example, energy budgets could be set on an industry basis, based upon the needs of the industry. There could be regulation to ensure that industries did not go beyond their allowance, and the different levels could be managed at a political level. Organisations could be incentivised to be sustainable through energy ratings that would ensure that standards were being maintained. An organisation might receive an 'A' energy rating for particularly energy efficient processes and a lower rating ('B') for a lower energy efficiency. As with energy efficiency levels used to rate buildings and appliances (ref), the energy expended for each project would need to be measured against the size and utility of the project.

A strong contribution of the Sustainability Budget approach is to provide a way to embed the topic of environmental sustainability into ML practices and organisations from the beginning. The urgent need to address climate change is injected directly in the project development phase of machine learning and AI applications. Whereas most of the time such topics are managed purely in an abstract way with review processes that are being applied after the fact and that are followed or not, sustainability budgets put this matter at the heart of the development process itself. With the aforementioned gamification implementation, there is a direct and accessible way to address these problems.

And most of all: sustainability budgets raise awareness about this topic. Next to incentivisation, sustainability budgets and the related gamification strategies are likely to

raise awareness of energy consumption and sustainability issues, which may influence the behaviour of individuals and the organisation. Just like privacy budgets in differential privacy, sustainability budgets need to be handled before starting a training run of a machine learning task and, because of that, they force developers, project managers and policy makers to think about the impact of the project early on and continuously. It does matter if a machine learning task is solved by using the next best large network trained on all available next best data or rather designed and developed with energy consumption and sustainability in mind, carefully selecting the optimal network architecture, data subset and iterative development approach. Being aware, after all, is the first and most important step when trying to change and influence things. Sustainability budgets help raise awareness of the climate change impacts Machine Learning development can have at and before the development process itself.

5. Limits of the Methodology

It is important to note that the notion of ‘budgeting’ energy consumption has its limitations. As with the concept of budgeting privacy, careful calculations need to be carried out to weigh up the balance between the amount of energy that should be expended and how useful it would be to expend that energy. If the correct ratio is not established, then it becomes useless to set sustainability limits in the first place. If energy limitations are hindering innovation and project development, then the budget needs to be reassessed and project priorities need to be reassigned. This, however, is something that would need considering at the management level.

Another limitation of the presented approach is the extent to which the privacy budget metaphor extends to the case of energy budgets. It is noted that there are some dissimilarities between the two notions, namely ‘privacy’ as a unit of measurement is an abstract concept whereas ‘energy’ has real tangible values (i.e., the amount of energy that is consumed). Differential Privacy is a mathematical concept, whereas sustainability budgets stem from an organisational or management background. However, focusing on such dissimilarities distracts from the focus of the approach that has been detailed, which is an emphasis on creating an awareness of how sustainable a project might be and putting in a framework to mitigate or manage this. In terms of the methodological approach and its consequences, namely the required attention to the data protection and sustainability factors, respectively, in the development process itself, there are useful similarities that can be taken from the management of differential privacy to encouraging sustainability.

Finally, one key limitation of the approach detailed in this paper is the extent to which this is an implementable process. What has been given is a conceptual approach and a sample case study, but by no means is it ready to be placed into the software development process. Ultimately, we see this as an area for future development, and if Sustainability Budgets would be pursued, an area that would require further research. The following areas need developing further:

- Calculation of the appropriate ratio for the performance vs. cost.
- Investigation for how organisations ought to be governed given energy efficiency levels.
- Development of a suitable gamification platform on which to record (and provide a game for) development energy consumption.

6. Conclusions

In this paper we offered a technique for managing sustainability in the development process of AI system creation, which utilises a notion we term ‘Sustainability Budgets’. This both empowers the developer creating the AI system and allows the management and governments to have appropriate oversight. The notion of ‘games’, whereby developers can engage in competitions to achieve the most sustainable product vs. its effectiveness, is introduced as a possible incentive to encourage developers to achieve their organisational and a nation’s sustainability goals. The conversation in this paper not only offers a new methodology for individuals to move closer towards achieving the SDG’s, it also inspires

debate in this area, and this may lead to even more practical ways to ensure that Artificial Intelligence aids (rather than hinders) a sustainable future.

Author Contributions: Conceptualization, J.B., R.R. and M.C.; methodology, J.B., R.R., M.C. and W.G.; validation, W.G., C.N.L. and P.C.; formal analysis, R.R.; investigation, R.R. and W.G.; resources, W.G. and J.B.; data curation, W.G.; writing—original draft preparation, R.R., M.C., J.B., W.G., P.C. and C.N.L.; writing—review and editing, R.R., M.C., C.N.L., P.C. and J.B.; visualization, W.G.; supervision, R.R. and M.C.; project administration, R.R.; funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partly funded by the FFG, “Österreichische Forschungsförderungsgesellschaft” (Austrian Research Funding Institution) as part of the research project “Ethische KI”, funded as an FFG “Basisprogramm” and by Gradient Zero. Open Access Funding by the University of Vienna.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Open Access Funding by the University of Vienna.

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

References

1. UN News. ‘No Time to Lose’ Curbing Greenhouse Gases: WMO. 2021. Available online: <https://news.un.org/en/story/2021/10/1103892> (accessed on 13 December 2021).
2. UK Met Office Online. Effects of Climate Change—Met Office. 2021. Available online: <https://www.metoffice.gov.uk/weather/climate-change/effects-of-climate-change> (accessed on 13 December 2021).
3. NBC News. Heat Wave 2021: Climate Scientists Warn about a New Normal. 2021. Available online: <https://www.nbcnews.com/science/environment/heat-wave-2021-climate-scientists-warn-new-normal-rcna1664> (accessed on 21 March 2022).
4. The Guardian. Fires Rage around the World: Where Are the Worst Blazes? 2021. Available online: <https://www.theguardian.com/world/2021/aug/09/fires-rage-around-the-world-where-are-the-worst-blazes> (accessed on 25 March 2021).
5. BBC News. Germany Floods: Dozens Killed after Record Rain in Germany and Belgium. 2021. Available online: <https://www.bbc.co.uk/news/world-europe-57846200> (accessed on 25 March 2021).
6. Globalgoals.org. The Global Goals. 2021. Available online: <https://www.globalgoals.org/> (accessed on 13 December 2021).
7. Stein, A.L. Artificial Intelligence and Climate Change. *Yale J. Reg.* **2020**, *37*, 890.
8. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019.
9. Schwab, K. The Fourth Industrial Revolution. 2017. Available online: <https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab> (accessed on 22 March 2022).
10. Wynsberghe, A.V. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 1–6. [CrossRef]
11. Coeckelbergh, M. AI for Climate: Freedom, Justice, and other Ethical and Political Challenges. *AI Ethics* **2021**, *1*, 61–72. [CrossRef]
12. Regulation (EU) 2016/679 of the European Parliament and of the Council, General Data Protection Regulation. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02016R0679-20160504&from=EN> (accessed on 23 February 2022).
13. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the TCC 2006: Theory of Cryptography Conference, New York, NY, USA, 4–7 March 2006; pp. 265–284.
14. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
15. Gao, J. Machine Learning Applications for Data Center Optimization. 2014. Available online: <https://www.google.com/url?q=https://research.google/pubs/pub42542/&sa=D&source=docs&ust=1645625964864718&usq=AOvVaw2sF6awAp8KZDfvcvpuLU5i> (accessed on 23 February 2022).
16. Hussain, A.; Gupta, A.; Time, H.-W.; Buchanan, W.; Bergman, S.; Knight, V.; Lloyd-Jones, C.; Srinivasan; Kariya, M.; Lewis-Toakley, D. Software Carbon Intensity (SCI) Specification (v.Alpha). 2021. Available online: https://github.com/Green-Software-Foundation/software_carbon_intensity/blob/main/Software_Carbon_Intensity/Software_Carbon_Intensity_Specification.md (accessed on 2 February 2022).
17. Justus, D.; Brennan, J.; Bonner, S.; McGough, A.S. Predicting the computational cost of deep learning models. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 3873–3882.
18. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1–9.

19. Howard, A.; Menglong, Z.; Chen, B.; Kalenichenko, D.; Wang, W.; Wayand, T.; Andreeto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
20. Han, S.; Mao, H.; Dally, W. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv* **2015**, arXiv:1510.00149.
21. LeCun, Y.; Denker, J.; Solla, S. *Advances in Neural Information Processing Systems*; Massachusetts Institute of Technology Press: Cambridge, MA, USA, 1989; pp. 598–605.
22. Tanaka, H.; Kunin, D.; Yamins, D.; Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *arXiv* **2020**, arXiv:2006.05467.
23. Arora, S.; Du, S.; Li, Z.; Salakhutdinov, R.; Wang, R.; Yu, D. Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks. *arXiv* **2019**, arXiv:1910.01663.
24. Tartaglione, E.; Barbano, C.A.; Berzovini, C.; Calandri, M.; Grangetto, M. Unveiling COVID-19 from chest X-ray with deep learning: A hurdles race with small data. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6933. [[CrossRef](#)] [[PubMed](#)]
25. Donahue, J.; Yangqing, J.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 647–655.
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
27. Audet, C.; Kokkolaras, M. Blackbox and derivative-free optimization: Theory, algorithms and applications. *Optim. Eng.* **2016**, *9*, 100011. [[CrossRef](#)]
28. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
29. Deterding, S.; Dixon, D.; Khaled, R.; Nacke, L. From game design elements to gamefulness: Defining “gamification”. In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, New York, NY, USA, 28–30 September 2011; pp. 9–15.
30. Blohm, I.; Leimeister, J.M. Gamification. *Bus. Inf. Syst. Eng.* **2013**, *5*, 275–278. [[CrossRef](#)]

EKIP: designing a practical framework for embedding ethics in AI software development

Rebecca Raper¹ and Mark Coeckelbergh¹

¹ The University of Vienna, Vienna, Austria
rebecca.raper@univie.ac.at

Abstract. As Artificial Intelligence (AI) systems are used more within society, there is the pressing need to ensure that they are developed in an ethical way.

Several notable attempts to ensure that AI is developed ethically have been advanced. Though each approach has its merit in terms of offering guidance for ethical AI development, none, yet offer a rigorous, practical approach to ensure that AI systems are developed in the right way. Furthermore, these approaches seem to be out of sync with current actual software development methodologies.

EKIP is a three-year-long FFG-sponsored project which aims to address this issue by offering a *practical* framework for software developers. Inspired by *Ethical by Design* (Mulvenna et al., 2017) – that ethics should be placed into the design phase of software development – EKIP’s purpose is to explore more practical approaches for AI software developers to embed ethics into their systems.

The aim of this paper is to outline the motivation behind EKIP, by giving an overview of the attempts that have been made to embed ethics in AI so far, and by introducing a ‘methodological gap’. It is argued that if we are to effectively embed ethics in AI systems, we need to consider new frameworks that integrate with the traditional software development techniques, such as ‘waterfall’ and ‘Agile’. An *ethical requirements*-based approach is described which presents a twist on traditional software development methodologies. To conclude, suggestions are made for future work, and there is a plea for further research to advance this topic.

Keywords: AI Ethics, AI Frameworks, Ethical AI, EKIP, Software Development Lifecycle, Software Development, Waterfall, ML-Ops, Kanban, Requirements Engineering

1 Background and Introduction: Ethical AI

Defining *Ethical Artificial Intelligence* (AI) as the pursuit to design, create and implement AI systems that adhere to ethical standards (Coeckelbergh, 2020), the discipline has gained an increasing amount of traction in recent years. This is partially owing to various high-profile incidents concerning the use (or misuse) of AI systems. For instance, in March 2020, it was reported that Cambridge Analytica was collecting and processing vast quantities of data to form decisions about the types of advertisements an individual might be targeted with, or what articles they see online (BBC News, 2020). The controversy and objections concerned the fact that AI profiling was being used to target and manipulate individuals with material relating to a political referendum campaign, highlighting the vast amount of influence AI decision-making can have. Also in 2020, A-Level result allocation in the UK had a similar story, with algorithms used to assign grades to students showing to be prejudiced against those from certain demographics or poorer backgrounds (The Guardian, 2020) – the implication being that if you were from a poorer background, you were less likely to get into the university of your choice – ultimately reducing social mobility and career fulfilment. There have been similar cases in the United States, where AI profiling used by the police to forecast criminal behaviour, has been shown to be prejudiced against those from minority groups (Wired, 2020). Similarly, facial recognition technology used in procedures such as interviews, or for identity verification, has been shown to be less effective for users with darker skin (The Independent, 2021). As AI is being used more and more in society, to deal with economic, social, and technological problems, there is the concern that these identified types of incidents are widespread and that they are having a significant effect on individual lives and communities.

Other implications of AI systems concern issues such as *privacy* – if an AI system is making inferences about our personality characteristics, to what extent do we still have privacy? Another issue is *trust* – with algorithms having such bad press (whether this be qualified or not), how can (and should) public perception of AI be developed so that they can realise the true benefits that AI can bring? *Autonomy* and *control* are other issues – who gets to decide how AI algorithms are applied, and then decide the eventual fate of individuals based upon the algorithmic decisions? This then leads to questions of delegated *power*, and *political influence* (Coeckelbergh, 2022). There is also a piece for *education*. With AI systems affecting everybody's lives (in some way), and with Artificial Intelligence itself due to become increasingly powerful (Bostrom, 1998) everybody should be equipped to understand how it works and the impact it has on their day-to-day lives, so they can appropriately interrogate, ask questions and engage in social and political debates.

It is these such issues that have led to calls for more Ethical AI and have led to increased academic attention to the area. Institutes such as 'The Ethical AI Institute' in the UK, and 'The Institute for Ethical AI and Machine Learning' in Germany, have been established in recognition of this problem - the idea being that interdisciplinary collaboration (i.e. between science and the humanities (Tasioulas, 2021)) will help to resolve some of these issues. There are calls for immediate action to be taken.

2. Attempts to deal with the issues so far

Various attempts have been made to address the ethical issues pertaining to AI, without the need to call a moratorium on AI use/development in general. One of the first attempts made was to create a set of guidelines or principles for AI developers to follow. Different bodies have different principles (see Floridi and Cowls, 2021) but broadly speaking, the principles can be reduced to themes such as *Transparency*, *Trust*, *Autonomy* and *Explainability*, namely, an AI is only deemed ethical if it matches these principles - so is Trustworthy, Grants Human Autonomy etc. In line with the principles, Ethical Standards have been developed to highlight what needs to be done to ensure that the principles are adhered to. For instance, a system might only be regarded as 'explainable' if it can be explained to stakeholders in a certain way. Under this approach, only systems that meet these standards can be said to be ethical, and therefore worthy of approval. The problem is: with such standards in place, it's difficult for developers to understand (without extensive training) (a) what the specific standards mean or (b) what *actually* needs to be done to meet them. Requests such as 'explain to stakeholders' can be ambiguous when notions such as 'explainability' are still very much open to philosophical debate (Arrieta et al, 2020). With the very pressing need to ensure that AI systems are developed in the right way now (rather than in 10 years' time after the terms have gone through rigorous philosophical discussion) another methodology is required that prescribes to developers how they can create their AI systems to meet such standards.

'Ethical by Design' (Mulvenna et. al, 2017) is one such methodology that has been proposed to ensure this, with the suggestion being that ethics be integrated into the AI software development process from the very start. This gained a significant amount of traction, with the IEEE adopting an 'Ethically Aligned Design' (How, 2018) manifesto to aim to achieve this. However, how to create AI systems that meet the various list of standards is still left very much open. Still – developers know they must do something to make their systems more ethical, but it is not obvious what.

One proposal - ‘Value-sensitive design’ (see Umbrello, 2018, Friedman, 2004 and Van den Hoven, 2007) - aims at determining a set of values at the beginning of the development process from which to work upon. Stakeholders (and end users) undergo a series of interviews to assess *what is important to them*, and then this information is used to inform the design of the AI system.

However, there are problems with such an approach. Namely, no end of user interaction will be able to fully capture every individual value, and even if they were, how do we then deal with the fact that individual values may be conflicting? An additional mechanism is required on top of this approach to determine which values take priority, and without strict oversight, this then faces the possibility of becoming nothing more than a series of prejudices. More fundamentally, however, the Value by Design approach neglects the foresight and academic rigor that has gone into the development of principles and standards such as those developed by the IEEE in the first place. There needs to be some way to bridge the gap between AI system design and the need for adherence to principles such as transparency, fairness etc.

More recently, AI Auditing (Mokander and Floridi, 2021) has been offered as a verification method to ensure that AI systems meet ethical standards previously set out. So, AI systems can be assessed against a checklist of criteria for satisfaction of the standards. However, though a good mechanism to assess that AI systems are doing what they should, there is still a *methodological gap* insofar as something is required to outline to developers how to make their AI systems ethical.

3. A ‘methodological gap’ in Ethical AI

In practice, we can see that we have *top-down* principles or standards that prescribe how AI systems ought to be developed (i.e. with *transparency* etc). We also have *horizontal* methodologies that advise that such principles should be incorporated into the AI system at the start of AI system creation (i.e. Ethical by design). We also have a technique to assess that AI systems have been created at the appropriate ethical level (AI auditing), and some techniques to show us that stakeholder values are important in the design of AI systems. Schematically, we can represent the interplay of all the different approaches by a diagram as per Figure 1:

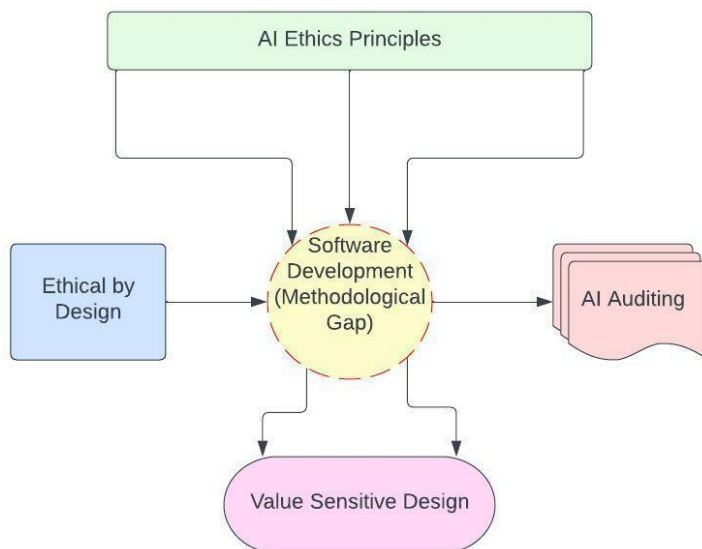
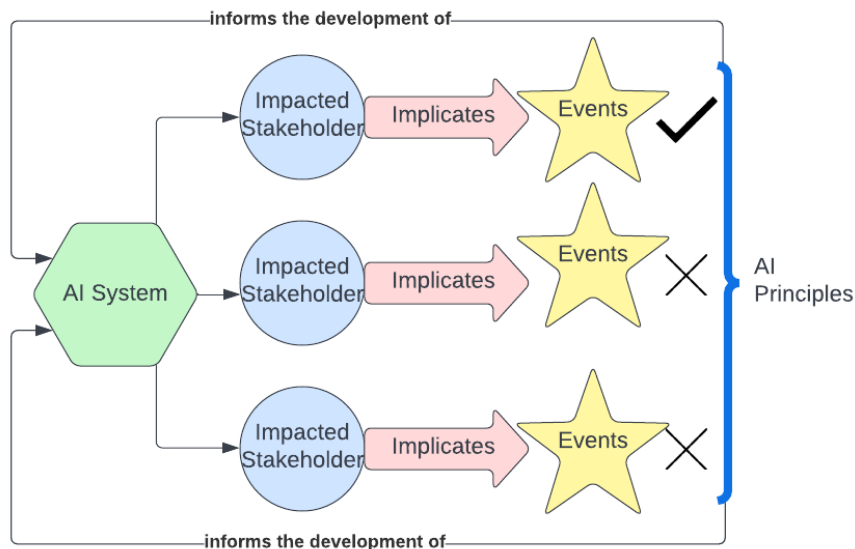


Fig. 1. The current state of AI Ethics

We can see that although we have mechanisms to provide oversight of AI systems (namely, AI Ethical Principles), and a technique to prescribe when we ought to be incorporating ethics into our systems (Ethical by Design), and how to assess them (AI Auditing), there is still space to understand how we ought to be developing the AI systems to ensure that both the principles are adhered to, and that the relevant appropriate standards are met. **This** is the methodological gap.

Describing this problem in terms of a ‘methodological gap’ allows us to begin to envisage what needs to be done to rectify the problems associated in the first chapter of this paper: namely, AI bias, discrimination, privacy and broad impacts to humans. What is required is a methodology that takes note of the already determined AI principles (i.e. that it’s important not to infringe upon autonomy), assess these against real-life potential impacts to stakeholders (i.e. my autonomy might be infringed if AI decides what I’m watching on TV tonight), and then puts mechanisms in place to protect against these. Schematically again, this need can be represented as per Figure 2.



current Ethical AI techniques.

Fig. 2. A need for integration of

As can be seen from the diagram, this approach still appeals to the ‘Ethical by Design’ methodology, since what ultimately informs the development of the AI system, is either an infringement or adherence to AI Ethical Principles (indicated by the cross or tick on the diagram): the system being ethical insofar as it adheres to Ethical Principles (such as those previously discussed). However, the way in which this approach differs is that the principles act as a *guide* as opposed to enforcement for how AI systems ought to be developed. In other words, the AI developer’s autonomy is preserved, since they are not forced to develop their systems in a certain way (See Raper et al, 2022, for some benefits of preserving developer autonomy). Up until recently, a top-down regulation-style approach has been used (see Madiega, 2019), but it has been criticised for potentially inhibiting innovation. The new approach allows AI systems to be developed ethically, whilst preserving AI developer autonomy.

In understanding this approach, it’s important to realise that it is not an impact-based model. There have been recent discussions about using impact assessments to determine how to develop an AI system (Kazim, 2021) (the ethical system being that which does not lead to negative consequences). Ultimately, however, it is not possible to know the full impact a system might have until it has been developed and put into place. We might reflect afterwards (based upon impact) that an AI system should not be the way it is (i.e. the negative consequences of the Cambridge Analytica situation), but this approach would be *reactive* as opposed to *proactive*. We want to ensure that there are no negative consequences for the AI systems that are developed, and therefore, that they are ethical before they are integrated into the world.

As opposed to assessing an AI system based upon its (previously unknown) impacts, the AI system is assessed according to its (previously known) implications. The implications can be determined through case studies, interviews and workshops with impacted stakeholders. For example, I might assess how a new medical diagnostic device will implicate the end stakeholders (i.e. patients) by asking a series of questions to determine how the system (in principle) will affect their lives. A conversation such as the following might be used in a user workshop to ascertain this:

AI Analyst: We are developing an AI tool to assess how likely somebody is to develop diabetes based upon their current diet and lifestyle.

Patient: That sounds good, but some days I am healthy, other days I eat unhealthy because it depends upon how busy I am...

This is a very short interview, but it highlights the types of conversations that need to be held to truly ascertain the human implications an AI system might have. As this example demonstrates, after discussion with a patient we see that such a tool would need to take account of variable diets and lifestyles (since some patients eat mixed diets), therefore, we are able to determine that there is a requirement to ensure that any AI system used to predict diabetes risk, considers diets that change. After significant further analysis, we begin to develop a requirements-based model for Ethical AI Development of this specific system.

There are many ways to satisfy the requirements specified during the analysis phase, but what is important to note is that doing so should be an exercise in *requirements-based design*, rather than harm/bias etc. mitigation. In this instance, the new system has the need to accommodate for changing diets. It then becomes the AI developer's creative agency to design a system which satisfies this.

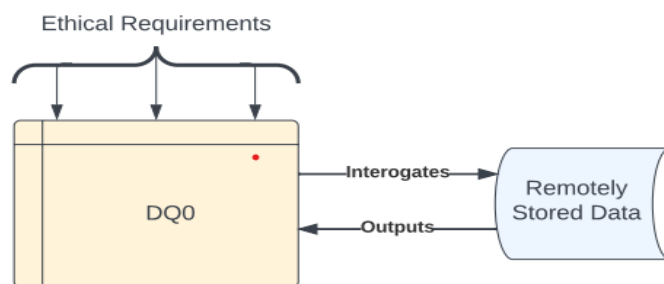
5. EKIP: A practical methodology for Ethical AI Software Development

The requirements-based approach to Ethical AI software development is one that has been proposed previously (Guizzardi et al, 2020), however there has not yet been a thorough articulation for why this approach is needed, how it significantly differs to current methodologies, and – most importantly - what an implementation of the solution might look like.

EKIP (shorthand for “Ethische KI Plattform” = “Ethical AI Platform”) - an Austrian, FFG-sponsored project - was established in an attempt to provide a very practical way to integrate ethics into software development by proposing that ethics be integrated into an already-established AI development platform created by the company Gradient Zero, called ‘DQ0’.

DQ0 offers a privacy-preserving platform for data scientists to develop Machine Learning algorithms to run queries against sets of (sensible) data. The aim of EKIP then is to extend the platform so that when AI developers create their systems, it facilitates creation of them in an ethical way.

Referring to the Ethical Requirements Framework approach detailed in the previous section, therefore, DQ0 best enables this facilitation (and in turn, preserves developer creativity and agency) by providing a platform to integrate the elicited requirements. Therefore, the most effective way to practically integrate ethics that resolves the ‘methodological gap’ is to create a new software development process that allows for *Ethical Requirements Elicitation*. Figure 3 shows how this should be done in respect to the DQ0 platform.



As mentioned previously, the ethical requirements would be ascertained through case studies, interviews, or general analysis techniques. For instance, in the (brief) conversation scenario outlined above, we determine that the ethical requirement is that: ‘The AI system must accommodate for variable diets’, with the development *problem to be solved* being left to the freedom of the AI developer. It may be that a new role is required within a project team to satisfy this new type of requirement – an individual with the capacity to facilitate such workshops and map implications for affected stakeholders. The key to this phase is in ensuring that the analysis assesses stakeholder implications rather than just looking toward impacts.

6. Integrating ethics into pre-existing software development methodologies (i.e. Waterfall and Kanban)

It is important in understanding how to execute this framework, that it integrates well into the pre-existing software development processes, and that it can also be distinguished from these. Where we are able to know the outcomes of a traditional piece of software, when it comes to AI software development, because AI acts autonomously, the outcomes are not always known.

Previously, before AI software development became more commonplace, methodologies such as ‘The Waterfall Method’ (see Figure 4) were traditionally used to ensure that software systems were designed according to the needs of the customer (or business). Traditionally, this included several phases, the first being ‘project scoping’ and ‘requirements elicitation’, where it was determined (1) how useful the new system might be and then (2) what design requirements the system should have to satisfy the overarching business objectives. Though The Waterfall Method has evolved since its initial inception (i.e. more agile software development approaches such as Kanban (Huang, 1996) are now frequently used), fundamentally, the same procedural steps remain (albeit in a more dynamic way). However, though still useful for broad software development, these are insufficient for AI software development. Because an AI system acts on its own, it is not possible to design it to meet specific conditions (i.e. have a blue screen) - since these will be variable. Instead, mechanisms are required to ensure that the AI system acts in an *ethical way* - hence it must meet ethical design requirements.

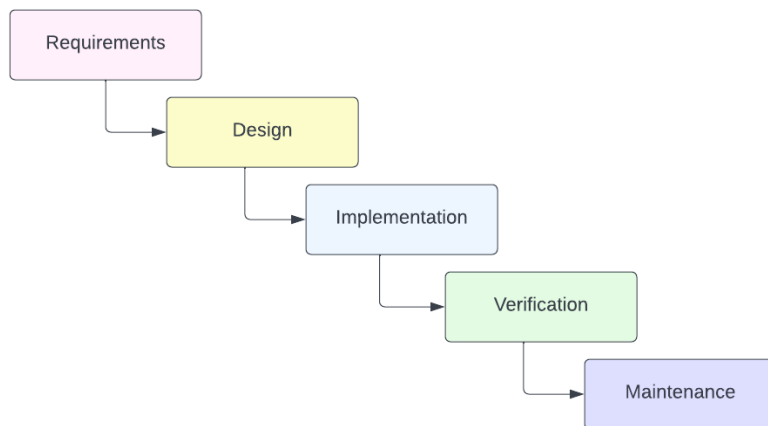


Fig.4. The ‘Waterfall’ software development methodology

Considering ethical requirements elicitation alongside traditional software development approaches, therefore the new process for Ethical AI Software Development should look as per Figure 5.

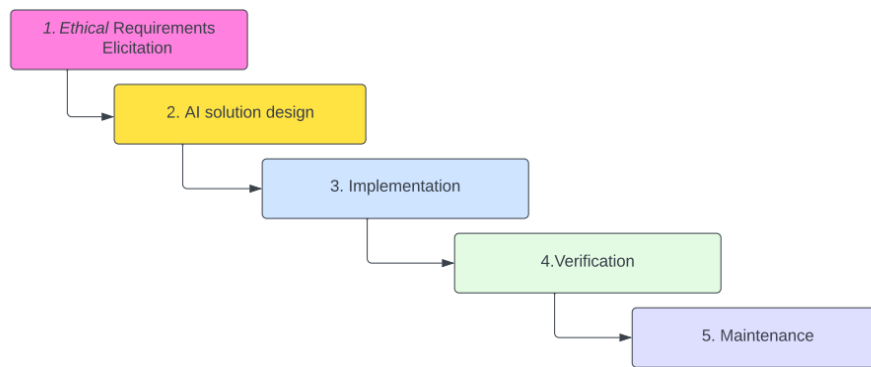


Fig.5. The Ethical AI Software development cycle.

Though the models are very similar, as already mentioned, the type (and methodology) for the requirements elicitation (and thereby solutions design) are very different. In Ethical AI case, workshops are used to find out not just what needs to be done to meet business needs, but what needs to be done to ensure that human beings have a fulfilling and flourishing life. As already mentioned, this might require a new role within the project team – a psychologist or social scientist? – but it is a twist on the original requirements elicitation framework.

As with traditional software development, it is important that there is still verification of the AI system in the end (i.e. it should match the earlier prescribed requirements) along with constant re-evaluation and maintenance of the AI system to ensure issues such as ‘AI drift’ (the susceptibility to later become biased) don’t creep in. In the case of AI systems this means periodic checking to ensure that the system still meets its initial ethical requirements (or requirements for the current stakeholders) and an iterative design process to constantly improve and evaluate the AI system.

As a final point, it is worth mentioning processes that have already been established to try to address the differences between traditional software and AI software development. For example, Microsoft’s ‘ML-Ops’ (Alla and Adari, 2021) was designed to accommodate for these very differences. However, though offering an agile environment for AI development, they neglect the need for ‘human-centricity’ in AI software development. This can only be achieved through thorough analysis of humanity and human lives.

7. Conclusion

Much current academic literature and industry conversations talk about integrating ethics into software, but few practical methodologies are available to do this in specific software development and business contexts in a way that is useful to developers and respects the way they work and solve problems. In this paper we have described a motivation for, and given a broad description of the EKIP project, which enables us to work on precisely this problem through exploring how ethics can be integrated into software development methodologies in the context of an AI platform.

We propose that the best approach to achieving the Ethical AI aims set out by bodies such as the IEEE and The Ethical AI Institute, is by integration of an *ethical requirements gathering* technique into pre-existing software development processes.

In terms of next steps, to progress this research further work is required to develop the finer details of the ethical requirements elicitation process. There need to be tools to facilitate this methodology as well as specific roles and responsibilities. There is also space for further discussion about what constitutes a fulfilling, flourishing life, in the context of Technology Ethics and Ethical AI more broadly, and a plea to work more towards this *practical* type of approach.

References

- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, pp.82-115.
- Alla, S. and Adari, S.K., 2021. What is mlops?. In *Beginning MLOps with MLFlow* (pp. 79-124). Apress, Berkeley, CA.
- BBC News, 2020, ‘Your data and how it is used to gain your vote’: <https://www.bbc.co.uk/news/technology-54915779> (Last Accessed: 27/04/2022)

- Bostrom, N., 1998. 'How long before superintelligence?'. *International Journal of Futures Studies*.
- Coeckelbergh, M., 2020. *AI ethics*. MIT Press.
- Floridi, L. and Cowls, J., 2021. A unified framework of five principles for AI in society. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 5-17). Springer, Cham.
- Friedman B, 2004. Value sensitive design. In: Bainbridge WS (ed) *Berkshire encyclopedia of human-computer interaction*. Berkshire Publishing Group, Great Barrington
- Guizzardi, R., Amaral, G., Guizzardi, G. and Mylopoulos, J., 2020, May. Ethical requirements for ai systems. In *Canadian Conference on Artificial Intelligence* (pp. 251-256). Springer, Cham.
- How, J.P., 2018. Ethically aligned design [From the Editor]. *IEEE Control Systems Magazine*, 38(3), pp.3-4.
- Huang, C.C. and Kusiak, A., 1996. Overview of Kanban systems.
- Kazim, E., Denny, D.M.T. and Koshiyama, A., 2021. AI auditing and impact assessment: according to the UK information commissioner's office. *AI and Ethics*, 1(3), pp.301-310.
- Madiega, T.A., 2019. EU guidelines on ethics in artificial intelligence: Context and implementation.
- Mökander, J. and Floridi, L., 2021. Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, 31(2), pp.323-327.
- Mulvenna, M., Boger, J. and Bond, R., 2017, September. Ethical by design: A manifesto. In *Proceedings of the European Conference on Cognitive Ergonomics 2017* (pp. 51-54).
- Tasioloas, J. (2021) The role of the arts and humanities in thinking about artificial intelligence (AI) | Ada Lovelace Institute
- The Independent, 2021, 'How racist robots are being used in recruitment': <https://www.independent.co.uk/news/world/americas/robots-racism-algorithms-job-hiring-b1860835.html> (Last Accessed: 27/04/2022)
- The Guardian, 2020, 'The Guardian view on A-Level algorithms: failing the test of fairness.': <https://www.theguardian.com/commentisfree/2020/aug/11/the-guardian-view-on-a-level-algorithms-failing-the-test-of-fairness> (Last Accessed: 27/04/2022)
- Raper, R., Boeddinghaus, J., Coeckelbergh, M., Gross, W., Campigotto, P. and Lincoln, C.N., 2022. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. *Sustainability*, 14(7), p.4019.
- Shahriari, K. and Shahriari, M., 2017, July. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)* (pp. 197-201). IEEE.
- Umbrello, S. and De Bellis, A.F., 2018. A value-sensitive design approach to intelligent agents. *Artificial Intelligence Safety and Security* (2018) CRC Press (.ed) Roman Yampolskiy.
- Van den Hoven MJ, 2007. ICT and value sensitive design. In: Goujon P et al (eds) *The information society: innovation, legitimacy, ethics and democracy*. Springer, Dordrecht, pp 67–73
- Wired, 2020, 'Police built an AI to predict violent crime: it was seriously flawed': <https://www.wired.co.uk/article/police-violence-prediction-ndas> (Last Accessed: 27/04/2022)

Acknowledgements

As this is an FFG-sponsored project, we would like to thank our funders for giving us the opportunity to develop this programme of work. We would also like to thank the team at Gradient Zero who have been keen and enthusiastic partners in this pursuit, in particular Jona Boeddinghaus who gave guidance and insightful comments on a late version of this paper.

Imagination, Morals and Machines: The role of imagination in moral decision-making and its implications for the project of creating moral machines

1. Introduction

Moral Machines (that is, machines with the capacity to tell between *right and wrong* (Wallach and Allen, 2008)) have recently received an increasing amount of attention. Where the challenge is to create a machine that can reason in moral situations, this has proven to be a difficult feat, with recent attempts resulting in controversial outcomes (ref: *Delphi* (Jiang et al., 2021)). The challenge of creating machines with moral capacity - we assume - is related, but distinct from the challenge of ensuring machines are designed (created and integrated) in an ethical way (Moor, 2006), which for the purpose of this paper will be left under the term *Ethical AI*.

There have been various previous attempts to create Moral Machines. Notable early attempts include those made by Anderson et al. (2006) who aimed to design moral machines by integrating philosophical ethical theories with computer science techniques in a healthcare setting. Their machine is moral insofar as it can follow a philosophical doctrine (namely: *Deontological Ethics*) and apply this to moral situations, with a machine's course of action being dictated by prioritised duties. However, one reason why this is problematic is that what ultimately counts as a *prioritised duty* comes down to a crowdsourced consensus on opinion. I may not appeal to my duty to preserve autonomy because it conflicts with my duty to be honest, but the significance of this in the machine designed by Anderson et al, comes down to training on examples that only capture opinions. So, if people intuit that honesty is more important than autonomy, then this is the option the machine will take. The problem is that, in reality, the *right thing to do* does not always correspond to the majority opinion. As Kenward and Sinclair (2021) outline, the *minority* opinion is often the one that drives moral and social change. More recent attempts to create moral machines have come across similar issues. This was particularly evidenced by *Delphi* (2021) which, designed on a similar crowdsourced-based architecture, came to make some controversial and seemingly 'immoral' judgments (deteriorating in much the same fashion as Microsoft's Tay (see Davis, 2016)).

A more fundamental philosophical problem is the question whether morality can be reduced to opinions and rules at all, and whether the ethical frameworks used here (for example Deontological Ethics) can sufficiently deal with the challenges of real-world moral situations, in which it is not always clear, for example, what one's duty is or what doing one's duty implies in the specific situation.

The aim of this paper, however, is not to criticise previous attempts to create moral machines, nor is it to provide an argument for or against their possibility, or whether we should be building them. Neither is our aim to apply theories or philosophical doctrines to the Machine Ethics problem. We are not attempting to *solve Machine Ethics*. Instead, we aim to use empirical evidence concerning the use of the imagination in moral decision-making (as per Coeckelbergh and Mesman, 2007), along with a detailed analysis, to specify some requirements for moral capacity in a machine that we believe have yet to be addressed in the Machine Ethics literature.

Thus, we do not offer a pathway to resolving *Machine Ethics*, but a clarification piece on what moral capacity should consist of, with specific requirements derived from the need for imagination in moral decision-making, and the hope of informing future research in this area (whether this be in

providing further evidence to suggest that moral machines are not possible, or in aiding their appropriate development).

2. Imagination, and its role in moral decision-making

Whereas many modern philosophical accounts of morality may focus on rational capacities for individuals to resolve moral conflicts (tracing back to the works of Kant, 1908 and Bentham, 1843), many arguments suggest that morality is much more than just rational thinking. For example, according to Virtue Ethics, the *virtues*, ultimately linked to the promotion of well-being, are also associated with complex emotional and social features (Kraut, 2022). This means, in terms of 'moral thinking' we cannot simply appeal to rationally deducible rules or laws to make moral decisions.

Given the complex social and emotional (and rational) nature of morality, it is often argued that *the imagination* is one capacity that is needed at least in moral decision-making. Imagination can be conceptualised in various ways. One conception is imagination as empathy. For example, it has been argued that in order to *socially empathise* with another individual (and thereby make a *rational* moral decision) some capacity for 'putting oneself in another's position' is required. This echoes philosophical theories of empathy (i.e. Gibson, 2015) as well more psychological accounts concerning *Theory of Mind* (For example, Baron-Cohen, 1999). In Applied Ethics and the Philosophy of Technology, the idea has also been asserted with regard to value-sensitive design (Friedman and Hendry, 2019).

Coeckelbergh and Mesman (2007) argue that not only is imagination required for driving empathy, but it is also required in the planning phase of moral decision-making - for example - when making moral decisions with future implications or when deciding upon the best course of action to take in a moral dilemma. Here imagination means envisaging future consequences.

The particular case we draw upon in this paper involves this imagination about the future, drawing from a particular case study involving parent decisions in neonatal intensive care units. When making decisions that will have implications for their unborn child, parents are required to exercise the imagination to envisage what the different outcomes will be when faced with multiple options. But the case also involves other uses of the imagination. Outcomes of the case study suggest that the following are required for moral decision-making:

- 1) being able to imagine possible futures (future planning)
- 2) improvisation (imagining oneself in different scenarios)
- 3) empathy/sympathy (having an appropriate affective response to such imaginings)

3. A case study: neonatal intensive care units

The case study that is highlighted in this paper is that of parents making decisions that will implicate the mother and their unborn child within a neonatal intensive care unit (NICU). As per Coeckelbergh and Mesman (2004), with NICU's providing treatment for some of the most unwell newborn infants, there are frequent instances where parents are forced to make quick moral decisions that could have life-affecting outcomes for both the mother and baby. Though regulation and protocols exist to guide medical practitioners in the delivery of care in such situations, often moral dilemmas present themselves that are not covered by regulation, and important judgments need to be made.

In a particular highlighted example, two parents are presented with such a dilemma. With the unborn child's health rapidly deteriorating inside the mother, the parents must make a quick decision concerning whether to administer medication to the baby (the result of which would be terminal) or to carry out an emergency caesarean, which has a (roughly) 10% chance of being

terminal, an 80% chance of resulting in the baby having severe disabilities, or a 10% chance of no problems at all. It is important to note that in both instances, the mother should have no resulting problems.

When asked to consider which option should be taken, the father gives the following statement:

"So, it is a choice between 'taking a risk and giving it a chance' or 'avoid any risk and take away every chance. I think that in spite of her slim chances we have to give her the benefit of the doubt and go for it. But I don't want her to live a life as a severely handicapped human being. Not for my own sake but for her. But if she has a chance, no matter how small, I think we should go for it." (pg. 7)

Similarly, but with different considerations, the mother asserts the following:

"At some point the heart beat slowed down and at times even disappeared completely. You can't imagine how it is to hear how the heartbeat of your child stops suddenly [...]. So something had to be done quickly. I didn't want the child to die in my belly. To have a dead child in my belly was such an unbearable idea. It had to go out." (pg. 7)

Given the thoughts of both parents, a decision is made to carry out the caesarian. However, although the motivations for coming to this decision somewhat vary amongst the mother and father, it's obvious that the imagination plays an active role in each decision. In the father's case, it is used to make projections about the future of the child - namely, having a life after a successful caesarean. Statements such as 'I don't want her to live a life as a severely handicapped human being' indicate use of the imagination to consider possibilities that have not yet happened. Two possibilities are *imagined* and a moral decision is made from this imagination. In the mother's case, it is her imagining of the child inside her, along with the associated emotion it evoked, that leads to her decision.

What is important to note, is that where forward-thinking is the main contributing factor in the moral decision made by the father, it is the imagery invoked by the prospect of no decision that leads to an emotive response, and thereby moral decision by the mother. Though both decisions use the *imagination* to facilitate their decisions, different cognitive processes lead up to both.

In addition, the doctors - and more general medical staff - make decisions within the space left by the protocols and the regulation. The specialist projects scenarios based on their own experience. But sometimes a quick decision and improvisation is necessary. In staff meetings difficult decisions must be made. There are rules and general categories to guide such decision-making, but there is room for improvisation and tuning into a social context when dealing with the concrete situation at hand. A new action is improvised in that particular situation.

4. Understanding the role of The Imagination

According to Liao and Gendlar (2020):

"To *imagine* is to represent without aiming at things as they actually, presently, and subjectively are. One can use imagination to represent possibilities other than the actual, to represent times other than the present, and to represent perspectives other than one's own."

Both of the examples described in the case study above would be considered an instance of *imagining*, as both the mother and father represent possibilities that are *other* than the actual regarding their unborn child.

Currie (2005) draws the distinction between two different types of *Imagination*: what he calls *Recreative Imagination* - the capacity to recreate things, i.e. to copy or *simulate* - and *Creative Imagination* - which is the capacity to put ideas together to create something new, or 'make a [cognitive] leap' (pg. 254). The examples described above could be regarded as mainly constituting instances of recreative imagination (the parents are essentially 'simulating' different outcomes). But some creative imagination was also used: it came when the doctors improvised, using their skill and experience and harmonising the decision. *Both* the capacity to 'recreate things' and 'create things' could be regarded as instances, or exercises of, the *imagination*.

With this in mind, we see that the imagination has a four-fold effect on the moral decision-making of its owners:

- 1) *it allows the owners to make forecasts about what might happen in the future*

Both parents are able to envisage what the outcomes will be given in different moral decisions. In the case of the father, he is able to realise that there is a chance of there being no problems if an intervention is made. The mother is able to envisage what will happen if there is no intervention, and due to not liking the imaging, makes a decision.

- 2) *it allows to owners to imagine how they might react in different situations*

The father is able to envisage the various outcomes and what the case might be if the intervention is not entirely successful. Similarly the mother emotes how she would feel if no intervention were made, and it is her response to this which leads to her decision.

- 3) *the owners have an affective/emotive response to the imaginings*

Though not explicitly stated by the father, there is an obvious disliking of the situation where the intervention is not wholly successful implied by 'in spite of her slim chances we have to give her the benefit of the doubt and go for it'. The role of the emotions is explicitly described by the mother who describes one possibility as 'unbearable'.

- 4) *the owners can use their creative imagination to improvise when it comes to decisions about action*

Based on the parents' recreative, forward-looking imagination, but also on their own scenarios, drawn from past experience, the doctors then make consensus-based decisions that involve the creative use of the imagination to improvise when rules are not enough, next to using technologies and scientific knowledge and social tuning. What the rules say has to be aligned with the particularities of the moral situation. This is what the doctors did: they needed to improvise their position and find new action options with the social process of trying to reach a consensus. Imagination thus helped to harmonise between the general and the particular, between the staff members, and between the various elements available (including evidence.) (pg. 16)

5. Implications for Machine Ethics

If the imagination is required in these morally relevant ways, then this means that we are unable to create a moral machine unless this imaginative faculty can sufficiently be replicated. It is not simply sufficient to create a machine capable of *mimicking* moral decision-making behaviours, but for a machine to be sufficiently regarded as moral it must exercise the same capacities that are present in the example above, and we must be able to see evidence of these capacities through testing and verification afterwards.

For instance, we see that part of moral decision-making is the ability to make forecasts about what might happen in the future. If we are to accurately recreate a machine with moral capacity we must also recreate this ability; the machine must be able to have some ability to forecast future outcomes based upon speculative decisions that are made.

Though Vanderelst and Winfield (2018) discuss the need for a 'simulation faculty' in their approach to creating moral machines, the capacity they give to their machine does not allow the machine to speculate about what will happen given certain decisions, and it also does not affect an emotive response in the machine. Though simulation appears to equate to the recreative imaginative faculty described above, more needs to be added to machines to enable them to make moral decisions.

Delphi and similar crowd-source based attempts also seem to lack any imaginative faculty on behalf of the machine. In training a machine to optimise behaviour towards expected human behaviour, there is no forward-planning ability *at all*. The machine is not envisioning a future, and then making a decision to enable that future, it is simply copying how others behave. As outlined in Coecklebergh (2010) such moral behaviour can only ever be considered as 'quasi-moral' in the respect that it gives the impression of being moral:

Can we build 'moral robots'? If morality depends on emotions, the answer seems negative. (...) at most, robots could be programmed to follow rules, but it would seem that such 'psychopathic' robots would be dangerous since they would lack full moral agency. However, I will argue that in the future we might nevertheless be able to build quasi-moral robots that can learn to create the appearance of emotions and the appearance of being fully moral."

Thus, the challenge for Machine Ethicists changes focus, where more work needs to be carried out to understand how these cognitive capacities can be recreated, rather than merely imitated.

6. Reflections and future research

As the imagination is so fundamental to moral decision-making in the various ways outlined in this paper, it seems that a greater understanding of this capacity and its use in situations and contexts needs to be reached before we can truly begin to create machines with the ability to make moral decisions. However, the imagination is not the only capacity that seems fundamental to moral decision-making; and more research needs to be carried out to understand (a) what other capacities are required for moral decision-making and (b) how these might be recreated in a machine.

The task of creating moral machines thus becomes much more complicated than understanding human value systems and directly applying these to a machine; it becomes an endeavour in understanding and then recreating the cognitive capacities underlying moral decision-making.

So far the emphasis in Machine Ethics has been on replicating moral capacity that has been viewed in humans, whereas, this approach demonstrates a shift in approach and a new challenge for researchers working in the field: how to create an AI system that is able to cognitively carry out moral decision-making rather than just replicating it?

The pursuit is likely to require the involvement of researchers from across cognitive and computer science as well as philosophy. Moreover, though this paper has introduced some examination of the imagination as a cognitive function, perhaps greater understanding can be gained by examining not only the cognitive but also neuroscientific mechanisms underlying imaginative thinking and in understanding the conditions in which it operates (psychology). If this is true, then if research in Machine Ethics is to progress, more work needs to be undertaken in understanding cognitive functions in the brain; one could try to model these.

7. Conclusion

This paper has demonstrated the key role the imagination plays in moral decision-making through reintroduction of a practical case study involving prenatal infants and the roles parents play in making decisions concerning their outcomes.

We have demonstrated that several variations of the imagination are at play when it comes to making moral decisions, including the ability to *imagine* oneself in future potential scenarios (for example on the part of the parents) and the ability to creatively problem-solve (the Doctors). Various capacities seem necessary to exercise the moral imagination: the capacity to have emotional responses to future imaginings and the capacity to imagine 'what might be' given various moral decisions, but also the capacity to imagine action options and find social tuning and alignment when existing rules are not enough and opinions differ.

After reflection and analysis of recent Machine Ethics literature, it seems that those who are attempting to build moral machines are neglecting to include such criteria in their moral machine designs. As such, the current moral machines fall short in several respects on account of the fact that they are only merely mimicking human morality rather than recreating moral capacity, in this case the capacity of imagination.

We call for more research into the cognitive capacities underlying moral decision-making, with the help from philosophy next to psychology and perhaps neuroscience, and suggest that Machine Ethics research advances in understanding whether and how these can be replicated, rather than explicitly trying to recreate human value systems.

References

- Anderson, M., Anderson, S.L. and Armen, C., 2004, July. Towards machine ethics. In *AAAI-04 workshop on agent organizations: theory and practice*, San Jose, CA.
- Anderson, M., Anderson, S.L. and Armen, C., 2006, August. MedEthEx: a prototype medical ethics advisor. In *AAAI* (pp. 1759-1765).
- Anderson, M. and Anderson, S.L., 2007. The status of machine ethics: a report from the AAAI Symposium. *Minds and Machines*, 17(1), pp.1-10.
- Baron-Cohen, S., 1999. *The evolution of a theory of mind* (pp. 261-277). na.

- Bentham, Jeremy. *The Works of Jeremy Bentham*. Vol. 7. W. Tait, 1843.
- Coeckelbergh, M. 2004. *Imagination and Principles: An Essay on the Role of Imagination in Moral Reasoning*. Palgrave.
- Coeckelbergh, M. and Mesman, J., 2007. With hope and imagination: imaginative moral decision-making in neonatal intensive care units. *Ethical theory and moral practice*, 10(1), pp.3-21.
- Coeckelbergh, M., 2010. Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12(3), pp.235-241.
- Currie, G., 2005. Imagination and make-believe. In *The Routledge companion to aesthetics* (pp. 355-366). Routledge.
- Davis, E., 2016. AI amusements: the tragic tale of Tay the chatbot. *AI Matters*, 2(4), pp.20-24.
- Friedman, B. and Hendry, D.G., 2019. *Value sensitive design: Shaping technology with moral imagination*. Mit Press.
- Gibson, J., 2015. Empathy.
- Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M. and Choi, Y., 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.
- Kant, I., 1908. Critique of pure reason. 1781. *Modern Classical Philosophers*, Cambridge, MA: Houghton Mifflin, pp.370-456.
- Kraut, Richard, "Aristotle's Ethics", *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2022/entries/aristotle-ethics/>.
- Kenward, B. and Sinclair, T., 2021. Machine morality, moral progress, and the looming environmental disaster. *Cognitive Computation and Systems*.
- Liao, Shen-yi and Tamar Gendler, "Imagination", *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/imagination/>.
- Moor, J.H., 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4), pp.18-21.
- Vanderelst, D. and Winfield, A., 2018. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48, pp.56-66.
- Wallach, W. and Allen, C., 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Imagination, Morals and Machines: The role of imagination in moral decision-making and its implications for the project of creating moral machines

Rebecca Raper¹ and Mark Coeckelbergh²

Abstract

The field of Machine Ethics has long grappled with issues concerning how we can create machines with the capacity to tell between right and wrong. However, research in the field, particularly recent research, seems to ignore a body of empirical evidence concerning what might be required for moral capacity. In this paper we consider empirical evidence concerning the role the imagination plays in moral decision-making and use it to argue for its role in making moral decisions, and that if we are to truly create machines with morals, we need to focus research on trying to recreate the imaginative faculty. The argument is extended more generally, and it is suggested that for future research within machine ethics, that research should focus, not on trying to mimic moral behaviour, but on trying to understand the cognitive features underlying moral decision-making.

Key words: Imagination, Machine Ethics, Cognitive Robotics, AI, Morals, Moral Decision-making.

1. Rebecca Raper, Oxford Brookes University, Oxford, UK, 17113822@brookes.ac.uk (corresponding author)
2. Professor Mark Coeckelbergh, The University of Vienna, Vienna, Austria, mark.coeckelbergh@univie.ac.uk

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

11.2. Other Media

Throughout the course of my PhD I have delivered many public-facing presentations and featured on podcasts and interviews. Some of these features are listed below:

- **Moral Machines Podcast** where I discuss the topic of my PhD: <https://www.machine-ethics.net/podcast/moral-machines-with-rebecca-raper/>
- **PyData Berlin Presentation:** https://youtu.be/xeGR_VZXJcl